Human-centered Explainable Artificial Intelligence for Natural Language Processing

Von der Fakultät Informatik, Elektrotechnik und Informationstechnik der Universität Stuttgart zur Erlangung der Würde eines Doktors der Naturwissenschaften (Dr. rer. nat.) genehmigte Abhandlung.

Vorgelegt von

Hendrik Schuff

aus Wiesbaden

Hauptberichter	Prof. Dr. Ngoc Thang Vu
Mitberichterin	Prof. Dr. Elisabeth André

Tag der mündlichen Prüfung: 01.12.2023

Institut für Maschinelle Sprachverarbeitung
der Universität Stuttgart

2024

Erklärung (Statement of Authorship)
Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig verfasst habe und dabei keine andere als die angegebene Literatur verwendet habe. Alle Zitate und sinngemäßen Entlehnungen sind als solche unter genauer Angabe der Quelle gekennzeichnet.

Contents

Acknowledgments

I would like to thank my supervisors Heike Adel and Thang Vu for their exceptional support throughout my entire PhD time. I am grateful for the research skills they taught me, the excellent discussion and brainstorming sessions we had, and the scientific freedom they gave me to pursue my own research interests. Most of all, I would like to thank them for the trusting and friendly working atmosphere which I genuinely enjoyed working in over the last years.

Within in my PhD, I had the chance to collaborate with a number of great researchers of whom I would particularly want to thank Alon Jacovi, Yoav Goldberg, Lindsey Vanderlyn, Peng Qi, Diego Frassinelli, Hsiu-Yu Yang, and Wei Zhou. I am grateful for the breadth of research methods, styles of working, and ways of thinking I could learn from them.

Throughout my PhD time, I could rely on the support of the Digital Phonetics/SaLT group at the University of Stuttgart as well as the R26/AIR3 group at the Bosch Center for Artificial Intelligence. I am happy to have met great PhD colleagues and friends in both groups and would in particular like to thank Lukas Lange, Stefan Grünewald, Sophie Henning, Subhash Chandra Pujari, Youmna Ismaeil, Lindsey Vanderlyn, Dirk Väth, Pascal Tilli, Daniel Ortega, Florian Lux, and Maximilian Schmidt with whom I could share great discussions, coffee breaks, pizza, and travels. I also would particularly like to thank Jannik Strötgen, Annemarie Friedrich, and Daria Stepanova for their help, feedback, and advice at the BCAI.

In addition, I am grateful for the over thousand anonymous participants of this thesis' user studies and greatly appreciate their ratings, feedback, and ideas, which fueled this thesis. Also, I thank Luigi Bezzera and Angelo Moriondo for providing engaging input and motivation.

Finally, I thank my family for always supporting me and enabling me to complete my studies after a total of ten years at the University of Stuttgart. I particularly thank Daniela Nürnberg for always being there for me and helping me to put things into perspective.

List of Abbreviations

ADM automated decision making

AI artificial intelligence

ANOVA analysis of variance

CE cross-entropy

CFA confirmatory factor analysis

CLD compact letter display

CLMM cumulative link mixed model

EM exact match

FOL feeling of learning

FOST facets of system trustworthiness

GAM generalized additive model

GAMM generalized additive mixed model

GLM generalized linear model

GLMM generalized linear mixed model

HCI human-computer interaction

IAA inter annotator agreement

KI künstliche Intelligenz

KL Kullback-Leibler

LLM large language model

LRT likelihood-ratio test

MLP multilayer perceptron

MSE mean squared error

MSV maschinelle Sprachverarbeitung

MTurk Amazon Mechanical Turk

NFC need for cognition

NLG natural language generation

NLI natural language inference

NLP natural language processing

PSP perceived system predictability

QA question answering

RL reinforcement learning

RNN recurrent neural network

ROC receiver operating characteristic

SELU scaled exponential linear unit

SGD stochastic gradient descent

SIPA subjective information processing awareness

List of Figures

List of Tables

Abstract

With the ongoing advances in artificial intelligence (AI) systems, their influence on our private, professional, and public life is expanding. While these systems' prediction performance increases, they often rely on opaque system architectures that hide the reasons for the systems' decisions. The field of explainable AI thus seeks to answer *why* a system returns its prediction.

In this thesis, we explore explanatory methods for natural language processing (NLP) systems. Instead of focusing on the technical aspects of explainability in isolation, we take a human-centered approach and additionally explore users' perception of and their interaction with explainable NLP systems. Our contributions thus range on a spectrum from technology-centered machine learning contributions to human-centered studies of cognitive biases.

On the technical end of the spectrum, we first contribute novel approaches to integrate external knowledge into explainable natural language inference (NLI) systems and study the effect of different sources of external knowledge on fine-grained model reasoning capabilities. We compare automatic evaluation with user-perceived system quality and find an equally surprising and alarming disconnect between the two. Second, we present a novel self-correction paradigm inspired by Hegel's dialectics. We apply our resulting *thought flow network* method to question answering (QA) systems and demonstrate our method's ability to self-correct model predictions that increase prediction performance and additionally find that the corresponding decision sequence explanations enable significant improvements in the users' interaction with the system and enhance user-perceived system quality.

Our architectural and algorithmic contributions are followed by an in-depth investigation of explanation quality quantification. We first focus on explainable QA systems and find that the currently used proxy scores fail to capture to which extent an explanation is relevant to the system's answer. We thus propose the two novel model-agnostic scores FARM and LOCA, which quantify a system's internal explanation-answer coupling following two complementary approaches. Second, we consider general explanation quality and discuss its characteristics and how they are violated by current evaluation practices at the example of a popular explainable QA leaderboard. We provide guidelines for explanation quality evaluation and propose our novel "Pareto Front leaderboard" method to construct system rankings to overcome challenges in explanation quality evaluation.

In the last part of the thesis, we focus on human perception of explanations. We first investigate how users interpret the frequently used heatmap explanations over text. We find that the information communicated by the explanations differs from the information understood by the users. In a series of studies, we discover distorting effects of various types of biases and demonstrate that cognitive biases, learning effects, and linguistic properties can distort users' interpretation of explanations. We question the use of heatmap visualizations and propose alternative visualization methods. Second, we develop, validate, and apply a novel questionnaire to measure perceived system predictability. Concretely, we contribute the novel perceived system predictability (PSP) scale, demonstrate its desirable psychometric properties, and use it to uncover a dissociation of perceived and objective predictability in the context of explainable NLP systems.

Overall, this thesis highlights that progress in explainable NLP cannot rely on technical advances in isolation, but needs to simultaneously involve the recipients of explanations including their requirements, perception, and cognition.

Zusammenfassung

Die Fortschritte im Bereich künstlicher Intelligenz (KI) dehnen den Einfluss von KI-Systemen auf unser Privat-, Arbeits- und Gesellschaftsleben kontinuierlich aus. Obwohl die Vorhersagegenauigkeit dieser Systeme kontinuierlich ansteigt, liegen ihnen häufig undurchsichtige Systemarchitekturen zugrunde, die es nicht ermöglichen, den Entscheidungsprozess der Systeme nachvollziehen zu können. Forschung im Bereich erklärbarer KI befasst sich deswegen mit der Frage *wie* ein System zu seiner Vorhersage gelangt.

In dieser Arbeit setzen wir uns mit Methoden der erklärbaren KI im Kontext maschineller Sprachverarbeitung (MSV) auseinander. Anstatt den Fokus auf rein technische Aspekte erklärbarer KI zu legen, verfolgen wir einen nutzerzentrierten Ansatz und schließen zusätzlich die Wahrnehmung erklärbarer MSV Systeme durch deren Nutzer sowie deren Interaktion mit diesen Systemen in unsere Analyse ein. Die Forschungsbeiträge dieser Arbeit decken ein Spektrum von technologiezentrierten Methoden im Bereich des maschinellen Lernens bis hin zu nutzerzentrierten Analysen kognitiver Verzerrungen ab.

Am technologiezentrierten Ende des Spektrums stellen wir neue Methoden zur Integration externer Wissensquellen in erklärbare Systeme zur natürlichsprachigen Inferenz (NLI) vor und untersuchen die Auswirkungen verschiedener Quellen auf spezifische Inferenzfähigkeiten der entsprechenden Systeme. Wir vergleichen die Ergebnisse einer automatisierten Systemevaluation mit der durch Nutzer wahrgenommen Systemqualität und entdecken eine gleichermaßen überraschende und alarmierende Diskrepanz zwischen den beiden Evaluierungssansätzen. Darüber hinaus stellen wir ein neues, von Hegels Dialektik inspiriertes Paradigma zur Systemselbstkorrektur vor. Wir wenden unsere abgeleitete *Thought Flow Network* Methode auf natürlichsprachige Antwortsysteme (QA Systeme) an und zeigen, dass unsere Methode effektiv Systemvorhersagen korrigieren kann, die die Genauigkeit des System erhöhen, und die resultierenden Erklärungen in Form von Entscheidungssequenzen signifikante Verbesserungen der Mensch-System Interaktion und der Systemwahrnehmung durch Nutzer ermöglichen.

Unserer Diskussion verschiedener Systemarchitekturen und Algorithmen folgt eine tiefgehende Untersuchung der quantitativen Messung von Erklärungsqualität. Wir konzentrieren unsere Analyse auf erklärbare QA Systeme und stellen fest, dass herkömmliche Proxymetriken nicht ausreichend erfassen, inwieweit eine Erklärung für die Antwort des Systems relevant ist. Wir stellen deshalb die zwei neuen modellagnostischen Proxyscores FARM und LOCA vor, die die systeminterne Erklärungs-Antwort Kopplung eines erklärbaren QA Systems anhand zweier komplementärer Ansätze quantifizieren.

Anschließend widmen wir uns der Qualität von Erklärungen im Allgemeinen und beschreiben deren generelle Merkmale und die Verletzung dieser durch herkömmliche Evaluierungspraktiken am Beispiel eines populären Leaderboars für erklärbare QA Systeme. Um einige der Herausforderungen bei der Messung von Erklärungsqualität zu bewältigen, stellen wir Leitlinien zur Evaluation von Erklärungsqualität vor und präsentieren unsere neue “Pareto Front

Leaderboard” Methode zur Konstruktion von Ranglisten.

Im letzten Teil der Arbeit befassen wir uns mit der menschlichen Wahrnehmung von Erklärungen. Wir untersuchen zunächst, wie Nutzer die häufig verwendeten Heatmap-Erklärungen über Text interpretieren. Wir stellen fest, dass die durch die Erklärungen kommunizierte Information sich von der von den Nutzern verstandenen Information unterscheidet. In einer Reihe von Studien decken wir den Einfluss verschiedener verzerrender Einflüsse auf und demonstrieren, dass kognitive Verzerrungen, Lerneffekte und linguistische Merkmale die Nutzerinterpretation von Erklärungen verfälschen können. Wir stellen die Verwendung von Heatmap-Visualisierungen infrage und schlagen alternative Visualisierungsmethoden vor. Zusätzlich entwickeln, validieren und verwenden wir einen neuartigen Fragebogen zur Messung wahrgenommener Systemvorhersagbarkeit. Konkret stellen wir die neue PSP Skala vor, weisen deren gute psychometrischen Eigenschaften nach und verwenden die Skala, um eine Diskrepanz zwischen wahrgenommener und objektiver Vorhersagbarkeit im Kontext erklärbarer NLP Systemen offenzulegen.

Zusammenfassend verdeutlicht diese Arbeit, dass Fortschritte auf dem Gebiet erklärbarer KI nicht allein auf technischen Neuerungen beruhen können, sondern gleichzeitig die Nutzer der Erklärungen einschließlich ihrer Anforderungen, Wahrnehmung und kognitiver Eigenschaften einbeziehen müssen.

Prepublications

Parts of this thesis' content have been published in peer-reviewed journals and international conferences. Additionally, preprints of these articles as well as preprints of manuscripts under review are available on the `https://arXiv.org` repository. Further, patent applications have been filed for parts of the presented methods. Table 1 displays the content mapping of this thesis' sections to the respective prepublications. In the following, we list all prepublications along with available code, data, and presentation resources.

Published Papers

- **EMNLP'20**: Schuff, H., Adel, H., and Vu, N. T. (2020). F1 is Not Enough! Models and Evaluation Towards User-Centered Explainable Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7076–7095, Online. Association for Computational Linguistics,
  code and data, presentation.

- **BlackboxNLP'21**: Schuff, H., Yang, H.-Y., Adel, H., and Vu, N. T. (2021b). Does external knowledge help explainable natural language inference? automatic evaluation vs. human ratings. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 26–41, Punta Cana, Dominican Republic. Association for Computational Linguistics,
  experiment data.

- **FAccT'22**: Schuff, H., Jacovi, A., Adel, H., Goldberg, Y., and Vu, N. T. (2022b). Human interpretation of saliency-based explanation over text. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 611–636, New York, NY, USA. Association for Computing Machinery,
  code and data, short presentation, long presentation.

- **NLE**: Schuff, H., Vanderlyn, L., Adel, H., and Vu, N. T. (2023b). How to do human evaluation: A brief introduction to user studies in nlp. *Natural Language Engineering*, page 1–24.

- **Findings of ACL'23**: Jacovi, A., Schuff, H., Adel, H., Vu, N. T., and Goldberg, Y. (2023). Neighboring words affect human interpretation of saliency explanations. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics,
  code and data.

	EMNLP’20	**BlackboxNLP’21**	**FAccT’22**	**JNLE**	**F. of ACL’23**	**arXiv’21**	**arXiv’22**	**USPTO**
Section 2.3				×				
Section 2.4				×				
Section 3.1		×						
Section 3.2						×		×
Section 4.1	×							
Section 4.2							×	
Section 4.3							×	
Section 4.4							×	
Section 5.1			×		×			

Table 1.: Relation of this thesis’ sections to prepublished articles.

Preprints Under Submission

- **arXiv’21**: Schuff, H., Adel, H., and Vu, N. T. (2021a). Thought flow nets: From single predictions to trains of model thought. *CoRR*, abs/2107.12220 (under submission).

- **arXiv’22**: Schuff, H., Adel, H., Qi, P., and Vu, N. T. (2022a). Challenges in explanation quality evaluation. *CoRR*, abs/2210.07126 (under submission).

Published Patent Applications

- **USPTO**: Schuff, H., Adel, H., and Vu, N. T. (2023a). Device and method for classifying a signal and/or for performing regression analysis on a signal. US Patent App. 17/831,750.

1. Introduction

1.1. Motivation

AI and NLP Systems. AI systems affect nearly every aspect of our digital lives: Recommender systems determine which media content we are shown (Covington et al., 2016; Schedl et al., 2018; Raza and Ding, 2022), advertisement systems decide which products we are offered (Gharibshah and Zhu, 2022), and dating apps decide which user profiles we get to see (Courtois and Timmermans, 2018). While these examples primarily concern our voluntary interaction with automated systems in the sense that we can opt out by not using them, AI systems can also affect us indirectly and without our consent. For example, clinical decision support systems are used in medical diagnosis (Sutton et al., 2020), passenger screening systems affect determine the probability that we are being searched at an airport (Waldman, 2019), grade prediction systems can decide on which A-Level grade students receive (Smith, 2020), and welfare fraud prediction system are used to decide which individuals are investigated[1].

A large part of today's AI systems processes natural language and chatbots, QA systems, or content filters already affect the daily work as well as social lives of millions of users. For example, in Helsinki, a maternity clinic chatbot is offered to citizens[2] and in Amsterdam, the municipality uses NLP to assign categories to public space issue reports[3]. Further, social media platforms decide which content we are exposed to by, e.g., using automatic content moderation systems (Jaki and Smedt, 2019; Zhou and Zafarani, 2021) to detect and remove harmful content, such as disinformation.

When these systems are poorly designed or even misused on purpose, they pose serious risks to users, patients, citizens, and societies. These risks include discrimination, arbitrary decisions, misinformation, and manipulation. However, systems that are developed and evaluated carefully can provide benefits to the ones who are affected by their operations. These benefits include improved quality of public service, reduction of harmful content, and increased end user utility.

[1] https://www.wired.com/story/welfare-state-algorithms/

[2] https://ai.hel.fi/en/maternity-clinic-chatbot-nero/

[3] https://algoritmeregister.amsterdam.nl/en/reporting-issues-in-public-space/

The Need for Explainability. How can a malfunctioning or malicious system be distinguished from a genuinely useful system? Clearly, the system's decisions/predictions in a test scenario should be correct, i.e., we have to assess whether the system takes the right decisions. However, this only scratches the surface of the system's behavior. While its predictions might be perfectly correct, we also want them to be "right for the right reasons" (Ross et al., 2017). Investigating *why* a system takes the decisions it does, can allow us to detect model failure beyond prediction behavior. An illustrative real-world example can be found in the work of Zech et al. (2018) who analyze a system developed to detect pneumonia from X-ray scans. The system successfully detected pneumonia cases, however, an analysis of the system's activation maps revealed that — instead of focusing on the patient — the system actually learned to focus on small metal tags in the scans' corners. These tags allowed the system to identify the X-ray system that was used to make the scan. As, e.g., the system used in the emergency unit and an inpatient floor differ, this information allows to make strong predictions regarding a patient having pneumonia without considering the patient. This example shows that monitoring system performance alone is not enough and we, additionally, have to investigate how a system arrives at its decision. In contrast to this computer vision example, a key aspect that makes explainability for NLP challenging is that language is discrete and highly structured. While images contain continuous (color channel) information, a natural language sentence contains discrete elements (in the sense that two words cannot be averaged) which are highly structured (following the grammatical rules of its language).

User-centered Explainability. Explanations for a system's behavior can serve multiple stakeholders. First, explanations can help developers to debug and improve systems (Bhatt et al., 2020). For example, the finding that the pneumonia detection system described above focused on identifying a specific X-ray system can drive the collection of more diverse datasets to enable performance improvements and system robustness. Second, researchers can use explanations to derive findings regarding the modeled task or domain. For example, Watson (2021) describes how explainability methods can be (and are) used for knowledge discovery in genetics. Third, explanations can serve end users. When users are subject to an algorithmic decision, explanations can empower them to detect and challenge erroneous or unfair algorithmic treatment. When users make use of an explainable system as a tool or in human-AI collaboration, explanations can calibrate users' trust in the system's abilities such that they neither under- nor over-rely on its decisions. Each user group will have different requirements and there is no one-size-fits-all solution to explainability. However, as researchers and developers are the ones developing explainability methods, there is a risk of having *the inmates running the asylum*

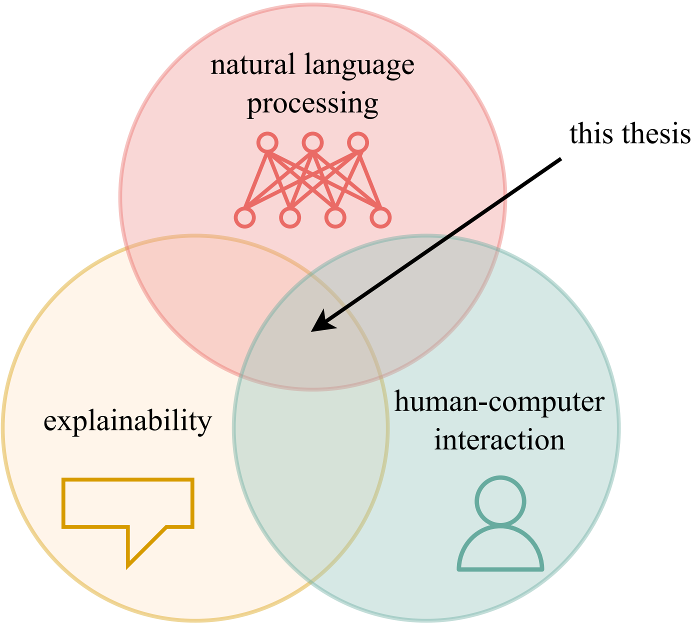


Figure 1.1.: This thesis combines explainability and NLP with a human-centered perspective.

(Miller et al., 2017), i.e., researchers and developers developing methods that suit *their* needs and neglect the needs of the users intended to use these methods. As argued by Miller et al. (2017), developing explainability methods should thus incorporate findings from the social and behavioral sciences.

This Thesis: User-centered Explainable NLP. This thesis connects explainability, NLP, and human-centered approaches (as illustrated in Figure 1.1). We are convinced that only by studying all three aspects in combination, we can move towards meaningful and effective explanations. Concretely, we study explainable NLP along a spectrum ranging from technical aspects of explainability on the one hand to human perception on the other hand (as shown in Figure 1.2). This spectrum is reflected in the thesis' three-folded structure: The first part of this thesis proposes new techniques for integrating external knowledge into explanation generation as well as a novel model self-correction method to produce sequences of model decisions ("thought flows"). The second part addresses evaluation of explanation quality and raises awareness of the shortcomings in today's evaluation practices. In particular, it provides empirical evidence for a disconnect between automatic evaluation and human evaluation across various explainability tasks in NLP. The third part centers on humans and their perception of NLP explanations. We demonstrate that users' perception of explanations can be biased and develop and validate a questionnaire to measure user-perceived system predictability. We present a detailed outline of the structure of this thesis in Section 1.3.

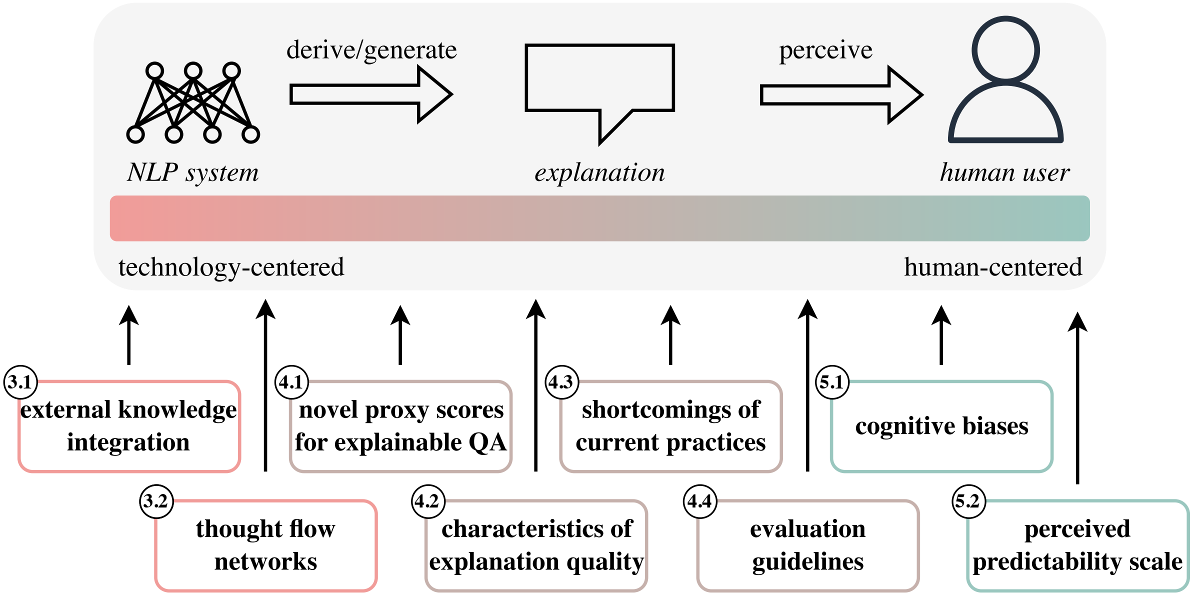


Figure 1.2.: Overview of this thesis' contributions. Contributions are ordered along a spectrum ranging from technology-centered contributions (such as novel system architectures) to human-centered contributions (such as measures of human perception). Circled numbers refer to the respective sections within this thesis.

1.2. Main Contributions

In this thesis, we contribute to explainability research with a focus on NLP. Our contributions cover a broad spectrum ranging from technical machine learning contributions to human-centered contributions. Figure 1.2 depicts our main contributions ordered along this spectrum. In the following, we partition this spectrum into three groups reflecting the three-fold structure of this thesis.

Architectures and Algorithms

On the technology-centered end of the spectrum (depicted on the left in Figure 1.2), we first contribute (i) various approaches to integrate external knowledge into explainable NLI systems. Prior work showed that external knowledge can improve NLP systems across a wide range of tasks (Shi et al., 2016; Seyler et al., 2018; Pan et al., 2019; Lin et al., 2019a) including NLI systems (Chen et al., 2018; Wang et al., 2019; Li et al., 2019; Faldu et al., 2021; Bauer et al., 2021). We explore how the integration of external knowledge affects *explainable* NLI systems. Concretely, we compare various sources of external knowledge for which we propose different integration methods and investigate how the choice of knowledge source/integration method affects model performance, fine-grained model reasoning capabilities, and explanation generation. In addition, we complement our automatic evaluation with human evaluation and find that improvements in proxy scores do not transfer to quality improvements in user ratings.

Further, we (ii) propose a novel self-correction paradigm based on Hegel's dialectics that we call "thought flow networks". Related work explored various approaches for model self-correction, however, existing approaches are either task-specific (Mori et al., 1973; Katupitiya and Gock, 2005), are not applicable to pre-trained models (Hopfield, 1982; Koller and Friedman, 2009; Barra et al., 2018; Ramsauer et al., 2020) or cannot be applied iteratively (Wei et al., 2022). In contrast, we derive a new system architecture along a novel training/prediction paradigm, that is task-agnostic, can be applied on top of existing models, and can iteratively correct model predictions. We apply our method to QA models and demonstrate our model's ability to correct its own predictions and its potential to notably boost model performance, and find promising improvements in user performance and user-perceived system quality.

Evaluation Methodology and Proxy Scores

In the middle of the spectrum, we first (iii) propose novel proxy scores to evaluate explainable QA systems. While prior development of explainable QA system focused on F_1-scores to quantify the models' explanation quality (i.a., Yang et al., 2018; Fang et al., 2020; Tu et al., 2020; Nishida et al., 2021; Li et al., 2022) by comparing ground truth explanations to system explanations, we introduce two scores that quantify answer-explanation coupling *without* ground truth explanations.

Further, we (iv) discuss general explanation quality based on insights from behavioral sciences. We formulate characteristics of explanation quality, and (v) demonstrate how today's explanation quality evaluation approaches violate them. Most importantly, we discover an alarming disconnect between automatic and human evaluation of explanation quality. While the use of automatic proxy scores has been questioned frequently within natural language generation (NLG) (e.g., regarding BLEU) (Callison-Burch et al., 2006; Liu et al., 2016; Novikova et al., 2017; Sulem et al., 2018; Reiter, 2018) and the need for human-centered evaluation approaches in explainable AI have been stressed by (i.a. Ribera and Lapedriza, 2019; Chu et al., 2020; Gonzalez et al., 2021; Colin et al., 2021; Schlegel et al., 2022; Liao et al., 2022), the relation between proxy scores and human ratings received little attention in the context of explainable AI (Kayser et al., 2021; Clinciu et al., 2021). In particular, this relation was — to the best of our knowledge — not studied for explainable QA. We thus investigate explainable QA models and compare the respectively used *de facto* standard proxy scores to numerous human self-reports of explanation quality. We find that automatic evaluation poorly reflects the explanations' utility to users and the users' perceived explanation quality.

To address some of the main challenges that today's explanation quality evaluation is facing, we (vi) provide guidelines and practical recommendations including a new leaderboard construction method that we call "Pareto Front leaderboards". In contrast to previously proposed leaderboard alternatives (Chaganty et al., 2017; Ethayarajh and Jurafsky, 2020; Linzen, 2020) our method can combine multiple quality dimensions into a joint system ranking without condensing score dimensions into a single score.

Cognitive Biases and Human Perception

On the human-centered end of the spectrum (depicted on the right in Figure 1.2), we first contribute (vii) the discovery that cognitive biases can affect human understanding of explanations. Concretely, we study saliency (heatmap) explanations over text, find that the information communicated by the explanation differs from the information understood by users, and propose alternative visualization methods that mitigate the effect of the respective cognitive biases. In contrast to prior work that explored belief bias in users' decision behavior (Gonzalez et al., 2021), we investigate and discover a broad range of biasing influences, such as visual properties of an explanation or learning effects and focus on the users' perception of explanations.

Finally, we (iix) develop and validate a questionnaire to provide a solid foundation to study perceived system predictability. While prior work only includes predictability as a facet of higher-level constructs (Schrills et al., 2022) or explored related constructs, such as trust (Cramer et al., 2008; Ribes et al., 2021; Khurana et al., 2021) or perceived usefulness (Khurana et al., 2021; Bansal et al., 2021) our scale is the first validated instrument to measure perceived predictability along with subordinate dimensions of predictability. Concretely, we propose a theory of perceived predictability based on uncertainty theory, construct a 6-item Likert scale, demonstrate its desirable psychometric properties, and apply it in a large-scale user study to explore how explanations and system stochasticity affect perceived predictability and how perceived predictability is related to objective predictability and other subjective dimensions, such as trust.

1.3. Structure

The mapping of the described contributions to sections is depicted in Figure 1.2. Coarsely, this thesis is structured into five chapters following this introduction:

- Chapter 2 provides the reader with background on NLP tasks and model architectures (Section 2.1.1), explainability (Section 2.2), user study design (Section 2.3), and statistical methods (Section 2.4).

- Chapter 3 presents our technical system contributions regarding external knowledge integration for explainable NLI (Section 3.1) and our self-reflective thought flow networks (Section 3.2).

- Chapter 4 addresses explanation quality evaluation by proposing new proxy scores for explainable QA (Section 4.1), discussing fundamental characteristics of explanation quality (Section 4.2), arguing how today's evaluation approaches violate them (Section 4.3), and proposing remedies to re-orient explainable NLP system development towards more effective explanation quality evaluation (Section 4.4).

- Chapter 5 focuses on human perception of explanations. In a series of studies, we demonstrate that human interpretation of heatmap explanations over text is distorted by cognitive biases and present alternative visualization methods (Section 5.1). Additionally, we develop and psychometrically validate a questionnaire to measure general user-perceived system predictability and explore the impact that different types of explanations and levels of system stochasticity have on users' perceived ability to be able to predict a system's behavior (Section 5.2).

- Chapter 6 concludes the thesis and discusses future work.

2. Background

This chapter introduces this thesis' context with respect to the addressed NLP tasks and typical model architectures (Section 2.1), explainable AI and explainability methods within NLP (Section 2.2), core concepts of user studies (Section 2.3) as well as statistical methods that we use in the following chapters (Section 2.4).

2.1. Tasks, Datasets, and Systems

In this section, we present the NLP tasks that we address within this thesis (Section 2.1.1) along with common system architectures that are used to model them (Section 2.1.2).

2.1.1. Tasks and Datasets

Throughout this thesis, we focus on two tasks: explainable QA and explainable NLI. In the following, we introduce the "classic" as well as the explainable versions of these tasks and clarify what the inputs and outputs of the respective systems are.

2.1.1.1. (Explainable) Natural Language Inference

Natural Language Inference. NLI tasks require systems to decide how two sentences are related regarding the relation of their meanings (Jurafsky and Martin, 2023). More concretely, the systems receive two sentences (the *premise* and the *hypothesis*) and predict a label corresponding to entailment, contradiction, and (in some datasets) a neutral relation. Table 2.1 displays three examples of premise-hypothesis pairs corresponding to the three relations. Figure 2.1 depicts an example in which the correct relation class is entailment. As shown in Figure 2.1, NLI systems receive the premise and the hypothesis sentences and predict an inference relation label. The yellow box on the bottom right of Figure 2.1 shows an explanatory extension and is not part of the classic NLI task. Popular NLI datasets include SNLI (Bowman et al., 2015), MultiNLI (Williams et al., 2018), and XNLI (Conneau et al., 2018).

Relation	Premise	Hypothesis
entailment	A soccer game with multiple males playing.	Some men are playing a sport.
contradiction	A man inspects the uniform of a figure in some East Asian country.	The man is sleeping.
neutral	An older and younger man smiling.	Two men are smiling and laughing at the cats playing on the floor.

Table 2.1.: NLI examples with the three relations entailment, contradiction, and neutral. Examples are selected from the SNLI dataset (Bowman et al., 2015).

Explainable Natural Language Inference: e-SNLI. We will work on the e-SNLI dataset (Camburu et al., 2018) in the following chapters of this thesis. The e-SNLI dataset is the SNLI dataset with an additional explanation annotation layer that textually justifies why the entailment/contradiction/neutral label is correct. In the example shown in Figure 2.1, the textual explanation relates “dog” to “animal”, “snow” to “cold weather”, and “jumping for a frisbee” to “plating with a plastic toy”. The dataset is split into 549k training instances, 9842 validation instances, and 9824 test instances and contains crowdsourced free-text explanations.

2.1.1.2. (Explainable) Question Answering

Question Answering. In the QA task, the system receives a question text along with a textual context, such as Wikipedia articles (depending on the specific task setup, the system has to retrieve the relevant context itself). The system’s task is to map these two inputs to an answer. Depending on how the answer is constructed, one distinguishes extractive QA, abstractive QA, and multiple choice QA. We depict an exemplary (extractive) QA instance in Figure 2.2. In *extractive* QA, the answer is chosen as a text span from the context, i.e., we assume that the question can be answered by determining a start position and a stop position within the given context. Popular datasets for extractive QA include SQuAD (Rajpurkar et al., 2016), TriviaQA (Joshi et al., 2017), NewsQA (Trischler et al., 2017) and Natural Questions (Kwiatkowski et al., 2019). In *abstractive* QA, the answer is generated based on the question and context texts. This allows an abstractive system to generate answers that do not exist as a text span in the context. Popular datasets for abstractive QA include NarrativeQA (Kočiský et al., 2018), ELI5 (Fan et al., 2019), and TweetQA (Xiong et al., 2019). In *multiple choice* QA, the system receives — in addition to the question and the context — a set of answer candidates. The system’s task is to select the correct answer candidate. Popular datasets for multiple choice QA include RACE (Lai et al., 2017), CommonsenseQA (Talmor et al., 2019), and Cosmos QA (Huang et al., 2019). We refer to Cambazoglu et al. (2020) for an overview of QA datasets.

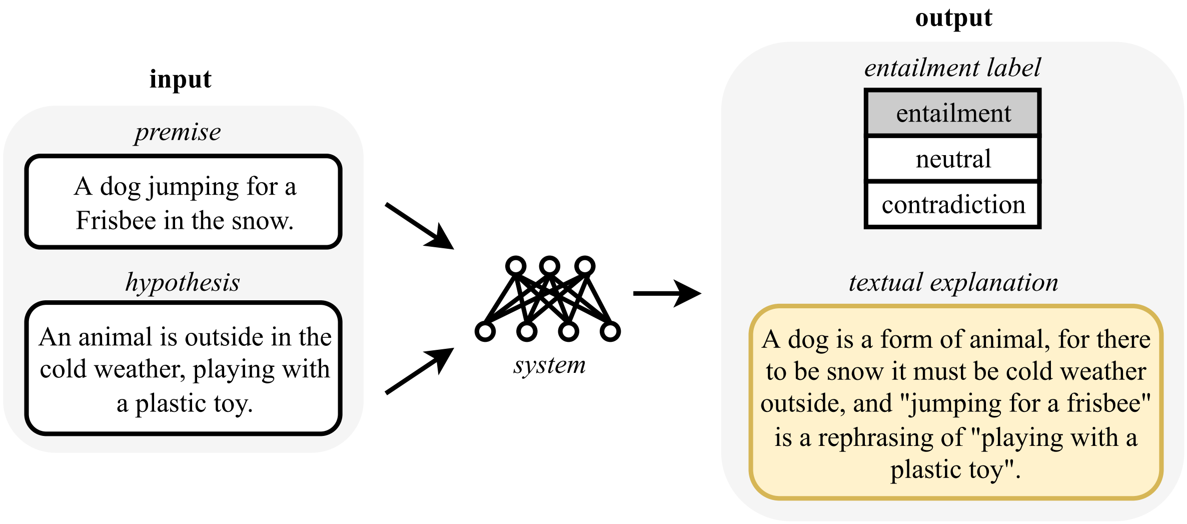

output
input
premise
A dog jumping for a Frisbee in the snow.
hypothesis
An animal is outside in the cold weather, playing with a plastic toy.
system
entailment label
entailment
neutral
contradiction
textual explanation
A dog is a form of animal, for there to be snow it must be cold weather outside, and "jumping for a frisbee" is a rephrasing of "playing with a plastic toy".

Figure 2.1.: Example of an explainable NLI task. The NLI system receives a premise text and a hypothesis text and classifies the hypothesis to either (a) be entailed by the premise, (b) be neutral with respect to the premise, or (c) contradict the premise. Additionally, the system generates a textual explanation to support its prediction. Example texts are taken from the e-SNLI dataset (Camburu et al., 2018).

Explainable Question Answering: HotpotQA. Similar to the explainable extension of NLI task described above, the QA task can also be extended with an additional explanatory output. HotpotQA (Yang et al., 2018) is an extractive QA dataset and extends the answer annotations with supporting facts. A supporting fact is a sentence from the input context and supports the model's answer prediction and thereby serves as an explanation of the model's behavior. Figure 2.2 shows a supporting fact explanation within the bottom right yellow box. We will consider a more complex explainable QA example in Chapter 4.

2.1.2. Deep Learning Models for NLP

While the previous discussed what the inputs and outputs for each task are, we now focus on how the mapping between input and output is established from a system perspective.

2.1.2.1. Representing the Input: From Text to Vectors

The first step in any NLP system considered in this thesis is to represent a textual input numerically. More specifically, we map a text containing $n \in \mathbb{N}$ tokens to a sequence of vectors represented by a matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$ where $d \in \mathbb{N}$ is the embedding dimension, i.e., the number of dimensions we use to represent a token. We demonstrate this first step at the example of the

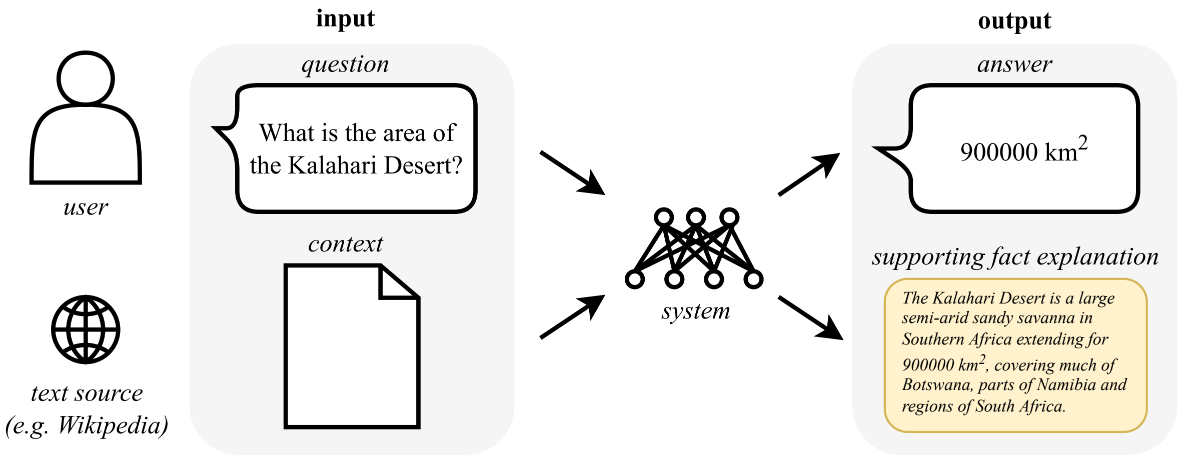

input
output
user
question
What is the area of the Kalahari Desert?
context
text source (e.g. Wikipedia)
system
answer
900000 km²
supporting fact explanation
The Kalahari Desert is a large semi-arid sandy savanna in Southern Africa extending for 900000 km², covering much of Botswana, parts of Namibia and regions of South Africa.

Figure 2.2.: Example of an explainable question answering task. The QA system receives a question text and a context text as inputs and returns an answer text as output. Additionally, the system provides an explanation by predicting which parts of the input context are supporting its answer prediction. Context and answer texts are taken from the HotpotQA dataset (Yang et al., 2018).

text "A dog jumping for a Frisbee in the snow". Note that, in our NLI example (as shown in Figure 2.1), the input consists of two sentences, i.e., the input would, e.g., be the concatenation of the premise and the hypothesis "A dog jumping for a Frisbee in the snow. An animal is outside [...]". In the following, we focus on the processing of a single sentence or parts of a sentence to introduce the several processing stages. We discuss different approaches to how two inputs can be combined into one text in Chapter 3.

Tokenization. First, the input sentence is split into *tokens* in a process called *tokenization*. Figure 2.3 depicts how the text string is mapped to a sequence of tokens. Note that today's systems usually do not break sentences into words and punctuation but apply *subword tokenization* meaning that, e.g., the word "jumping" might be further split into the subword tokens "jump" and "ing". The set of subword tokens that are used by a tokenizer is determined empirically, for instance, using byte pair encoding (Gage, 1994; Sennrich et al., 2016).

Embeddings. The next step is to represent the sequence of tokens in vector space. For this, each token is mapped to a vector representation. Figure 2.4 shows how each token of our example sentence is "embedded" into vector space and, by this, the example sentence is mapped to a sequence of vectors.

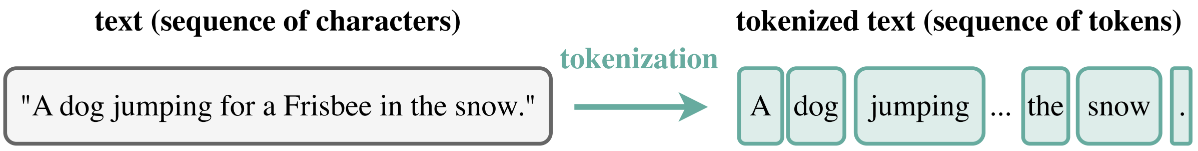

text (sequence of characters)
"A dog jumping for a Frisbee in the snow."
tokenization
tokenized text (sequence of tokens)
A
dog
jumping
...
the
snow
.

Figure 2.3.: Tokenization maps a text string to a sequence of tokens. Tokens can be (sub)words or punctuation.

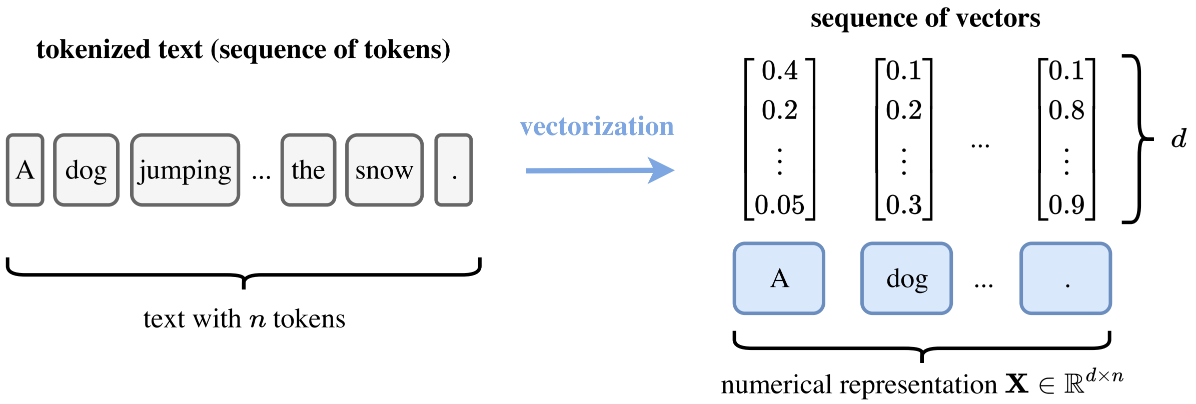


Figure 2.4.: Each token is mapped to a (learned) vector representation.

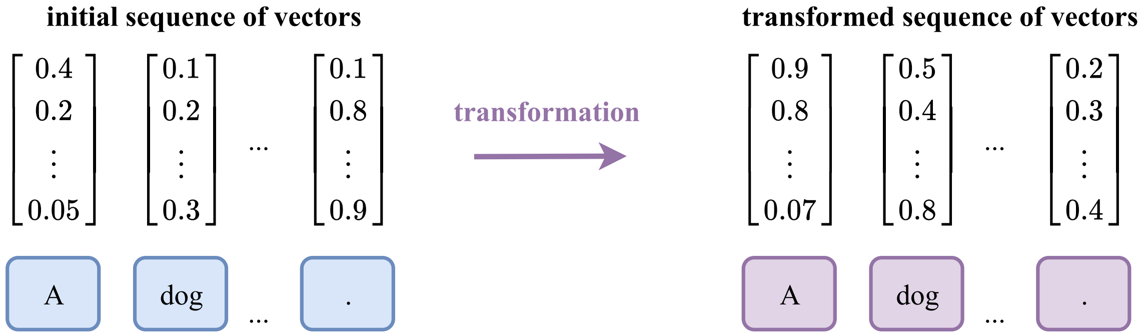


Figure 2.5.: Each token's vector representation is transformed into a contextualized vector representation.

2.1.2.2. Mapping Inputs to Outputs: From Vectors to Vectors

The previous paragraphs describe how a text is mapped to an initial sequence of vectors. Before this sequence is related to, e.g., a predicted class label or a predicted sequence of class labels (which we will discuss in Section 2.1.2.3), the sequence of vectors is transformed into a representation that allows to numerically access the, e.g., entailment information within a text containing two sentences, i.e., the sequence of vectors is transformed into another sequence of vectors as illustrated in Figure 2.5.

In the following, we review model components that we will use within the following chapters. For more details on deep learning models and their components we refer to the textbook by Goodfellow et al. (2016).

Linear Layers. Linear layers (or fully-connected layers) implement affine transformations. Let $\mathbf{x} \in \mathbb{R}^m$ denote the input vector. A linear layer determines the transformed representation $\mathbf{y} \in \mathbb{R}^n$ as

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b}, \tag{2.1}$$

where $\mathbf{W} \in \mathbb{R}^{n \times m}$ is a weight matrix and $\mathbf{b} \in \mathbb{R}^n$ is a bias vector. Intuitively, multiplying with $\mathbf{W}$ defines each dimension of $\mathbf{y}$ as a linear combination of dimensions of $\mathbf{x}$. Adding $\mathbf{b}$ adds a fixed offset to each new dimension. Linear layers are typically used as the very last layer of a model, mapping the last-layer vector representation to, e.g., class probabilities.

Attention. A critical component that can be found within the architectures of merely every recent large language model (LLM) is the attention mechanism. It was initially introduced to align words in neural machine translation using recurrent neural networks (RNNs) (Bahdanau et al., 2015). While previous RNN-based translation systems aggregated the input sentence into a fixed vector representation and subsequently used a decoder to, step by step, generate the predicted translation, Bahdanau et al. (2015) propose to equip the decoder with a mechanism that allows the system to *dynamically* aggregate the input token representations during decoding. Figure 2.6 shows a motivating example in which different decoding steps need to pay "attention" to different parts of the input representation within an English-to-German translation setting. Technically, given a sequence of input vectors $(\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_N)$, the layer's output $\mathbf{A}_t$ at decoding position t is defined as

$$\mathbf{A}_t = \sum_{j=1}^{N} \alpha_{t,j} \mathbf{X}_j \tag{2.2}$$

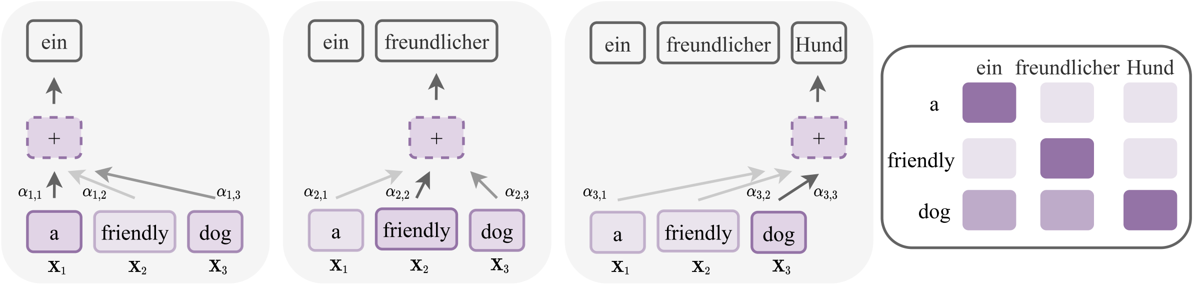


Figure 2.6.: Example showing the motivation behind the attention mechanism. The three decoding steps for the German translation of "a friendly dog" require information from different parts of the English sentence. For example, the translation of "friendly" needs to integrate information about the gender of "dog". As the German translation of "dog" is "Hund" (male), "friendly" translates to "freundlich*er*". In contrast, the German translation of "a friendly cat" is "eine freundlich*e* Katze" as "Katze" is female. Similarly, the German translation of the article "a" depends on the noun's gender. $\alpha_{t,j}$ indicate the attention weight that decoding step t puts on input token j. The rightmost box summarizes the attention weights in a matrix.

with

$$\alpha_{t,j} = \frac{\exp(\mathbf{e}_{t,j})}{\sum_{k=1}^{n} \exp(\mathbf{e}_{t,k})}. \tag{2.3}$$

where $\mathbf{e}$ depends on $\mathbf{X}_j$ (e.g., via dot product attention).

While this attention mechanism laid the conceptual foundation of attention (in transformers), it differs from the succeeding self-attention and masked attention mechanisms, which we will introduce in the following.

Self-Attention. So far, we introduced attention as a mechanism that relates two token sequences as shown in the rightmost box of Figure 2.6. However, attention can also be applied within a single sequence, which is referred to as self-attention (or, originally, intra-attention) introduced by Cheng et al. (2016). Intuitively, self-attention allows a transformation of a token vector sequence into another (same width) sequence that integrates contextual information. Figure 2.7 depicts an example of self-attention in left-to-right encoding. In particular, the fourth row of the left part of the figure (corresponding to the right part of the figure), shows how self-attention can be used to contextualize a token's representation. Concretely, the representation of the token "its" is updated to integrate information about "the dog", to which it refers. As mentioned, different variations of attention scores can be used. In their seminal introduction of the transformer architecture, Vaswani et al. (2017) propose scaled dot product self-attention, which they express in terms of key, value, and query matrices as follows. First, the input vector sequence (i.e., matrix) $\mathbf{X}$ is multiplied with a weight matrix $\mathbf{W}^{\mathbf{Q}}$ to obtain the query matrix $\mathbf{Q}$.

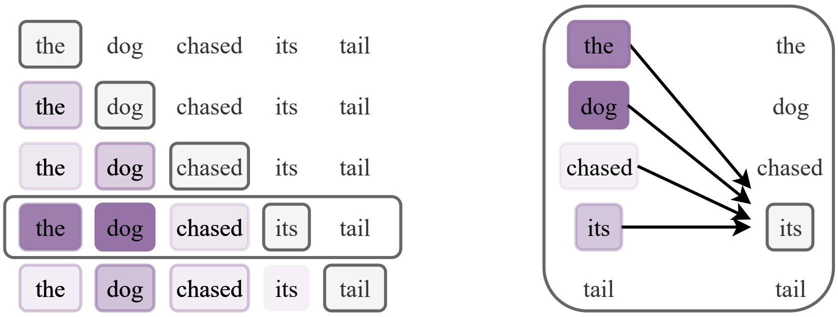


Figure 2.7.: Illustration of the self-attention mechanism. While encoding the sentence “a dog chased its tail”, self-attention allows the model to integrate contextual information. In the example, the representation of the token “its” is updated with a weighted average of the previous step’s representations with emphasis on the sentence’s subject “the dog”. The right part of the figure depicts the third row of the left part of the figure in more detail. Note that this example depicts masked self-attention. Figure 2.8 makes the difference between masked attention and non-masked attention explicit.

Similarly, the key, and the value matrices $\mathbf{K}$ and $\mathbf{V}$ are calculated as the matrix-matrix products of $\mathbf{X}$ with a weight matrix $\mathbf{W}^{\mathbf{K}}$ and $\mathbf{X}$ with a weight matrix $\mathbf{W}^{\mathbf{V}}$ respectively. The result of scaled dot product self-attention is then defined as

$$\text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\mathbf{V}\right), \tag{2.4}$$

where d_k is the embedding dimension of $\mathbf{Q}$ and $\mathbf{K}$. Vaswani et al. (2017) add the normalization $\sqrt{d_k}^{-1}$ to stabilize gradients during training.

Masked Attention. The example in Figure 2.7 shows how a sequence of tokens is processed left-to-right and each token’s new representation is a composition of its preceding, i.e., left tokens. In contrast to such a *masked self-attention*, self-attention can be also applied without restricting a token’s context to its preceding tokens. Figure 2.8 compares how the same example sentence shown in Figure 2.7 can be transformed with masked self-attention (left) and without masking (right). While it might seem to be an unnecessary restriction to mask tokens, we will discuss how the two types of attention are fundamental ingredients of the transformer architecture in the following.

Transformers. One of the driving motivations behind the transformer architecture, as it was originally proposed in the seminal work of Vaswani et al. (2017), is that the previously dominant RNN architectures suffered from the practical disadvantage of not allowing fully-

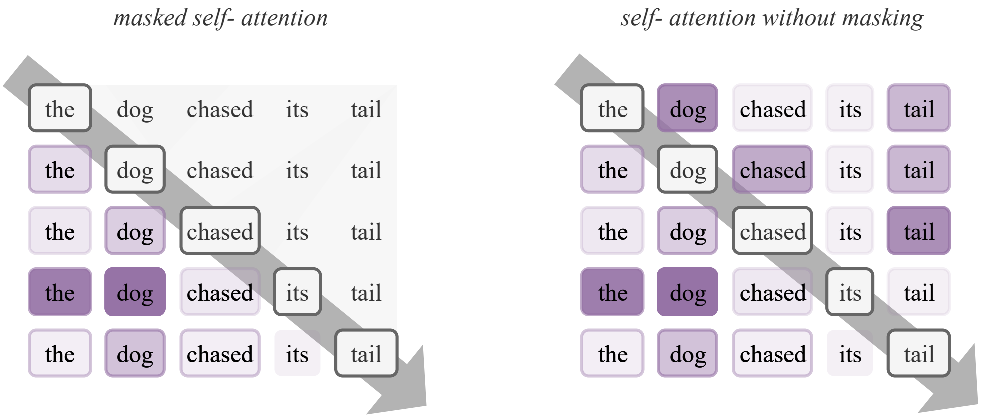


Figure 2.8.: Masked self-attention (left) restricts the information that is available to encode a word to its preceding context (left side in English). In contrast, self-attention without masking can access the full context (right).

parallelized training with batches that contain token sequences with different lengths. The transformer architecture, in contrast, contains no more recurrent model components and fully relies on self-attention to capture long-distance relationships between tokens. Importantly, the transformer follows (like its recurrent network predecessors) an encoder-decoder approach in which an input is first encoded and then, this encoding is used to generate an output sequence step-by-step. A detailed description of the fine-grained components (such as multi-headed attention or positional embeddings) of the transformer is outside the scope of this thesis. We refer to the original paper by Vaswani et al. (2017) for further information.

2.1.2.3. Representing the Output: From Vectors to Outputs

The previous paragraphs describe how a text is mapped to a vector representation and how this vector representation can be transformed into a new representation. Ultimately, this representation has to be mapped to a system output in order to reflect, e.g., a class prediction or a token generation. In the following, we present typical approaches to model outputs for (a) text classification, (b) span extraction, and (c) text generation.

Text Classification. For text classification, the model's input is mapped to a categorical output. In our NLI example, these classes are "entailment", "contradiction", and "neutral". Figure 2.9 depicts a NLI classification example in which the input text is mapped to an "entailment" label. The model's prediction is as a discrete probability distribution over the three

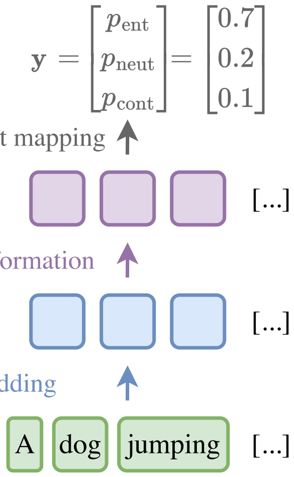

$\mathbf{y} = \begin{bmatrix} p_{\text{ent}} \\ p_{\text{neut}} \\ p_{\text{cont}} \end{bmatrix} = \begin{bmatrix} 0.7 \\ 0.2 \\ 0.1 \end{bmatrix}$
output mapping
[...]
transformation
[...]
embedding
A
dog
jumping
[...]

Figure 2.9.: Example of a binary classification output. The system returns a predicted probability distribution over a set of classes. The example depicts a NLI classification returning a distribution that corresponds to an "entailment" class label prediction with 70% confidence.

classes "positive", "neutral", and "negative" which is represented as a three-dimensional vector

$$\mathbf{y} = \begin{bmatrix} p_{\text{pos}} \\ p_{\text{neut}} \\ p_{\text{neg}} \end{bmatrix} . \tag{2.5}$$

Typically, the model returns a vector of class *logits* which can be related to class probabilities using the softmax function, that maps an input $\mathbf{z} \in \mathbb{R}^n$ to a probability distribution over classes using

$$\mathbf{y}_i = \frac{\exp(\mathbf{z}_i)}{\sum_{j=1}^{n} \exp(\mathbf{z}_j)} . \tag{2.6}$$

The resulting vector $\mathbf{y}$ is normalized and corresponds to a probability distribution over classes where $\mathbf{y}_i$ is the probability estimate of the i-th class.

Span Extraction. As we detailed in Section 2.1.1, extractive QA systems return an answer by selecting a span (i.e., a text snippet) from a given context. Assuming that answers are continuous spans from the next, the output can be described by the combination of (a) the predicted span's start position and (b) the predicted span's end position. The example shown in Figure 2.10 shows a part of the input context ("[...] extending for 900000 km²,") to the question "What is the area of the Kalahari desert?" (the full context sentence can be found in Figure 2.2). The shown model returns two probability distributions. In the example, the value 0.9 in the

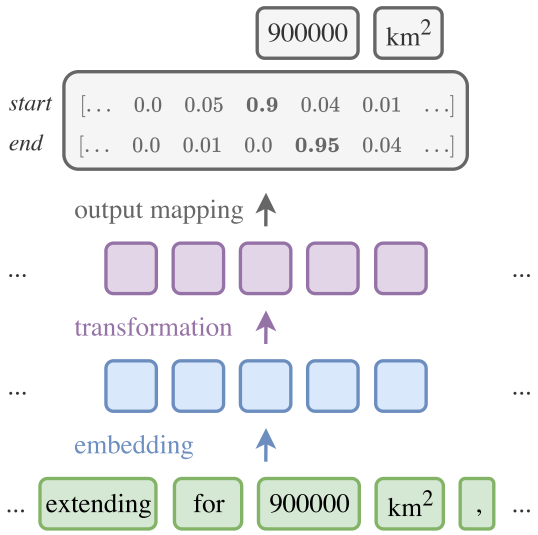


Figure 2.10.: Example of a span extraction output. The system returns two distributions, one regarding the start token position and one regarding the end token position. The shown example refers to the question answering example introduced in Figure 2.2 in which the question asks “What is the area of the Kalahari desert?”. In the example, the system correctly returns “900000 km²” as the most probable span.

predicted start position distribution reflects an estimated 90% probability that the answer starts at the token “90000”. Similarly, the value 0.04 in the end distribution reflects a 4% probability of the answer stopping at the “,” punctuation symbol.

To derive a combination of start and stop positions, a common approach is to first assume the two distributions to be independent and then account for implausible position combinations. Concretely, one first calculates the outer product of the start distribution with the end distribution:

$$\mathbf{p}_{\text{start}} \otimes \mathbf{p}_{\text{end}} = \begin{bmatrix} \mathbf{p}_{\text{start}_1}\mathbf{p}_{\text{end}_1} & \mathbf{p}_{\text{start}_1}\mathbf{p}_{\text{end}_2} & \mathbf{p}_{\text{start}_1}\mathbf{p}_{\text{end}_3} & \cdots & \mathbf{p}_{\text{start}_1}\mathbf{p}_{\text{end}_n} \\ \mathbf{p}_{\text{start}_2}\mathbf{p}_{\text{end}_1} & \mathbf{p}_{\text{start}_2}\mathbf{p}_{\text{end}_2} & \mathbf{p}_{\text{start}_2}\mathbf{p}_{\text{end}_3} & \cdots & \mathbf{p}_{\text{start}_2}\mathbf{p}_{\text{end}_n} \\ \mathbf{p}_{\text{start}_3}\mathbf{p}_{\text{end}_1} & \mathbf{p}_{\text{start}_3}\mathbf{p}_{\text{end}_2} & \mathbf{p}_{\text{start}_3}\mathbf{p}_{\text{end}_3} & \cdots & \mathbf{p}_{\text{start}_3}\mathbf{p}_{\text{end}_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{p}_{\text{start}_n}\mathbf{p}_{\text{end}_1} & \mathbf{p}_{\text{start}_n}\mathbf{p}_{\text{end}_2} & \mathbf{p}_{\text{start}_n}\mathbf{p}_{\text{end}_3} & \cdots & \mathbf{p}_{\text{start}_n}\mathbf{p}_{\text{end}_n} \end{bmatrix} \tag{2.7}$$

Next, combinations for which the end position is in front of the start position are eliminated by setting all $\mathbf{p}_{\text{start}_i}\mathbf{p}_{\text{end}_j}$ with $i < j$ to zero, i.e., taking the element-wise product with an upper diagonal masking matrix resulting in:

$$\begin{bmatrix} \mathbf{p}_{\text{start}_1}\mathbf{p}_{\text{end}_1} & \mathbf{p}_{\text{start}_1}\mathbf{p}_{\text{end}_2} & \mathbf{p}_{\text{start}_1}\mathbf{p}_{\text{end}_3} & \cdots & \mathbf{p}_{\text{start}_1}\mathbf{p}_{\text{end}_n} \\ 0 & \mathbf{p}_{\text{start}_2}\mathbf{p}_{\text{end}_2} & \mathbf{p}_{\text{start}_2}\mathbf{p}_{\text{end}_3} & \cdots & \mathbf{p}_{\text{start}_2}\mathbf{p}_{\text{end}_n} \\ 0 & 0 & \mathbf{p}_{\text{start}_3}\mathbf{p}_{\text{end}_3} & \cdots & \mathbf{p}_{\text{start}_3}\mathbf{p}_{\text{end}_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{p}_{\text{start}_n}\mathbf{p}_{\text{end}_n} \end{bmatrix} \tag{2.8}$$

Additionally, one can constrain the maximum span length to exclude implausibly long answer spans by applying an element-wise product with a band matrix to the matrix in Section 2.1.2.3. For our example, we restrict the answers to a maximum length of two tokens (for better visualization) and multiply our matrix with a bidiagonal mask matrix:

$$\begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 1 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \odot \begin{bmatrix} \mathbf{p}_{\text{start}_1}\mathbf{p}_{\text{end}_1} & \mathbf{p}_{\text{start}_1}\mathbf{p}_{\text{end}_2} & \mathbf{p}_{\text{start}_1}\mathbf{p}_{\text{end}_3} & \cdots & \mathbf{p}_{\text{start}_1}\mathbf{p}_{\text{end}_n} \\ 0 & \mathbf{p}_{\text{start}_2}\mathbf{p}_{\text{end}_2} & \mathbf{p}_{\text{start}_2}\mathbf{p}_{\text{end}_3} & \cdots & \mathbf{p}_{\text{start}_2}\mathbf{p}_{\text{end}_n} \\ 0 & 0 & \mathbf{p}_{\text{start}_3}\mathbf{p}_{\text{end}_3} & \cdots & \mathbf{p}_{\text{start}_3}\mathbf{p}_{\text{end}_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{p}_{\text{start}_n}\mathbf{p}_{\text{end}_n} \end{bmatrix} \tag{2.9}$$

The resulting matrix thus reads:

$$\begin{bmatrix} \mathbf{p}_{\text{start}_1}\mathbf{p}_{\text{end}_1} & \mathbf{p}_{\text{start}_1}\mathbf{p}_{\text{end}_2} & 0 & \cdots & 0 \\ 0 & \mathbf{p}_{\text{start}_2}\mathbf{p}_{\text{end}_2} & \mathbf{p}_{\text{start}_2}\mathbf{p}_{\text{end}_3} & \cdots & 0 \\ 0 & 0 & \mathbf{p}_{\text{start}_3}\mathbf{p}_{\text{end}_3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{p}_{\text{start}_n}\mathbf{p}_{\text{end}_n} \end{bmatrix} \tag{2.10}$$

Note that this masking step includes the previous masking step, however, we introduce both to illustrate the underlying motivations. In our example from Figure 2.10, the corresponding matrix of the shown part of the context would be:

$$\begin{bmatrix} 0 \\ 0.05 \\ 0.9 \\ 0.04 \\ 0.01 \end{bmatrix} \otimes \begin{bmatrix} 0 \\ 0.01 \\ 0.0 \\ 0.95 \\ 0.04 \end{bmatrix} \odot \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \odot \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0.0005 & 0 & 0 & 0 \\ 0 & 0 & 0 & \boxed{0.855} & 0 \\ 0 & 0 & 0 & 0.038 & 0.0016 \\ 0 & 0 & 0 & 0 & 0.0004 \end{bmatrix}$$

The value 0.855 is the maximum and its position reflects the correct answer “900000 km²”.

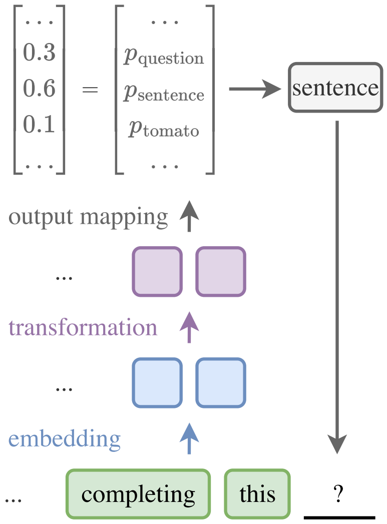


Figure 2.11.: Example of text generation. The system is asked to complete the text "[...] completing this" and returns a probability distribution over its vocabulary. In the example, the vocabulary element "sentence" receives the highest probability and the system thus completes the text with "sentence".

Text Generation. For text generation systems, we consider the example shown in Figure 2.11. The system receives the text "[...] completing this" and should predict the next token. This is usually modeled as a probability distribution over the system's vocabulary. In our example, the token with the highest probability is "sentence". Thus, "sentence" would be chosen to complete the sentence and the subsequent text reads "[...] completing this sentence". This left-to-right completion procedure can be iterated to predict one token after the other. This procedure, also known as *autoregressive* language modeling, is typically approached with transformer architectures (see Section 2.1.2.2).

2.1.2.4. Training

So far, we specified how an input can be represented in vector space, how different NLP tasks can be modeled in terms of a model's output, and how the mapping between input and output can be modeled in terms of various components (e.g., matrix-vector products). In the following, we provide a high-level overview of the process of optimizing model parameters in order to approximate a desired quality measure, i.e., model *training*.

Loss Functions. In order to decide how to determine the parameters of a model, one has to formalize how model "badness" — and thereby model "goodness" is measured. When the model output corresponds to one (or multiple) classification decisions, one typically seeks to minimize the cross-entropy (CE) loss, which quantifies the distance between the predicted probability distribution $\hat{\mathbf{y}}$ and the ground truth distribution $\mathbf{y}$. When training on annotated data, the ground truth distribution boils down to a one-hot vector, i.e., a vector that has a one at the index of the ground truth class and zeros at all other indices. In this case, the CE loss for distributions with M classes then reads

$$\mathrm{CE}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{i=1}^{M} \mathbf{y}_i \log(\hat{\mathbf{y}}_i). \tag{2.11}$$

In the case of a binary classification problem, i.e. $M = 2$, Equation (2.11) can be simplified to

$$\begin{aligned} \mathrm{CE}(\mathbf{y}, \hat{\mathbf{y}}) &= -\sum_{i=1}^{2} \mathbf{y}_i \log(\hat{\mathbf{y}}_i) && (2.12) \\ &= -\left(\mathbf{y}_1 \log(\hat{\mathbf{y}}_1) + \mathbf{y}_2 \log(\hat{\mathbf{y}}_2)\right) && (2.13) \\ &= -\left(\mathbf{y}_1 \log(\hat{\mathbf{y}}_1) + (1 - \mathbf{y}_1) \log(1 - \hat{\mathbf{y}}_1)\right). && (2.14) \end{aligned}$$

Note that minimizing CE is equivalent to minimizing the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951). We use the CE loss across all systems trained within this thesis. When the model does not perform classification but regression, i.e., the output is not a probability, a typical loss function is the mean squared error (MSE) loss. With a scalar prediction $\hat{\mathbf{y}}$ and a scalar target value $\mathbf{y}$, the MSE reads

$$\mathrm{MSE}(\mathbf{y}, \hat{\mathbf{y}}) = (\mathbf{y} - \hat{\mathbf{y}})^2. \tag{2.15}$$

We make use of the MSE loss within our thought flow network method presented in Section 3.2.

Optimizers. Once a model architecture as well as an optimizer, are determined, the model parameters have to be trained towards decreasing loss, i.e., better model fit. While closed-form solutions exist for, e.g., linear regression (Bishop, 2006), there is no closed-form solution for typical neural, and, in particular, transformer models, which thus have to be optimized iteratively using (typically) gradient descent. While basic stochastic gradient descent (SGD) has been extended to a large variety of optimizers, we follow the typical choices of ADAM (Kingma and Ba, 2015) with decoupled weight decay (Loshchilov and Hutter, 2019) for LLMs and plain SGD for vision models. Note that learning rate scheduling (in particular warm-up (Goyal et al., 2017)), batch sizes, and weight initialization can cause fundamental increases/decreases in model performance.

2.2. Explainability

So far, we discussed systems that map an input to an output, i.e., how a system determines *what* to output. For explainability, we ask *why* it returned that output. In the following, we first define explainability and dimensions along which explainability methods can be categorized (Section 2.2.1). Second, we present popular types of explanations in NLP (Section 2.2.2).

2.2.1. Definition and Taxonomy of Explainability

2.2.1.1. Explainability, Explanation, Interpretability, and Justification

The terminology in explainability research is largely inconsistent and notions of what explainability means and, e.g., whether it is different from interpretability differ (i.a., Rudin, 2019; Miller, 2019). In this thesis, we follow the terminology of Miller (2019), which in turn builds upon the definition of Biran and Cotton (2017). Concretely, Miller (2019) equates *interpretability* and *explainability* and defines them as "*the degree to which an observer can understand the cause of a decision*". *Explanation* is considered one way to achieve this understanding. In contrast to an explanation, a *justification's* primary goal is not to increase understanding of the model decision process, but, instead, to provide evidence why its decision should be correct. Many of the methods described in this thesis could foremost be categorized as justifications. For consistency with related work and to ease readability, we will, however, refer to explanations throughout the remainder of this thesis.

2.2.1.2. Explanandum, Explanans, Explainer, Explainee, and Explanation

Going back to the deductive nomological model of scientific explanation by Hempel and Oppenheim (1948), an explanandum can be distinguished from an explanans. The *explanandum* is the phenomenon or event to be explained, the *explanans*, in turn, explains the explanandum (Overton, 2012). *Explainer* and *explainee* refer to two agents of which the explainer explains (its own) decisions to the explainee. Within this thesis, the explainer is an automated system and the explainee is a user. Miller (2019) argues for a three-fold notion of *explanation* extending the dual process-product distinction of Lombrozo (2006): (i) the cognitive process to determine an explanans, (ii) the product of this cognitive process, and (iii) the social process of communicating information between the explainer and the explainee. Within this thesis, we predominantly refer to *explanation* in terms of its product notion. Figure 2.12 shows an overview of the discussed terminology.

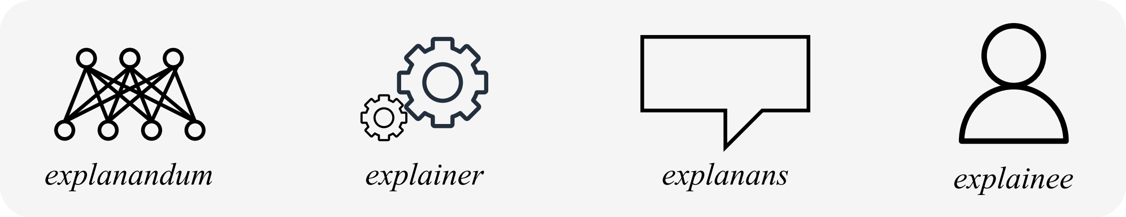


Figure 2.12.: Overview over explanation-related terminology used in this thesis. The system's behavior is what should be explained (*explanandum*), an automatic method explains the system's behavior (*explainer*) using a NLP explanation modality, such as saliency explanations (*explanans*), which in turn is received by a user (*explainee*).

2.2.1.3. Taxonomies of Explanation Methods

Explanations in the context of explainable AI can be categorized along multiple dimensions. Speith (2022) reviews several taxonomies of explainable AI methods and proposes a unified taxonomy. The unified taxonomy distinguishes explanations along five high-level dimensions (and an additional "other" dimension) which are shown in Figure 2.13. We refer to Speith (2022) for a detailed discussion of each dimension and provide an overview in the following.

Scope. An explanation can either explain a single system decision (e.g., why a system assigned an entailment label to a particular pair of sentences) or explain a system's overall decision behavior (e.g., why a system fails to detect cows when there is snow in the background of an image). The former situation is often referred to as a *local* explanation, while the latter is referred to as *global* (Guidotti et al., 2019; Sokol and Flach, 2020; Vilone and Longo, 2021; Speith, 2022). All explanation methods discussed in this thesis correspond to local explanations.

Stage. Various taxonomies distinguish between *post hoc* and *ante hoc* explanations (Guidotti et al., 2019; Sokol and Flach, 2020; Vilone and Longo, 2021; Speith, 2022). Post hoc methods are methods that explain model behavior after the decision is made (e.g., by tracing back what inputs affected the model to take that decision). Speith (2022) further distinguish model-agnostic and model-specific methods as a subordinate dimension of post hoc methods (not depicted in Figure 2.13). Ante hoc explanations refer to inherently-interpretable models, i.e., models that are expected to be explainable by design. Typical model architectures that are argued to fall into that category include linear regression, decision trees, or k-nearest neighbors classification. However, Speith (2022) note that presumable interpretable architecture can become uninterpretable with a high number of parameters.

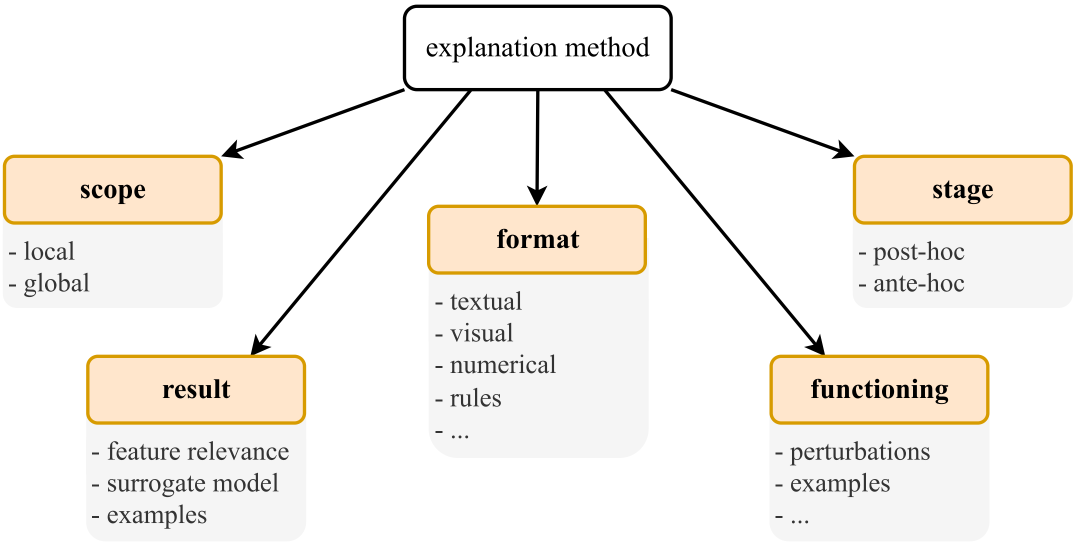


Figure 2.13.: Taxonomy of explainable AI methods replicated from Speith (2022) who reviewed and unified existing taxonomies.

Format. Explanation methods can further be categorized with respect to their output format (Vilone and Longo, 2021). While visual explanations, such as saliency maps or numeric importance scores might be among the most well-known explanation output formats, explanations can also be given in the form of, e.g., extracted rules. In this thesis, we focus on textual explanations as well as saliency explanations over text and additionally explore a novel method that explains a system's final decision by means of a sequence of initial and intermediate decisions (see Section 3.2).

Result. Orthogonal to the output format, explanation methods can be distinguished along their result (McDermid et al., 2021; Speith, 2022). Speith (2022) distinguish feature relevance, surrogate models and examples. While two methods might both yield the same result (e.g., feature relevance), their format can be different (e.g., one method uses visual representations of the feature importance while the other reports numeric importance values).

Functioning. Functioning refers to the approach an explanation method takes to derive an explanation (Arrieta et al., 2020; Speith, 2022). For example, perturbation methods can modify the input and derive an explanation from the observed changes in a model's output while example-based methods can operate, e.g., on example instances from the training set.

2.2.2. Types of Explanations in NLP

While Section 2.2.1.3 discussed general properties of explanation methods, the following introduces explanation methods that are used to explain text or explain with text. Concretely, we detail the three classes of NLP explanations that are used in this thesis: rationals/saliency explanations, supporting fact explanations, and free-text explanations. Table 2.2 provides an overview of these explanation types including examples as well as typical automatic evaluation scores which we will discuss in detail in Chapter 4.

2.2.2.1. Saliency and Rational Explanations

Saliency (or heatmap) explanations indicate how strongly a part of the input (i.e., a token) influences a model's output. Saliency explanations are typically communicated via heatmaps and make use of an importance measure, such as Integrated Gradients (Sundararajan et al., 2017) or attention weights (Wiegreffe and Pinter, 2019) which we detail in the following. While general saliency explanations communicate graded importance which allows to interpret them in a relative mode (e.g., "movie" being more important than "this" but less important than "like" in the example shown in the first row of Table 2.2), rational explanations communicate binary relevant/irrelevant information (e.g., "this" as well as "movie" are irrelevant in the rational explanation shown in Table 2.2).

So far, we discussed the saliency/rational explanations output format. Next, we discuss their result as well as their functioning. For this, we have to distinguish between the underlying attribution scores and discuss attention scores, Shapley values, and SHAP as well as Integrated Gradients in the following.

Attention as Explanation. As we described in Section 2.1.2.2, the attention mechanism and its variants are fundamental components of today's NLP systems. Attention scores can be used as explanations by averaging each word's attention score over the last layer's attention heads and visualizing the respective scalar using color-coding as shown in the first row of Table 2.2. Its (apparent) analogy to human attention makes attention scores, at first sight, promising measures of how important a token was to a model. Note that there is an ongoing debate on whether attention can serve as an explanation (i.a., Wiegreffe and Pinter, 2019) or whether it can not (i.a., Jain and Wallace, 2019). We refer to Bastings and Filippova (2020) for a detailed argumentation of why saliency scores should be chosen over attention scores and to Pruthi et al. (2020) for a practical demonstration of how attention weights can be used to deceive explainees. Note that we will use "saliency explanations" to refer to any explanation method

Type	Description	Example	Proxy Scores
Rationals, saliency maps	Input tokens are highlighted to reflect what was most important to the model (e.g., based on attention or saliency scores, such as Integrated Gradients).	*Input*: I like this movie. *Prediction*: positive sentiment *Rational explanation*: I like this movie. *Saliency explanation*: I like this movie .	**Overlap to human rational annotations** (e.g., via F1), **Removal analysis** (e.g., quantifying the drop in performance when highlighted input parts are removed) (i.a., Atanasova et al., 2020), **Student model accuracy gains** when trained on the explanations (Pruthi et al., 2022)
Supporting facts	A set of facts (i.e., sentences) extracted from a given context is provided as evidence for the prediction.	*Question*: What is the area of the desert that Ghanzi is in the middle of? *Answer*: 900000 km² Fact 1: Ghanzi is a town in the middle of the Kalahari Desert the western part of the Republic of Botswana in southern Africa. Fact 2: The Kalahari Desert is [...] extending for 900000 km².	**Overlap to human annotations** of supporting facts (e.g., via F1) (Yang et al., 2018) **Removal and consistency analysis** in Section 4.1.2
Free text	Generated textual explanation that supports the prediction.	*Premise*: A man in an orange vest leans over a pickup truck. *Hypothesis*: A man is touching a truck. *Predicted label*: entailment *Explanation*: Man leans over a pickup truck implies that he is touching it. (from Camburu et al. (2018))	**Overlap to human-written references** (e.g., via BLEU or BLEURT) (Camburu et al., 2018; Kayser et al., 2021)) and Section 3.1.2

Table 2.2.: Three examples of different explanation types in NLP along with proxy scores that are used to quantify their quality (see Chapter 4 for a detailed discussion of explanation quality evaluation).

that communicates saliencies (i.e., referring to the format of the explanations) regardless of how the underlying scores were obtained for the remainder of this thesis. Attention scores can only be obtained from models using an attention mechanism and require access to the model's internal state.

Shapley Values and SHAP. Two important saliency methods are Shapley values (Shapley, 1953) and the derived SHAP values (Lundberg and Lee, 2017). Shapley values originate from game theory and answer the question of how much — for a set of players (a coalition) — each player contributed individually to an outcome. In the context of explainable NLP, the players are the input tokens and the outcome is, e.g., a class probability. Following the notation of Mosca et al. (2022b), we denote the full set of tokens with $F = \{1, 2, ..., p\}$ where each number represents the token at the respective position, and use S to refer to a subset of tokens. The value contribution of a coalition is denoted with val(S). The marginal contribution of a token i to a coalition is defined as

$$\Delta_{\text{val}}(i, S) = \text{val}(S \cup \{i\}) - \text{val}(S). \tag{2.16}$$

The marginal contribution of a token is used to define its overall contribution by summing and normalizing $\Delta_{\text{val}}(i, S)$ for all coalitions $S \subseteq F \setminus \{i\}$

$$\phi_{\text{val}}(i) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \Delta_{\text{val}}(i, S) \tag{2.17}$$

which defines the Shapley value $\phi_{\text{val}}(i)$. Based on Shapley values, Lundberg and Lee (2017) proposed SHAP values along with various approximation methods that make their computation feasible. SHAP values have a unique solution and fulfill several desirable properties. Notably, Lundberg and Lee (2017) also demonstrate SHAP values to have a stronger agreement to human explanations compared to LIME (Ribeiro et al., 2016) and DeepLift (Shrikumar et al., 2016, 2017). Shapley values and SHAP values both are model-agnostic explainability methods.

Integrated Gradients. Another important saliency method are Integrated Gradients (Sundararajan et al., 2017). Following the authors' original notation, F denotes the model function and is bound to $[0, 1]$ and $\mathbf{x}$ denotes the input vector and $\mathbf{x}'$ refers to a baseline input (which usually is chosen as $\mathbf{0}$ for embedded text inputs). Integrated gradients are now defined as the

Integrated Gradients along the path connecting $\mathbf{x}'$ and $\mathbf{x}$:

$$\mathrm{IG}_i(\mathbf{x}) = (\mathbf{x}_i - \mathbf{x}'_i) \cdot \int_{\alpha=0}^{1} \frac{\partial F(\mathbf{x}' + \alpha \cdot (\mathbf{x} - \mathbf{x}'))}{\partial \mathbf{x}_i} d\alpha. \tag{2.18}$$

Integrated gradients are easy to implement and satisfy desirable properties. We refer to Sundararajan et al. (2017) for an in-depth discussion. Integrated gradients are a model-specific explanation method in the sense that they require a differentiable model to calculate gradients and require access to the model's internal state.

2.2.2.2. Supporting Fact Explanations

Besides saliency explanations, another popular class of explanations in NLP are supporting fact explanations. In a supporting fact explanation, a set of facts (i.e., sentences) is provided on top of the predicted model output. It thus resembles search engine interfaces which — in answer to a search query — return links to websites along with text snippets with highlighted search query terms. The middle row of Table 2.2 depicts such a supporting fact explanation taken from the HotpotQA explainable QA dataset (Yang et al., 2018) which we will introduce in detail in Chapter 4.

Referring back to the distinction between justification and explanation in Section 2.2.1, supporting facts can be regarded to be justifications that can provide explanatory value. Concretely, a supporting fact explanation that is consistent with the predicted output, can indicate successful processing of the input while a supporting fact explanation that is inconsistent with the predicted output can signal model failure. We discuss this consistency property and approaches to measuring it in detail in Section 4.1.

2.2.2.3. Free-text Explanations

While supporting text explanations stem from a textual context, such as a Wikipedia article or a website, supporting evidence for the correctness of a model prediction can also be freely generated. In that sense, free-text explanations can be regarded to be — tying back to the distinction between abstractive and extractive QA in Section 2.1.1 — the abstractive equivalent to extractive supporting fact explanations. As for supporting fact explanations, free-text explanations can be considered justifications with a potential for explanatory value. The bottom row of Table 2.2 shows an example of a free-text explanation from the e-SNLI explainable NLI dataset (Camburu et al., 2018). We introduce the dataset and various model architectures to approach its modeling in Section 3.1.

2.3. Designing User Studies

The previous sections provided this thesis' background in NLP and explainability. This section complements the background chapter by presenting the necessary background on human evaluation, and, in particular the design of user studies which we will build upon in the following chapters. We introduce essential concepts of variables (Section 2.3.2), metrics (Section 2.3.3), and experimental designs (Section 2.3.4). For in-depth introductions to the respective topics, we refer to textbooks by, i.a., Field and Hole (2002) and MacKenzie (2013).

2.3.1. The Need for Human Evaluation in NLP

Over the past years, the NLP community (beyond explainable NLP) has increasingly expressed the need for and the importance of human evaluation to complement automatic evaluation (Belz and Reiter, 2006). Tasks, such as machine translation (Graham et al., 2013), explanation generation (Nguyen, 2018; Narang et al., 2020; Clinciu et al., 2021), text-to-speech generation (Cardoso et al., 2015; Clark et al., 2019), question answering (Chen et al., 2019), and automatic summarization (Owczarzak et al., 2012; Paulus et al., 2018) still rely heavily on automatic measures like BLEU Papineni et al. (2002) or F_1-scores. However, these scores have been shown to correlate only loosely with human perception of such systems (Callison-Burch et al., 2006; Liu et al., 2016; Mathur et al., 2020; Iskender et al., 2020; Clinciu et al., 2021) and do not necessarily reflect how a system might perform with respect to extrinsic evaluations, such as downstream tasks (Gaudio et al., 2016).

For example, BLEU scores are commonly used to quantify the similarity of a generated sentence to a ground truth sentence, e.g., in machine translation. In this thesis, we use (and question the use of) BLEU to quantify the quality of generated textual explanations in Section 3.1. BLEU scores rely on the n-gram overlap between the generated and the reference text. However, this approach has two important shortcomings: (i) relying on "ground truth" reference texts ignores the breadth of possible correct translations (in the context of translation), and (ii) assuming that similarity of meaning can be inferred from n-gram overlap discounts, e.g., that different words in the sentence contribute to its meaning differently. Consider an explainable NLI instance for which the premise "the boy went for a walk with his dog yesterday" contradicts the hypothesis "the boy did not do anything yesterday". The contradiction label is further explained with a reference explanation "walking one's dog means doing something". Consider two candidate explanations: (a) "walking one's pet means doing something' and "eating one's dog means doing something". Both receive the same BLEU-2 scores, however, from a human perspective, sentence (a) reflects the reference explanation

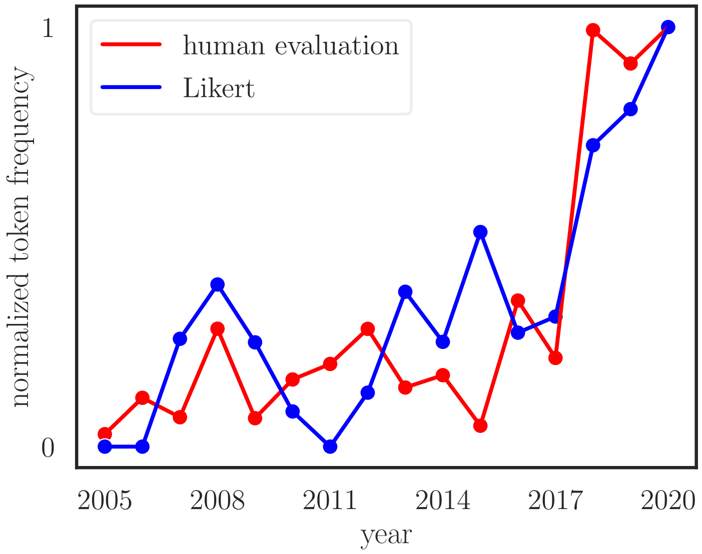


Figure 2.14.: Normalized frequencies of “human evaluation” and “Likert” (as in the Likert scale questionnaire type) in the ACL anthology from 2005 to 2020 indicating the growing attention on human evaluation.

much better.[1] Similarly, automatic evaluation measures used by other NLP tasks face the same problem (Callison-Burch et al., 2006; Liu et al., 2016; Mathur et al., 2020; Iskender et al., 2020; Clinciu et al., 2021). Therefore, human evaluation has begun to gain more and more attention in the NLP community (especially in the context of natural language generation tasks, including machine translation (Belz and Reiter, 2006; Novikova et al., 2018; van der Lee et al., 2019)). This trend is indicated in Figure 2.14.

2.3.2. Variables

Before discussing experimental designs and evaluation methods, it is important to distinguish, which variables are intentionally being changed, which variables are being measured, and which variables one cannot control. In order to support a repeatable experiment that reliably answers a research question, one also has to choose an *operationalization*, i.e., a clear, measurable definition for each of these variables.

2.3.2.1. Independent

The independent variable(s) are those which are controlled within the study, also called *factors* (MacKenzie, 2013). Experimental designs involving a single independent variable are referred

[1] There exist different versions of BLEU, e.g., BLEU-2 refers to the score that considers unigrams and bigrams.

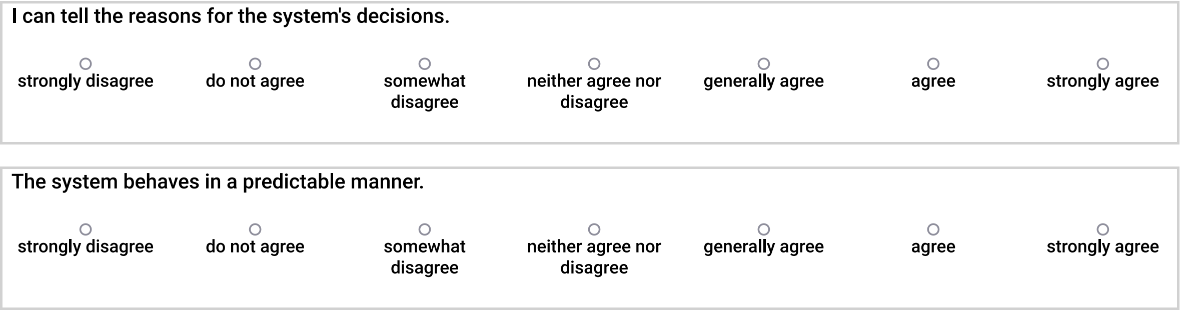
to as *unifactorial* and experiments involving multiple independent variables are referred to as *multifactorial*. The values a variable can take are called *levels*. For example, if the variable is “explanation method”, levels might be “no explanation”, “SHAP value heatmap”, and “new explanation method”. Here it is important to be deliberate about the changes between the two systems so it is clear that any changes observed are a result of the independent variable in question, e.g., the explanation method. For our example of explanation method comparison, it would be important that all explanation methods (including no method) are evaluated on the same dataset. Otherwise, one might not be able to attribute an observed difference in the dependent variables to a difference in the factor of interest (explanation method), but only be able to conclude this difference as the result of the combined effects (explanation method and dataset) without being able to disentangle the effects of each variable.

2.3.2.2. Dependent

The dependent or *response* variable(s) are those which are measured and whose changes are a result of the independent variable (MacKenzie, 2013). For this, it is important to consider not just the general concept (*construct*), but also what concrete measurement to take. This process is known as *operationalization*. For example, in order to evaluate the hypothesis that “the new explanation method will enable users to better detect false predictions”, it is necessary to first operationalize the construct ”better" into a dependent variable, which can be concretely measured. In this case, one could decide, for example, that “better” refers to a lower user acceptance rate of incorrect model predictions.

2.3.2.3. Confounding

A confounding variable or *confounder* is a variable that affects the dependent variable, but cannot be controlled for, e.g., age, gender, or education of the participants. Education, for example, might affect the users’ acceptance rate (e.g., when users are familiar with the topic of the system predictions), but one cannot deliberately change the education level of participants. Potential confounding variables should either be accounted for in the experiment design or in the statistical evaluation of the collected responses. One option is to include confounding variables as random effects, as discussed in Section 2.4. Therefore, it is important to consider which variables might be confounding variables and to measure these when conducting a study.



I can tell the reasons for the system's decisions.

○ strongly disagree ○ do not agree ○ somewhat disagree ○ neither agree nor disagree ○ generally agree ○ agree ○ strongly agree

The system behaves in a predictable manner.

○ strongly disagree ○ do not agree ○ somewhat disagree ○ neither agree nor disagree ○ generally agree ○ agree ○ strongly agree

Figure 2.15.: A subset of Likert items from our novel perceived system predictability scale we present in Chapter 5.

2.3.3. Metrics

Depending on the choice of dependent variable(s), there are different means to concretely quantify user responses. We focus on Likert scales as a measure of self-reported user responses (Section 2.3.3.1), but depending on the research question at hand, other quantitative (Section 2.3.3.2) or qualitative measurements (Section 2.3.3.3) may be equally important. For quantitative measurements, it is crucial to be aware of the measurement level of the collected responses (Section 2.3.3.4) as it will directly affect which statistical tests can be applied to the collected data.

2.3.3.1. Likert Scales

While it is clear how to collect objective measures, e.g., the length of a dialog, it is less straightforward how to collect scores of trust, cognitive load, or even creepiness. For such subjective metrics, one usually obtains scores via a validated scale (Körber, 2018; Hart and Staveland, 1988; Langer and König, 2018), e.g., in the form of a questionnaire. In the following chapters, we will use various Likert scales to score, i.a., usability, grammaticality, and mental demand. In Section 5.2, we develop our own scale to measure perceived system predictability. We deepen our discussion of scale development and validation in Section 5.2 and Section 5.3. Figure 2.15 shows a subset of the 7-point Likert items of our novel system predictability scale.

2.3.3.2. Other Useful Metrics for NLP

As an alternative to Likert scales, continuous rating scales like the visual analog scale (VAS) can be used to measure a construct. Santhanam and Shaikh (2019) found that continuous rating scales can yield more consistent results than Likert scales for dialog system evaluation. In tasks like generating text or speech, direct comparisons or ranked order comparisons (ranked output from multiple systems best to worst) can be a good option (Vilar et al., 2007; Bojar et al., 2016). Another option for tasks involving text generation is error classification, which involves users annotating text output from a set of predefined error labels (Secară, 2005; Howcroft et al., 2020). Other measurements of interest to NLP research include completion time, and bio-signals, such as gaze, EEG, ECG, and electrodermal activity. Bio-signals may provide insight into, e.g., emotional state (Kim and André, 2008), engagement (Renshaw et al., 2009), stress (McDuff et al., 2016), and user uncertainty (Greis et al., 2017a).

2.3.3.3. Qualitative Analysis

In addition to quantitative analysis, qualitative analysis can provide valuable insights into users' perspectives by allowing them more freedom of expression than metrics like a Likert scale. For example, in order to understand a user's perception of a chatbot, free response questions can be used alongside, e.g., Likert scales, allowing the user to express which aspects of the chatbot had the largest impact on them. These responses can then be analyzed with techniques, such as content/theme analysis (Hsieh and Shannon, 2005; Braun and Clarke, 2006), where user responses are "coded" using a set of labels generated from the collected data, to identify similar themes across responses. These codes can then be quantified and patterns can be analyzed about how often certain codes/themes, appeared and under which conditions. For example, one code might be "smart", then all user responses that indicated that they found the chatbot to be intelligent could be marked with this label. Researchers could then, for example, analyze that 76% of users found the chatbot to be intelligent, and that this correlated highly with users who reached their goal. We revisit qualitative evaluation for explanation quality evaluation in Sections 4.3 and 4.4.

2.3.3.4. Level of Measurement

It is important to consider the scale on which a variable is measured in order to choose a correct statistical test (Section 2.4) and measures of central tendency (i.e., mode, median, and mean). Typically, four types of measurement scales are considered: *nominal*, *ordinal*, *interval*, and *ratio* (Stevens, 1946; Borgatta and Bohrnstedt, 1980; MacKenzie, 2013).

Nominal. On a nominal (categorical) scale, items are simply named, with no concept of order or distance between them. An example is emotions perceived in a generated voice ("happiness", "sadness", "fear", etc.). If the scale only contains two choices, it is called *dichotomous*. The only measure of central tendency applicable to such data is the mode.

Ordinal. An ordinal scale adds order to the elements. However, the distance between them cannot be assumed to be equal. An example is measuring intelligibility using the values "very low", "low", "medium", "high", and "very high". In addition to the mode, ordinal data also enables the derivation of a median.

Interval. On an interval scale, the elements are ordered with an equal distance between them, allowing one to additionally take the mean. Scores obtained from multi-item Likert scales (as shown in the perceived system predictability scale in Figure 2.15) are frequently considered interval data. There has been a long debate between *ordinalists* who claim that Likert scales should be treated as ordinal data and non-parametric statistics have to be used, and *intervalists* who argue for an interval interpretation and thus support parametric approaches (Jamieson, 2004; Carifio and Perla, 2008; De Winter and Dodou, 2010). For a deeper discussion as well as practical recommendations, we refer to Harpe (2015).

Ratio. A ratio measurement adds the property of a true zero point making ratios of interval measurements sensible. An example are interaction times with an interactive explanation generation system or the number of dialog turns for a chat bot.

2.3.4. Experimental Designs

When the independent and dependent variables are chosen and operationalized, the question of how to assign participants to *conditions*, i.e., to levels of the independent variable(s), has to be addressed. The choice of the assignment determines applicable statistical tests and can mitigate confounding effects. To illustrate experiment design choices, we will use the example of investigating the perceived naturalness of a text-to-speech system with the independent variable "system", the levels "old" and "new", and the confounding variable "native speaker", i.e., that some participants are native speakers while others are not.

2.3.4.1. Within-Subject

In this study design, also called a *repeated-measures* design, participants are exposed to all study conditions and can thus make comparisons between them (Charness et al., 2012; MacKenzie, 2013). With a fixed number of participants, this allows to collect more samples than a between-subjects design. However, a within-subject design cannot be scaled to an arbitrary number of conditions both because users are often unwilling to participate in longer studies and because they will be affected by fatigue after too many conditions. Additionally, repeated measures may cause participant responses for later conditions to be affected by their responses to earlier ones due to *carry-over effects* and learning. One way to account for carry-over effects is to control the order of conditions the participants are exposed to. Typical approaches are *randomization* (i.e. participants are shown conditions in random order), *blocking* (i.e., participants are grouped into blocks regarding a participant characteristic, such as age), and *Latin square* designs. For details, we refer to Dean et al. (1999). Within-subject designs require a statistical comparison of differences per subject which is accounted for using *paired* tests. In our example, we could use a within-subject approach and mitigate carry-over effects by sampling all possible four combinations[2] equally often. We could account for the possible confounding effect of being a native speaker by balancing the number of native/non-native speakers per condition.

2.3.4.2. Between-Subject

In this design, each participant is only exposed to one condition (Charness et al., 2012; MacKenzie, 2013). While collecting a fixed number of samples requires a higher number of participants than a within-subject design, a between-subject design can easily be scaled to an arbitrarily high number of conditions, assuming the research budget supports this. Participant responses collected with a between-subject design must use unpaired tests as there are no paired responses, but rather two (or more) independently-sampled groups. In our example, it could be preferable to use a between-subject approach if the interaction of the users with the system takes a long time and, thus, users could become fatigued when being exposed to both conditions (i.e., old and new system).

2.4. Statistical Evaluation

The previous section discussed user study design. Although van der Lee et al. (2019) found that, in their review of INLG and ACL papers, from the papers that conduct human evaluation, only 33% report statistical analyses, statistical evaluation is the epistemic backbone of empirical research. In this section, we thus address how the results of a user study can be analyzed with the appropriate statistical tools. Concretely, we provide an overview of how to choose an appropriate sample size, select an applicable statistical test and decide whether a post hoc test and a multiplicity adjustment need to be used.

[2] (i) native speaker: “old” first → “new” second, (ii) native speaker: “new” → “old”, (iii) not native speaker: “old” → “new”, (iv) not native speaker: “new” → “old”.

2.4.1. Sample Size, Effect Size, Significance, and Power

Before starting a user study, an important step is to consider what sample size will be necessary to make meaningful claims about the results. If, e.g., too few participants are chosen, it will reduce the *statistical power* of the study, and thereby the probability of recognizing a statistically significant difference between experimental groups if one occurs. In short, statistical power is important to consider because it represents the likelihood of not reporting a false negative. Therefore designing an experiment with enough power is critical to ensure that time, energy, and money are not wasted conducting a study only to report a false negative result because there were not enough participants. A power level of 0.80 or higher is generally recommended (Bausell and Li, 2002) as it represents that if an experimental design is carried out correctly, 80% of the time, a significant difference will be detected by the chosen statistical test if such a difference exists.

To ensure enough statistical power in an experiment, researchers can conduct a power analysis before starting their experiment to hypothesize what power they can expect given an estimated *effect size*, a number of participants (N), and a desired significance level. In the following, each of these factors is discussed in more detail and an example is provided to show how one can perform such an analysis.

2.4.1.1. Effect Size

The effect size refers to the size or magnitude of an effect (difference between experimental groups) which would be expected to be observed in a population. In general, there are three different ways to calculate effect size: (i) as a standardized result (e.g., standard deviation units from the mean) which allows for interpretation across applications, (ii) using the original units (e.g., difference of means) which may be useful for domain-specific interpretation of results, or (iii) as a unit-free result (e.g., a correlation coefficient) (Sullivan and Feinn, 2012).

For NLP system comparisons, the independent variable is typically categorical and one of the most common methods for calculating standardized unit effect sizes is Cohen's d. Cohen's d measures the difference between the mean from two Gaussian-distributed variables in standard deviation units. It can be calculated by taking the difference between the means of two groups and dividing this by the pooled standard deviation of both samples.

While estimating effect size before starting the actual experiment can be difficult, previous research in the field or the results from a pilot study can provide a good starting point. However if there is no prior information available on the expected effect size, the values 0.2, 0.5, and 0.8 are commonly used as Cohen's d values for small, medium, or large expected effect sizes

(Cohen, 1988). In a meta-study of 302 social and behavioral meta-analyses, Lipsey and Wilson (1993), found the average effect size to be exactly 0.5. As an important note, the smaller the effect size is, the more participants will be required to achieve the same statistical power.

2.4.1.2. Sample Size

The general goal of a power analysis is to identify the minimum sample size needed to achieve a desired level of power (normally 0.8). To this end, increasing the sample size will always increase the power of an experiment. In some cases, however, this may not be feasible. In these cases, it is advisable to try to reduce the number of experimental groups (levels of the independent variable) to as few as is scientifically defensible. The fewer groups there are, the higher the number of participants per group. Alternatively, a within-subject design, if applicable, can also greatly increase the statistical power of a study (Cohen, 1988).

2.4.1.3. Significance Level

Finally, it is important to consider what statistical test will be run on the data and what significance level α is appropriate for the study. Often, an alpha level of 0.05 is chosen which represents that 95% of the time if a statistically significant difference is observed, it is not due to random chance (Kennedy-Shaffer, 2019; Leo and Sardanelli, 2020). For more information on choosing the right statistical test, see Section 2.4.2.

2.4.1.4. Power analysis

Once all of these pieces of information have been decided, a power analysis can be performed to determine the expected power of the planned study. This is commonly used to determine what the minimum number of participants needed will be to ensure a study with sufficient power. For more information, including tables with the relationship between power, N, and hypothesized effect size as well details on calculating power with more complex study designs, Dean et al. (1999), Bausell and Li (2002), Sullivan and Feinn (2012), and Montgomery (2017) provide a solid introduction to the topic and VanVoorhis et al. (2007) discuss common rules of thumbs of sample size. Additionally, Faul et al. (2009) provide an open-source tool for performing power analysis including support for most common statistical tests.[3]

[3] www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower

2.4.2. Choosing the Correct Statistical Test

The (set of) applicable statistical test(s) is determined by the experimental setup including the choice of measurement scale (Section 2.3.3.4) and the experimental design (Section 2.3.4). To choose a test, one has to determine the number of levels (groups), if the samples were collected in a paired or unpaired design, the measurement scale of the dependent variable, and whether parametric assumptions apply. In the following, we discuss these aspects and present common tests. Figure 2.16 summarizes these tests within a flow chart, illustrating the conditions under which each test is applicable. We refer to Buckley (2006) for an extensive discussion of a broad range of tests along with SPSS and R code.

2.4.2.1. Paired and Unpaired Tests

Whether a paired or an unpaired test is the correct choice depends on the choice of experimental design (see Section 2.3.4) as different designs require accounting for the subject-dependent variances in the responses differently. A paired test is applicable if the samples were collected in a within-subject design (repeated measures), i.e., from one group. An unpaired test is applicable if the samples were collected in a between-subjects design, i.e., from different groups.

2.4.2.2. Parametric and Non-Parametric Tests

Parametric tests make assumptions on the underlying population distribution (such as normality), non-parametric tests do not make assumptions on the distributions but still can make other assumptions (Colquhoun, 1971). Therefore, the measurement scale of the dependent variable can directly determine whether a parametric test is applicable. For example, we cannot run a t-test (which is parametric) on ordinal responses from {“often”, “sometimes”, “never”}. It is often claimed that parametric tests offer higher statistical power. This statement has to be restricted to very specific conditions and Colquhoun (1971) argues to prefer non-parametric tests as long as there is no experimental evidence of the error distribution. We refer to Colquhoun (1971) for a discussion of the differences between parametric and non-parametric methods and to Sprent (2012) and Corder and Foreman (2014) for details on non-parametric statistics.

2.4.2.3. Frequently-used Tests for NLP

In the following, we present a selection of common statistical tests, highlight important assumptions they make, and provide examples of NLP applications they are relevant to. We do not exhaustively discuss all assumptions of each test here, but instead, offer first guidance in choosing the right test.

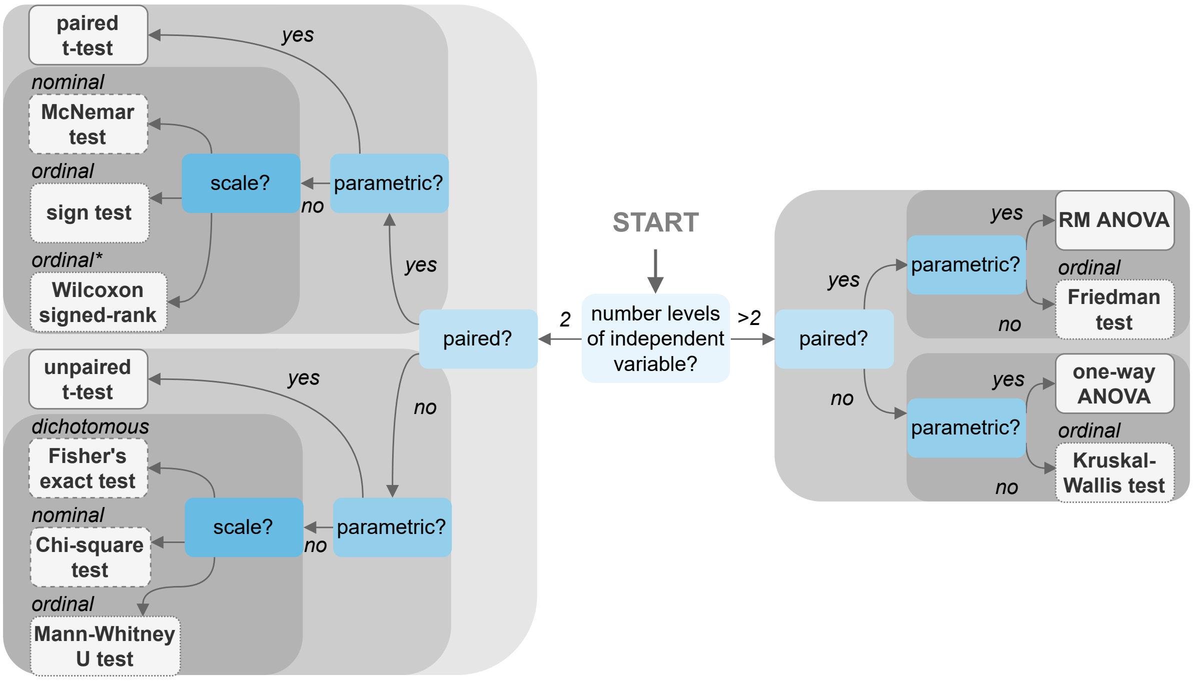


Figure 2.16.: A flow chart to help find an appropriate test to analyze collected responses. Starting from the middle, the chart shows tests suited to analyze experiments with two levels of independent variables (e.g., system A and system B) on the left and tests suited to analyze experiments with more than two levels of independent variables (e.g., systems A, B and C) on the right. A paired test needs to be used if, e.g., a within-subject design is used and the level of measurement determines whether a parametric test can be used. For example, yes/no ratings are nominal/dichotomous by definition and cannot be analyzed using a t-test. *The pairwise differences have to be on an ordinal scale, see Colquhoun (1971).

We first discuss tests that are applicable to experiment designs with one factor that has two levels (e.g., the factor chatbot system with the levels "system A" and "system B"). Thereafter, we consider tests involving one factor with more than two levels (e.g., the factor chatbot system with an additional third "system C"). These tests are called *omnibus tests*, which means that they only can detect that "there is a difference" but make no statement about pairwise differences. Therefore, pairwise post hoc tests are usually used after detecting a significant difference with an omnibus test. For a more detailed discussion, we refer to Buckley (2006).

Unpaired and Paired Two-Sample t-Test. In the context of user studies, the t-test is usually used to test if the means of two samples differ significantly, i.e., a *two-sample* t-test.[4] In NLG evaluation, the time a participant takes to read a sentence generated by one versus another

[4] A *one-sample* t-test compares a sample's mean with a predefined reference mean.

system could be compared using a t-test. For the two-sample test one further distinguishes an *unpaired* or *independent* test and a *paired* or *dependent* test. The t-test assumes that the errors follow a normal distribution which is usually decided subjectively by inspecting the quantile-quantile (Q-Q) plot of the data (Hull, 1993). When analyzing Likert scale responses, the choice of test depends on whether one regards the scale scores to be measures to be ordinal or interval measures (see Section 2.3.3.4). For more detailed recommendations on when and when not to apply parametric statistics to Likert responses we refer to Harpe (2015). However, De Winter and Dodou (2010) compare error rates between the non-parametric Mann-Whitney U test with the parametric t-test for five-point Likert items and find that both tests yield similar power. A typical situation to apply a t-test is to compare task completion times, e.g., the time it takes a participant to read a text or the times a user takes to engage with a chat bot.

Mann-Whitney U and Wilcoxon Signed-Rank. Although the t-test can be robust to violations of normality (Hull, 1993), non-parametric alternatives, such as the Mann-Whitney U test for unpaired samples and the Wilcoxon signed-rank test for paired samples are preferable for non-parametric data. The Mann-Whitney U test is the non-parametric counterpart to the unpaired t-test. In contrast to the t-test, which is restricted to interval data, it is additionally applicable to ordinal data as well as interval data that does not fulfill the parametric assumptions. For example, testing user acceptance of a voice assistant could involve asking participants how often they would use the system: "daily", "weekly", "monthly" or "never". The paired counterpart to the Mann-Whitney U test is the Wilcoxon signed-rank test which compares median differences between the two groups and can be applied as long as the pairwise differences between samples can be ranked. If this is not possible, a sign test should be used instead (Colquhoun, 1971). An application for the Mann-Whitney U test and the Wilcoxon Signed-Rank test are Likert ratings of, e.g., text fluency or coherence.

Fisher's Exact, χ^2, and McNemar Test. If the measurement scale is nominal, the Mann-Whitney U and the Wilcoxon signed rank test are not applicable. Instead, Fisher's exact test should be used for unpaired groups if the dependent variable is *dichotomous*, i.e., can only take two values like "yes" and "no", e.g. for rating the correctness of answers generated by a question answering system. If it can take more values, e.g. additionally "I do not know", a chi-square (χ^2) test can be used. When samples are paired, the test of choice should be a McNemar test. An exemplary NLP application of these two tests, are binary responses, to, e.g., "Is this sentence grammatically correct?" (Fisher's exact or chi-square test for unpaired samples and McNemar test for paired samples) or categorial responses to, e.g.,"For which tasks would you use this travel chat bot most likely: (a) searching for travel information, (b) booking a travel or (c) making a modification to a booked travel?" (chi-square test for unpaired samples and McNemar test for paired samples).

One-Way and Repeated-Measures ANOVA. So far, we only addressed tests that compare two groups, such as samples from "dialog system A" to samples from "dialog system B". When we add a third or more conditions, the discussed tests are no longer applicable. Instead, if the samples are parametric, a one-way analysis of variance (ANOVA) can be applied to unpaired samples and a repeated-measures ANOVA can be applied to paired samples.
For example, when interaction times with three different explainability methods should be compared, one can use a one-way ANOVA when using a between-subjects design (i.e., each participant sees only one method) and a repeated-measures ANOVA if each participant sees each method (in a randomized order), i.e. a within-subject design.

Kruskal-Wallis and Friedman Test. Like the Mann-Whitney U and the Wilcoxon-signed rank test are the non-parametric counterparts to the paired and unpaired t-test, one can use the non-parametric Kruskal-Wallis test instead of a one-way ANOVA and the non-parametric Friedman test instead of a repeated-measures ANOVA. For further details, we refer to Ostertagova et al. (2014) and Pereira et al. (2015). In the above explainability methods example, these tests are appropriate choices if instead of measuring interaction times (interval scale), one, e.g., asks participants to rate trust on a single-item Likert scale (ordinal scale).

2.4.2.4. More Complex Tests

In addition to the tests above, there also are more general models and tests, which can be useful for NLP applications. If the response variable is, e.g., categorical (e.g., "dog", or "cat"), linear models can be extended to *generalized linear models* (Nelder and Wedderburn, 1972), where the (e.g., categorical) response scale is linked to a latent scale (e.g., logits) via a *link function*. If the experimental setup requires accounting for, e.g., subject-specific influences (e.g., mother tongue or literacy) or repeated measures of one factor within a mixed design (e.g., a design in which each participant uses one version of a dialog system, i.e. a between-subjects factor, but all participants perform the same set of tasks, i.e., a within-subject factor), generalized linear mixed models (GLMMs) can be an appropriate statistical model. The difference between a linear and a linear mixed model is that the latter is extended to include *random effects*, such as individual participant characteristics on top of *fixed effects*, such as "system type" resulting in a *mixed* model. Intuitively, the purpose of including random effects is to get a clearer picture of the fixed effects and not to falsely attribute, e.g., an effect of participant age to be a difference between two chatbots. An introduction to linear mixed models and their usage in R is provided by Winter (2013). More details can be found in McCulloch and Neuhaus (2005) and Jiang (2007). Howcroft and Rieser (2021) discuss ways to improve power in human evaluations in NLP and recommend to make use of ordinal mixed effects models. Other commonly used models are generalized additive models (GAMs) (Hastie and Tibshirani, 1990; Hastie et al., 2009) which model the response variable as a sum of general basis functions. We refer to Wood (2017) for an introduction using R. Two applications of (ordinal) GAMs can be found (a) in Divjak and Baayen (2017) who analyze grammaticality ratings and (b) in Section 5.1 of this thesis in which we study human perception of saliency explanations. As this model class is central to our analysis of explanation understanding, we introduce GAMs generalized additive mixed models (GAMMs) models in more detail in the following.

2.4.2.5. (Ordinal) Generalized Additive Mixed Models

For an intuitive understanding of GAMMs, we sketch how one arrives at ordinal GAMMs starting from linear models. We follow the notation of Wood (2017).

Linear Model. In a linear model, the response variable $\mathbf{y}$ (e.g., a numeric rating of importance) is modeled as a function of explanatory variables $\mathbf{X}$ which are related to y *linearly* via parameters β assuming additional noise ϵ:

$$y = \mathbf{X}\beta + \epsilon. \tag{2.19}$$

Linear Mixed Model. In many scenarios, there are *random effects* which one wants to account for in the model. For example, we collect 150 word importance ratings per participant, i.e., we collect *repeated measures* and are in danger of violating the independence assumption and introducing a confounding effect of the variable *participant ID* because specific participants might have a tendency to give overall higher ratings than other participants. Like the linear model, linear mixed models estimate *fixed effects* but in addition they also model *random effects* (e.g., of the participant ID) to disentangle their influence on the response variable and thereby offer a clearer view on the fixed effects. The general formulation of a linear mixed model reads

$$y = \mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \epsilon, \tag{2.20}$$

where $\mathbf{Z}$ corresponds to the random effects and $\mathbf{b}$ to the respective weights.

Generalized Linear Model (GLM). While linear models require the response distribution to be normal, generalized linear models (GLMs) Nelder and Wedderburn (1972) generalize to non-normal (exponential family) response distributions, such as categorical responses (e.g., dog or cat) or ordinal responses (e.g., Likert item ratings). To achieve this generalization, GLMs link values on the response scale (e.g., categorical ratings) to a latent scale (e.g., logits) via a *link function* $g(\cdot)$ (e.g., logit function). For a row i, the general formulation reads:

$$g(\mu_i) = \mathbf{X}_i\beta. \tag{2.21}$$

Generalized Additive Model (GAM). While a generalized linear model only allows to model linear relationships between the explanatory variables and $g(\mu_i)$, a GAM Hastie and Tibshirani (1990) generalizes the linear relationship to a *sum of smooth functions* of explanatory variables using:

$$g(\mu_i) = \mathbf{X}^*{}_i\theta + f_1(x_{1i}) + f_2(x_{2i}, x_{3i}) + ... \,, \tag{2.22}$$

where f_1 and f_2 are smooth functions that typically are chosen to be a sum of basis functions, such as splines, and $\mathbf{X}^*$ corresponds to strictly parametric model components. A regularized estimation of these functions allows GAMs to model complex functions, but also to fall back to simpler, e.g., constant or linear functions when an increase in model complexity is not sufficiently warranted by improved model fit.

Ordinal Generalized Additive Mixed Model (ordinal GAMMs). Having introduced the previous models, an ordinal GAMM can be described as a generalized additive model that additionally accounts for random effects and models ordinal ratings via a continuous latent variable that is separated into the ordinal categories via estimated threshold values. For further details, Divjak and Baayen (2017) provide a practical introduction to ordinal GAMs in a linguistic context and Wood (2017) offers a detailed textbook on GAM(M)s including implementation and analysis details.

2.4.2.6. Post Hoc Tests

The presented omnibus tests do not allow to make statements about pairwise differences between conditions. For example, an ANOVA might detect a significant difference within the groups {"system A", "system B", "system C"} but makes no statement if there is for example a significant difference between "system A" and "system B". In such cases, one needs to use a post hoc test (Buckley, 2006; MacKenzie, 2013). The respective post hoc test is typically only applied if the omnibus test found a significant effect and — depending on the method

— requires a multiple testing adjustment. Commonly used tests are Tukey HSD, Scheffé, Games-Howell, Nemenyi, and Conover.

2.4.2.7. The Multiple Comparisons Problem

The intuition behind the multiple comparisons problem is that every time a statistical test is run, it bears the risk of a Type I error, i.e., falsely reporting a positive result. When one considers the standard significance level, α of 0.05, this represents 95% confidence in a reported significant difference or a 5% chance that there was a type I error. However, if multiple hypotheses are tested, the chance for a type I error over the entire experiment increases. For example, if two hypotheses are tested each with a 95% confidence level, the confidence for the entire experiment drops to 0.9 (i.e., 0.95^2, the likelihood that both tests were not falsely positive), and thus α equals 0.1.

Thus, when multiple hypotheses are tested at once, the individual α levels need to be adjusted. A simple and well-known adjustment method is the Bonferroni correction, which divides the α level per test by the number of tests to ensure a given family-wise error rate – error rate across the entire experiment – is achieved. Less conservative methods, such as the Benjamini-Hochberg technique or the Holm procedure, also called the Holm-Bonferroni method can provide more power for an experiment (Bender and Lange, 2001; Streiner and Norman, 2011). Alternatively, if the data in an experiment was suitable for an ANOVA test, the Tukey HSD, also called the Tukey test, can be a good choice. When and when not to apply α adjustments is discussed by Rothman (1990); Ottenbacher (1998); Moyé (1998); Bender and Lange (2001); Streiner and Norman (2011).

2.4.2.8. Further Analysis Methods for NLP

As NLP systems are frequently evaluated in side-by-side comparisons, the collected variables can also be ranks or preferences (Callison-Burch et al., 2007; Grundkiewicz et al., 2015). For example, participants can be asked to rank pairs of translations or generated speech snippets. TrueSkill™ (Herbrich et al., 2006; Sakaguchi et al., 2014) can be used to construct ranks from pairwise preferences. Pairwise preferences can be analyzed statistically using models, such as the (log-linear) Bradly-Terry model (Bradley and Terry, 1952; Dras, 2015) or approaches based on item response theory (Sedoc et al., 2019; Sedoc and Ungar, 2020). Further, hybrid approaches that combine ranking with scale ratings (Novikova et al., 2018) or human judgments with automatic evaluation (Hashimoto et al., 2019) have been proposed for NLG.

3. System Architectures and Explanations

In this chapter, we present our contributions to integrating external knowledge into explanation generation for NLI (Section 3.1) as well as our novel self-reflective thought flow architecture, that provides explanations in the form of decision sequences, which we demonstrate at the example of QA (Section 3.2).

3.1. Explanations as Output: External Knowledge Improves Explainable NLI

We introduced the task of NLI in Section 2.1.1. In essence, given two sentences (premise and hypothesis), systems are trained to decide whether (a) the first sentence entails the second sentence, (b) the two sentences contradict each other or (c) they have a neutral relation. As discussed, the NLI task can be extended to an explainable NLI task in which the system needs to provide an additional textual explanation of why the predicted answer should be the correct answer. Figure 2.1 on page 11 in Section 2.1.1 shows an example of an explainable NLI instance. Solving the task requires models to not only reason over the provided information but also to link it with commonsense knowledge.

Integrating external knowledge was shown to improve NLI systems (Jijkoun and de Rijke, 2005; Chen et al., 2018; Li et al., 2019; Faldu et al., 2021). However, the following question remains: *Does the positive effect of external knowledge on the inference ability transfer to the generation of explanations?* Figure 2.1 shows an NLI example for which external knowledge potentially helps to infer the correct label and explanation. In the example, the system needs to link “dog” to “animal”, “jumping for a Frisbee” to “playing”, “Frisbee” to “plastic toy”, and “snow” to “outside” as well as to ‘cold weather”. The predicted explanation needs to explicitly state this reasoning chain and would be expected to benefit from external knowledge.

Pre-trained language models, such as BERT (Devlin et al., 2019) or GPT-2 (Radford et al.,

2019) have been shown to be able to learn and store commonsense knowledge implicitly (Petroni et al., 2019). However, an open question is: *How effective is the implicit commonsense knowledge of language models compared to symbolic sources of knowledge, such as knowledge base triplets?*

To evaluate NLI models, mainly automatic measures, such as accuracy, are used. However, model weaknesses can stay unnoticed using automatic scores alone and automatic scores are not necessarily correlated to human-perceived model quality (we will revisit this topic in Chapter 4). Thus, human evaluation is a crucial step in the development of user-centered explainable AI systems. Therefore, we ask the question: *How do humans perceive explanation quality of state-of-the-art natural language inference models?*

In this section, we investigate the three mentioned research questions. To answer them, we analyze the impact of external knowledge from multiple sources, such as knowledge graphs, embeddings, and language models, and propose novel architectures to include and combine them into explainable NLI systems. Further, we conduct an extensive automatic analysis as well as a user study. To the best of our knowledge, our study exceeds previous human evaluations of explainable NLI models regarding the number of participants as well as the variety of rated explanation criteria.

For our first research question, we find that the positive effect of external knowledge on label accuracy in the standard NLI setting can also be observed in the explainable NLI setting and external knowledge can improve the BLEU scores of the generated explanations. In regard to our second research question, we observe that pre-trained language models are the most promising source of commonsense knowledge but at the same time identify weaknesses with respect to negations and numerical reasoning abilities which, however, can be mitigated through combination with additional knowledge sources. Despite the improvements in accuracy, BLEU, or BLEURT scores, our user study shows that, for our third research question, these do not reflect in human ratings of explanation correctness, commonsense inclusion, nor grammar or label correctness. Our results provide initial evidence for caution to solely rely on automatic scores for explainability and motivate our in-depth analysis of the relation between proxy scores and human ratings discussed in Chapter 4.

To facilitate future work, we make our model's predictions as well crowdsourced human ratings available at `https://github.com/boschresearch/external-knowledge-explainable-nli`.

3.1.1. Knowledge Integration Methods

In the following, we describe our base model and present the models we analyze.

3.1.1.1. Base Model

We combine a state-of-the-art attention-based inference model with an explainable NLI model that predicts entailment labels and generates explanations. In particular, we use the encoder part of the enhanced sequential inference model (ESIM), which has a cross-attention layer to capture relevant semantics between premise and hypothesis (Chen et al., 2017), and the prediction part of the PRED-EXPL model of Camburu et al. (2018). We refer back to our introduction to different varieties of attention in Section 2.1.2.2. In contrast to the self-attention used in transformers, the cross-attention in the ESIM model is used to capture dependencies between the hypothesis and the premise. We represent the input sentences with BERT embeddings (Devlin et al., 2019) which we fine-tune on the SNLI dataset. We pass inputs of the form "*[CLS] premise [SEP] hypothesis*" to BERT and use a softmax layer on top of the CLS token's embedding to predict the entailment label and fine-tune the model for up to two epochs. Throughout this chapter, we refer to this model as VANILLA.

3.1.1.2. Integration of Knowledge Sources

External knowledge can be found in various formats. We aim to cover a possibly broad variety and focus on state-of-the-art sources and methods. We include the natural language knowledge base COMET (Bosselut et al., 2019), the ConceptNet Numberbatch embeddings (Speer et al., 2017), and the GPT-2 language model (Radford et al., 2019).

Background Knowledge from COMET. As our example in Figure 2.1 in Section 2.1.1 showed, resolving natural language entailment can require reasoning over multiple concepts and relations, such as inferring *cold weather* and *outside* from *snow*. We seek to facilitate this resolvement by providing the model with related words (and phrases) that can be seen as a natural language extension of the premise and the hypothesis. We use the COMmonsEnse Transformers (COMET) (Bosselut et al., 2019) as a natural language knowledge base to query background knowledge for the premise and the hypothesis. COMET is based on a transformer language model that is fine-tuned on a knowledge base completion task on ConceptNet. Given an input sentence and a ConceptNet relation, it generates a phrase to complete the object in a knowledge statement expressed in the (subject, relation, object) format. Instead of feeding in the whole premise and hypothesis, we find that chunking them into noun and verb subphrases based on POS tags patterns yields better object phrase generations.[1] Thus, for each

[1] We manually find that feeding in the whole sentence predominantly relates the output to the last tokens of the sentence and fails to include information from tokens earlier in the sentence.

sentence (premise/hypothesis) we generate #chunks × #relations object phrases.[2] Afterward, we embed each object phrase (with the respective relation string prepended) with Sentence-BERT (Reimers and Gurevych, 2019) and quantify its similarity to the embedding of the source sentence using cosine similarity. For each relation, we keep the object phrase with the highest similarity score. Given the relation *HasA* and the chunked sentence *The dog* | *is walking in the snow*, for example, COMET will generate *bone* and *effect of freeze* for the two sub-phrases, respectively. We only preserve the object phrase *effect of freeze* as it has a higher similarity to the source sentence. To condense the object phrases into a fixed-length vector representation, we average the respective Sentence-BERT embeddings. This procedure yields one vector representing the background knowledge regarding the premise and one regarding the hypothesis. We combine them with the local inference vector representation of Chen et al. (2017). Following Camburu et al. (2018), this vector is passed to the label prediction module as well as the explanation decoder. We refer to this model as COMET.

Modified Attention with ConceptNet. Following Li and Srikumar (2019), we use knowledge-driven rules to modify the attention weights within the cross-attention layer between premise and hypothesis in the encoder. This supports the attention mechanism to align word pairs p_i and h_j from premise and hypothesis based on world knowledge. The rules proposed by Li and Srikumar (2019) are shown in Equation 3.1 and 3.2. In R_1, the antecedent K_{p_i,h_j} indicates that a word pair p_i and h_j is of a certain relation within ConceptNet. If the condition of the antecedent is true, the consequent A'_{p_i,h_j} that aligns the word pair follows. R_2 is a relatively conservative rule that additionally takes the model's own decision into account. The antecedent $K_{p_i,h_j} \wedge A_{p_i,h_j}$ in R_2 is a conjunctive condition that becomes true if a word pair is both in a relation and aligned by a model's original attention. If such a conjunctive condition is true, the word pair must be aligned which results in a new alignment as the consequent A'_{p_i,h_j} indicates.

$$R_1 : K_{p_i,h_j} \rightarrow A'_{p_i,h_j} \tag{3.1}$$

$$R_2 : K_{p_i,h_j} \wedge A_{p_i,h_j} \rightarrow A'_{p_i,h_j} \tag{3.2}$$

Different from the approach of Li and Srikumar (2019) that checks a word pair's relation in a binary fashion, we hypothesize that knowledge-aware embeddings might capture more fine-grained word relationship that exists in multi-hop relational edges. Considering *playground* and *playroom*, for example, the former is usually located outdoors whereas the latter is located indoors. We generalize the binary relational inclusion from Li and Srikumar (2019) to continu-

[2] We consider the relations AtLocation, CapableOf, DefinedAs, HasA, HasProperty, HasSubevent, InheritsFrom, InstanceOf, IsA, LocatedNear, MadeOf, PartOf, SymbolOf, UsedFor, and LocationOfAction.

ous relation scores. For this, we replace the binary rule antecedent with the absolute cosine similarity between the ConceptNet Numberbatch (Speer et al., 2017) vector representations of p_i and h_j. We empirically confirm that our continuous formulation outperforms the binary version regarding label accuracy as well as explanation correctness. In the following, we refer to these modified rules as continuous constraints and use CONT to refer to the respective model.

All-text Prediction with GPT-2. Similar to Kumar and Talukdar (2020), we fine-tune a pre-trained GPT-2 language model on the e-SNLI dataset. In contrast to Kumar and Talukdar (2020), we use a single GPT-2 model to generate explanations for all three entailment labels instead of training a separate model for each of them. This allows us to directly integrate the label prediction into the language model instead of training an additional model which predicts the label on top of the three explanations. Therefore, we propose two models, which both are GPT-2-large models, but differ regarding their training setting. In the label-first setting (GPT-LF), the model is trained on text following the structure:
Premise: <premise> Hypothesis: <hypothesis> [LAB] [label] [EXP] <explanation> EOS.
In the explanation-first setting (GPT-EF) it is trained on text following the structure:
Premise: <premise> Hypothesis: <hypothesis> [EXP] <explanation> [LAB] <label> EOS.

3.1.1.3. Combined Models

COMET and ConceptNet. We combine COMET with CONT to benefit from both integrated background information from COMET and a knowledge-enhanced attention mechanism based on ConceptNet Numberbatch. We expect this to help the model focus on important relations between premise and hypothesis.

Knowledge-enhanced Ensembles. We combine the world knowledge of BERT embeddings (VANILLA), ConceptNet Numberbatch (CONT), COMET (COMET) and the combined model COMET+CONT with the language model abilities of GPT-2 (GPT-LF and GPT-EF). For this, we propose an ensemble that not merely aggregates label votes but combines the models with respect to their different strengths. The label predictions of VANILLA, CONT, COMET, COMET+CONT as well as GPT-LF are passed to a majority voting. In the *basic ensemble*, the GPT-LF model is then conditioned on the voted label and generates the final explanation. We refer to this model as ENSEMBLE. In the *filtered ensemble*, the majority voting only allows models to vote if their generated explanation lets the GPT-EF model predict the same label prediction as the original model. In other words, we fix the input as well as the generated explanation and only let the GPT-EF model predict the label. This step can be interpreted as a

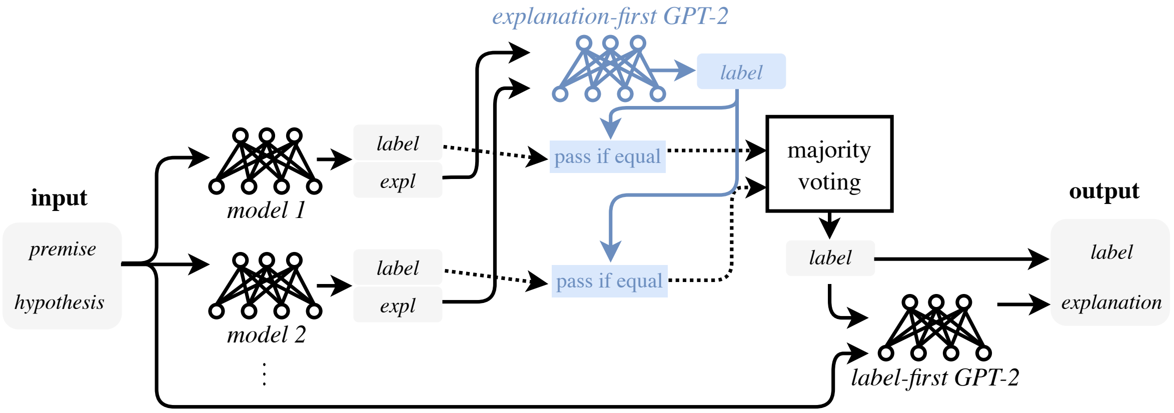


Figure 3.1.: Schematic depiction of the two proposed ensemble architectures ENSEMBLE and FILTERED-ENS. The blue components correspond to the consistency-filter.

consistency filter that prevents models from voting if their label prediction does not match their explanation prediction. In the following, we refer to this model as FILTERED-ENS. Figure 3.1 depicts the corresponding ENSEMBLE and FILTERED-ENS ensemble architectures.

3.1.2. Automatic Evaluation

First, we evaluate the discussed knowledge-enhanced models with respect to commonly used scores on e-SNLI and a stress test evaluation. In addition to our constructed models, we also include PRED-EXPL (Camburu et al., 2018), which is basically our VANILLA baseline without cross-attention and with GloVe embeddings instead of fine-tuned BERT embeddings. Further, we include two more recent models proposed for e-SNLI: NILE:post-hoc, which is the highest performing model from Kumar and Talukdar (2020), and WT5-11B from Narang et al. (2020), which held the state-of-the-art performance at the time of our study. While NILE:post-hoc is based on GPT-2 as well, WT5-11B is a fine-tuned version of the T5 language model (Raffel et al., 2020). We train all non-LM models with five random seeds and report scores of the median model based on label accuracy. Table 3.1 shows predicted explanations for the subset of models that we investigate within the human evaluation in Section 3.1.3. Further examples are provided in Appendix A.1.

3.1.2.1. Performance on e-SNLI

Following prior work on e-SNLI, we report label accuracy as well as BLEU scores (Papineni et al., 2002) for explanations. We additionally evaluate BLEURT scores (Sellam et al., 2020), which is a reference-based learned evaluation metric to model human judgments of text

Model	**Predicted Explanation**
GROUND-TRUTH	a man is either playing the accordion or performs a mime act while happy people pass by or angry people glare at him.
VANILLA	a man can not be playing and a mime at the same time
COMET	the man is either playing the accordion or a mime
CONT	people can not be playing and angry at the same time
COMET+ CONT	the man can not be playing the accordion and the mime at the same time
GPT-LF	Happy people are not angry people.
WT5-11B	The man cannot be playing the accordion and performing a mime act at the same time.

Table 3.1.: Explanation predictions of the models used within the human evaluation for the premise "*A man on a sidewalk is playing the accordion while happy people pass by*" and the hypothesis "*A man on the sidewalk performs a mime act while angry people glare at him*". All models correctly predict the class *contradiction* but generate different explanations. The predicted explanation of the FILTERED-ENS model is identical to the explanation of the GPT-LF model as GPT-LF is used to predict the ensemble's explanation. Missing punctuation reflects exact model generations.

generation. BLEURT is of particular interest for explanation evaluation as Clinciu et al. (2021) compare how various automatic scores, such as BLEU, ROUGE, and METEOR correlate to human ratings of generated explanations and find that embedding-based methods and particularly BLEURT scores show distinctly higher correlations than, e.g., BLEU.

Table 3.2 shows the respective scores for all considered models. For NILE:post-hoc (Kumar and Talukdar, 2020) and WT5-11B (Narang et al., 2020) we report the label accuracy from their paper and calculate BLEU/BLEURT scores based on the explanation predictions provided by the authors. Narang et al. (2020) calculate BLEU scores using SacreBLEU v1.3. (Post, 2018) leading to a higher reported score of 33.7. The upper block lists models that share or extend the PRED-EXPL architecture. Compared to PRED-EXPL, the VANILLA model achieves a notable increase in label accuracy as well as BLEURT scores. Surprisingly, COMET reduces all scores and even decreases the BLEU score below the PRED-EXPL score. In contrast, knowledge-enhanced cross attention (CONT) improves BLEU and BLEURT scores and reaches a label accuracy close to VANILLA. Combining CONT with COMET retains the CONT label accuracy but again slightly decreases BLEU and BLEURT scores. The lower block contains models that are or include language models. All language model-based models increase BLEU and BLEURT scores. All except GPT-EF outperform all non-language model models.

To analyze whether the performance differences of models can be really attributed to a better reasoning and commonsense knowledge ability instead of merely different model capacity, we next evaluate our models on the NLI stress test evaluation.

Type	Model	Label Accuracy	BLEU	BLEURT
non-LM	PRED-EXPL	84.21	19.77	-0.871
	VANILLA	89.20	19.71	-0.820
	COMET	88.97	18.84	-0.822
	CONT	89.02	20.10	-0.799
	COMET+CONT	89.07	19.66	-0.809
LM-based	GPT-EF	87.89	21.70	-0.624
	GPT-LF	89.70	26.90	-0.577
	ENSEMBLE	90.24	27.10	-0.576
	FILTERED ENS	90.24	27.09	-0.577
	NILE:POST-HOC	91.49	26.26	-0.577
	WT5-11B	**92.30**	**29.01**	**-0.511**

Table 3.2.: Automatic evaluation metrics on the e-SNLI test set. Label accuracy quantifies NLI performance. BLEU and BLEURT score the similarity between predicted and ground truth explanation texts. BLEURT is a learned score which predicts scores given the text to be evaluated and a reference text. Higher values are better.

3.1.2.2. Stress Test Evaluation

Table 3.3 shows the results of our models on the NLI stress test evaluation proposed by Naik et al. (2018). The dataset contains multiple subsets of which each subset is used to evaluate the robustness of the system against a specific type of perturbation, e.g., spelling errors, negations, numerical reasoning, and more. On average, all models distinctly improve performance compared to the PRED-EXPL baseline. With respect to VANILLA, all models except GPT-EF improve average performance. Further, both COMET and CONT improve average label accuracy, while their combination decreases performance. Surprisingly, GPT-LF outperforms the ensemble methods on average. While COMET+CONT reaches the best performance in terms of e-SNLI label accuracies, it performs worst on the stress tests. The same effect can be observed for the FILTERED-ENS. While it reaches top performance for the spelling error test, its performance drops for numerical reasoning, where it performs worse than any other model. These results show that combining different knowledge sources does not result in a consistent combination of their weaknesses and strengths. Instead, the sources of external knowledge have to be carefully adjusted to the target domain and our results paint a rather pessimistic picture regarding a cure-all solution. Further, a model's reasoning capabilities have to be assessed in detail as evaluation across different reasoning types easily masks model weaknesses.

Finally, we investigate whether language models reach their higher performance due to better reasoning: For most of the assessed reasoning types — with the exception of numerical reasoning and negation — the best non-ensemble model in fact is GPT-LF. Also, GPT-LF reaches

Type	**Model**	**Total**	**Competence Test**		**Distraction Test**			**Noise Test**
			Antonymy	Numerical	Word Overlap	Length Mismatch	Negation	Spelling
non-LM	PRED-EXPL	48.69	36.36	36.55	47.17	53.44	45.31	52.42
	VANILLA	56.94	37.94	32.24	55.46	65.21	52.03	62.90
	COMET	57.05	34.54	35.48	57.31	64.15	52.85	62.33
	CONT	57.09	32.50	**40.28**	52.10	64.35	**53.38**	62.77
	COMET+CONT	56.26	44.43	34.16	51.34	64.39	49.36	63.03
LM-based	GPT-EF	52.74	51.81	31.33	55.91	60.97	38.44	58.20
	GPT-LF	**59.28**	**54.84**	28.80	**64.06**	**68.72**	42.82	67.07
	ENSEMBLE	59.19	37.97	34.03	58.13	67.45	52.51	65.92
	FILTERED-ENS	58.99	52.53	28.54	63.70	68.02	42.18	**67.10**

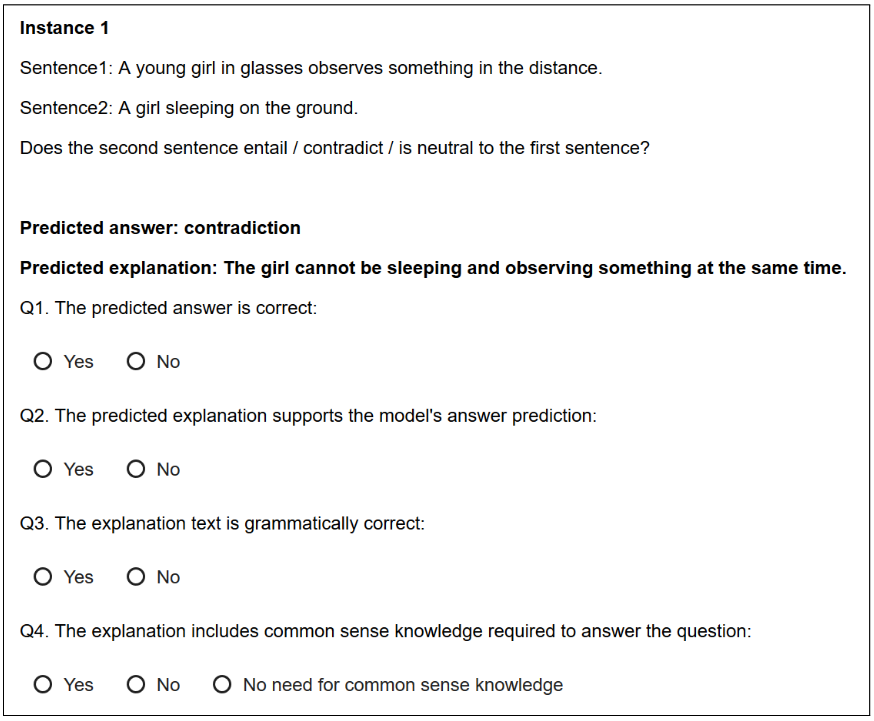
Table 3.3.: Label accuracies (higher is better) for all categories in the NLI stress test tasks (Naik et al., 2018). The six rightmost columns show (i) the model's reasoning abilities (competence), (ii) how sensitive it is to lexical distractors (distraction), and (iii) how robust it is against noise from different perturbations (noise). Each column corresponds to one dataset. For datasets with matched and mismatched subsets, we report the accuracy over all labels within the group. Similarly, the total accuracy is calculated over all labels.

the highest accuracy on average. Therefore one could generally recommend to include external knowledge in the form of a pre-trained language model as the foremost option. However, our results also show that language models are not necessarily the best choice for all reasoning needs and can, e.g., severely decrease performance for numerical reasoning and negations, where models based on language models perform worse than all other models.

3.1.3. Human Evaluation

While automatic scores, such as BLEU, provide a valuable starting point for evaluating explanations, they fall short of capturing the model's real explanation capabilities. We, therefore, conduct a large-scale crowdsourcing study to complement our automatic evaluations on e-SNLI and the stress tests. Following related work (Narang et al., 2020), we assess explanation quality based on ratings from crowdworkers on Amazon Mechanical Turk (MTurk).[3] While previous work limited evaluation to rating explanation correctness, we additionally ask participants to provide fine-grained ratings of commonsense inclusion and grammatical correctness. A screenshot of the interface is shown in Figure 3.2.

[3] We provide an in-depth discussion of explanation quality evaluation in Chapter 4.



Instance 1

Sentence1: A young girl in glasses observes something in the distance.

Sentence2: A girl sleeping on the ground.

Does the second sentence entail / contradict / is neutral to the first sentence?

Predicted answer: contradiction

Predicted explanation: The girl cannot be sleeping and observing something at the same time.

Q1. The predicted answer is correct:

○ Yes ○ No

Q2. The predicted explanation supports the model's answer prediction:

○ Yes ○ No

Q3. The explanation text is grammatically correct:

○ Yes ○ No

Q4. The explanation includes common sense knowledge required to answer the question:

○ Yes ○ No ○ No need for common sense knowledge

Figure 3.2.: Screenshot of the study interface presented to crowdworkers on MTurk.

3.1.3.1. Conditions

In order to evaluate effects across the discussed sources of external knowledge, we include seven models in our human evaluation: VANILLA, COMET, CONT, COMET+CONT, GPT-LF, FILTERED-ENS and WT5-11B. Additionally, we evaluate the e-SNLI ground truth labels and explanations as a representation of a hypothetical perfect model. Table 3.1 displays the different explanations the models predict for an exemplary input as well as the ground truth explanation.

3.1.3.2. Dependent Variables

We evaluate the models' predicted labels and explanations along four self-reported dimensions.

Label Correctness. Following Kumar and Talukdar (2020) and Narang et al. (2020), we ask participants to rate if the predicted label answer is correct or not.

Explanation Correctness. Similar to Camburu et al. (2018), Kumar and Talukdar (2020), and Narang et al. (2020), we collect subjective yes/no explanation correctness ratings.

Grammatical Correctness. We ask participants to rate if the explanation is grammatical.

Commonsense Inclusion. We ask participants whether the explanation includes commonsense knowledge that is needed to answer the question. We collect responses on an item with the options *yes*, *no*, and *no need*.

3.1.3.3. Study Design

In order to evaluate the effect of the level of required external knowledge, we compile, like Kumar and Talukdar (2020) and Narang et al. (2020), a set of 100 premise-hypothesis pairs. In contrast to them, we compose the 100 pairs to contain 50 pairs that require a low level of external knowledge and 50 pairs that require a high level. To gather pairs of both categories, we let two annotators rate 250 premise-hypothesis pairs from the e-SNLI test set. We sample 50 pairs per level of external knowledge from the 179 pairs on which the annotators agree. We provide details on the annotation criteria in Appendix A.1. During the study, we, like Narang et al. (2020), collect five crowdsourced ratings for each condition and for each of the ten input pairs per batch, i.e., 500 ratings per model and a total of 4000 ratings for each variable. Table 3.4 displays various model predictions along with examples of corresponding human ratings (non-averaged individual ratings).

3.1.3.4. Analysis

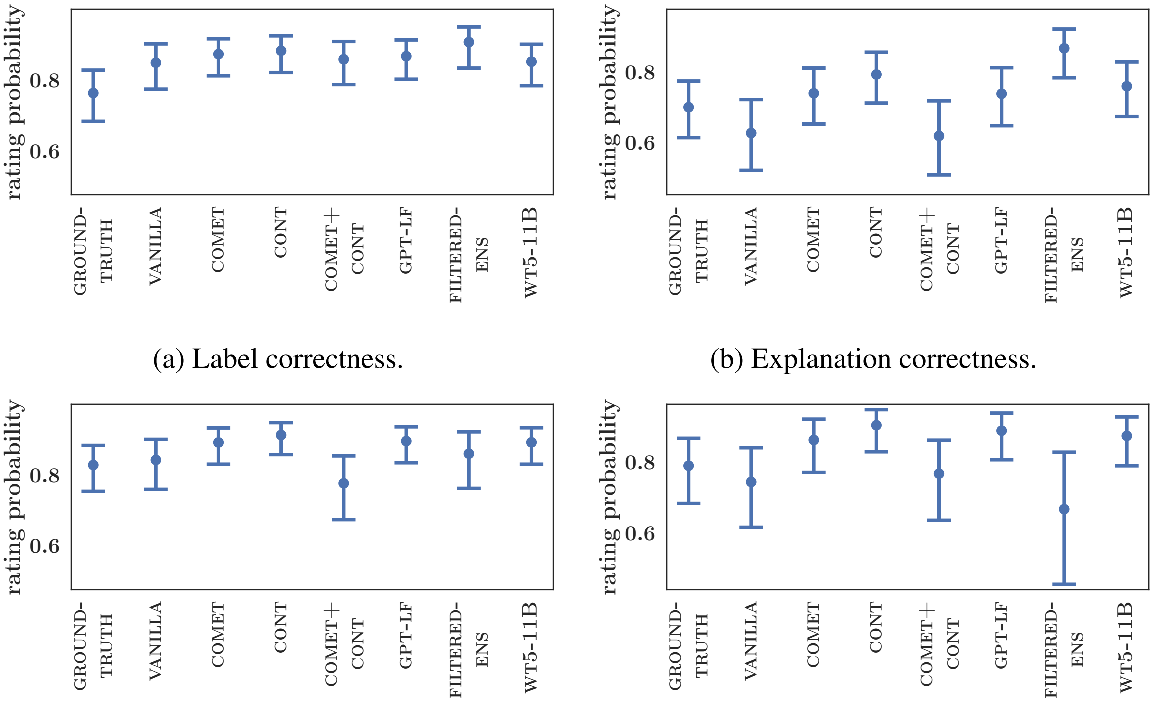
We collect responses from 290 crowdworkers and discard those that were entered in less than five minutes (31%) as this might indicate arbitrary answer selection. Note that the repeated measures design of our study possibly introduces inter-dependencies within ratings as, e.g., certain participants can have a tendency to rate explanations as correct more often than others or a certain question might elicit more label correctness ratings. Thus, we use GLMMs to account for the potentially confounding variables (worker ID, question ID, and level of required commonsense knowledge). We refer to our brief introduction to GLMMs in Section 2.4.2. As our response variables are binary,[4] we use binomial GLMMs. We include fixed effects (model and commonsense level) as well as random intercepts (worker and question IDs). Figure 3.3 shows effect displays for the collected ratings in relation to the predictor *model type*.

We conduct likelihood-ratio tests (LRTs) between the full model and the model without the evaluated predictor to test the effects of *model type* and *commonsense level* on all four rating variables. As *model type* contains more than two factors, we additionally conduct single-step corrected Tukey HSD post hoc tests for all four variables.

[4] We do not consider "no need" commonsense ratings during the respective model estimation.

Premise	**Hypothesis**	**CoS**	**Model**	**Pred. Label**	**Predicted Explanation**	**Human Ratings**			
						Label	Expl.	Gramm.	CoS
By a table with several boxes and two coffee cups on it, a woman in a green sweater is opening a small box and another person (whose head cannot be seen) is operating a small handheld electronic device.	Members of congress opening care packages.	high	WT5-11B	neutral	Not all women are members of congress. Not all boxes are care packages.	✓	✓	✓	✓
In front is a guy singing and playing guitar and a white shirted drummer in the back.	Some people perform at a concert.	high	GPT-LF	neutral	Just because a guy singing and playing guitar and a white shirted drummer in the back does not mean they are performing at a concert.	✓	✓	✓	✓
A man in a black coat and hat and a woman in a blue coat with pink hat snowshoeing on the side of a road with a stoplight in the background.	A man and woman are outside in the snow.	high	COMET +CONT	entailment	in the snow is outside	✓	✓	✓	✓
Two individuals dressed up like animals are posing for the camera.	Two young boys dressed like cows are posing .	low	COMET +CONT	neutral	not all people are young	✗	✗	✓	✓
A man playing a piano	The man's hands are on the keys of a piano.	high	VANILLA	entailment	playing a piano requires one to be on the hands	✓	✓	✗	✓
Two women are hugging on a path through a grassy area with a cow visible past them.	There are women outside in this picture	high	WT5-11B	entailment	A grassy area is outside.	✓	✓	✓	✓
A woman in a spaghetti strap tank wearing a flower in her hair is staring towards the right.	The woman is sleeping on a pillow.	low	COMET +CONT	contradiction	the woman can not be looking towards a sleeping she is sleeping	✓	✗	✗	✓

Table 3.4.: Examples of model predictions and individual human ratings of the predicted labels and explanations for various model types. "CoS" denotes common sense requirements/ratings.



(a) Label correctness.

(b) Explanation correctness.

(c) Grammatical correctness.

(d) Commonsense correctness.

Figure 3.3.: Effect displays for user ratings of label, explanation, grammatical, and commonsense correctness depending on *model type* following Fox (2003). The rating probability is the probability that a prediction of a respective model type is perceived to be correct by a human considering fixed effects. Error bars mark 95% confidence limits.

Label Correctness. We do not observe a significant main effect of *model type* ($\chi^2(7) = 13.00$, $p = 0.0723$) but a significant main effect of *commonsense level* ($\beta = 0.28$, $\chi^2(1) = 4.54$, $p < 0.0331$). β refers to the estimate of a *high* commonsense level.

Explanation Correctness. We observe a main effect of *model type* (χ^2(7)=24.06, p<0.0012) and *commonsense level* ($\beta = 0.27$, $\chi^2(1) = 7.79$, $p < 0.0053$). For *model type*, a post hoc Tukey test showed significant differences between FILTERED-ENS and VANILLA ($p < 0.0055$) as well as FILTERED-ENS and COMET+CONT ($p < 0.0029$).

Grammatical Correctness. We observe a main effect of *model type* ($\chi^2(7) = 14.20$, $p < 0.0479$). However, a post hoc Tukey test did not reveal significant differences between any model type pair. No significant main effect of *commonsense level* was observed ($\beta = 0.02$, $\chi^2(1) = 0.02$, $p = 0.8803$).

Commonsense Correctness. We observe a main effect of *model type* ($\chi^2(7) = 20.63$, $p < 0.0044$). However, a post hoc Tukey test did not reveal significant differences between any model type pair. No significant main effect of *commonsense level* was observed ($\beta = 0.07$, $\chi^2(1) = 0.25$, $p = 0.6163$).

Overall, these results show surprisingly few significant differences between the different model types and conflict with the large differences within automatic evaluation scores.

3.1.4. Overall Discussion

Effect of External Knowledge. We showed that external knowledge can increase label accuracies on e-SNLI as well as on the stress tests. In addition, we found external knowledge to increase BLEU(RT) scores and thus help explanation generation in terms of proxy scores.

Implicit Knowledge in Language Models. While language models achieve the best scores on general e-SNLI performance, the stress tests showed that they do not succeed in all reasoning types. Thus, for choosing the best way of integrating commonsense knowledge, the final reasoning goal of the model needs to be considered.

Perceived Explanation Quality by Humans. We expected the large differences in e-SNLI label accuracy (up to 3.23%), BLEU (up to 10.17), and BLEURT (0.31) to reflect in human ratings, but none of these maximal differences in scores leads to a statistically significant difference in ratings for any dependent variable. Regarding the observed significant differences, FILTERED-ENS is not the best model included in the study with respect to e-SNLI (WT5-11B reaches distinctly higher values for all scores) and, similarly, neither VANILLA nor COMET+CONT are the worst models on any score in Table 3.2. Thus, large accuracy gains do not necessarily imply better models when used in real-world applications with users. In the following, we will further discuss these results.

Superhuman Model or Noisy Ground Truth? It is particularly remarkable that the ground truth ratings do not significantly differ from any other model's ratings. In fact, the ground truth condition ranks in the lower half across all four rating dimensions and yields the lowest probability of receiving label correctness ratings as shown in Figure 3.3a. Similarly, Narang et al. (2020) note that in their experiment the WT5-11B model reaches a 12%-higher explanation correctness rating than the ground truths. This indicates that e-SNLI might not be suitable to distinguish performances of today's high-performing models. While it remains valuable for training, models should be scored on specifically designed evaluation sets, for example, an explainable extension of the NLI stress test dataset.

Limitations and Future Directions. Although we evaluated a total of eleven different model architectures and various different sources of external knowledge, this clearly does not exhaust all possible knowledge sources or architectures. While our analysis provides insight into the most common knowledge sources integrated into representative model architectures, future work should confirm our findings for additional sources and architectures. Although our user study already is — to the best of our knowledge — the largest and most fine-grained evaluation of explainable NLI, future work should further expand the set of dependent variables to potentially reveal effects that are not visible through the lens of our experimental setup. In addition, our results raise the question of whether the observed evaluation disconnect also holds for other explainable NLP tasks. We address this question in detail in Chapter 4 and, i.a., observe a similar disconnect for explainable QA.

3.2. Decision Processes as Explanations: Thought Flow Networks

While the previous section addressed the generation of textual explanations, this section explores decision sequences as a novel explanation format. Inspired by Hegel's dialectics, we propose the concept of *thought flows*, formalize it in terms of gradient-based optimization within the model's decision space, and demonstrate its application to QA models.

Our method builds upon the observation that today's classification models map a specific input $\mathbf{x}$, e.g., a token or a sentence, to an output $\hat{y}$ (Bishop, 2006) where $\hat{y}$ can be, e.g., a class, a sequence (e.g., a generated text) or an answer span extracted from a context. This mapping $\mathbf{x} \rightarrow \hat{y}$ might involve various modulations and abstractions of $\mathbf{x}$ in a latent space, e.g., hidden layers of a neural network, but typically does not allow variations or trajectories of $\hat{y}$. Humans, on the other hand, rarely come to a single decision right away but follow a complex thought process that involves reflecting on initial decisions, comparing different hypotheses, or resolving contradictions. While humans' trains of thought are extensively studied in cognitive sciences and philosophy — one particular example being Hegel's dialectics (Maybee, 2020) — such theories are rarely explored in machine learning. However, with increasingly complex tasks that have large output spaces, such as QA[5], or tasks that require multiple reasoning steps,

[5] A Longformer QA model can output 16M possible spans.

such as multi-hop QA, learning to directly hit the right prediction in one shot might be more difficult than to learn to iteratively self-correct an initial prediction.

In this section, we propose the concept of a *thought flow* as a sequence of inter-dependent probability distributions. Thought flows thereby can offer explanatory value by relating a model decision to a sequence of preceding, intermediate decisions, that open a novel perspective on the model's prediction behavior. To implement the concept of thought flow, we propose a simple *correction module* which can be used on top of any model that provides output logits of one or multiple distributions. In particular, it is inspired by the three moments of Hegel's dialectics which it relates to forward and backward passes of the model and is trained to judge whether the predicted class distribution corresponds to a correct prediction.

We apply our method to QA and conduct experiments on the HotpotQA dataset (Yang et al., 2018). We demonstrate our method's ability to self-correct flawed answer span predictions and identify qualitative patterns of self-correction, such as span reductions/extensions. Figure 3.4 shows a real example of a thought flow that corrects a prediction ($\mathbf{y}^{(0)}$), that would be the output of a standard model, to a new prediction ($\mathbf{y}^{(2)}$) within two steps. Concretely, two gradient updates using our method result in a shrinkage of the answer span followed by a cross-sentence answer jump. We find that our method can achieve performance improvements up to 9.6% F_1-score (absolute) on HotpotQA.

Finally, we assess the impact of thought-flow predictions on human users within a crowdsourced study. We find that thought-flow predictions are perceived as significantly more correct, understandable, helpful, natural, and intelligent than single-answer predictions and/or top-3 predictions and result in the overall best user performance without increasing completion times or mental effort.

To sum up, this section presents our contributions on (i) a formalization of a thought flow inspired by human thinking and Hegel's dialectics, (ii) a novel correction module and a corresponding gradient-based update scheme to generate a thought flow in a state-of-the-art transformer network, (iii) experiments on QA that demonstrate its strong correction capabilities and reveal qualitative patterns of self-correction, (iv) a crowdsourced user study that demonstrates that thought flows can improve perceived system performance as well as actual user performance using the system.

3.2.1. Thought Flow Networks

In the following, we present background on Hegel's dialectics (Section 3.2.1.1), formalize thought flows based on it (Section 3.2.1.2), and present a concrete implementation for QA (Section 3.2.1.3).

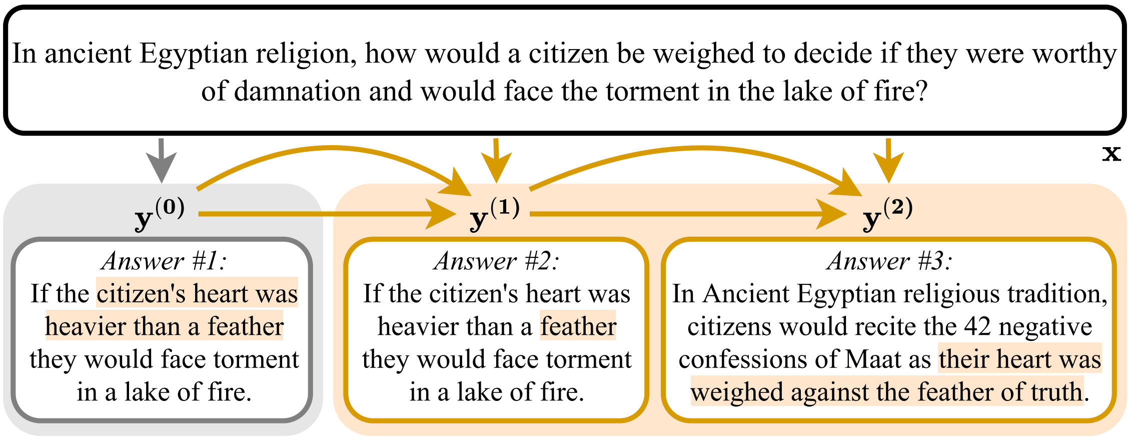

In ancient Egyptian religion, how would a citizen be weighed to decide if they were worthy of damnation and would face the torment in the lake of fire?
$\mathbf{x}$
$\mathbf{y}^{(0)}$
$\mathbf{y}^{(1)}$
$\mathbf{y}^{(2)}$
Answer #1:
If the citizen's heart was heavier than a feather they would face torment in a lake of fire.
Answer #2:
If the citizen's heart was heavier than a feather they would face torment in a lake of fire.
Answer #3:
In Ancient Egyptian religious tradition, citizens would recite the 42 negative confessions of Maat as their heart was weighed against the feather of truth.

Figure 3.4.: In contrast to the standard approach of mapping an input to an output in a single step (grey box), we propose a method that allows models to sequentially "reconsider" and update their predictions, i.e., the *thought flow*. In this (real) question answering example, the orange box marks our thought flow extension, which corrects a flawed answer in two steps.

3.2.1.1. Inspiration: Hegel's Dialectics

To give models the opportunity to reflect and refine their predictions, we take inspiration from Hegel's dialectics. Dialectics, in general, describes an argumentative method involving opposing sides (Maybee, 2020). What distinguishes Hegel's dialectics from other dialectics is that in his dialectics, the opposing sides are views or definitions while, e.g., in Platon's dialectics the opposing sides are people (Maybee, 2020). Besides its philosophical relevance, Hegel's dialectics has been related to various fields before, such as cognitive sciences (Riegel, 1973), neuroscience (Boonstra and Slagter, 2019), or optimization (Kadioglu and Sellmann, 2009).

In the following, we will briefly introduce the three *moments* of Hegel's dialectics and distinguish them from the thesis-antithesis-synthesis triad before we use them to derive our thought flow concept in the following section.

Three Moments. Hegel's dialectics distinguishes three moments: (i) the *moment of understanding*, (ii) the *dialectical moment*, and (iii) the *speculative moment*. The moment of understanding refers to the initial, "seemingly stable" view. In the second moment, this supposed stability is lost due to the view's one-sidedness or restrictedness and the initial determination *sublates* itself into its own negation. The speculative moment unifies the first two determinations by negating the contradiction (Maybee, 2020).

Thesis-Antithesis-Synthesis Triads. The three moments are often compared to a thesis-antithesis-synthesis triad, which was popularized by Heinrich Moritz Chalybäus, but *cannot* necessarily be equated to it as argued by, e.g., Mueller (1958). While the thesis-antithesis-synthesis triad can suggest the notion of a "one pass" process, the dialectical process in Hegel's dialectic does not have to end after a single iteration but can go through several iterations (Maybee, 2020).[6] The possibility for iteration is an essential property of our thought flow.

3.2.1.2. Formalization of Thought Flow Concept

We now translate the abstract description of these three moments into a simplified formalized setting that can be implemented in any differentiable model that uses a vector-valued representation of the input (such as an embedding) and outputs (tuples of) logits. In particular, we embed Hegel's dialectics in a framework of obtaining an initial "thought" vector and iteratively updating it in the three "moments". Note that our formalization is not to be understood as an accurate reflection of Hegel's dialectics. Instead, Hegel's dialectics serves as a useful inspiration to enable the development of a novel machine learning method.

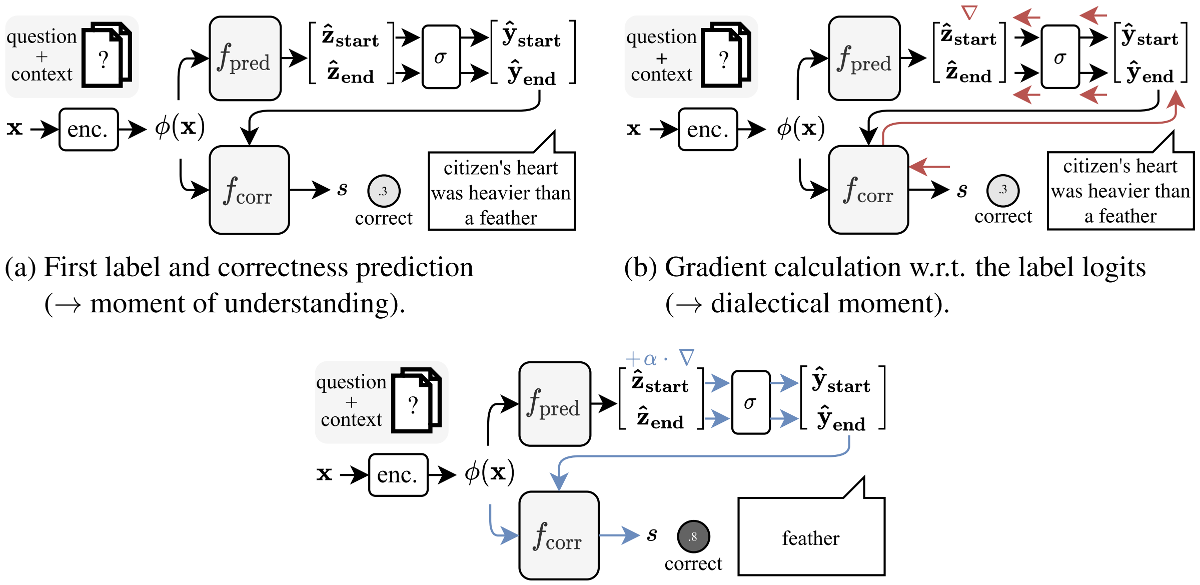
Thought. We model a *thought* with $\hat{\mathbf{z}} \in Z$, the logits corresponding to a model's prediction and $Z \subseteq \mathbb{R}^c$ being the logit space.[7] This $\hat{\mathbf{z}}$ serves as a representation of the model's "decision state" between c classes and captures information including the most probable output as well as possible alternatives and uncertainty.

Moment of Understanding. The first moment relates to an initial, seemingly stable view. We model this with the initial value of $\hat{\mathbf{z}}^{(\mathbf{0})}$, obtained from applying the prediction function $f_{\text{pred}} : \Phi \rightarrow Z$ to the model to the encoded input $\phi(\mathbf{x})$ with an encoding function $\phi : \mathbb{R} \rightarrow \Phi$ and the encoding space $\Phi \subseteq \mathbb{R}^e$ (see Figure 3.5a).

Dialectical Moment. In the second moment, the stability breaks down due to the view's one-sidedness or restrictedness. To model this, we first introduce a new function $f_{\text{corr}} : Z \times \Phi \rightarrow \mathbb{R}$ that differentiably maps $\hat{\mathbf{z}}^{(\mathbf{0})}$ to a correctness score $s \in \mathbb{R}$ that is an estimate of the quality of the model prediction corresponding to $\hat{\mathbf{z}}^{(\mathbf{0})}$ conditioned on $\phi(\mathbf{x})$. Intuitively, $f_{\text{corr}}(\hat{\mathbf{z}}^{(\mathbf{0})}, \phi(\mathbf{x}))$ scores how good the current decision state $\hat{\mathbf{z}}^{(\mathbf{0})}$ is given the model input, which is represented using $\phi(\mathbf{x})$. Next, we formalize the dialectical moment with the gradient of the correctness

[6] A particular example of such an iterative process within Hegel's work can be found in the dialectical development of Hegel's logic regarding the concepts of "Abstract Purpose" and "Realized Purpose" (Maybee, 2020).

[7] We choose $\hat{\mathbf{z}}$ over $\hat{\mathbf{y}}$ because we can modify logits in energy space without normalization.


question + context
?
x
enc.
$\phi(\mathbf{x})$
f_{pred}
$\hat{\mathbf{z}}_{\text{start}}$
$\hat{\mathbf{z}}_{\text{end}}$
σ
$\hat{\mathbf{y}}_{\text{start}}$
$\hat{\mathbf{y}}_{\text{end}}$
f_{corr}
s
.3
correct
citizen's heart was heavier than a feather
∇
$+\alpha \cdot \nabla$
.8
feather

(a) First label and correctness prediction (→ moment of understanding).

(b) Gradient calculation w.r.t. the label logits (→ dialectical moment).

(c) Update logits and correctness score (→ speculative moment).

Figure 3.5.: The steps of the prediction update scheme and their relation to the three moments of Hegel's Dialectics. The depicted example corresponds to the first answer change from Figure 3.4.

score with respect to $\hat{\mathbf{z}}^{(0)}$, i.e. $\nabla^T_{\hat{\mathbf{z}}^{(0)}} s$ (see Figure 3.5b). Thus, we ask "How does the thought $\hat{\mathbf{z}}^{(0)}$ have to change in order to be more correct?" This gradient represents the view's instability. As it creates a tension away from the current $\hat{\mathbf{z}}^{(0)}$ towards a new one, it destroys its stability and thus negates the initial view.

Speculative Moment. The third moment unites the initial view with the negation from the dialectical moment. We formalize this by modifying $\hat{\mathbf{z}}^{(0)}$ with a step into the gradient's direction that yields

$$\hat{\mathbf{z}}^{(1)} := \hat{\mathbf{z}}^{(0)} + \alpha^{(0)} \cdot \nabla^T_{\hat{\mathbf{z}}^{(0)}} s \tag{3.3}$$

where $\alpha^{(0)}$ is a, potentially dynamic, step width and $\hat{\mathbf{z}}^{(1)}$ again constitutes the subsequent first moment of the next iteration (see Figure 3.5c).

Iteration. Iterative application of the dialectical and the speculative moment yields a sequence of logits $\left(\hat{\mathbf{z}}^{(\mathbf{k})}\right)_{k=0}^{N}$ and corresponding predictions $\left(\hat{\mathbf{y}}^{(\mathbf{k})}\right)_{k=0}^{N}$.

In the following, we detail this abstract formalization for the example of QA.

3.2.1.3. Implementation in Transformers for Question Answering

Figure 3.5 visualizes our formalization for the question answering example introduced in Figure 3.4. We now discuss QA-related implementation details.

Choosing Parameters and Functions. To apply our abstract thought flow method to a real model we have to (a) determine how to structure the model prediction logit vector $\hat{\mathbf{z}}$, (b) choose an input representation $\phi(\mathbf{x})$ (that is passed to f_{pred} as well as f_{corr}), (c) choose a parametrization of the correctness score prediction function f_{corr} and (d) define what the correctness score s measures. In the following, we describe how these aspects can be realized in a transformer-based QA model.[8]

Composing $\hat{\mathbf{z}}$: In extractive QA, a typical approach to model answer span extraction from a context of L tokens is to use two probability distributions: (i) $\hat{\mathbf{y}}_{\text{start}} \in [0,1]^L$ that assigns a probability of being the start of the answer to each token in the context and (ii) a respective end token distribution $\hat{\mathbf{y}}_{\text{end}} \in [0,1]^L$.[9] To match our previously defined formalization, we define $\hat{\mathbf{z}}^{(\mathbf{i})} := \begin{bmatrix}\hat{\mathbf{z}}^{(\mathbf{i})}_{\text{start}} & \hat{\mathbf{z}}^{(\mathbf{i})}_{\text{end}}\end{bmatrix}^{\mathrm{T}}$ which is linked to the respective probabilities via the softmax function σ:

$$\begin{aligned}\hat{\mathbf{y}}^{(\mathbf{i})} :&= \begin{bmatrix}\hat{\mathbf{y}}^{(\mathbf{i})}_{\text{start}} & \hat{\mathbf{y}}^{(\mathbf{i})}_{\text{end}}\end{bmatrix}^{\mathrm{T}} \\ &= \begin{bmatrix}\sigma(\hat{\mathbf{z}}^{(\mathbf{i})}_{\text{start}}) & \sigma(\hat{\mathbf{z}}^{(\mathbf{i})}_{\text{end}})\end{bmatrix}^{\mathrm{T}}.\end{aligned}$$

Input Representation $\phi(\mathbf{x})$: In contrast to transformer-based classification models that conventionally rely on the embedding of the [CLS] token, typical transformer-based QA models apply a linear function on top of each token's embedding that maps the embedding to a start and an end logit. We follow this convention and define

$$\phi(\mathbf{x}) := [\mathbf{e_1}, \mathbf{e_2}, ..., \mathbf{e_L}] \in \mathbb{R}^{d \times L} \tag{3.4}$$

i.e., as the sequence of L contextualized embeddings with embedding dimension d.

Choosing f_{corr}: To represent the input within f_{corr}, we need a representation of $\phi(\mathbf{x})$ that focuses on the relevant parts of the (potentially very long) input that were relevant to the start and end logit predictions. We thus choose a weighted average over all token embeddings to retain as much as possible of the important information from the input while heavily reducing

[8] For a background on the QA task and transformer models, we refer to our brief introduction in Section 2.1.

[9] See Section 2.1.2.3 for a detailed description and step-by-step example of span extraction modeling.

its available representation dimensionality to a single vector. As weights, we choose the element-wise product of the predicted start and end probabilities. We thus define a modified input encoding $\tilde{\phi}^{(i)}(\mathbf{x}) \in \mathbb{R}^d$ where d denotes the dimension of the embeddings (e.g., 768 for BERT-base (Devlin et al., 2019)) as follows:

$$\tilde{\mathbf{w}}^{(\mathbf{i})} := \left(\hat{\mathbf{y}}^{(\mathbf{i})}_{\text{start}} \odot \hat{\mathbf{y}}^{(\mathbf{i})}_{\text{end}} + \epsilon \cdot \mathbf{1}\right) \in \mathbb{R}^L \tag{3.5}$$

$$\tilde{\phi}(\mathbf{x})^{(i)} := \phi(\mathbf{x}) \cdot \frac{\tilde{\mathbf{w}}^{(\mathbf{i})}}{\Sigma_j \tilde{\mathbf{w}}^{(\mathbf{i})}_j} \in \mathbb{R}^d \tag{3.6}$$

where ϵ is a small constant that ensures that we do not divide by zero, e_i is the embedding of the i-th token, $\odot$ is element-wise multiplication, and L is the maximum number of tokens in the context. This modified input representation $\tilde{\phi}(\mathbf{x})^{(i)}$ can be regarded to be a dynamic perspective onto $\phi(\mathbf{x})$ that highlights these parts of $\phi(\mathbf{x})$ that are most important to the model's answer prediction. The intuition behind this is that the correction module should have access to all information about the context that the prediction model focused on. Based on initial empirical findings, we choose to use a two-layer MLP with SELU activation (Klambauer et al., 2017) to map the concatenated vector

$$\left[\text{dropout}(\tilde{\phi}^{(i)}) \quad \hat{\mathbf{z}}^{(\mathbf{i})}_{\text{start}} \quad \hat{\mathbf{z}}^{(\mathbf{i})}_{\text{end}}\right]^{\mathrm{T}} \in \mathbb{R}^{d+2\cdot L} \tag{3.7}$$

to an estimated correctness score s. Note that f_{corr} does not receive the decoded answer text but uses the start and end logits directly to provide differentiability.

Correctness Score s: Following standard QA evaluation metrics, we use the F_1-score of the predicted answer as the correctness score that f_{corr} is trained to predict.

Training. To train f_{corr}, we freeze the parameters of f_{pred}. Then, we pass the training instances through the whole model (including ϕ, f_{pred}, and f_{corr}) as shown in Figure 3.5a to obtain the target of the predicted correctness score s (i.e., f_{corr} predicts an F_1-score without access to the ground-truth answer span). We determine the ground-truth correctness score by calculating the F_1-score between the ground truth answer and the answer prediction from f_{pred}. We define the correctness estimate prediction loss as the mean squared error between the calculated score, and the predicted s and train f_{corr} to minimize it. Overall, we thus train f_{corr} to score how correct a model prediction (represented by the start and end logits) is given a model input (represented by the condensed input encoding $\tilde{\phi}(\mathbf{x})$) and use the model's predictions on the training set to generate ground truth correctness scores (using F_1-score).

Inference. At inference time, we encode a new input and predict (i) the answer start and end logits using f_{pred} and (ii) an estimated F$_1$-score s of the predicted answer span using the correction module f_{corr} as shown in Figure 3.5a. Instead of directly using the initial logits as the model's prediction — as would be done in a standard model — we iteratively update the logits w.r.t. the estimated correctness score's gradient following our formalization from Section 3.2.1.2 as shown in Figures 3.5b and 3.5c.

Update Rule: As described in Section 3.2.1.2, we aim at modifying $\hat{\mathbf{z}}^{(i)}$ such that the correction module assigns an increased correctness (i.e., F$_1$-score in this application to QA). To apply Equation (3.3), we have to define how the step size α is chosen in our QA application. We choose a time-independent α such that a predefined probability mass δ is expected to move. To this end, we first take a probing step of length one, calculate the distance as the L_1 norm between the initial distribution and the probe distribution and choose the step width $\alpha \in \mathbb{R}^+$ such that it scales the linearized distance to the hyperparameter δ using

$$\alpha := \left[\frac{\delta}{||\sigma(\hat{\mathbf{z}}^{(i)}) - \sigma\left(\hat{\mathbf{z}}^{(i)} + \nabla^T_{\hat{\mathbf{z}}^{(i)}} s\right)||_1} + \epsilon \right] \tag{3.8}$$

with the softmax function $\sigma(\cdot)$ and a small constant $\epsilon \in \mathbb{R}^+$ needed for numerical stability.

Monte Carlo Dropout Stabilization: The gradient $\nabla_{\hat{\mathbf{z}}^{(i)}} s$ is deterministic but can — as we find in preliminary experiments — be sensitive to small changes in the input representation $\phi(\mathbf{x})$. We, therefore, stabilize our correction gradient estimation by *sampling* and averaging gradients instead. For this, we use the dropped-out input encoding from Equation (3.7) and sample five gradients for every step using MCDrop (Gal and Ghahramani, 2016).

3.2.2. Question Answering Experiments

3.2.2.1. Data, Model, and Training

Dataset. We choose the HotpotQA dataset (distractor setting) (Yang et al., 2018) to evaluate our models as it contains complex questions that require multi-hop reasoning over two Wikipedia articles. In the distractor setting, the model is "distracted" by eight irrelevant articles that are passed to the model in addition to the two relevant articles. In addition to yes/no/answer span annotations, HotpotQA also provides explanation annotations in the form of binary relevance labels over the paragraphs of the relevant articles which we do not use when training our models. As the public test set is secret, we use the official validation set as test set and sample a custom validation set of size 10k from the training set leaving 80,564 training instances.

Base model. We use a Longformer-large (Beltagy et al., 2020) model[10] with a linear layer on top that maps token embeddings to start and end logits as our underlying QA model. The model reaches 63.5% F_1-score (SD=0.6) on the HotpotQA validation set averaged over three random seeds and can handle input lengths up to 4096 tokens which enables us to feed in the entire context as a single instance without truncation. The model's input is a single token sequence that contains the question followed by the answer context (i.e., the ten concatenated Wikipedia articles). The model's output are two distributions over the input tokens (i.e., two 4096-dimensional distributions), one for the answer start position and one for the answer end position. We prepend a "yes" and a "no" token to the context, which offers the advantage of modeling these answer options within the same distributions as the text span answers. In total, this model has 435M parameters compared to the additional 331k parameters our multilayer perceptron (MLP) implementation of f_{corr} adds.

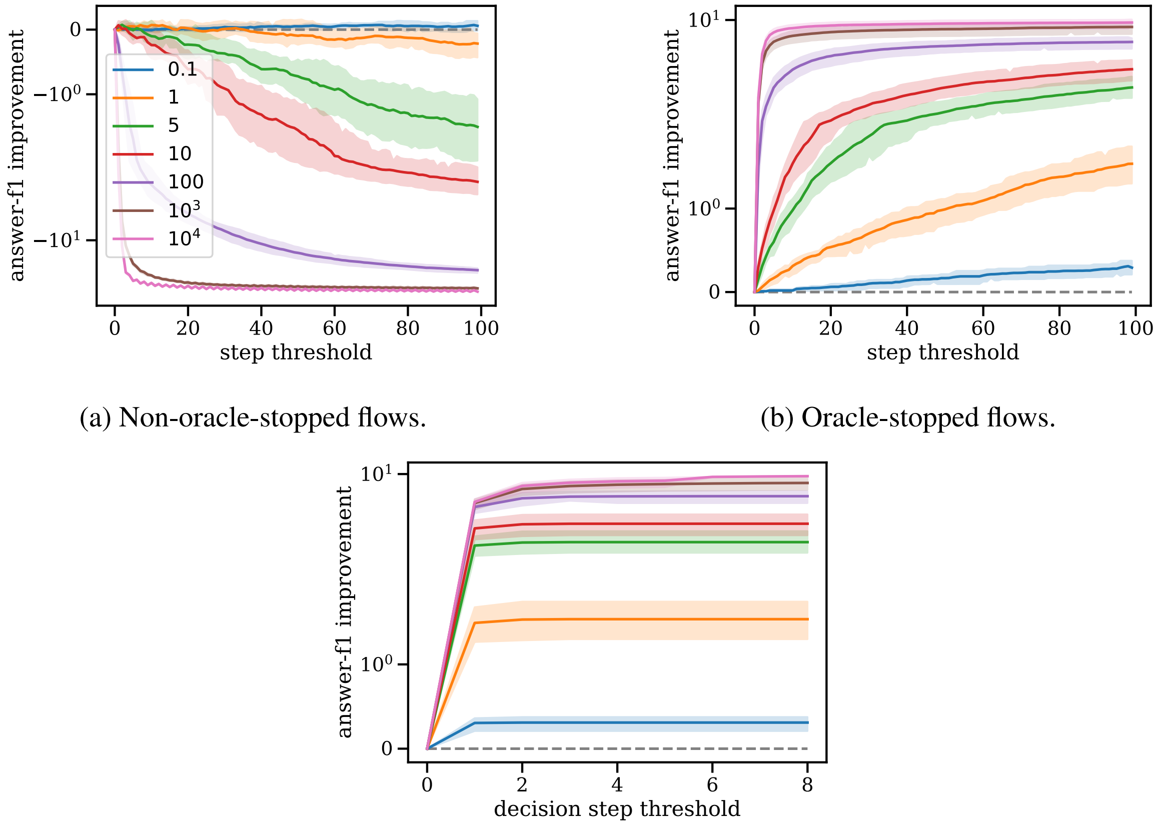
Training Details. We first train the base models for five epochs on a single V100 GPU using a learning rate of 10^{-5}, an effective batch size of 64, the ADAM (Kingma and Ba, 2015) optimizer with decoupled weight decay (Loshchilov and Hutter, 2019), early stopping and a CE loss on the start/end logits. We subsequently train the correction modules using the same setting but the MSE loss function for F_1-score prediction training. Training models took approximately three days each. In the following, we report all results as averages over three random seeds including standard deviations.

3.2.2.2. Performance Improvements

Performance Over Steps. Figure 3.6a shows how F_1-scores per gradient scaling target δ evolve over 100 steps. We observe that small δ values enable small F_1-score improvements. While $\delta = 0.1$ consistently improves F_1-scores, all other δ values eventually deteriorate F_1-scores. The higher the δ value, the quicker the F_1-score decreases. We conclude that (i) very small δ values are not sufficient to reach notable performance gains and that (ii) larger δ can initially improve performance but then "overshoot" with their corrections. We hypothesize that a remedy to this trade-off is to use larger δ values but stop the flows at the right time.

Stopping Oracle. To test this hypothesis, we introduce an oracle-stopping function that stops the thought flow where it achieves its best F_1-score performance. Figure 3.6b shows that, with this oracle function, thought flows can reach performance improvements up to 9.6% (absolute) F_1-score (SD=0.61).

[10] `https://huggingface.co/allenai/longformer-large-4096`



(a) Non-oracle-stopped flows.

(b) Oracle-stopped flows.

(c) Oracle-stopped flows per decision change.

Figure 3.6.: Thought flows with different gradient scaling targets δ averaged over three seeds of a QA model. Higher values for δ correspond to more aggressive decision changes. Without a stopping oracle that stops when the thought flow does no longer improve an answer (top left), only $\delta = 0.1$ provides consistently stable, but very small F_1-score improvements. With an oracle (top right), higher values for δ reach higher and faster F_1-score improvements up to $>9\%$. Nearly all performance gains are achieved by the first decision change (bottom). y axes use a symlog scale. Improvements are reported as absolute F_1-scores (not relative).

Figure 3.6c shows that almost all performance improvements are due to the first decision change within the thought flows and that answer spans constantly improve and do not randomly shift across the context. This observation shows that single thought flow changes are highly effective and can reach substantial corrections fast.

3.2.2.3. Thought Flow Patterns

In a qualitative evaluation, we identify various thought flow patterns. We randomly sample 150 instances from the subset of the official validation split for which the thought flow changed the initial answer prediction. We identify six (non-exclusive) correction patterns and show selected examples in Table 3.5. In addition, Table 3.6 shows thought flow examples using three correction steps.

Cross-sentence. With 52.7%, this is the most frequent type of correction. The thought flow shifts the predicted answer from one sentence to another.

Span Reduction. Thought flows can shorten the predicted answer span to correct it.

Span Extension. Similarly, thought flows can enlarge a predicted answer span to correct it.

In-Sentence. On top of in-sentence span reduction/extension, the thought flow can also jump between non-overlapping spans within a sentence.

Entity Refinement. In this correction pattern, the thought flow keeps predicting the same entity but jumps to an alternative mention of the entity.

Logic Hops. The thought flow performs a step-wise reasoning that first resolves the first step of HotpotQA's two-step reasoning structure before jumping to the second step, i.e., the correct answer.

Combinations. We observe various combinations of the aforementioned patterns. A model can, for instance, jump between sentences, refine entities and reduce the answer span.

Sequential Corrections. Corrections can also occur sequentially as shown in the examples in Table 3.6. While the example in the upper part of Table 3.6 demonstrates a combination of a cross-sentence correction followed by a span reduction correction, the example in the lower part

Pattern	**Frequ.**	**Example**
cross-sentence	52.7%	Question: Who is older Danny Green or James Worthy? (1) Daniel Richard "Danny" Green, Jr. (born June 22, 1987) is an American professional basketball player for the San Antonio Spurs of the National Basketball Association (NBA). (2) **James Ager Worthy** (born February 27, 1961) is an American professional basketball coach and former player, commentator, television host, and analyst.
span reduction	23.3%	Question: What philosophy related to creationism is Paul Nelson noted for? (1) Paul A. Nelson (born 1958) is an American philosopher of science noted for his advocacy of young earth creationism and **intelligent design** (2) Paul A. Nelson (born 1958) is an American philosopher of science noted for his advocacy of young earth creationism and **intelligent design**
span extension	21.3%	Question: Ronald Reagan and George H. W. Bush both held which position in office? (1) The presidency of Ronald Reagan began on January 20, 1981, when Ronald Reagan was inaugurated as **President of the United States**, and ended on January 20, 1989. (2) The presidency of Ronald Reagan began on January 20, 1981, when Ronald Reagan was inaugurated as **President of the United States**, and ended on January 20, 1989.
in-sentence	7.3%	Question: When was the stadium that held the 2015 Magyar Kupa demolished? (1) The stadium was closed in 2016 and demolished in **2017** to give place to the new Ferenc Puskas Stadium. (2) The stadium was closed in 2016 and demolished in **2017** to give place to the new Ferenc Puskas Stadium.
entity ref.	8%	Question: Which host of Sunday Night Safran has the hebrew first name Yehoshua? (1) John Michael Safran (Hebrew: "Yehoshua Safran" ; born 13 August 1972) is an Australian radio personality, satirist, documentary maker and author, known for combining humour with religious, political and ethnic issues. (2) It was hosted by **John Safran** and Catholic priest, Bob Maguire.
logic hops	4%	Question: Is the Pakistan fast bowler who joined the Kent County Cricket Club in June, 2011 a left-hand or right-hand batsmans? (1) Wahab Riaz (Punjabi, Urdu: ; born 28 June 1985) is a Pakistani cricketer. (2) He is a left-arm fast bowler and a **right-hand** batsman.
combined	9.3%	Question: Who was born in 1922 and published a book in 1985 by Delacorte Press? (1) Kurt Vonnegut Jr. (November 11, 1922; April 11, 2007) was an American writer. (2) Galapagos is the eleventh novel written by American author **Kurt Vonnegut** .

Table 3.5.: Correction patterns identified in 150 randomly sampled thought flows using $\delta = 1$. The **correct answer** is marked bold, the predicted answer per flow step is marked in orange. For each example, the wrong (1) and the corrected (2) prediction steps are shown.

Examples
Question: How many times did the man who coached the 1986-87 UNLV Runnin' Rebels fail to win 20 games in a season? (1) He spent the majority of his career coaching with the UNLV Runnin' Rebels, leading them four times to the Final Four of the NCAA Men's Division I Basketball Tournament, winning the national championship in 1990. (2) Overall, he won over 700 games in his career, and only **twice** failed to win 20 games in a season. (3) Overall, he won over 700 games in his career, and only **twice** failed to win 20 games in a season.
Question: Why did the CEO of the football team based in Denver, Colorado step down in 2014? (1) He served as the Broncos CEO from his purchase of the club in 1984 until July 2014, when he stepped down as Broncos' CEO **due to the onset and progression of Alzheimer's disease** . (2) He served [...], when he stepped down as Broncos' CEO **due to the onset and progression of Alzheimer's disease** . (3) He served [...], when he stepped down as Broncos' CEO **due to the onset and progression of Alzheimer's disease** .

Table 3.6.: Multi-step correction examples ($\delta = 1$).

illustrates how a span extension correction can iteratively correct a prediction. We additionally observe flow patterns with a very high number of decision changes. These typically correspond to two- or three-cycles between answer spans or exhibit a seemingly chaotic behavior.

3.2.3. Human Evaluation

While the previous experiments showed that our thought flow implementation can enable complex self-correction and can reach promising performance gains, we now investigate how the respective thought flow predictions affect human users in an AI-assisted QA task.

3.2.3.1. Experiment Design

We choose a within-subject design in which each participant is exposed to three variations of a QA system.[11]

Conditions. We aim at assessing the effect of the thought flow concept on users and present the outputs of the oracle-stopped thought flow in one condition (**TF**) and compare it to two baseline conditions. As baselines, we use top-1 predictions (**SINGLE**) (to compare against standard models) and top-3 predictions (**TOP-3**) (to compare to an alternative approach to show several predictions). For all conditions, we present the predicted answer(s) along with the sentence in which they appear in the context.

[11] We refer to our introduction to user studies and experiment designs in Section 2.3.

Dependent Variables. We study the effect of the condition (SINGLE, TF and TOP-3) on a set of dependent variables. We include variables on a per-question level (after each question) and on a per-system level (after all questions of one condition). The per-question variables include: (i) human answer correctness, (ii) perceived model correctness, (iii) perceived understanding, (iv) perceived helpfulness, and (v) completion time. The per-system variables include: (vi) usability using the UMUX questionnaire (Finstad, 2010, 2013), (vii) mental effort using the Paas scale (Paas, 1992), (viii) anthropomorphism using the respective subscale of the Godspeed questionnaire (Bartneck et al., 2009)[12], (ix) perceived intelligence using the subscale from the same questionnaire, (x) average completion time. We provide a full list of all questionnaires in Appendix A.2.

Apparatus. We sample 100 instances from the HotpotQA validation instances for which a thought flow using $\delta = 1$ causes at least one prediction change.[13] From these, we sample 30 instances per participant and randomly assign the instances to three bins of ten questions (one bin per condition).[14] We balance the six possible condition orders across participants and include three attention checks per participant. Figure 3.7 shows our user study interface for the TF condition. We provide screenshots of all conditions' interfaces in Appendix A.2.

3.2.3.2. Quantitative Results

We use MTurk to recruit US crowdworkers with >90% approval rate and the MTurk Masters qualification and collect responses from 55 workers.[15]

Statistical Models. We evaluate the collected responses using appropriate statistical tests.

Per-System Ratings: We analyze the per-system ratings using Friedman tests to account for the paired responses due to the within-subject design.[16] We use Holm-corrected Conover post hoc tests to identify significant pairwise differences.

Per-Item Ratings: Note that the within-subject design of our study possibly introduces interdependencies within ratings that we have to account for using an appropriate statistical model (see Section 2.4.2 for a deeper discussion). Additionally, our dependent variables are measured on different levels, e.g., completion time is measured on a ratio scale while human answer

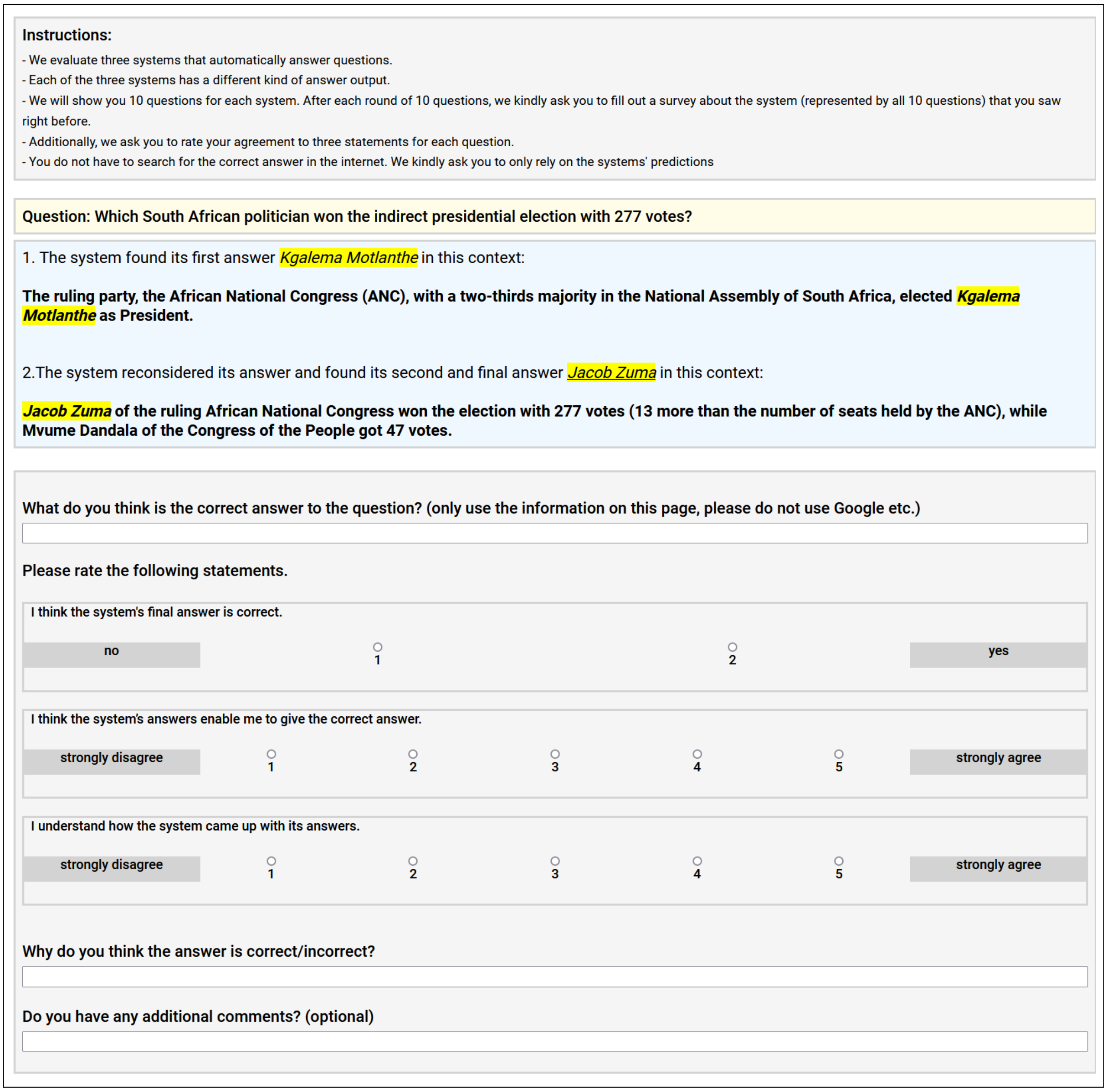
[12] We drop the robotics-specific item regarding "moving rigidly/elegantly" as it is not applicable to QA.

[13] If there is no prediction change, TF is identical to SINGLE.

[14] We statistically account for random effects of individual questions.

[15] We filter out two participants that did not pass the attention checks and collect two additional responses.

[16] Although aggregated Likert item scores are commonly considered interval responses, we use a Friedman test that only requires ordinal responses and is more conservative than its parametric counterpart RM-ANOVA.



Instructions:

- We evaluate three systems that automatically answer questions.
- Each of the three systems has a different kind of answer output.
- We will show you 10 questions for each system. After each round of 10 questions, we kindly ask you to fill out a survey about the system (represented by all 10 questions) that you saw right before.
- Additionally, we ask you to rate your agreement to three statements for each question.
- You do not have to search for the correct answer in the internet. We kindly ask you to only rely on the systems' predictions

Question: Which South African politician won the indirect presidential election with 277 votes?

1. The system found its first answer *Kgalema Motlanthe* in this context:

The ruling party, the African National Congress (ANC), with a two-thirds majority in the National Assembly of South Africa, elected *Kgalema Motlanthe* as President.

2.The system reconsidered its answer and found its second and final answer *Jacob Zuma* in this context:

***Jacob Zuma* of the ruling African National Congress won the election with 277 votes (13 more than the number of seats held by the ANC), while Mvume Dandala of the Congress of the People got 47 votes.**

What do you think is the correct answer to the question? (only use the information on this page, please do not use Google etc.)

Please rate the following statements.

I think the system's final answer is correct.

no ○ 1 ○ 2 yes

I think the system's answers enable me to give the correct answer.

strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 strongly agree

I understand how the system came up with its answers.

strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 strongly agree

Why do you think the answer is correct/incorrect?

Do you have any additional comments? (optional)

Figure 3.7.: User study interface showing the TF condition (ours).

Condition	Perceived Quality							User Performance	
	correct*	understand*	helpful*	usability	mental effort	humanlike*	intelligent*	time*	answer F1*
SINGLE	A	A	A	$\boxed{A}$	$\boxed{A}$	A	A	$\boxed{A}$	A
TOP-3	A	$\boxed{B}$	$\boxed{B}$	$\boxed{A}$	$\boxed{A}$	$\boxed{AB}$	$\boxed{B}$	B	B
TF	$\boxed{B}$	$\boxed{B}$	$\boxed{B}$	$\boxed{A}$	$\boxed{A}$	$\boxed{B}$	$\boxed{B}$	$\boxed{AB}$	$\boxed{C}$

Table 3.7.: Statistical results of our human evaluation ($N = 55$). "*" marks dependent variables on which a significant effect of the system condition was observed (Friedman tests and LRT tests for GLMMs/CLMMs). Pairwise differences between conditions (Holm-adjusted Tukey/Conover tests) are reported as compact letter displays (CLDs) (Piepho, 2004). E.g., the "humanlike" column shows that the post hoc test detected a significant difference between SINGLE and TF but no significant difference between any other pair. Similarly, the last column shows pairwise differences between all conditions and the TF condition reaches significantly higher human answer F_1-scores than any other condition. Variables for which TF is among the best performing models are marked cyan, variables for which it is found to be the sole superior system are marked green.

correctness is measured on a nominal (dichotomous) scale.[17] We, therefore, use GLMMs and cumulative link mixed models cumulative link mixed models (CLMMs) to (i) account for random effects of question and subject IDs, and (ii) account for the variables' respective measurement scales. We use GLMMs to analyze continuous and dichotomous responses (Gamma/binomial link) and CLMMs to analyze ordinal ones. We use a LRT between the full model and the model without the condition variable to identify main effects of the condition variable and conduct Holm-corrected Tukey post hoc tests.

Results. We find significant differences for all dependent variables except usability and mental effort. We summarize the results of our statistical analysis in Table 3.7 using compact letter displays (CLDs) (Piepho, 2004). Table 3.8 provides the p values for main effects and each pairwise comparison. In the following, we discuss our findings for each dependent variable for which we found a significant main effect.

Perceived Answer Correctness: While there is no statistically significant difference between showing users single answers or top-3 predictions, displaying thought flows leads to significantly higher answer correctness ratings.

Understanding: Top-3 as well as thought flow predictions significantly increased the feeling of understanding how the system came up with its answer compared to single predictions.

[17] We follow related work and treat Paas mental effort, UMUX, and Godspeed subscale responses as interval data but analyze single-item perceived understanding and helpfulness on an ordinal level.

	Perceived Quality							User Performance	
	correct*	understand*	helpful*	usability	mental effort	humanlike*	intelligent*	time*	answer F1*
Main effect	<**0.0001**	<**0.0001**	<**0.0001**	0.07968	0.6282	**0.03575**	**0.00124**	<**0.0001**	<**0.0001**
TF – SINGLE	<**0.0001**	<**0.0001**	<**0.0001**	0.13116	1	**0.03431**	**0.00586**	0.15304	<**0.0001**
TF – TOP-3	**0.00891**	0.8867	0.9994	0.84254	1	0.30556	1	0.06207	<**0.0001**
TOP-3 – SINGLE	0.51897	<**0.0001**	<**0.0001**	0.13653	1	0.25097	**0.00586**	**0.00012**	<**0.0001**

Table 3.8.: Detailed p values for all main effects and pairwise comparisons shown in Table 3.7. Significant p values are marked **bold**. The cell colors follow the color coding of Table 3.7.

Helpfulness: Similarly, top-3 and the thought flow predictions significantly improve perceived system helpfulness compared to single predictions.

Anthropomorphism: While we observe no significant difference in anthropomorphism ratings between single and top-3 predictions, the thought flow predictions are perceived as significantly more human-like/natural than the single answers.

Perceived Intelligence: Both, the top-3 and the thought flow predictions, lead to a significantly increased perceived system intelligence.

Completion Time: We observe that the top-3 predictions significantly improve completion times compared to single answers, but there is no significant increase for thought flows.

User Performance: While top-3 predictions already improve user performance in terms of F_1-score of the user's answer, thought flow predictions enable even higher performances, that are significantly higher compared to the single answer or top-3 conditions. We additionally analyze user answers using exact match scores and find the same effects and model orders.

Overall, our results indicate that *thought flows are better or equally good than single answer or top-3 predictions regarding all evaluated dimensions.* In particular for perceived answer correctness, humanlikeness, and user performance, thought flows are significantly better than both, the single answers and the top-3 predictions. While comparable (statistically indistinguishable) improvements of understanding, helpfulness, naturalness, and intelligence can also be achieved using top-3 predictions, these come at the cost of significantly increased completion times compared to single answers. In contrast, we do not find a significant time increase using thought flows.

3.2.4. Application to Image Classification

So far, we explored our thought flow method in the context of QA systems. As our method only requires a model to provide a vector representation of the model input and a differentiably-linked model output, it can be applied to the vast majority of classification models within as well as outside NLP. In the following, we demonstrate an application to image classification.

3.2.4.1. Vision Transformers on CIFAR

We use a pre-trained vision transformer model (Dosovitskiy et al., 2020) as base model and fine-tune the model on the CIFAR-10 and CIFAR-100 image classification datasets (Krizhevsky, 2009). We use the ViT-L-32 model variant pre-trained on the ILSVRC-2012 ImageNet and the ImageNet-21k datasets (Deng et al., 2009) as described by Dosovitskiy et al. (2020).[18]

As for our QA implementation discussed in Section 3.2.1.3, we have to specify our choice of logit vector $\hat{\mathbf{z}}$, input representation $\phi(\mathbf{x})$, correctness score s, and correctness score prediction function f_{corr}. While our QA span extraction model did yield two probability distributions (one for the start position and one for the end position), we now only have to consider a single distribution over image classes. Following our notation in Section 3.2.1.3, we thus define $\hat{\mathbf{z}}$ to be the predicted class logits. As input representation $\phi(\mathbf{x})$, we use the vision transformer's embedding of the [CLS] token as — in contrast to our QA model which used each token's embeddings — our image classifier only relies on the [CLS] embedding when predicting the image class. While we used F_1-score as correctness score in our QA experiments, we use a probability score s now, i.e., the correction module predicts a probability estimate that the label prediction is correct.[19] As for our QA implementation, we implement f_{corr} as a two-layer MLP with scaled exponential linear unit (SELU) activation. We train the correction module using CE loss. Overall, we train five models for each of the datasets using different random seeds.

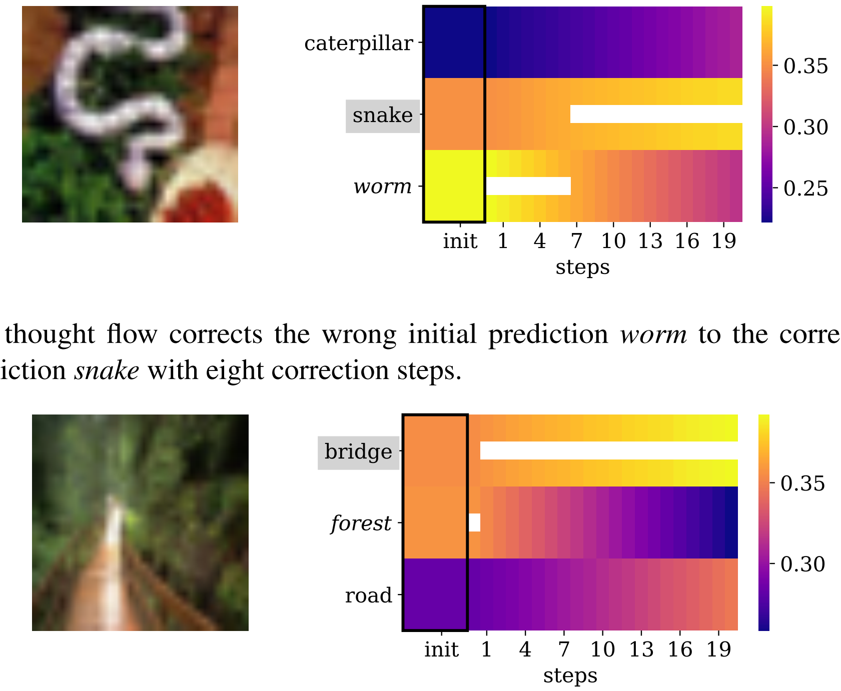
3.2.4.2. Error Correction Capability

We observe that applying our thought flow can successfully correct erroneous predictions. Figure 3.8 shows two examples. In Figure 3.8a, the wrong prediction *worm* is corrected to *snake* after eight gradient steps. Similarly, Figure 3.8b shows a correction from *forest* to *bridge*. While the probability mass is redistributed over the course of the thought flow, the class *road* gains probability as well which can be interpreted as a sensible "change of mind" as the central object could be a road on a bridge as well.

In terms of accuracy, our models yield consistent but small performance gains (<0.3% for both datasets). However, as our baseline models reach 98.7% (SD=0.7) accuracy on CIFAR-10

[18] The models are available via `https://github.com/google-research/vision_transformer`.

[19] We also experimented with predicting the label module's true class probability instead of correctness probability, similar to Corbière et al. (2019), but did not observe improvements over our setting.



(a) The thought flow corrects the wrong initial prediction *worm* to the correct prediction *snake* with eight correction steps.

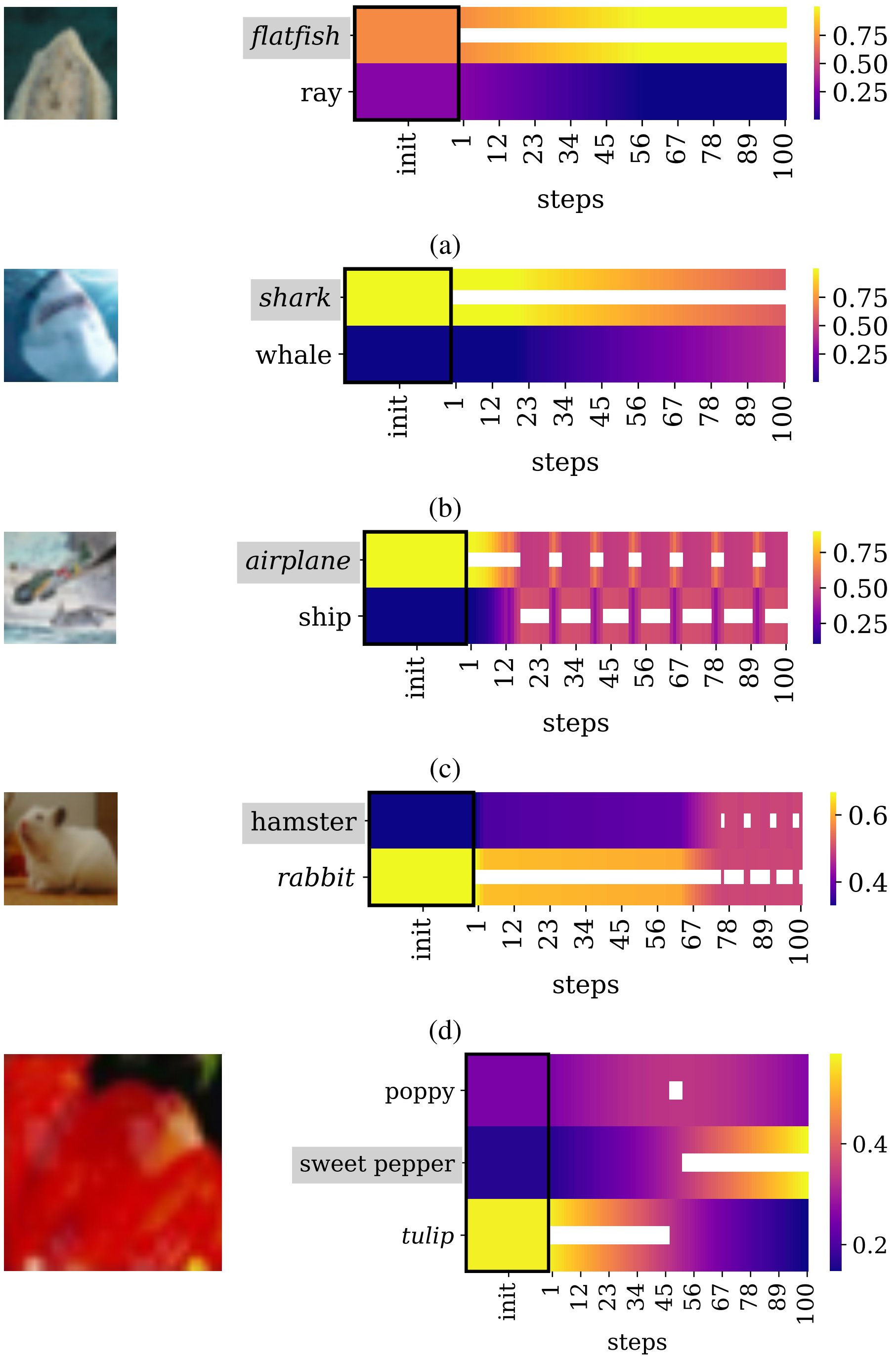
(b) The wrong (but plausible) label *forest* is corrected to *bridge*. Notably, the probability of *road* increases with the probability of *bridge*.

Figure 3.8.: Exemplary thought flows on CIFAR-100 instances. The black rectangle shows the initial class probabilities from the base model (step 0), i.e., the unmodified prediction, from a bird's eye perspective. The corresponding predicted label is marked in *italics*. On the right side of the black rectangle, the thought flow is depicted. The white lines mark the maximum probability across classes for each step. The ground truth label is marked with a gray box . For readability, we only show classes that reach a probability of at least 1% within the thought flow.

and 92.5% (SD=0.7) accuracy on CIFAR-100, there is much less room for improvement than in our QA experiments for which our base model reached 63.5% F_1-score.

3.2.4.3. Thought Flow Patterns

Similar to the qualitative analysis of flow patterns in our QA experiments (see Section 3.2.2.3), we now investigate the dynamics of the generated image classification thought flows. While Figure 3.8 shows thought flows that gradually transition from one class to another and then converge to that class, we observe diverse flow patterns which we display in Figures 3.9 and 3.10. Figure 3.9a shows an example for which our method does not change the (correct) label prediction but increases the model's confidence in its prediction. In Figure 3.9b, the



(a)

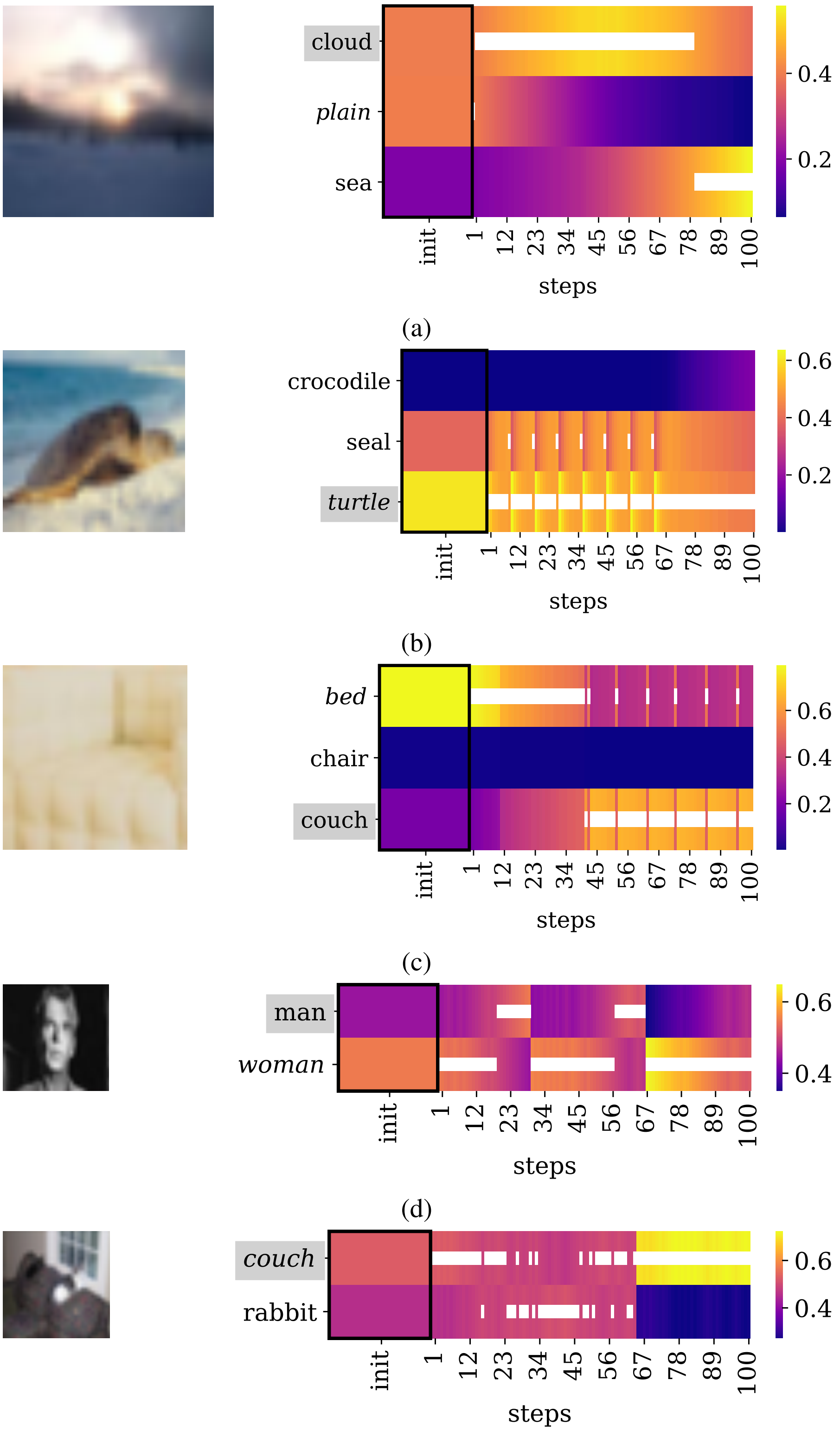
(b)

(c)

(d)

(e)

Figure 3.9.: Exemplary thought flows from different models on CIFAR demonstrating the diverse range of correction dynamics. A detailed description of the plots is provided in Figure 3.8.



(a)

(b)

(c)

(d)

(e)

Figure 3.10.: More exemplary thought flows from different models on CIFAR demonstrating the diverse range of correction dynamics. A detailed description of the plots is provided in Figure 3.8.

thought flow does not change the predicted label but decreases the model's confidence. Thus, the flows in Figures 3.9a and 3.9b can be interpreted as a form of neural network calibration (Guo et al., 2017). Figure 3.9d shows a smooth transition from one class to a gradual back-and-forth between two classes. In Figure 3.9e one can see a transition from the class *tulip* to the class *sweet pepper* via the class *poppy*. In Figure 3.10a, the thought flow quickly changes from *plain* to *cloud*. While the predicted class remains *cloud*, the probability of plain decreases continuously until the flow changes its prediction to *sea* which we interpret as *overthinking* (Kaya et al., 2019). Figures 3.9c and 3.10b to 3.10d show different periodic behaviors including the transition from a cycle to a fixed class in Figure 3.10b, smooth cycles in Figure 3.9c and longer sawtooth-like cycles in Figure 3.10d. Importantly, Figures 3.10b and 3.10e are examples for flows that explore an alternative class prediction but "return" to the initial class prediction and thus show that our method can be used to explore alternatives without necessarily neglecting a correct prediction.

Overall, we observe that our thought flow method is applicable beyond QA and can correct model predictions of image classifiers. As for the QA thought flow patterns discussed in Section 3.2.2.3, we observe numerous correction patterns that exhibit a surprisingly high complexity and motivate a deeper study of the correction dynamics in future work.

3.2.5. Overall Discussion

In this section, we introduced a task-agnostic self-correction formalism that turns a model's single output prediction into an evolving sequence of predictions — the *thought flow*. We take inspiration from Hegel's dialectics and propose a correction module along with a gradient-based update rule that sequentially updates a model's output distributions in the direction of an increasing self-estimate of correctness. We apply our method to QA models and conduct extensive experiments including human evaluation. We find that thought flows (i) can increase F_1-scores up to 9.3%, (ii) exhibit complex self-correction patterns, and (iii) provide significant improvements in human interaction and system perception including task performance and perceived system correctness and naturalness. Finally, we apply our thought flow method to image classifiers and (vi) demonstrate that it can correct model predictions using non-trivial correction patterns across input modalities. A potential next step to further improve performance is learning to stop.

3.3. Related Work

In the following, we discuss (i) prior work related to external knowledge, explanation generation, and (human) evaluation of (explainable) NLI systems (Section 3.3.1) and (ii) work on cognitive modeling, confidence estimation, (sequential) prediction, and model correction (Section 3.3.2).

3.3.1. External Knowledge and Explainable NLI

External Knowledge for NLI. External knowledge was shown to help a variety of NLP tasks (Shi et al., 2016; Seyler et al., 2018; Pan et al., 2019; Lin et al., 2019a). While early sources for external knowledge are WordNet and NomBank (Jijkoun and de Rijke, 2005; MacCartney et al., 2008), today, a large number of sources exists: From COMET (Bosselut et al., 2019) over ConceptNet (Speer et al., 2017) to language models. Chen et al. (2018) show that enriching an NLI system with external lexical-level semantic knowledge increases accuracy scores on SNLI and enhances transfer to MultiNLI. Wang et al. (2019) show the potential of knowledge from ConceptNet for NLI systems. Li et al. (2019) find that external knowledge from pre-training helps NLI and suggest to combine it with external knowledge from human-curated resources. Li and Sethy (2019) propose knowledge-enhanced attention modifications for Transformers and decomposable methods and show that their methods improve model robustness. Faldu et al. (2021) extend BERT by extracting entities from the input text and adding their projected KG embeddings derived from ConceptNet and WordNet as sequential input to a modified BERT layer. Bauer et al. (2021) present ERNIE-NLI, a modified ERNIE Zhang et al. (2019) model using NLI-specific knowledge embeddings and find that it improves performance over a non-adapted ERNIE model using general-domain TransE embeddings. We propose various models to compare different possibilities of integrating external knowledge and address the question of whether external knowledge also improves explanation generation.

Explainable NLI. The task of explainable NLI consists of (i) predicting the correct entailment label and (ii) providing an explanation that allows the user to assess the model's reasoning. In general, such explanation can take various forms, such as weights and gradients over the input (Simonyan et al., 2014; Ribeiro et al., 2016; Lundberg and Lee, 2017) and text spans or snippets from the input or external text (Zaidan and Eisner, 2008; Lei et al., 2016; Yang et al., 2018). Beyond that, there exist various resources and approaches designed to generate textual explanations. Rajani et al. (2019) present a dataset that contains free-text explanations for multiple-choice commonsense reasoning and Bhagavatula et al. (2020) provide a dataset for abductive multiple-choice answering as well as abductive NLG. Camburu et al. (2018) provide the e-SNLI dataset, which adds free-text explanations as an additional layer on the SNLI dataset (Bowman et al., 2015). As numerous models with and without external knowledge have been developed on the SNLI dataset, we use its explainable extension e-SNLI to conduct our analysis and train our models. Various models have been proposed on e-SNLI including systems based on alignment (Swanson et al., 2020), label-specific explanation generators (Kumar and Talukdar, 2020), and fine-tuned text-to-text models (Narang et al., 2020). In contrast to those, our focus is not on proposing a new architecture or paradigm to develop a high-scoring system. Much more, we seek to conduct a broad comparison across knowledge sources and isolate their effect on automatic scores as well as human perception.

Evaluation and Human Ratings of Explainable NLI. Explainable NLI system performance is typically scored using (i) accuracy with respect to annotated gold labels on a reference dataset and (ii) BLEU scores (Papineni et al., 2002) between the generated explanations and the ground truth explanations (Camburu et al., 2018; Kumar and Talukdar, 2020; Narang et al., 2020). BLEU scores can only quantify explanation quality loosely (Narang et al., 2020). Therefore, previous work evaluates explanation quality either by manual annotation (Camburu et al., 2018; Kumar and Talukdar, 2020) or crowdsourcing (Narang et al., 2020). However, previous human evaluations regarding explainable NLI are limited to assessing label and/or explanation correctness. In contrast, we additionally evaluate commonsense inclusion as well as grammatical correctness of explanations. As Clinciu et al. (2021) find automatic BLEURT scores to have distinctly stronger correlations to human ratings of generated explanations than BLEU, we investigate whether BLEURT is a viable replacement for a user study.

3.3.2. Thought Flow Networks

Cognitive Modeling and Systems. Regarding our thought flow methodology, the fields of cognitive modeling and cognitive systems provide numerous models of human thinking (Rupert, 2009; Busemeyer and Diederich, 2010; Levine, 2018; Lake et al., 2017). While work in these fields often orients towards accurate descriptions of human cognition, our method does not aim to provide a plausible description of cognitive process but, instead, applies a philosophical concept to machine learning to improve system performance and user utility.

Confidence Estimation and Model Corrections. Estimating a model's confidence and the correctness of its predictions is addressed with various methods, including the training of secondary models for predicting the main model's uncertainty (Blatz et al., 2004; DeVries and Taylor, 2018). Among those, *ConfidNet* is particularly related to our approach as it predicts the true-class probability of the main model (Corbière et al., 2019). In contrast, our correction module receives the class probabilities of the main model as input and predicts a correctness score. In contrast to methods aiming at estimating accurate confidence scores, we predict such scores only as an auxiliary signal in order to generate a gradient that allows us to update the model prediction. Regarding model correction, the arguably most established approach to learn corrections of model predictions is gradient boosting (Friedman, 2001) including its popular variant XGBoost (Chen and Guestrin, 2016). In contrast to those works, we do not use an ensemble of weak learners but propose a lightweight correction module that is applicable on top of any existing classification model. Further, in our method, the correction module receives the main model's predictions and is able to directly adapt them.

Sequences of Predictions. The idea of iteratively predicting and correcting has been explored for a long time. Early work includes Mori *et al.* who present a non-neural iterative correction method tailored to estimate elevation maps from aerial stereo imagery (Mori et al., 1973). Katupitiya and Gock (2005) propose to iterate two neural networks to address the problem of predicting inputs of a mechanical process given the outputs of the process. While their method is specifically designed for the task of input prediction, our work presents a general-purpose classification model that iterates class label predictions. Besides those task-specific methods, there are models and inference methods that make use of an iterative prediction process by design, such as Hopfield networks (Hopfield, 1982) and their modern variants (Barra et al., 2018; Ramsauer et al., 2020), or Loopy Belief Propagation, Markov Chain Monte Carlo or Gibbs sampling (Bishop, 2006; Koller and Friedman, 2009). While these techniques can be linked to our work conceptually, they all require to train a new model. In contrast, our approach can be applied to an existing neural model as well. Another related approach is chain-of-thought prompting (Wei et al., 2022) in which a language model is prompted with demonstrations of problem decomposition/reasoning in a few-shot manner and subsequently can be observed to show similar behavior in its answer. While this method yields impressive model answers, it predicts *one* answer that contains information on its deduction without changing or correcting its answer. In contrast, our method is not targeted towards decomposition/reasoning but predicts a sequence of answers with the goal of iteratively improving it.

Learning to Stop. A further line of work, including Graves (2016) and Banino et al. (2021), trains RNNs to learn when to stop applying recurrent transformations within the model. While their approaches require the model to contain recurrent components and to retrain the model, our method only requires the model to yield output logits and leaves the base model unchanged.

4. Evaluating and Quantifying Explainability

In this chapter, we discuss the limitations of current evaluation scores used to quantify explanation quality for explainable QA systems and present two novel proxy scores (Section 4.1), propose general characteristics of explanation quality (Section 4.2), demonstrate how current evaluation practices violate them resulting in an alarming disconnect between automatic evaluation and human evaluation (Section 4.3), and propose general guidelines and a novel ranking approach to alleviate the challenges that explanation quality evaluation faces (Section 4.4).

4.1. Proxy Scores to Quantify Explanation Quality

In this section, we introduce how explainable QA systems are evaluated to date and why the respective explanation evaluation is insufficient (Section 4.1.1), present two novel proxy scores that quantify answer-explanation coupling (Section 4.1.2), and demonstrate that these scores can reflect an explanation's utility to users better than current scores do (Section 4.1.3.2).

4.1.1. Limitations of Current Evaluation Scores

The performance of explainable QA systems is quantified regarding two aspects: (i) the QA performance and (ii) the explanation quality. For QA performance, typical QA proxy scores, such as F_1-score, exact match (EM), precision, or recall are used. These scores quantify the word-level overlap between a predicted answer with a ground truth answer annotation. For explanation quality, the same proxy scores have been applied on a sentence level (Yang et al., 2018), i.e., the scores quantify the overlap of binary relevant/irrelevant decisions over the sentences in a given context. The HotpotQA leaderboard aggregates the QA scores with explanation scores and uses joint-F_1 as the leaderboard ranking criterion. In the following, we discuss why this choice can be problematic.

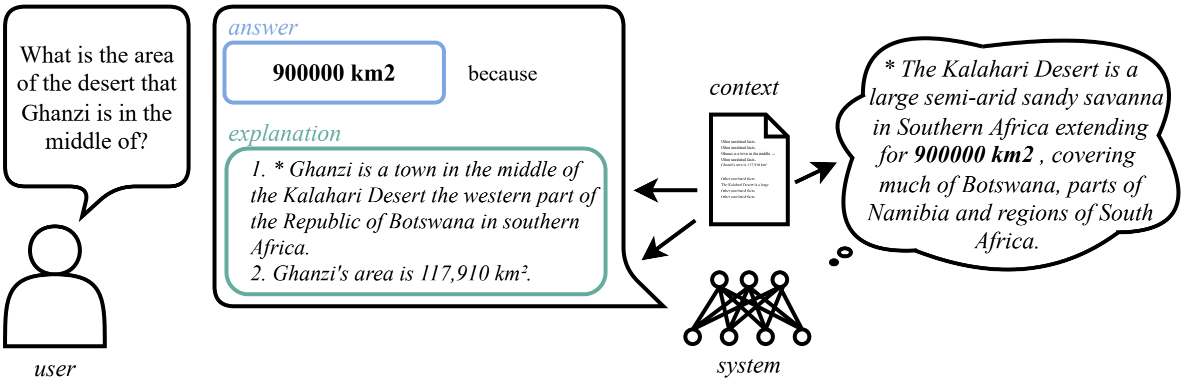

What is the area of the desert that Ghanzi is in the middle of?
user
answer
900000 km2
because
explanation
1. * Ghanzi is a town in the middle of the Kalahari Desert the western part of the Republic of Botswana in southern Africa.
2. Ghanzi's area is 117,910 km².
context
system
* The Kalahari Desert is a large semi-arid sandy savanna in Southern Africa extending for 900000 km2 , covering much of Botswana, parts of Namibia and regions of South Africa.

Figure 4.1.: Example prediction from the HotpotQA explainable QA dataset. The model returns the correct answer (blue box) but its predicted explanation, i.e., selection of supporting facts (green box), is only partially correct as it (a) reports an irrelevant fact about the size of Ghanzi and (b) fails to report the relevant fact containing the predicted answer. "$*$" marks facts within the human-annotated ground truth explanation. How can the resulting (lack of) explanation quantity be evaluated meaningfully? Example data from Yang et al. (2018).

No Empirical Evidence. There is no empirical evidence that joint-F_1 is related to user performance or experience regarding explainable QA. While F_1-score is a well-established score across NLP, there is — to the best of our knowledge — no demonstration of a strong relation between F_1-scores and human-perceived quality in explainable NLP. In fact, as we will demonstrate in Section 4.3.2, there only is a moderate correlation between joint-F_1 and various dimensions of human-perceived quality or utility.

Rewarding Poor Explanations. Figure 4.1 shows an example prediction that is rewarded with a joint-F_1 of 0.5 although its explanation provides no value to the user. The reward stems from the overlap of the explanation with the ground truth but does not consider that the predicted answer is not contained in any of the predicted relevant facts.

Punishing Good Explanations. Consider a model output in which the predicted answer is wrong but the explanation perfectly explains this wrong answer, showing to the user why the model has selected it. Standard F_1-scores compare the model output to the ground truth annotations and will, therefore, score both the answer and the explanation with an F_1 of 0. However, we argue that an explanation should be evaluated with a score higher than zero if it is able to explain the reasoning process of the model to the user and, thus, lets the user identify the failure of the model.

4.1.2. Novel Scores: FARM and LOCA

In the following, we propose two new proxy scores to quantify answer-explanation coupling within explainable QA systems. Our FARM and LOCA scores build upon and unify the scores proposed in our prior work (Schuff, 2020). We review the relevant fine-grained scores and describe how they lead to our final FARM and LOCA scores in the following.

4.1.2.1. Fact-Removal Score (FARM)

Ideally, the explanation of the model includes all facts that the model uses within its reasoning chain but no additional facts beyond that. Note that even for a wrong model answer, this assumption should hold so that the relevant facts provide explanations for the (wrongly) predicted answer.

To quantify the degree of answer-explanation coupling, we propose to iteratively remove parts (individual facts) from the explanation, re-evaluate the model using the reduced context, and track how many of the model's answers change. For a model with perfect coupling of answer and explanation, the answer will change with the first fact being removed (assuming no redundancy) but will not change when removing irrelevant facts not belonging to the explanation. We remove facts in order of decreasing predicted relevance as more relevant facts should influence the model's reasoning process the strongest.

In the following, we denote an instance of the data set by $e \in E$ with its corresponding question e_{ques} and context e_{con}. We use $\text{answer}(\cdot, \cdot)$ to denote the answer that a model predicts for a given question and context. The functions $\text{reduce}_{\text{rel}}(\cdot, k)$ ($\text{reduce}_{\text{irr}}(\cdot, k)$) return a context from which up to k facts the model predicts to be relevant (irrelevant) have been removed.[1] We re-evaluate the model on this reduced context and calculate the fraction of changed answers $c_{\text{rel}}(k)$ and $c_{\text{irr}}(k)$, respectively.

$$a(e) = \text{answer}(e_{\text{ques}}, e_{\text{con}}) \tag{4.1}$$

$$\hat{a}_{\text{rel},k}(e) = \text{answer}(e_{\text{ques}}, \text{reduce}_{\text{rel}}(e_{\text{con}}, k)) \tag{4.2}$$

$$\hat{a}_{\text{irr},k}(e) = \text{answer}(e_{\text{ques}}, \text{reduce}_{\text{irr}}(e_{\text{con}}, k)) \tag{4.3}$$

$$c_{\text{rel}}(k) = \frac{|\{e \in E : a(e) \neq \hat{a}_{\text{rel},k}(e)\}|}{|E|} \tag{4.4}$$

$$c_{\text{irr}}(k) = \frac{|\{e \in E : a(e) \neq \hat{a}_{\text{irr},k}(e)\}|}{|E|} \tag{4.5}$$

[1] If the number of facts predicted as (ir)relevant is less or equal to k, we remove all (ir)relevant facts.

Finally, we condense $c_{\text{rel}}(k)$ and $c_{\text{irr}}(k)$ into a single fact-removal score:

$$\text{F}\textsc{a}\text{RM}(k) = \frac{c_{\text{rel}}(k)}{1 + c_{\text{irr}}(k)} \in [0, 1] \tag{4.6}$$

$\text{F}\textsc{a}\text{RM}(k)$ ranges between zero and one and a higher score reflects a better explanation.

4.1.2.2. Location-of-Answer Score (LocA)

A second important indicator for the degree of a model's answer-explanation coupling is the location of the answer span: As shown in Figure 4.1, the model can predict answers that are located outside the facts it predicts to be relevant, i.e., outside the explanation. We argue that this is confusing for users. Therefore, we consider the fractions of answer spans that are inside the explanation of the model and the fraction of answer spans that are outside. For an ideal model, all answer spans would be located inside the explanation. We use I and O to denote the number of answers inside/outside of the set of facts predicted as relevant. A denotes the total number of answers.[2]

Based on these counts, we propose the answer-location score that we define as

$$\text{L}\textsc{oc}\text{A} = \frac{\frac{I}{A}}{1 + \frac{O}{A}} = \frac{I}{A + O} \in [0, 1]. \tag{4.7}$$

LocA ranges between zero and one, larger values indicate better answer-explanation coupling.

4.1.3. Comparison with Established Scores and Human Evaluation

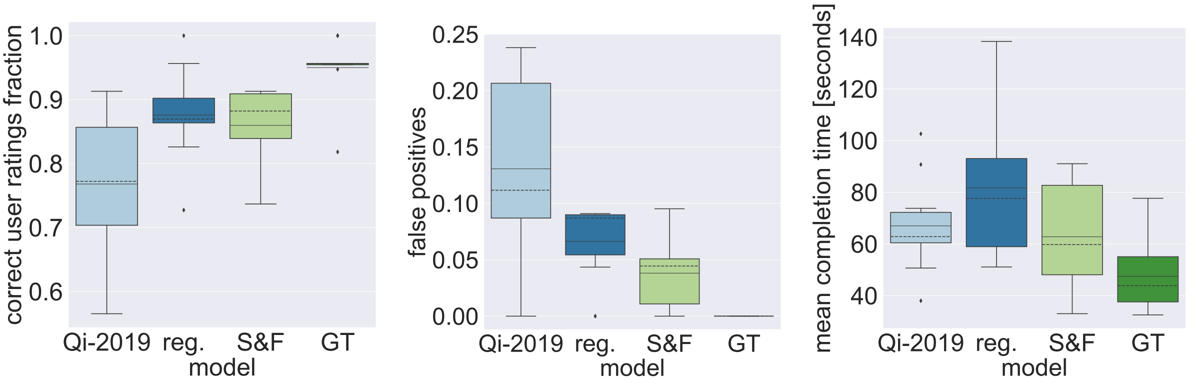
4.1.3.1. Comparison with Established Scores

We compare our proposed proxy scores to the widely adopted proxy scores SP-EM, SP-F$_1$, SP-recall, SP-precision, and the respective joint answer-explanation scores used by Yang et al. (2018). We evaluate all scores on three question answering models: the model proposed by Qi et al. (2019) (Qi-2019) and two models we proposed in prior work (Schuff, 2020). Concretely, these two models are a "select and forget" model (S&F), that performs supporting fact selection before predicting the answer on a reduced context, and a model that was regularized with an answer-explanation coupling term during training (REG).

[2] In HotpotQA, answers can stem from article titles although titles are never used as relevant facts. Thus, $A > I + O$ is possible. Our score is still applicable in this case.

	Proxy score	**Model**		
		Qi-2019	S&F	REG
established	SP-EM	39.81	**42.16**	25.98
	SP-F_1	79.34	**80.07**	75.60
	SP-P	78.01	**78.84**	66.79
	SP-R	85.26	85.45	**93.26**
	Joint-EM	22.28	**22.78**	14.56
	$\boxed{\text{Joint-F}_1}$	**52.51**	50.71	49.66
	Joint-P	**53.33**	51.40	45.64
	Joint-R	57.92	55.61	**62.09**
ours	FaRM(4)	66.20	**75.54**	73.32
	LocA	60.49	**70.60**	67.92

Table 4.1.: Comparison of three explainable QA models regarding established explainable QA proxy scores (upper part) and our proposed scores (lower part). **Bold** numbers mark highest (best) values. Scores are calculated on the distractor dev set of the HotpotQA dataset. The $\boxed{\text{box}}$ marks the main leaderboard score used to rank models in the official HotpotQA leaderboard Yang et al. (2018). We observe that the established proxy scores and our novel scores yield contradicting conclusions regarding what the best model is. SP refers to supporting facts.



(a) Correct user ratings.

(b) False positive ratio.

(c) Completion times.

Figure 4.2.: Box plots showing the results of the model comparison from the user study we conducted in prior work (Schuff, 2020). Boxes mark quartiles, whiskers mark 1.5 inter-quartile ranges, outliers are plotted separately. Horizontal solid/dashed lines within boxes mark means and medians, respectively. GT refers to ground truth answers and explanations. Plots are reproduced from prior work (Schuff, 2020).

Table 4.1 shows that both of our novel scores rank the S&F model in the first place. In contrast, joint-F_1, which is the main ranking criterion in the official HotpotQA leaderboard, ranks the S&F model second and yields the highest score for the QI-2019 model, which in turn is ranked clearly last following our scores. So far, we only can conclude that the established scores and our novel scores yield contradicting conclusions. Without additional information quality information about the four models, we cannot argue that either of the established/novel scores captures some aspect of explanation quality better than the other. For this, we additionally consider human quality ratings.

4.1.3.2. Comparison with Human Evaluation

Preceding User Study. We re-analyze the human evaluation we conducted in prior work (Schuff, 2020). We previously collected ratings and interaction signals from 40 participants. Each participant was exposed to one of the three models or the ground truth "model" in a unifactorial between-subjects experiment design and was tasked to answer 25 questions using the respective model's answer and explanation predictions. We manually assessed whether participant answers were correct and derived multiple dependent variables from the participants' answers and interactions. Concretely, we collected completion times, several performance variables indicating how well they judged the correctness of the model (fraction of correct ratings, false positive ratio (FP), false negative ratio (FN), true positive ratio (TP), true negative ratio (TN), precision (P), recall (R) and F_1-scores), agreement (fraction of model predictions

Human Eval.	Established Scores								Proposed Scores		
	Supporting Facts				Joint				FARM		LocA
	EM	F_1	P	R	EM	F_1	P	R	FARM(1)	FARM(4)	LocA
correct decision				+		-	-		+		
overestimation										-	-
completion time	-	-	-		-			+			
human-FP										-	-
human-TP				+		-	-		+		
human-FN	+	+	+		+			-			
human-TN				+		-	-		+		
human-P										+	+
human-R				+		-	-		+		
human-F_1				+		-	-		+		

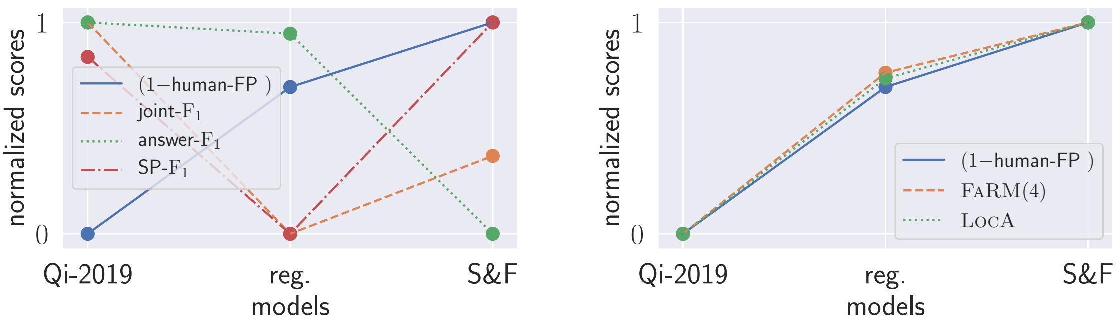
Table 4.2.: The table shows whether sorting the conditions by a human score (rows) and an automatized score (columns) result in the same order (+), the inverse order (-), or a different order (blank cell). Green (🟩) cells with boxes mark desirable relations, red (🟥) cells without boxes mark undesirable relations.

that the users rate as correct (Bussone et al., 2015)), and overestimation (difference between agreement and true model accuracy (Nourani et al., 2019)). Furthermore, we collected the following variables in self-reports with five-point Likert scales: certainty of the participants (Greis et al., 2017a), completeness and helpfulness of the explanations (Nourani et al., 2019), trust of the participants in the model (Bussone et al., 2015), and satisfaction (Kulesza et al., 2012; Greis et al., 2017b).

Relevant Results. Figure 4.2 displays the resulting fractions of correct user ratings, erroneous answer acceptance, and mean completion times per model. Notably, we observed that compared to the QI-2019 model, the S&F model (a) increased the fraction of correct user answers by 9.17%, (b) decreased (i.e., improved) the fraction of erroneous answer acceptance (FPs) by 9.25%, and (c) lowers the completion time by 4.2 seconds per questions on average.[3] These results indicate that the model of choice following our scores (i.e., the S&F model) would have been a better choice than the model choice following the joint-F_1-score (i.e., the QI-2019 model).

Relation Between Proxy Scores and Human Ratings. While the previous observations only consider the relation between the QI-2019 model and the S&F model, we now

[3] We refer to our discussion in Schuff (2020) for a detailed analysis of the additional dependent variables.



(a) Established explainable QA scores.

(b) Our proposed scores.

Figure 4.3.: Model score comparisons between human-false positives (FPs) and model scores. All scores are normalized to $[0, 1]$. We show $(1-$human-FP$)$ as fewer FPs are better. The left figure shows that F_1 poorly correlates to human performance. The right figure shows a much stronger correlation for our proposed scores.

extend the analysis to all four models (including the ground truth predictions). For this, we investigate the correlation of human ratings with model evaluation scores based on the model ranks they produce. Concretely, we rank the models by (i) human measures obtained in the user study and (ii) model evaluation scores. In Table 4.2, a cell is marked with a "+" if the ranking with respect to the human measure and the model score is identical (e.g., the ranks regarding human-FPs and answer-F_1 are identical). If the ranks are exactly reversed, we mark the cell with a "-". All other cells are left empty. "+" and "-" both indicate a perfect rank-wise correlation and do not imply that one is preferable over the other.

Next, we consider whether selecting a model based on the different model scores would result in a desired change in human evaluation scores or not. This depends on whether a high score (e.g., F_1-score) or a low score (e.g., the fraction of answers outside the predicted relevant facts) is aimed for. We indicate desired model selection with green boxed cells and undesired model selection with red cells (e.g., choosing a model with a higher answer-F_1 would result in a model with more human-FPs, which is not desired).

All F_1-scores show at least one undesirable rank relation. Notably, joint-F_1 is among the least aligned scores. In contrast, our scores have only desirable relations. In particular, FARM(4) and LOCA lead to a model ranking that is inverse to the ranking by human-overestimation and human-FPs. This is also confirmed in Figure 4.3, which shows how the human-FP ratio varies in comparison to the three F_1-scores (left plot: no correlation) and to our proposed scores (right plot: correlated). We provide the respective plots for all dependent variables in Appendix A.3.

4.1.4. Overall Discussion

Overall, we find that (i) our newly introduced FARM and LOCA scores contradict established proxy scores, such as joint-F_1 regarding model preference decisions, and (ii) our results indicate that our proposed scores predict user behavior better than the established scores regarding, i.a., the correctness of AI-assisted user decisions and false user trust into model predictions. Our findings highlight that developing new evaluation scores to capture explanation quality can be equally or even more important than developing new systems to increase existing scores. We investigate characteristics of explanation quality evaluation in Section 4.2 and extend our study of LOCA, joint-F_1, etc. in an in-depth analysis of HotpotQA leaderboard system submissions to explore the challenges current evaluation practices face in Section 4.3.

4.2. Characteristics of Explanation Quality

The two proxy scores introduced in the previous section are motivated by task-specific properties of explanation quality for the *specific* case of explainable (extractive) QA models. In this section, we explore which *general* properties explanation quality has.

Criteria for high-quality explanations have mainly been discussed in social sciences so far. Besides desirable explanation features, such as coherence (Thagard, 1989; Ranney and Thagard, 1988; Read and Marcus-Newhall, 1993), soundness, or completeness (Kulesza et al., 2013), literature has pointed out the importance of the explainees (Miller, 2019; Wang and Yin, 2021) and their goals (Vasilyeva et al., 2015). Based on these, we discuss characteristics of explanation quality in NLP in this section. We assume the faithfulness of an explanation and only focus on characteristics for its *perceivable quality*.[4] In the following, we focus on two characteristics: user-dependence (Section 4.2.1) and multidimensionality (Section 4.2.2).

4.2.1. Explanation Quality Is User-dependent

We argue that in AI, an explanation exists only in relation to a system that should be explained (the *explanandum*) and the human that receives the explanation (the *explainee*). This statement is in line with the social process function of an explanation described by Miller (2019) referring to the conversational model of explanation of Hilton (1990). Hilton argues that an explanation should be considered a conversation and emphasizes that "*the verb to explain is a three-*

[4] We consider explanation characteristics that can be judged without access to the underlying model. We refer to Jacovi and Goldberg (2020) for a discussion of faithfulness evaluation and to Liao et al. (2022) for a distinction between model-intrinsic and human-centered explanation properties.

place predicate: Someone explains something to someone" (Miller, 2019, p. 29). Given that explanations are always targeted towards a specific user group, we argue that their quality needs to be assessed accordingly. In the following, we detail how user goals, individual user characteristics as well as general properties of human perception impact the definition of how *good* an explanation is and why such a definition can never be universal.

Goals of Target Users. Vasilyeva et al. (2015) showed that users' perception of explanation quality depends on their goals. Similarly, Liao et al. (2022) found that users' usage context affects which explanation quality properties they consider to be important. While, for example, an explanation in the form of a heatmap over a text (as shown in the first row of Table 2.2) might be sufficient for an NLP developer or researcher who aims at analyzing and improving the system, it might not fit the needs of an end-user who has no machine-learning background but uses the system in practice. Although the explanation contains the same information, its perceived quality might be considered lower by end-users compared to developers because, for example, the mental effort to process the explanation could be higher for end-users that are unfamiliar with such visualizations.

Individual Differences of Target Users. In addition to the users' goals, their background knowledge affects which type and extent of explanations are most useful for them (Preece et al., 2018; Yu and Shi, 2018; Suresh et al., 2021). As a trivial but illustrative example, a perfect explanation in Spanish is clearly useless to a monolingual English speaker, and an "explanation" as it is provided by the coefficients of a linear model is useless to a user with dyscalculia. Concretely, prior work showed that, i.a., (a) an increase in users' education levels and technical literacy corresponds to an increased algorithm understanding (Cheng et al., 2019), (b) users' need for cognition (NFC) (i.e., their motivation to engage in effortful mental activities) impacts how much they benefit from interventions that increase analytical engagement with explanations (Buçinca et al., 2021), and (c) the effect of explanation on users strongly depends on the users' domain knowledge (Wang and Yin, 2021).

Intersubjective Quality within User Groups. While the individual goals and characteristics of each user make them perceive and use explanations in a unique way, certain groups of "similar" explainees (e.g., Spanish native speakers reading a Spanish text) will be affected by explanations similarly. We argue that explanation quality is an *intersubjective* construct. This has two immediate implications. First, it implies that every evaluation of explanation quality is limited to a specific group of explainees. However, it also implies that explanation quality can be *objectively* assessed within a suitable group of explainees. For example, an often-used categorization in explainability is to divide users into three groups: developers, domain experts, and lay users (Ribera and Lapedriza, 2019). Dividing users into such high-level groups can already help to identify important differences regarding their explanation needs, however, a more fine-grained categorization including, e.g., social and cognitive user properties — as suggested by Jacovi et al. (2022) — could further improve evaluation quality. We will revisit the influence of cognitive user properties in terms of a user's NFC in Section 5.2.

Cognitive Biases and Social Attribution. Hilton's conversational model of explanation distinguishes two stages: (a) the diagnosis stage in which causal factors of an event/observation are determined and (b) the explanation presentation stage in which this information is communicated to the explainee Hilton (1990) (we refer to Miller (2019) for a more detailed discussion of Hilton's work in the context of explainability). So even if the first stage is successful (i.e., the right "explanation information" has been identified), *communicating* the explanation information can fail (e.g., by relying on an inappropriate visualization to visualize the information). We empirically demonstrate that such problems in explanation communication can occur for heatmap explanations over text in Section 5.1 and that the information that users understand from these explanations is distorted by unrelated factors, such as word length. Similarly, Gonzalez et al. (2021) show that belief bias (i.e., a particular cognitive bias) affects which explanation method users prefer. More broadly, Jacovi et al. (2022) propose a framework of social attribution by the human explainee that describes which information an explainee is comprehending and thereby allows to identify failures of explainability methods.

4.2.2. Explanation Quality Has (Orthogonal) Dimensions

Explanation quality is commonly treated as a monolithic construct in which explanations can be ranked along a unidimensional range of explanation "goodness". We, in contrast, argue that there are different dimensions of explanation quality which also can be orthogonal to each other. Thus, explanations should be evaluated along *multiple* facets of explanation quality.

An example of two orthogonal quality dimensions are faithfulness and plausibility. Consider an explanation that explains the decision process of a system A in a way that (a) faithfully reflects the system decision process and (b) plausibly convinces a user of the correctness of the prediction. We then replace the system with a new system B while keeping the explanation constant. The explanation will still be plausible to the user (it did not change), however, if system B has a different decision process, the explanation is not faithful anymore as it no longer reflects the model's inner workings. Consequently, the two explanation quality dimensions faithfulness and plausibility can be independent and cannot be captured with the same score. We refer to Jacovi and Goldberg (2020) for a detailed comparison of faithfulness and plausibility.

Similarly, an explanation that is *perceived to be helpful* by explainees does not actually have to *be helpful* for them. Buçinca et al. (2020) showed that between two decision support systems, users preferred one system (in terms of rating it as more helpful and trusted), although their actual performance was significantly better with the less-favored system. Also, their subjective ratings were not predictive of their objective performance with the system. In their follow-up work, Buçinca et al. (2021) found a trade-off between subjective system quality ratings and effective human-AI performance for explainable AI systems. Related effects have been reported by Scharrer et al. (2012) who compared the impact of showing easy versus difficult scientific arguments to lay people and found that the easy arguments lead to participants being more convinced and underestimating their own knowledge limitations. These findings suggest that effective explanations have to combine or balance (a) *perceived* utility and (b) *actual* utility to their users. While an explanation that only subjectively seems to provide a benefit clearly is not desirable, an explanation that affects users to their own benefit but is disliked by them will not be used in practice, as, e.g., Nadarzynski et al. (2019) found AI acceptability to be correlated with, i.a., perceived utility, and trustworthiness. We explore the relation between perceived system predictability and objective system behavior prediction capability as a concrete instance of this objective-subjective relation in Section 5.2. Overall, effective explanation evaluation thus has to account for numerous, partially orthogonal dimensions of explanation quality.

4.3. Shortcomings of Current Evaluations

In the following, we present common evaluation practices and assess to which extent they conflict with the explanation quality characteristics presented in Section 4.2. Figure 4.4 provides an overview of the main challenges discussed in this section. Before we present our arguments on how current explainability quality evaluations fall short, we introduce our case study which we will refer back to throughout the remainder of this chapter.

4.3.1. Case Study on the HotpotQA Leaderboard

To support the following discussion with empirical evidence, we conduct a crowdsourcing study analyzing systems from *10 real models* submitted to the official HotpotQA (Yang et al., 2018) leaderboard that ranks explainable QA models.[5]

[5] We thank the HotpotQA maintainers for providing us with the predictions and the system submitters for giving us their consent to include their model in our case study.

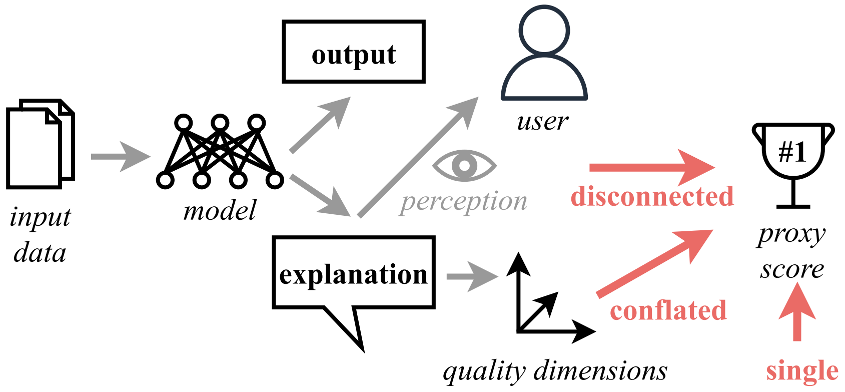


Figure 4.4.: Overview of the main drawbacks of current evaluation practices: (i) Disconnect of proxy scores and user perception, (ii) conflation of multiple dimensions into single proxy scores, and (iii) single-score leaderboards.

4.3.1.1. Task, Models, and Automatic Evaluation

In the following, we present the leaderboard models we analyze and list proxy metrics that we use to automatically quantify the models' explanation capabilities.

Evaluated Models. We obtained consent from submitters of 24 *real models* to include the system predictions in our analysis. From those 24 models, we choose ten models for our user study: AMGN (rank 16) (anonymous submitter), FE2H on ALBERT (3) (Li et al., 2022), HGN (Fang et al., 2020) (35), IRC (63) (Nishida et al., 2021), Longformer (25) (anonymous), S2G-large (31) (anonymous), Text-CAN (47) (Usyd NLP), GRN (65) (anonymous), SAE (48) (Tu et al., 2020), DecompRC (unranked[6]) (Min et al., 2019).[7]

Additionally, we derive *five synthetic models* using the ground truth annotations to include extreme cases of the potential space of systems: (i) *gold answers and gold facts* (plain gold annotations), (ii) *gold answers and random facts* (we sample the same number of facts as the gold annotations, but do not sample from the articles in which the gold facts are located), (iii) *random answers and gold facts* (we sample a random answer from the context while keeping the number of words the same as in the gold answer), (iv) *random answers and random facts* (both, answers and facts are sampled, as described before), (v) *gold answers and all facts* (gold answers but the predicted facts are *all* facts from the context, i.e. from ten Wikipedia articles).

Proxy Scores. As discussed in Section 4.1, the HotpotQA leaderboard reports the metrics EM, precision, recall, and F_1 for three levels: (i) answer, (ii) supporting facts (i.e., the expla-

[6] DecompRC reports answer metrics only.
[7] Ranks from 24th of February 2023.

	Proxy score	Description
leaderboard scores	**answer-precision, answer-recall, answer-F$_1$, answer-EM**	Overlap metrics that compare the predicted answer tokens and the ground truth answer tokens using precision, recall, F$_1$-score, and EM
	SP-precision, SP-recall, SP-F$_1$, SP-EM	Overlap metrics that compare the set of predicted supporting facts and the set of ground truth supporting facts using precision, recall, F$_1$-score, and EM on a sentence level
	joint-precision, joint-recall, $\boxed{\textbf{joint-F}_1}$, joint-EM	Joint versions of the answer and supporting facts metrics based on instance-wise products of EM, precision, and recall
additional scores	**LocA score**	A score that measures how well the predicted answer and explanation are coupled. It compares the fraction of answer tokens inside an explanation to the fraction of tokens outside an explanation.
	#facts	Number of facts (i.e., sentences) within the predicted explanation
	#words	Number of words over all facts inside the predicted explanation

Table 4.3.: Proxy scores that we use to automatically evaluate the explainable question answering systems. The upper part shows the scores that the HotpotQA leaderboard evaluates. The lower part shows additional metrics that are (a) two simple surface metrics related to the length of the predicted explanation and (b) one task-specific explanation quality score. $\boxed{\text{joint-F}_1}$ is used to rank models on the leaderboard.

nation), and (iii) on the answer and explanation jointly. Table 4.3 lists and describes all of these proxy scores in the upper part of the table. The leaderboard ranks the systems according to joint-F$_1$ on a non-public test set (breaking ties by using other measures like joint-EM and answer-F$_1$).

We consider three additional scores shown in the lower part of Table 4.3. The LocA score we proposed in Section 4.1 is a task-specific score that measures to which extent predictions and explanations are coupled and a higher LocA score corresponds to a better explanation-answer coupling. While our definition of the LocA score in Section 4.1 measures how many predicted answers are located in explanations by comparing offsets, we generalize this concept to general string matching because we do not have access to the answer offsets of the leaderboard models. As a positive side-effect, this makes the score applicable to every kind of model and not only to extractive question answering models. Furthermore, we include two additional surface scores that measure an explanation's length in terms of (a) the number of facts it includes and (b) the total number of words that these facts contain. Note that we do not include our proposed

FARM score (see Section 4.1) as evaluating it requires to have direct access to the model to track system prediction changes with modified prediction contexts. We provide all F_1, LOCA and explanation length values of the models for which we got permission to include them in our analysis as well as our five synthetic models in Table A.3 in Appendix A.4.1.

4.3.1.2. Human Evaluation

To obtain a clearer perspective on (i) the relation between the described proxy scores and human ratings and (ii) the model ranks regarding various human ratings, we conduct a human evaluation. While our analysis described in Section 4.1 was limited to three models and the ground truth predictions on the public validation set only, we now analyze *test set predictions* of the ten *real* model submissions as well as the five synthetic models we discussed in Section 4.3.1.1. We evaluate the models in a crowdsourced user study with 75 participants, collecting subjective quality ratings of *utility*, *consistency*, *usability*, *answer correctness*, and *mental effort* as well as objective *completion time* measures. In the following, we detail the conducted user study.

Experiment Design and Participants. We make use of a between-subject experiment design, i.e., each participant is exposed to model predictions from exactly one model. The participants are distributed to models such that each model receives ratings from five different participants. We include two attention checks to filter out participants that are likely to not have read the question or the explanations.

For each model, we collect ratings from five crowdworkers who each rate a sample of 25 questions drawn from a pool of 100 questions.[8] For each participant, we present the individual sample of 25 questions in a randomized order to avoid potential carry-over effects between questions. We make use of this approach to (i) cover a large number of questions to better reflect the dataset and at the same time (ii) restrict the user's workload to evade fatigue effects.

We recruit a total of 75 crowdworkers from the US using MTurk. We require workers to have a $>90\%$ approval rate and an MTurk Master qualification and ensure that each worker participates no more than once in our experiments.

Collected Human Ratings. We collect the human ratings/scores listed in Table 4.4. We collect *per-instance* participant ratings of perceived *explanation utility*, *explanation consistency*, and *answer correctness*. In addition, we track the *completion time*, the participants take to finish

[8] To support our assumption that a pool of 100 questions is sufficiently representative, we simulate experiments with various question subsets. We find that correlations stabilize for as few as 20 questions and report details in Appendix A.4.2.

	Quality Dimension	Description
instance	Explanation utility	"The explanation helps me to decide if the answer is correct."
	Explanation consistency	"The explanation helps me to understand how the model came up with its answer" (similar to Nourani et al. (2019) and Schuff et al. (2020))
	Answer correctness	"The answer is correct" (similar to Bussone et al. (2015), Camburu et al. (2018), Schuff et al. (2020), Kumar and Talukdar (2020), and Narang et al. (2020))
	Completion time	Time per instance (similar to Lim et al. (2009), Lage et al. (2019), Cheng et al. (2019), and Schuff et al. (2020))
system	Usability	UMUX system usability questionnaire (Finstad, 2010, 2013)
	Mental effort	Paas mental effort scale (Paas, 1992)

Table 4.4.: Human ratings/scores collected in our crowdsourcing study.

each question. Further, we collect *per-system* ratings within a post questionnaire at the end of the experiment where we ask participants to rate *usability* using the UMUX scale (Finstad, 2010, 2013) and *mental effort* using the Paas scale (Paas, 1992).

Results. We discuss our results and, in particular, the relation between the proxy scores and the collected human ratings in the context of the respective shortcomings in the evaluation of explanation quality in the following section. We provide the detailed averaged ratings over all 15 models in Appendix A.4. Further details on the collected proxy scores over all real and five synthetic models are provided in Table A.3 in Appendix A.4.1. Details on the exact human ratings over all models included in our human evaluation are provided in Appendix A.4.

4.3.2. Disconnect Between Automatic and Human Evaluation

The underlying assumption of using proxy scores for evaluating explanation quality is that an improvement in proxy scores implies an increase in user benefits. However, to the best of our knowledge, there is no established view to which extent those scores actually reflect the value of explanations to users (i.e., to which extent they are *valid* and measure what they should measure). This practice conflicts with both, the user-dependence characteristic (Section 4.2.1) and the multidimensionality characteristic (Section 4.2.2) of explanation quality. In the following, we discuss different aspects of the relation between proxy scores and human ratings. Concretely, we investigate (i) the pairwise relation between proxy scores and human ratings, (ii) their overall relation in terms of their underlying factor structure, and (iii) the dynamics of their relation over time.

4.3.2.1. Low Correlations Between Proxy Scores and Human Ratings

If the assumption that higher values on a specific proxy score correspond to higher user benefits holds true, this benefit should also reflect in one or multiple human subjective ratings or objective performance markers.

One of the few studies that study the strength of the link between proxy scores and human ratings in the context of explanations is conducted by Kayser et al. (2021). They analyze the correlation between NLG metrics (i.a., BLEU, BERT-Score and BLEURT) and human quality ratings to quantify free-text explanations for three visual QA datasets within a large crowdsourcing study. They find that (over the three datasets), the highest Spearman correlation across ten different proxy scores only reaches 0.29. Similarly, earlier work of Camburu et al. (2018) finds that BLEU does not reliably reflect textual explanation quality for a NLI task which is supported by our observation that even high performance differences regarding proxy scores of explainable NLI models are not reflected in human ratings of explanation quality we discussed in Section 3.1. Clinciu et al. (2021) study human ratings of explanation clarity and informativeness in the context of natural language explanations of Bayesian networks. Averaged over different scenarios, they find that, across eleven proxy scores, the highest Spearman correlation still only reaches 0.39 while the correlation between the two human ratings reaches a much higher value of 0.82.

Overall, to the best of our knowledge, all (of the few) available studies indicate that *proxy scores and human ratings correlate weakly.*

Case Study. We provide an additional analysis for explainable QA and exceed previous studies in the diversity of human rating dimensions. Additionally, our results are evaluated using predictions of real leaderboard models instead of evaluating correlations on ground truth datasets (Camburu et al., 2018; Kayser et al., 2021; Clinciu et al., 2021) and thus cover a distribution of scores and ratings that better reflects the scores' usage within leaderboards.

Figure 4.5 shows Kendall's τ correlation coefficients between (a) the automatic scores included in the leaderboard and (b) the human ratings we collected in our study. A heatmap visualization of all pairwise correlations including statistical significance markers and coefficients can be found in Appendix A.4.

While we observe moderate correlations between, e.g., joint-F_1 and explanation consistency, the majority of correlations is under 0.5 and thus the previously described weak relation between human ratings and proxy scores is *supported by our case study.*

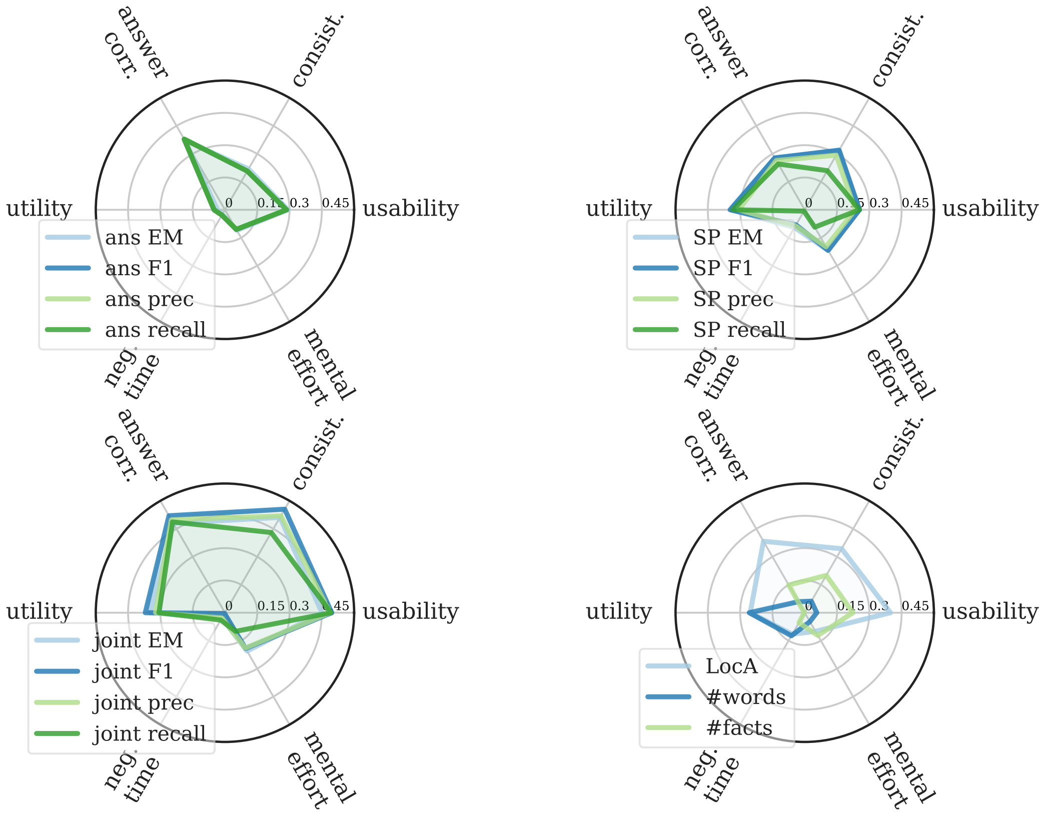


Figure 4.5.: Kendall's τ correlation coefficients for the correlation of different automatic scores and user-rated quality dimensions. The correlations illustrate the weak and conflated connection between proxy scores and human assessment (from left to right and top to bottom: scores evaluating answer correctness, scores evaluating correctness of supporting facts, scores jointly evaluating answer and fact correctness, additional scores including LocA and surface scores). Axes cropped at 0.6.

4.3.2.2. Proxy Scores Conflate Different Dimensions

We argue that the currently used explanation quality proxy scores can, and often will, conflate different dimensions of explanation quality, and, consequently, information about the individual independent dimensions is lost and cannot be recovered. For example, given two systems with similar proxy scores, it cannot be determined which one was superior in terms of individual explanation quality aspects, such as consistency or understandability. Therefore, it is not possible to identify an isolated improvement of a model in some of those aspects using the proxy score. For example, when we improve the proxy score, we cannot assess whether we actually improved all quality aspects or only a subset of them (and possibly decreased the performance on others). Similarly, a targeted improvement of particular quality aspects (e.g., for a particular use case) is not possible.

Case Study. In order to assess the degree to which proxy scores are related to *multiple* human ratings/scores, we also analyze their underlying factor structure. While the correlations shown in Figure 4.5 demonstrate that the *pairwise* correlations are weak, they do not tell us how a score/rating is associated with the "big picture" in terms of the latent structures behind the scores and ratings. A factor analysis is used to describe each score/rating in terms of its association with empirically-derived latent factors. In order to determine the number of factors, we follow the *method agreement procedure* in which the optimal number of factors is chosen based on the largest consensus between numerous methods. We find four dimensions to be supported by the highest number of methods as, from a total of 14 methods, four agree on a number of four factors (i.e., beta, Optimal coordinates, Parallel analysis, and Kaiser) which has higher support than every other number of factors (ranging between 1 and 19). We, therefore, conduct a factor analysis using a varimax rotation that maps each score/rating to one of the four latent factors such that the resulting factors describe the data as well as possible. Table 4.5 displays the respective loadings (i.e., correlations of the score/rating with the latent factor).

Factor F1 contains all answer-related scores including human ratings of answer correctness. Factor F2 only contains the fact-related automatic scores as well as joint-F_1 and joint-EM. Factor F3 contains all explanation-related human ratings. Interestingly, F3 also contains the automatic scores joint-recall, LOCA, and joint-precision. Factor F4 contains the explanation-length-related automatic scores #words and #facts. We observe that the answer-related proxy scores and the human answer-correctness ratings form a cluster and have strong loadings on their joint factor. This can be interpreted as evidence that perceived answer correctness can — to a moderately strong extent — be measured via the answer-related proxy scores.

If (one of) the evaluated proxy scores for explanation quality would have an equally strong association with any human rating, we would expect a factor, that, e.g., contains joint-F_1 and perceived utility along with strong factor loadings of both scores/ratings onto this factor. However, Table 4.5 demonstrates that all of the different explanation-related human ratings can be found in one factor along with the proxy scores joint-recall, LOCA, and joint-precision. This shows that our explanation quality measurements cannot be grouped into distinguishable groups (as we observe for the answer-related scores and ratings). Instead, they form a hardly-interpretable diffuse factor 3 that mixes up all kinds of human ratings and yields much lower factor loadings and — in addition — does not contain the leaderboard ranking score joint-F_1.

Overall, our factor analysis suggests that (a) *answer-related proxy scores reflect human answer correctness ratings*, (b) *no explanation-related proxy score can be associated to a particular human rating*. In particular, joint-F_1 does shares no factor with any human rating.

Score/rating	**Type**	**F1**	**F2**	**F3**	**F4**
answer-EM	🖥️	0.98			
answer-F_1	🖥️	0.97			
answer-recall	🖥️	0.97			
answer-precision	🖥️	0.97			
correctness rating	🧑‍⚖️	0.80			
sp-EM	🖥️		0.97		
sp-F_1	🖥️		0.89		
sp-precision	🖥️		0.83		
sp-recall	🖥️		0.82		
joint-EM	🖥️		0.58		
joint-F_1	🖥️		0.55		
consistency rating	🧑‍⚖️			0.79	
usability rating	🧑‍⚖️			0.74	
joint-recall	🖥️			0.64	
utility rating	🧑‍⚖️			0.63	
LOCA score	🖥️			0.56	
joint-precision	🖥️			0.53	
completion time	🧑‍⚖️			(0.24)	
#words	🖥️				0.95
#facts	🖥️				0.95

Table 4.5.: Factor loadings of the collected scores/ratings onto four factors (F1-F4). Proxy scores are marked as 🖥️, human ratings/scores are marked as 🧑‍⚖️. We observe that F3 contains all explanation-related human ratings as well as the three three proxy scores joint-recall, LOCA, and joint-precision. This suggests that these three scores can be better suited to capture perceived explanation quality compared to the currently used joint-F_1 which loads onto factor F2.

4.3.2.3. Goodhart's Law: Validity Can Change Over Time

Even if we had a score that is valid, i.e., it measures one dimension of explanation quality in a decent way, using this score as the sole ranking criterion of a leaderboard can subvert its validity over time. This effect is described in *Goodhart's Law* that is commonly stated as *when a measure becomes a target, it ceases to be a good measure* (Goodhart, 1975; Campbell, 1979; Strathern, 1997; Manheim, 2018; Manheim and Garrabrant, 2018). Thomas and Uminsky (2022) discuss this in the context of AI and highlight the field's problematic reliance on (single) metrics including the issue of metrics being gamed (Bevan and Hood, 2006).

Assume that an initial investigation of some systems showed that a particular proxy score can be considered to be valid (in a certain use case for a certain user group). If now more and

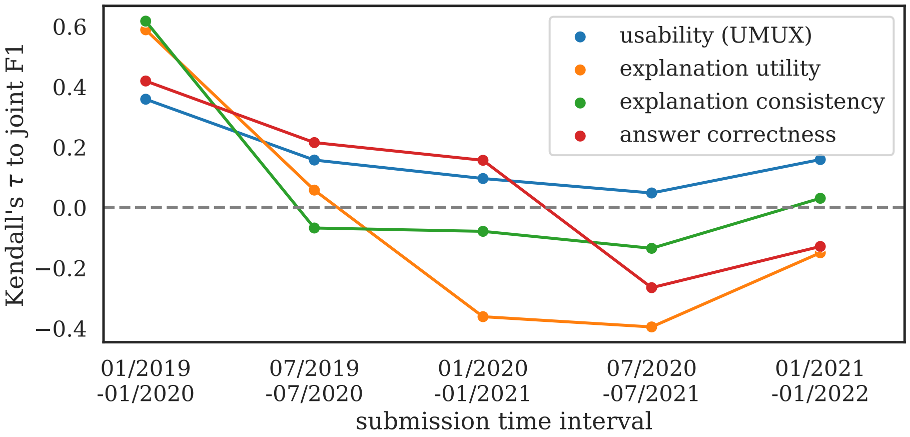


Figure 4.6.: Kendall's τ correlations over time between different human ratings and the official leaderboard metric joint-F_1. The gradual decline of the relation between joint-F_1 and human ratings indicates that joint-F_1 looses validity over time and thus supports Goodhart's law.

more systems are developed with the primary goal of reaching higher values on that score, the initial set of models no longer represents the new model population. As a result, it cannot be ensured that the original strong relation between the (initially valid) proxy score and the measured quality dimension still holds. Consequently, the score's validity can "wear off" over time as it is used in isolation.

Case Study. We investigate whether we can find such a temporal deterioration in the HotpotQA leaderboard. For this, we study the association of the leaderboard's target metric (i.e., joint-F_1) with the measured human ratings across different time windows. Figure 4.6 shows Kendall's τ correlation coefficients between joint-F_1 and human ratings for a 12-month sliding window over system submissions. We observe that correlations decrease from moderately positive to lower and even negative correlations. We hypothesize that this decrease could have been mitigated using multiple proxy scores.

Overall, our observations indicate that Goodhart's law affects today's leaderboards and *single target metrics lose their expressiveness over time*.

4.3.3. Neglecting Users

So far, we argued why the currently used proxy scores of explanation quality do not reliably reflect user-perceived quality properties. But if we had proxy scores that were shown to successfully reflect various aspects of explanation quality, could we stop human evaluation and

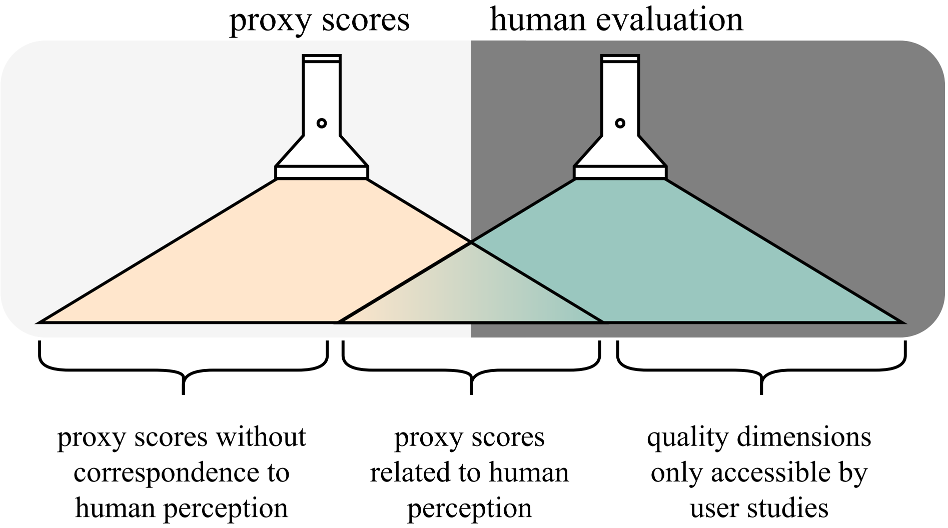

proxy scores
human evaluation
proxy scores without correspondence to human perception
proxy scores related to human perception
quality dimensions only accessible by user studies

Figure 4.7.: Visualization of the Streetlight Effect in the context of explanation quality evaluation. Searching where the light is, i.e., relying on proxy scores in isolation (left) does not allow to study the whole spectrum of explanation quality. Instead, we also have to study quality aspects that are only accessible via human evaluation (right).

rely on these scores alone? We, just like Thomas and Uminsky (2022), argue that we could not.

The predominant evaluation practice of relying on automatic scores is questioned in many contexts in NLP today, especially in NLG (Callison-Burch et al., 2006; Liu et al., 2016; Novikova et al., 2017; Sulem et al., 2018; Reiter, 2018). In the context of explainability, the need for *human-centered* evaluation is stressed by, i.a., Ribera and Lapedriza (2019), Chu et al. (2020), Gonzalez et al. (2021), Colin et al. (2021), Schlegel et al. (2022), or Liao et al. (2022). We argue that human evaluation always has to be part of explanation quality evaluation. User studies yield insights beyond proxy scores as they can comprise (i) a broader set of quantifiable dimensions than proxy scores can offer as well as (ii) dimensions of explanation quality that are inaccessible using quantitative methods at all but require *qualitative* approaches, such as mental model analysis (Schrills and Franke, 2020; Kulesza et al., 2013), or thematic analysis (Braun and Clarke, 2006) in which themes are extracted from textual responses or transcriptions via various steps (coding, theme generation, review, etc.).

We argue that searching for valuable systems based on proxy metrics alone can be regarded to be an instance of the *Streetlight Effect*, also known as the *Drunkard's Search Principle* (Kaplan, 1964; Iyengar, 1993). This effect describes a situation in which a drunken man lost his keys in a park, but instead of searching for them in the place where he lost them, he is *searching under a streetlight because this is where the light is*. We argue that we face a similar situation when we exclusively rely on proxy metrics as shown in Figure 4.7. Instead of focusing on what we ultimately are interested in, i.e., providing good explanations to users, we narrow our focus to increasing proxy scores instead. To shed light on a broader spectrum of explanation quality, our quantitative measures should include both, validated proxy scores and human ratings/signals.

Case Study. Our results reported in Section 4.3.2 already show that human ratings exceed the information we are able to get from our investigated proxy scores. In addition, there also is information that we cannot obtain using quantitative human ratings alone. To illustrate this, we collect voluntary free text feedback from our participants:

- "*I see why the model thought it, but it doesn't provide any useful info in reality*".
  This comment shows that users have the impression that a model "thinks", hinting at anthropomorphization. Concretely, this suggests to consider the inclusion of an anthropomorphism questionnaire in subsequent user studies.

- "*The question asks about two players but there is only a correct answer for one player and only one explanation*".
  This comment confirms that one type of model error is to provide answers that do not semantically match the question. Consequently, developing a new proxy score to quantify the semantic overlap between the predicted answer and the question could help to guide model development.

- "*It doesn't really state how it came up with this answer, as it only told about other fights. My default answer is incorrect, until the system proves it to be true.*"
  This comment informs us about the user's rating behavior and suggests that a re-wording of the question could allow us to capture the range of perceived correctness better.

Overall, the collected comments illustrate that *qualitative evaluation can yield insights beyond quantitative participant ratings* which in turn can help to improve proxy scores and human rating evaluation.

4.3.4. Single-score Leaderboards

The current practice in NLP leaderboards (and many NLP research work in general) is to score and compare of systems using a single score, such as accuracy, BLEU, or F_1. In Section 4.2.2, we already motivated that explanation quality has multiple independent dimensions. Therefore, it should be measured with multiple scores. Moreover, aggregating those scores (e.g., via averaging) to obtain a single measure will not be expedient either since the dimensions might be independently useful and/or scaled differently.

	ranking criterion				
	joint-F$_1$	**LocA**	**averaged proxy scores**	**factor-weighted proxy scores**	**human usability ratings**
1	*gold* (1.00)	*gold* (1.00)	*gold* (1.00)	*gold* (1.73)	FE2H on ALBERT (0.98)
2	FE2H on ALBERT (0.77)	gold-answers-all-facts (1.00)	FE2H on ALBERT (0.78)	FE2H on ALBERT (1.47)	HGN (0.90)
3	AMGN (0.74)	FE2H on ALBERT (0.98)	AMGN (0.76)	AMGN (1.43)	S2G-large (0.88)
4	Longformer (0.73)	HGN (0.97)	HGN (0.74)	HGN (1.4)	Longformer (0.87)
5	S2G-large (0.72)	AMGN (0.95)	Longformer (0.74)	Text-CAN (1.31)	SAE (0.87)
6	HGN (0.71)	Text-CAN (0.92)	Text-CAN (0.70)	Longformer (1.28)	Text-CAN (0.87)
7	Text-CAN (0.66)	GRN (0.89)	S2G-large (0.69)	SAE (1.25)	AMGN (0.87)
8	SAE (0.63)	SAE (0.86)	SAE (0.67)	gold-answers-all-facts (1.23)	gold-answers-all-facts (0.86)
9	IRC (0.59)	IRC (0.77)	GRN (0.64)	GRN (1.21)	*gold* (0.83)
10	GRN (0.58)	Longformer (0.72)	IRC (0.62)	IRC (1.17)	IRC (0.83)
11	gold-answers-all-facts (0.12)	random-answers-gold-facts (0.12)	gold-answers-all-facts (0.53)	S2G-large (0.94)	GRN (0.68)
12	random-answers-all-facts (0.02)	S2G-large (0.12)	random-answers-gold-facts (0.3)	random-answers-gold-facts (0.09)	random-answers-random-facts (0.23)
13	gold-answers-random-facts (0.00)	random-answers-random-facts (0.11)	gold-answers-random-facts (0.29)	random-answers-random-facts (0.06)	random-answers-gold-facts (0.21)
14	random-answers-random-facts (0.00)	gold-answers-random-facts (0.03)	random-answers-random-facts (0.01)	gold-answers-random-facts (0.02)	gold-answers-random-facts (0.16)

Table 4.6.: Ranking models with respect to different criteria. We construct leaderboards for (a) joint-F$_1$ (official leaderboard score), (b) the answer-explanation consistency measure LocA, (c) the average over all 14 proxy scores, (c) a factor-loading-weighted average over the three proxy scores which we found to be associated with human ratings within our factor analysis, and (d) human utility ratings. We mark S2G-large and *gold predictions* to demonstrate inconsistent model ranks across criteria.

Ranking systems using a single score can additionally lead to over-optimization of this one score (Thomas and Uminsky, 2022) and can lead to the deterioration of score validity as we argued and demonstrated in Section 4.3.2.3. This arguably could be prevented by using a diverse set of scores instead of only one score.

Case Study. We construct various leaderboards from the HotpotQA systems and evaluate how sensitive model rankings are with respect to the ranking criterion. Table 4.6 displays the respective model rankings of four criteria. We observe that *different scores and weighting schemes lead to contradicting model rankings*. For example, S2G-large (marked with a box in Table 4.6) is rated the fourth-best real model according to joint-F_1 and the third-best regarding usability ratings but rated the worst real model according to LocA. While all real models except FE2H on ALBERT differ with respect to their relative rankings, FE2H on ALBERT is ranked as the best real model across all criteria including human usability ratings, indicating that this model offers substantial benefits over the other models.

Further, *rankings regarding human ratings and proxy metrics disagree heavily* as we can see for the gold predictions that consistently are ranked top following the proxy score leaderboards, but are ranked eighth following human usability ratings. Interestingly, the gold answers along with *all facts* are ranked as more usable than the gold facts with only the relevant facts.

Overall, our results signal a disagreement between the user needs *assumed* within the HotpotQA dataset and the *actual* user needs within our participant sample.

4.4. Remedies

This section proposes guidelines to address the shortcomings described in Section 4.3.

4.4.1. Report Various Scores Without Averaging

As we argued in Section 4.3.4, using a single score for evaluation (regardless of proxy scores or human ratings) can be misleading. Thus, we propose to use various scores rather than weighting quality dimensions against each other to get a single score. This is in line with the recommendations by Thomas and Uminsky (2022). While prior work proposed alternative leaderboards using on-demand (crowdsourcing) evaluation (Chaganty et al., 2017) and personalized utility rankings (Ethayarajh and Jurafsky, 2020), we are — to the best of our knowledge — the first to provide a *leaderboard that does not condense multiple scores into a single one*.

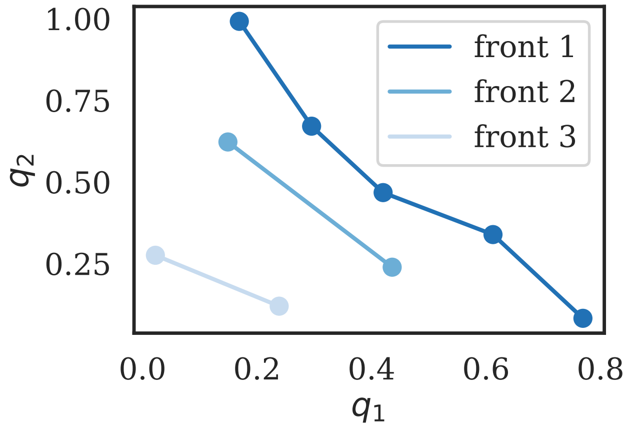


Figure 4.8.: Ranked Pareto fronts for two dimensions and nine (fictional) systems. Each point represents a system along two (higher-is-better) scores q_1 and q_2.

Pareto Front Leaderboards. To compare systems based on multiple scores, e.g., on a leaderboard, we propose to leverage the concept of *Pareto efficiency*. In the context of multidimensional leaderboards, a system is called Pareto efficient if the only way to select another system that is better regarding any score dimension is to worsen another score dimension. For example, system A is Pareto efficient if the only way to select another system to increase, e.g., the F_1-score, is to choose a system that has a lower, e.g., accuracy. Given a set of systems, multiple systems can simultaneously be Pareto efficient. Figure 4.8 shows a fictional example with nine systems (visualized by points) and two higher-is-better quality scores q_1 and q_2 (visualized by axes). All five systems on the so-called Pareto front (front 1) are Pareto efficient and thus have rank 1. To rank the remaining systems, we remove those five systems, calculate the next Pareto front (front 2), and repeat this until all systems are ranked. The resulting leaderboard of the example shown in Figure 4.8 would consequently have five models in the first place (i.e., front), two models in the second, and two models in the third.

Related Applications of Pareto Efficiency. We are not the first to leverage Pareto efficiency within NLP. Pimentel et al. (2020) use Pareto efficiency to propose a new probing approach that trades off probe accuracy and complexity. In contrast to their work, we use Pareto efficiency to construct leaderboards. Similar to our approach, Liu et al. (2022b) argue that, in the context of efficient NLP models, models should be judged in terms of how far they overstep the performance-efficiency Pareto front. In contrast to their work, we do not only consider the (first) Pareto front but extend the concept of Pareto efficiency to multiple fronts which form the ranks of our proposed leaderboard.

Advantages. Using multiple scores for evaluation offers the advantage of *capturing diverse aspects* of a system. If a sufficiently diverse set of scores is used, the *over-optimization of one score can be prevented* since other scores would likely be decreased at the same time. This is supported by *surrogation* effects (Choi et al., 2012, 2013) where, in the context of manager compensation, Choi et al. (2012) find that manager decisions can be improved when "managers are compensated on multiple measures of a strategic construct" instead of on a single one. We hypothesize that this observation also holds for AI practitioners that need to choose a system, e.g., from a leaderboard.

When using Pareto front leaderboards, we can *rank systems without weighting the different quality dimensions against each other*. In particular, the concept of Pareto efficiency allows us to choose systems that are *not worse* than others on all fronts. Note that Pareto fronts are robust to score re-scaling and are applicable to ordinal (e.g., Likert) ratings.

Limitations. With multiple scores, it can be hard to determine a "winning" system because *different models might rank best* on different scores. Pareto Front Leaderboards can mitigate this problem, however, they may result in a set of winning systems instead of a single winning system. We argue that this is not a real limitation though since the concept of Pareto efficiency ensures that a system on one front is not worse than other systems on the same front. In the extreme case when the number of scores is high in comparison to the number of systems that should be scored, the resulting leaderboard can collapse to a single front because the fronts' surface grows exponentially with the number of scores. We, therefore, recommend ensuring that the number of variables should only be increased along with a sufficient increase in the number of systems.

Further, Pareto Front leaderboards can be "attacked" by optimizing a single metric with the purpose of positioning a new system inside the first front. Although this allows the leaderboards to be gamed to a certain extent, a truly remarkable improvement is one that creates a new front that is, in turn, robust to the improvement of single metrics.

Case Study. We evaluate the 15 models described in Section 4.3.1.1 on numerous (i) human ratings and (ii) automatic scores. We construct two Pareto front leaderboards, one for human ratings and one for automatic scores.

Table 4.7 shows the Pareto front leaderboard based on human ratings (usability, mental effort, utility, correctness, consistency, and completion time). We observe that high-performing models, such as FE2H on ALBERT (official leaderboard rank 3), are located within the rank 1 Pareto front en par with the gold prediction system. Interestingly, previously lower-ranked

Rank	**Models** (original HotpotQA ranks in parentheses)
1	gold (*), random-answers-gold-facts (*), FE2H on ALBERT (3), Longformer (25), S2G-large (31), HGN (35), Text-CAN (47), IRC (63)
2	AMGN (16), SAE (48), GRN (65), DecompRC (unranked), random-answers-random-facts (*), gold-answers-all-facts (*)
3	gold-answers-random-facts (*)

Table 4.7.: Ranked Pareto fronts based on *human ratings*. "$*$" marks systems derived from the ground truth annotations.

Rank	**Models** (original HotpotQA ranks in parentheses)
1	gold (*)
2	gold-answers-all-facts (*), rand.-answers-gold-facts (*), FE2H on ALBERT (3), AMGN (16)
3	Longformer (25), HGN (35), IRC (63), gold-answers-random-facts (*)
4	S2G-large (31), Text-CAN (47)
5	SAE (48), GRN (65)
6	DecompRC (unranked)
7	random-answers-random-facts (*)

Table 4.8.: Ranked Pareto fronts based on *proxy scores*.

models, such as IRC (leaderboard rank 63) are also located in the first Pareto front which means that they also possess a combination of strengths that dominates the models in the other ranks.

Table 4.8 shows the leaderboard based on automatic proxy scores. The gold prediction system is the single winner in this leaderboard, followed by the two real models FE2H on ALBERT and AMGN. While the first models are ordered consistently with the HotpotQA leaderboard, the Pareto front leaderboard disagrees w.r.t. ranks for others, e.g., the IRC model (leaderboard rank 63), Longformer (leaderboard rank 25) or S2G-large (leaderboard rank 31). For the synthetic systems, we observe differences across the two Pareto front leaderboards. For example, the gold-answers-random-facts system is ranked last w.r.t. human ratings but ranked third w.r.t. automatic scores. Our results highlight, again, that proxy metrics do not reflect the quality dimensions probed in the human ratings sufficiently well. We provide details on the exact proxy scores and model ratings in Appendix A.4.1 and Appendix A.4.2.

4.4.2. Validate Proxy Scores Against Humans

While there is a lot of work on investigating the relation between automatic scores and human ratings in NLG (Belz and Reiter, 2006; Novikova et al., 2017; Dušek et al., 2019), only a few studies consider this aspect in the context of explanation evaluation (Jannach and Bauer, 2020;

Kayser et al., 2021; Clinciu et al., 2021). To address the problem of unvalidated proxy scores for explanation quality evaluation (Section 4.3.2), we advise to consistently validate the relation between proxy scores and human signals, such as human-AI performance, subjective ratings, completion times, or physiological measures like eye tracking. One straightforward approach to quantify these relations is a correlation analysis.

Advantages. Given proxy scores that yield sufficiently strong correlations to human ratings/signals, those scores can be used to develop systems that are actually useful for users.

Limitations. Given a new task or leaderboard, it is unlikely that we have access to a representable pool of models which can be used to validate the metrics. Therefore, we have to accept a certain *grace period* in which we can only assume that the chosen evaluation scores lead to reasonable results. Once there is a handful of models available, the proxy metrics should then be validated against human scores and revised if necessary.

Referring to our discussion of Goodhart's law in Section 4.3.4, any proxy metric has to be *periodically re-tested* for its validity. Concretely, the need for re-testing can be recognized by, e.g., monitoring demographic changes in the target population and/or changes in the correlations within user ratings.

Finally, each validity evaluation is limited to a *group of explainees* (see Section 4.2.1). Different groups of users will have different needs and, as a result, explanation quality evaluation will need different measures. For example, validity findings for the population of high-school students might not transfer to adult NLP researchers.

4.4.3. Do Human Evaluation

In Section 4.4.2, we already recommend user studies for the purpose of proxy score validation. Based on our discussion in Section 4.3.3, we also propose to conduct as much human evaluation as possible in order to gain *additional explanation quality indicators* from human rating scores directly. In the context of application-oriented model development, human evaluation can be conducted as the final evaluation step after model tuning. In the context of leaderboards, we propose to regularly conduct human assessments of (a subset of) system submissions.

Measures of Human Behavior and Perceived Quality. When choosing what to measure within a user study, we suggest to *collect objective measures of user behavior as well as subjective ratings*. Table 4.9 lists a selection of possible measures of (a) objective measures of human behavior (top) and (b) subjective human ratings (bottom) along with

	Measure	Description	References
objective scores	Time	Time measures of, e.g., task completion or interaction with the system	Lim et al. (2009), Lage et al. (2019), Cheng et al. (2019)
	Human performance	Task performance of users (e.g., accuracy in an AI-supported decision task)	Feng and Boyd-Graber (2019), Lage et al. (2019), Bansal et al. (2021)
	Simulatability	Measures related to how well explanations enable users to predict system performance (given an explanation when making predictions on new instances)	Hase and Bansal (2020), Wang and Yin (2021)
	Teachability	Measures related to how well explanations enable users to predict system performance (without having access to explanations when making predictions on new instances)	Goyal et al. (2019), Wang and Vasconcelos (2020)
	Agreement	Frequency of how often a user accepts a system decision	Zhang et al. (2020b), Bansal et al. (2021)
	Number of user interactions	Number of times a user, e.g., runs a model to predict an output	Pezeshkpour et al. (2022)
subjective ratings	Perceived performance	Subjective estimate of system performance	Nourani et al. (2019)
	Over- / underestimation	Difference between perceived system performance and actual system performance	Nourani et al. (2019)
	Trust	Trust in the model's abilities/correctness	Bussone et al. (2015), Ribes et al. (2021), Buçinca et al. (2021)
	Perceived usefulness	User-reported system usefulness	Khurana et al. (2021), Bansal et al. (2021)
	Subjective understanding	Self-reported degree of system understanding	Ehsan et al. (2019), Wang and Yin (2021), Ribes et al. (2021)
	Grammaticality	Ratings or grammatical correctness	Liu et al. (2022a)
	Perceived factuality	Ratings of factual correctness	Liu et al. (2022a)
	Mental demand	Self-reports of mental demand in processing the explanation	Buçinca et al. (2021)

Table 4.9.: Selection of (a) scores of objective human behavior (top) and (b) dimensions of subjective self-reports of perceived quality (bottom).

exemplary publications including the respective scores/ratings. We refer to Chromik and Schuessler (2020) as well as Nauta et al. (2022) for a review of (quantitative) human evaluation methods conducted in explainability research. Objective measures include, e.g., response time, human task performance, and human-AI agreement, but also more complex scores, e.g., Utility-k (Colin et al., 2021). Subjective ratings include, e.g., perceived accuracy, trust, perceived usefulness (Khurana et al., 2021), or mental demand (Buçinca et al., 2021). As noted by Buçinca et al. (2020), subjective ratings should complement objective user performance measures as the latter cannot necessarily be inferred from the former (Buçinca et al., 2020). We investigate the relation between objective and subjective system predictability in Section 5.2.

Following Jannach and Bauer (2020) and Thomas and Uminsky (2022), we further advocate to also *collect qualitative feedback* (e.g., participant comments within a user study or a focus group) to complement quantitative measures. We demonstrate how qualitative feedback can yield insights beyond quantitative evaluation within our case study in Section 4.3.3.

The study conducted by Cheng et al. (2019) is a good example of how objective measures can be combined with both qualitative as well as quantitative human evaluation. Additional examples of such *mixed-methods* evaluations can be found in the work of Bansal et al. (2021) in the context of complementary human-AI team performance and Sivaraman et al. (2023) in the context of clinical AI acceptance. Note, however, that collecting qualitative feedback via, e.g., the think-aloud method, can impact users' mental effort allocation which can potentially affect participant behavior (Buçinca et al., 2020) and, consequently, respective studies should be designed carefully.

Advantages. Human evaluation allows us to re-adjust the direction into which we develop systems by unveiling explanation quality dimensions that were previously hidden. For example, qualitative findings from user comments can help us to identify system qualities we did not think of before. Moreover, human evaluations can reward the development of systems that follow an unconventional approach and, as a result, have explanation qualities that might have been undetectable using proxy scores. This can motivate researchers to develop original models and can ultimately *diversify and accelerate* research.

Limitations. Each human evaluation is bound to noise w.r.t. the pool of participants and the way they approach the study (for example whether they carefully read the questions). However, in contrast to *annotation* (on an instance level), noisy human responses do not have to limit human *evaluation* (on a system level) using adequate statistical tools. Further, potentially high costs to compensate the participants and longer preparation times to recruit participants and conduct and carefully evaluate the studies might hinder the conduction of a user study. Additionally, proxy task evaluations (i.e., evaluations that are conducted on simplified human-AI tasks) do not necessarily lead to the same findings that real human-AI tasks yield and, in fact, can even contradict the latter (Buçinca et al., 2020).

Finally, user study interfaces have to be designed carefully as presumably minor design choices can heavily affect participant behavior. For example, Sullivan Jr. et al. (2022) find participants' rational explanation selections (i.e., marking relevant words in the text input) was greatly impacted by whether participants could mark multiple words at once or not. Introductory texts on designing and conducting user studies in NLP can be found in, e.g., Belz et al. (2020) (NLG), Iskender et al. (2021) (text summarization) or Sedoc et al. (2019) (chatbots). We additionally published an extended version of our general NLP user study background covered in Section 2.3 in Schuff et al. (2023b).

Case Study. We discuss the experiment design of our case study along with a description of collected ratings in Section 4.3.1.2. A detailed discussion of our results can be found in the "case study" paragraphs across Section 4.3.

Overall, our human evaluation allowed us to identify low correlations between human ratings and proxy scores, detect that correlations decreased over three years of system submissions, qualitative user feedback helps to spot shortcomings of proxy scores and human rating evaluation, and proxy-score-based systems ranks conflict with human-rating-based ranks.

4.4.4. Overall Discussion

In the previous sections, we discussed general characteristics of explanation quality (Section 4.2), described shortcomings of the current evaluation practices, and pointed out to which extent they violate the discussed characteristics (Section 4.3). We supported our arguments with empirical evidence of a crowdsourced case study that we conducted for the example of explainable QA systems from the HotpotQA leaderboard.

Concretely, we demonstrated that (i) proxy scores poorly reflect human explanation quality ratings, (ii) proxy scores can lose their expressiveness over time, (iii) human evaluation yields quantitative as well as qualitative insight beyond automatic evaluation, and (iv) single-score leaderboards fail to reflect the spectrum of explanation quality dimensions.

In (Section 4.4), we proposed (a) guidelines for a more effective and human-centered evaluation as well as (b) an alternative type of leaderboard that constructs ranks from multiple dimensions without averaging scores. We aim to inform and inspire future work and ultimately drive the field towards reliable and meaningful explanation quality evaluation.

4.5. Related Work

In the following, we discuss related criticism on F_1-score (Section 4.5.1), the relation of proxy scores to human evaluation (Section 4.5.2), and alternative leaderboards (Section 4.5.3).

4.5.1. Criticism on F_1-score

The F_1-score has been criticized regarding various aspects including theoretical considerations and applications (Hand and Christen, 2018; Chicco and Jurman, 2020; Sokolova et al., 2006).

Hand and Christen (2018) discuss F_1-scores for a record linkage task (i.e., linking entities to records across databases) and show that F_1-scores can be reformulated in terms of a weighted sum of precision and recall for which the weights depend on the evaluated system. They argue that, instead, these weights should depend on the system's usage context.

Further, Chicco and Jurman (2020) criticize the use of F_1-score in binary classification within genomics and recommend to use Matthews correlation coefficient instead as it advantages over F_1-score regarding, i.a., dataset imbalance, and label swapping. Similarly, Sokolova et al. (2006) criticize the usage of F_1-score, accuracy, and receiver operating characteristic (ROC). They argue that, while these measures focus on a classifier's ability to correctly predict a class label, desirable properties such as class discrimination or failure avoidance can provide deeper insight and propose to use alternative measures rooted in medical diagnoses, such as Youden's index (Youden, 1950) or discriminant power (Oddone et al., 1995).

Qian et al. (2016) further demonstrate that modifying F_1-scores for word segmentation based on insights from psychometrics can improve the scores' correlation to human ratings.

4.5.2. Relation Between Proxy Scores and Human Evaluation

4.5.2.1. NLP Systems

While the relation of proxy scores to human-rated quality has been extensively studied and criticized for NLG systems (i.a., Callison-Burch et al., 2006; Liu et al., 2016; Novikova et al., 2017; Sulem et al., 2018; Reiter, 2018), explainable systems received much less attention. In the following, we discuss important exceptions.

4.5.2.2. Explainable Systems

Kayser et al. (2021) investigate the correlation between ten NLG metrics (i.a., BLEU, BERTScore and BLEURT) and human quality ratings to quantify free text explanations for three visual

QA datasets. They find that among the ten proxy metrics, considering all three datasets, only BERT-Score (Zhang et al., 2020a) has a significant but small Spearman correlation ($r = 0.293$) with human ratings of how well the respective explanations justify the model answers.

Further Camburu et al. (2018) compare human-annotated textual explanations of the e-SNLI explainable NLI dataset (see Section 2.1) with model-generated explanations using BLEU. They observe that (a) the BLEU scores within multiple human explanation annotations are low and (b) when comparing the model-generated explanations to human-annotated ground truth explanations, the resulting BLEU scores are only slightly smaller than inter-annotator BLEU scores. They conclude that BLEU does not reliably reflect explanation quality.

Ultimately, Clinciu et al. (2021) investigate human ratings of explanation clarity and informativeness for natural language explanations of Bayesian networks. Averaged over different scenarios, they find that, across eleven proxy scores, BLEURT reaches the highest absolute Spearman correlation ($r = 0.39$), closely followed by BLEURT ($r = 0.37$). However, they find that correlations between the two human ratings reach a much higher value of $r = 0.82$.

In contrast to all described studies, our evaluation considers explanations for explainable QA. While the described studies only consider ground truth explanations, our evaluation includes predictions from ten systems that were submitted to the HotpotQA leaderboard.

4.5.3. Alternative Leaderboards

Numerous alternative leaderboards have been proposed in NLP. For example, Chaganty et al. (2017) introduce on-demand crowdsourcing evaluation to provide a fair comparison of knowledge base population systems. Concretely, their approach addresses the problem that new systems can predict previously unseen relations which is penalized even when it is correct.

Further, (Ethayarajh and Jurafsky, 2020) argue that the predominant focus of NLP leaderboards on system accuracy can neglect additional model qualities, such as efficiency, robustness, or fairness. They consider individual system utility and introduce personalized leaderboards in which system ranks are determined for each individual regarding, e.g., model size or robustness.

Similarly, Linzen (2020) argues that single-score task performance leaderboards do, i.a., not reflect sample-efficiency and calls for the development of metrics that incentivize this model asset as well as the evaluation of additional parallel leaderboards that score a model's linguistic generalizations abilities or its task performance on increasingly smaller datasets.

In contrast to the described alternative leaderboard approaches, we are — to the best of our knowledge — the first to provide a leaderboard that offers a joint ranking across all scoring dimensions without condensing multiple scores into a single one.

5. Human Perception and Explanations

In this chapter, we focus on human perception of explanations. First, we explore whether users understand heatmap explanations over text in the way they are intended (Section 5.1). We find that human interpretation of heatmap explanations is distorted by various word and sentence properties as well as assimilation and contrast effects within explanations. We propose bar charts and a model-based saliency correction method to mitigate biases in the users' interpretation of importance score explanations and demonstrate their effectiveness to reduce a distorting effect of word length and anticipate temporal changes in user perceptions, respectively. Second, we develop, validate, and apply a new scale to measure perceived predictability (Section 5.2). We motivate why perceived predictability should be measured, develop and validate our novel 6-item perceived system predictability (PSP) scale, and study the relation between perceived predictability and objective prediction correctness as well as trust and user's NFC in the context of NLP explanations. Our findings uncover orthogonal effects of explanations and system stochasticity on objective prediction correctness and subjective PSP ratings which we link to known cognitive biases. Our results underline the need to investigate subjective predictability in addition to objective user performance measures and demonstrate that our PSP scale is a valid instrument that can and should be used in future investigations of explanatory systems and within broader NLP and human-computer interaction (HCI) contexts.

5.1. Heatmaps Considered Harmful: Cognitive Biases and Saliency Explanations

Heatmap explanations are a popular class of explanation methods to explain model decisions by specifying the parts of the input which are most salient in the model's decision process (Burkart and Huber, 2021; Tjoa and Guan, 2021; Fel et al., 2021b). In NLP, this refers to which subwords, words, phrases, or sentences in the input contributed most to the model prediction

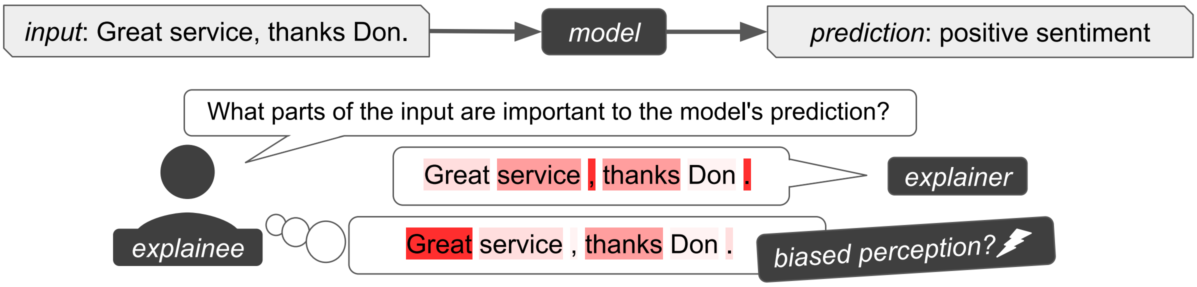

input: Great service, thanks Don.
model
prediction: positive sentiment
What parts of the input are important to the model's prediction?
Great service , thanks Don .
explainer
explainee
Great service , thanks Don .
biased perception?

Figure 5.1.: A saliency explanation is generated to answer the human's need to understand the model. We investigate whether the saliency explanation can be systematically misperceived by humans and which factors influence its perception.

(Madsen et al., 2023; Danilevsky et al., 2020). While much research exists on developing and verifying such explanations (Arras et al., 2017; Adebayo et al., 2018; Kindermans et al., 2019; Tuckey et al., 2019; Wang et al., 2020; Madsen et al., 2021), less is known about the information that human explainees actually understand from them (Miller, 2019; Dinu et al., 2020; Fel et al., 2021a; Arora et al., 2021).

In the explainable NLP literature, it is generally (implicitly) assumed that the explainee interprets the information "correctly", as it is communicated (Arras et al., 2017; Feng and Boyd-Graber, 2019; Fel et al., 2021a): e.g., when one word is explained to be influential in the model's decision process, or more influential than another word, it is assumed that the explainee understands this relationship (Jacovi and Goldberg, 2021). We question this assumption: research in the social sciences describes modes in which the human explainee may be biased — via some cognitive habit — in their interpretation of processes (Malle, 2003; Miller, 2019; Epley et al., 2007; Watson, 2020). Additional research shows this effect manifests in practice in AI settings (Gonzalez et al., 2021; Darling, 2015; Ehsan et al., 2021; Hartzog, 2015; Natarajan and Gombolay, 2020). This means, e.g., that the explainee may underestimate the influence of a punctuation token, even if the explanation reports that this token is highly significant (Figure 5.1), because the explainee is attempting to understand how the model reasons *by analogy to the explainee's own mind* which is an instance of *anthropomorphic bias* (Johnson, 2018; Dacey, 2017; Zlotowski et al., 2015) and *belief bias* (Evans et al., 1983; Gonzalez et al., 2021).

We identify four different such biases which may influence the explainee's interpretation: (i) *anthropomorphic bias* and *belief bias*: influence by the explainee's self projection onto the model, (ii) *visual perception bias*: influence by the explainee's visual affordances for comprehending information, (iii) *learning effects*: observable temporal changes in the explainee's interpretation as a result of interacting with the explanation over multiple instances, and (iv) *chunking effects*: influences within the perception of neighboring explanation elements that are understood as a unit by the explainee (such as "New York").

We thus address the following question in this section: *When a human explainee observes feature-attribution explanations, does their comprehended information differ from what the explanation "objectively" attempts to communicate? If so, how?* We propose a methodology to investigate whether explainees exhibit biases when interpreting feature-attribution explanations in NLP, which effectively distort the objective attribution into a subjective interpretation of it (Section 5.1.1.5). We conduct user studies in which we show an input sentence and a feature-attribution explanation (i.e., saliency map) to explainees, ask them to report their subjective interpretation, and analyze their responses for statistical significance across multiple factors, such as word length, total input length, or dependency relation, using a GAMM (Section 5.1.1.6).

In the first part of this section (Section 5.1.1), we find that word length, sentence length, the position of the sentence in the temporal course of the experiment, the saliency rank, capitalization, dependency relation, word position, word frequency as well as sentiment can significantly affect user perception. In addition to *whether* a factor has a significant influence, we also investigate *how* this factor affects perception. We find that, for example, short words decrease importance ratings while short sentences or intense sentiment polarities increase them.

In the second part of this section (Section 5.1.2), we explore the effect of *phrase-level* features. Concretely, we extend our analysis to investigate the effect of a word's *neighboring words* to the word's rating, conditioned on various a priori measures of bigram constructs, such as the words' syntactic relation or the degree to which they collocate in a corpus. We observe significant effects for (i) left-adjacency vs. right-adjacency, (ii) the difference in importance between the two words, and (iii) the phrase relationship between the words (common phrase vs. no relation) and discuss potential links of our observations to known effects from relevant literature.

Finally, in the third part of this section (Section 5.1.3), we propose two visualization interventions to mitigate learning effect and visual perception biases: model-based color correction and bar charts. We conclude that (a) model-based color correction can predict and mitigate distorting temporal effects and (b) bar charts can successfully remove the influence of word length.

5.1.1. Word- and Sentence-level Factors

5.1.1.1. Feature-attribution Explanations

Feature-attribution explanations aim to convey which parts of the input to a model decision are “important”, “responsible” or “influential” to the decision (Arras et al., 2017; Ribeiro et al., 2016; Carvalho et al., 2019; Madsen et al., 2023; Zhang et al., 2021). This class of methods is a prevalent mode of describing NLP processes (Madsen et al., 2023; Danilevsky et al., 2020; Kaur et al., 2020; Tenney et al., 2020), due to two main strengths: (i) it is flexible and convenient, with many different measures developed to communicate some aspect of feature importance and (ii) it is intuitive, with — seemingly, as we discover — straightforward interfaces of relaying this information. Here we cover background on feature-attribution on two fronts: the underlying technologies (Section 5.1.1.2) and the information which they communicate to humans (Section 5.1.1.3).

5.1.1.2. Attribution Methods

We consider feature-attribution explanations generally as scoring (or ranking) functions that map portions of the input to scores that communicate some aspect of importance about the aligned portion: $E_f(f(\mathbf{x})) : \Sigma^n \to \mathbb{R}^n$, where E_f is the explanation method with respect to f, f is the model and $\mathbf{x} \in \Sigma^n$ the input text to the model, i.e., the input consists of n tokens which are elements of an alphabet Σ. A high score implies high importance.

The loose definition proposed above for feature-attribution explanations as communicating “important” portions of the input (words, sub-words, or characters) is often interpreted with a causal lens: that by intervening on the tokens assigned a high score, the model behavior will change more than by intervening on the tokens assigned a low score (Jacovi and Goldberg, 2021; Grimsley et al., 2020; Arras et al., 2017). This perspective is relaxed in various ways to produce various softer measures of importance: for example, *gradient-based methods* measure the change required in the embedding space to cause a change in model output, while *Shapley-value methods* measure the change with respect to the “average case” in the data.

The granularity provided in the scoring function may vary greatly, from a binary measure — important or not important — to a complete saliency map, depending on the tokenization granularity, the method, and visualization. Most commonly, the explanation is given as a colorized saliency map over word tokens (e.g., Arras et al., 2017; Wang et al., 2020; Tenney et al., 2020; Arora et al., 2021). Note that this section is *not* concerned with a particular feature-attribution method, but rather how feature-attribution explanations generally communicate information to human explainees, and what the explainees comprehend from them.

5.1.1.3. Social Attribution: The Case of Text Marking

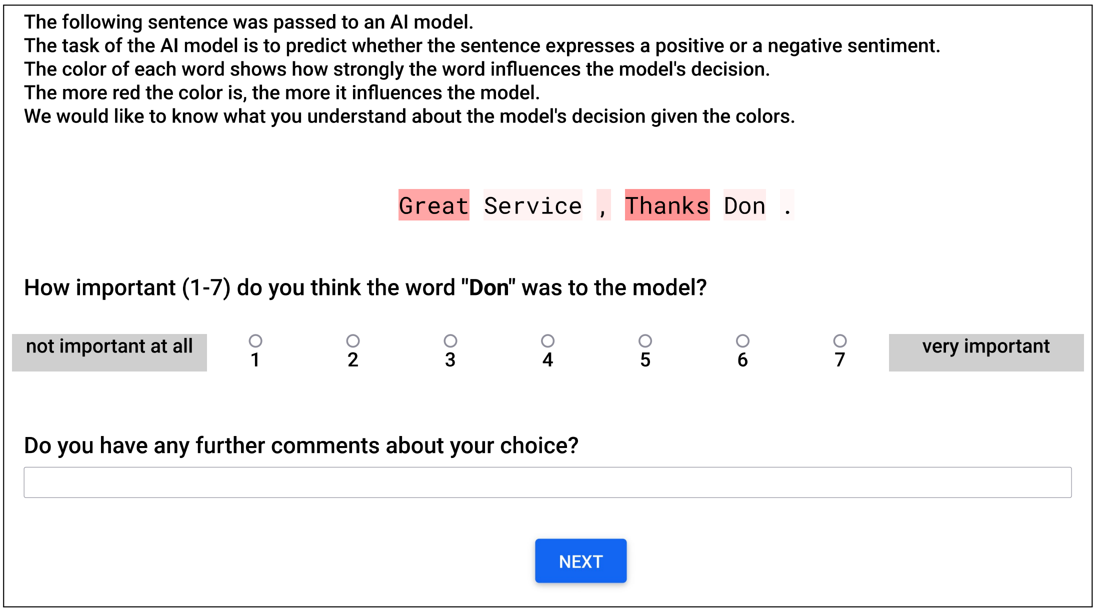
Is it really possible for the explainee to comprehend feature-attribution explanations differently from what they objectively communicate? What is the nature of any discrepancy in this perception?[1] As Miller (2019) writes, literature in the social sciences about how humans comprehend explanations and behavior can help illuminate this problem.

In particular, we assume that the human explainee comprehends the explanation with respect to their own reasoning. By assigning human-like reasoning to the model behavior being explained (Miller, 2019), the explainee may fill any incompleteness in the explanation with assumptions from their own priors about what is plausible to them (Dacey, 2017; Gonzalez et al., 2021). To demonstrate, consider the case of binary feature-attribution — marking parts of the input as "important" and "not important", also known as *highlighting* or *extractive rationalization* (Lei et al., 2016). Even this simple format of communicating information can be assigned human-like reasoning by the explainee, on account of "*who marked this text*" and "*for what purpose*": Marzouk (2018) identifies various objectives that humans follow when marking or observing marked text, e.g., marking forgettable sections (for memorization), marking as a summary (for subsequent reading), marking exemplifying text, marking contradicting or surprising text, etc. In the context of NLP models, Jacovi and Goldberg (2021) note two central objectives: reducing the input to a summary that comprehensively informs the decision, or identifying influential evidence in the input which non-comprehensively supports the decision.

These many different objectives can influence the choice of marking, and the information that it communicates. This means that both the marked text, and the choice of what text to mark, are information that the explainee comprehends when observing the explanation. Therefore, how the explanation is perceived is influenced by both factors.

Text marking is a special case of feature-attribution. The above demonstrates how the explainee's interpretation is potentially shaped by aspects of the explanation which are implicit or unintended — leading to an "erroneous" interpretation of the explanation. We identify four biases that may cause this effect, as motivation for our investigation: (i) anthropomorphic bias and belief bias, via the explainee's a priori opinion on human-like or plausible reasoning, (ii) visual perception bias, via characteristics of the explainee's visual affordances for comprehending information, (iii) learning effects, as observable influence in the explainee's interpretation by previous explanation attempts in-context, and (iv) assimilation and contrast effects, as influences of neighboring words and lexical chunks.

[1] This question is distinct from the question of whether the explanation faithfully communicates information about the model (Wiegreffe and Pinter, 2019; Jacovi and Goldberg, 2020): even if the feature-attribution information is entirely faithful, discrepancies may still arise in how humans comprehend this information.



The following sentence was passed to an AI model.
The task of the AI model is to predict whether the sentence expresses a positive or a negative sentiment.
The color of each word shows how strongly the word influences the model's decision.
The more red the color is, the more it influences the model.
We would like to know what you understand about the model's decision given the colors.

Great Service , Thanks Don .

How important (1-7) do you think the word "**Don**" was to the model?

not important at all ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 very important

Do you have any further comments about your choice?

Figure 5.2.: Screenshot of the importance rating interface for English sentiment sentences.

5.1.1.4. Study Overview

Research Question. The core research question of the following is to probe into which, if any, factors in the explanation process — aside from the saliency itself — may influence the explainee's interpretation of the saliency information. Formally, we view the saliency explanation as a process whose result is the explainee's interpretation of the saliency scores. The "input" to this process is the original text as well as the saliency information and the visualization method. Then, we ask which factors in the original text have statistically significant effects on the explainee's interpretation and how properties of the saliency score and the visualization method affect it. Notably, a key challenge in analyzing the explainees' saliency understanding is that we want to identify influencing factors on the explainee's ratings without the existence of an inherently correct ground truth perception.

Proposed Methodology. We propose a combination of study design and statistical analysis to quantify the influence of arbitrary factors, such as word length, sentiment polarity, or dependency relations. We collect explainees' subjective interpretations of the saliency scores in a crowdsourcing setup. We relate this interpretation to the original explanation considering various potentially influencing factors using an ordinal GAMM (we refer to our brief introduction to GAMMs in Section 2.4.2.5). The result from this comparison is an answer on *which* of the a priori candidate explanatory factors indeed have significant effect on the explainee's interpretation and *how* these factors functionally affect interpretation.

5.1.1.5. Study Methodology Specification

The study consists of two phases: collecting subjective importance interpretations (Section 5.1.1.5) and analyzing responses with an adequate statistical model (Section 5.1.1.5). We release the collected data and analysis code.[2]

Collecting Self-reported Importance Ratings. In our main study, we investigate the interpretation of color-coding saliency visualization of the feature-attribution by crowdsource laypeople (variations on this study will be described later). We measure the perceived importance of a word within a saliency score explanation by directly probing human self-reported word importance. In this instance, we ask "How important (1-7) do you think the word "X" was to the model?" (Figure 5.2). We collect answers on a single-item unipolar 7-point Likert scale ranging from *not important at all* to *very important*.

Texts: We use sentences from the Universal Dependencies English Web Treebank (Silveira et al., 2014).[3] This treebank contains comprehensive annotation, including dependency relations of sentences, stemming from various domains, such as newsgroups or online reviews. We use sentences from the reviews group for a plausible framing of a sentiment analysis task. We choose sentences without sub-token dependency relations (e.g., excluding "it's" because displaying it as two tokens breaks the orthography) and with unique word occurrences (i.e., excluding sentences that contain a word several times). From this subset we remove length outliers: sentences with a number of tokens longer than one standard deviation above the mean (concretely, eleven tokens). We randomly select 150 sentences to be used.

Saliency Scores: We assign random saliency scores to each token to uniformly sample the space of saliency intensities. We are, at this stage, not interested in using a "real" model or saliency score (e.g., attention or Integrated Gradients), as we investigate general perception of arbitrary scores. It is therefore useful to create saliency scores that "do not make sense" because a saliency score should reflect the model's reasoning which might very well not make sense at all. We study two "real" saliency scores later in Section 5.1.1.6 (Integrated Gradients) and Section 5.1.2 (SHAP).

Study Interface: See Figure 5.2 for the rating collection interface. We display all sentences using monospaced font and fixed whitespaces to obtain a direct mapping between the number of characters and the color area for each word (ligatures and other typographic attributes of non-monospaced fonts would break this mapping).

[2] https://github.com/boschresearch/human-interpretation-saliency

[3] https://github.com/UniversalDependencies/UD_English-EWT

Procedure: We ask participants to rate the importance of a randomly-selected word in the sentence.[4] We show all 150 sentences from the described review dataset to each participant, displayed in a randomized order per participant. Saliency scores for all tokens are randomized for each participant (such that we collect responses to many different saliency maps, rather than numerous responses for the same set). We do so because our aim is not to obtain accurate (mean) estimates of single ratings as one would do in a corpus annotation, but to collect rich data to build an accurate model describing the underlying general phenomenon. For each sentence, we collect the participant's importance rating, the completion time, and a voluntary free-text comment. We choose to not include a dedicated training phase, e.g., showing the participants ten explanation instances before starting the data collection as we explicitly want to study learning effects. These can be crucial in real-world applications: e.g., should we find a decaying learning effect, an effective model audit should ensure to include a sufficient number of model predictions. To filter out (a) careless responses and (b) noisy responses due to decreased participant attention towards the end of the study, we insert three trap sentences at random positions in the last two thirds of the real sentences. See an example and more integration details in Figure A.14 in Appendix A.5.1.

Participants: We recruit 50 crowdworkers on MTurk. One crowdworker failed all of the trap sentences, so we exclude this worker's responses and recruit one additional worker. All other participants successfully passed all trap sentences. In total, this yields 7500 importance ratings.

Factors of Saliency Perception. For our set of candidate factors, we model factors that are motivated by the three types of biases: anthropomorphic and belief biases, visual biases, and learning effects. Each factor is tested for statistical significance on the explainees' interpretations. Table 5.1 lists the factors we investigate in this subsection. Selected factors include: (i) *word length* as longer words correspond to a larger colored draw area, which we hypothesize influences visual perception bias, (ii) *word polarity* as we present participants a sentiment classification task and expect that the participants' own assessment of word importance influences their perception of how important it is to the model, which we hypothesize is an instance of belief bias, (iii) *display index* as we hypothesize that participant ratings are affected by temporal effects, such as learning, (iv) *word position* as we hypothesize that, e.g., words at the center of a sentence are perceived more strongly due to the center bias observed in various eye-tracking studies, i.a., for natural scenes (Tseng et al., 2009). We derive word

[4] Alternatively, one can imagine a setting in which participants rate all words within the sentence. We choose to ask for single-word ratings to (i) avoid carry-over effects from ratings of the first to the last words and (ii) collect ratings of more sentences within the same experiment time compared to splitting the set of sentences over participants which would introduce further dependencies in the statistical analysis.

Factor	**Description**	**Significant Effects**		
		EN	*DE*	*EN-IG*
Saliency	The color intensity specified as the saturation value ($S \in [0, 1]$) in a (H, S, V) color triple (Smith, 1978), e.g., (0°,0.5,1.0) (🟥) and (0°,0.25,1.0) (🟥).	✓	✓	✓
Word length	The number of characters in a word, e.g., 7 for “example”.	✓	✓	✓
Word frequency	The word’s normalized frequency, estimated on a large corpus.			✓
Sentence length	Number of words in the sentence.	✓	✓	
Display index	The sentence’s position within a sequence of sentences (e.g., the third sentence in the sequence of 150 sentences). This relates to temporal effects such as learning.	✓	✓	
Sentiment polarity	The sentiment polarity of a word (defined via its lemma) $\in [-1, 1]$.	✓	–	
Saliency rank	Normalized rank of a word’s saliency score (i.e. color intensity) in comparison to the other words in its sentence $\in [0, 1]$.	✓		✓
Word position	The index of the token’s position within its sentence.		✓	
Capitalization	The word’s capitalization, e.g., “example”, “Example” or “EXAMPLE”.		✓	
Dependency relation	Dependency relation to its parent within the dependency graph (36 types for *EN*).		✓	

Table 5.1.: List of factors that presupposedly affect saliency explanation perception along with the findings of our three user studies. EN refers to the English sentiment classification study, DE to the German fact checking study and EN-IG to the English sentiment classification study using Integrated Gradients as feature attribution method (without correction visualizations).

frequencies from the WikiMatrix corpus (Schwenk et al., 2021) and sentiment polarities from SentiWords (Gatti et al., 2016).

Statistical Analysis Using GAMMs. Given a set of inputs for which there are the feature-attribution scores and the interpreted importance scores, we describe the analysis methodology aiming to derive the possible input factors that cause a discrepancy between the two.

Ordinal Generalized Additive Mixed Model: We analyze the collected ratings of perceived importance using an ordinal GAMM. Its key properties are that it (i) models the ordinal response variable (i.e., the importance ratings in our setting) on a continuous latent scale (*ordinal generalized*), which is (ii) modeled as a sum of smooth functions of covariates (*additive*) and (iii) accounts for random effects (*mixed*). The continuous latent scale is linked to ordinal categories by estimating threshold values that separate neighboring categories. The smooth functions can comprise single covariates (*univariate* smooths), such as $f_1(x_1)$ or combinations of multiple covariates, such as $f_2(x_2, x_3)$. Random effects allow to account for, e.g., systematic differences in individual participants’ rating behavior. For example, a specific participant might have a tendency to give overall higher ratings than other participants. Including a *random effect* allows to disentangle this influence on the response variable from the influence of the covariates in question (such as word length) and thereby offers a clearer view on these *fixed effects*. The GAMM analysis enables us (i) to make statements about which factors significantly influence saliency perception, without prescribing any notion of "correct perception" and (ii) to study the relation between these factors and participants' importance ratings in detail, via an interpretation of the model's parametric terms (categorical factors) as well as smooth terms (numeric factors). We provide a brief introduction to ordinal GAMMs in Section 2.4.2.5.

Model Details: We include all factors listed in Table 5.1 into our model formula. We use smooth terms for numeric factors and parametric terms for categorical factors. Additionally, we include tensor product interactions for all pairs of smooth terms.[5] In order to statistically account for potentially confounding effects of individual participants or sentences, we include random intercepts as well as random slopes for each participant and each sentence. Before fitting the model, we remove a small number of outlier ratings. We remove outliers from the initial 7500 importance ratings by excluding words with 20 or more characters (8 ratings) and ratings with a completion time of 60 seconds or more (50 ratings), leaving 7442 ratings left for analysis. We apply the identical filters to the study described in Section 5.1.3. For the German study described in Section 5.1.1.6, we only apply the completion time filter. We use fast REML for smoothness selection and apply variable selection via double-penalty shrinkage (i.e., additionally penalizing the splines' null space). We fit the model using discretized covariates as described in Wood et al. (2017) and Li and Wood (2020).[6]

5.1.1.6. Study Results, Interpretation, and Generalizations

In the following, we conduct three user studies. The first study (Section 5.1.1.6) investigates saliency perception for English and a sentiment classification task. The second study (Section 5.1.1.6) extends the investigation to German language and a fact checking task to evaluate generalization of the findings. Since these two studies use random saliency scores so as to not prescribe a specific feature-attribution method, we report a third study (Section 5.1.1.6) which uses the wide-spread Integrated Gradient scores as a generalization to practically-used attribution methods.

[5] Such a functional ANOVA decomposition is supported by mgcv and allows to study, e.g., the interaction between word length and sentiment polarity in addition to the isolated main effects of word length and sentiment polarity.

[6] We use R and mgcv (Wood, 2011; Wood et al., 2016; Wood, 2004, 2017, 2003) (1.8-38) to fit all our models.

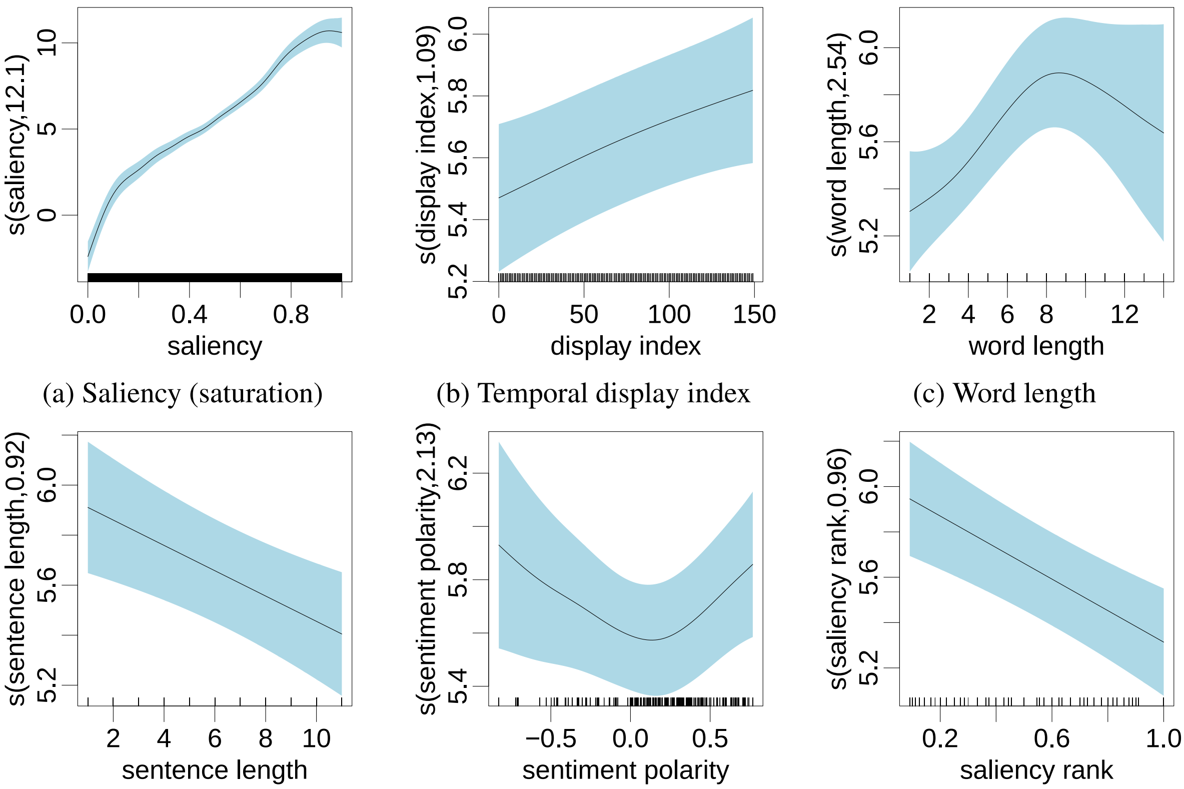
	edf	ref. df	F	p
s(saliency)	12.0967	19	728.8738	< **0.0001**
s(display index)	1.0921	9	2.0872	**0.0001**
s(word length)	2.5416	9	4.1826	< **0.0001**
s(sentence length)	0.9200	9	1.7531	**0.0001**
s(word frequency)	0.0011	9	0.0001	0.1082
s(sentiment polarity)	2.1281	9	1.6156	**0.0065**
s(saliency rank)	0.9580	9	4.4417	< **0.0001**
s(word position)	0.0005	9	0.0000	0.7882

Table 5.2.: Effective degrees of freedom (edf), reference df and Wald test statistics for the uniariate smooth terms of the first user study.

Sentiment Analysis in English. In the following, we discuss quantitative results based on the fitted GAMM as well as qualitative findings based on the participants' written feedback. Table 5.2 shows statistics for the univariate smooth terms in the fitted GAMM. Figure 5.3 shows partial effect plots of the respective significant smooth terms. Regarding the parametric terms, neither a word's capitalization (df=2, F=1.84, p=0.16) nor its dependency relation (df=35, F=1.17, p=0.24) show a significant effect on perceived importance. Regarding the smooth terms, we observe that saliency score, display index, word length, sentence length, word sentiment polarity, and saliency rank show significant effects on perceived importance. In addition to the significant partial effects, we also find numerous significant interactions. We provide the statistics of Wald tests for all pairwise tensor product interactions (following a functional ANOVA decomposition) as well as summed-effect plots of all significant pairwise interactions in Table A.5 and Figure A.15 in Appendix A.5.2. In the following, we discuss each partial effect in detail.

Saliency (Figure 5.3a): The saliency (i.e., the color saturation) has the strongest impact on perceived importance as the graph spans the by-far widest y-axis range of all plots in Figure 5.3. Except for the saliency scores around 1, the entire graph shows a monotonous relation between saliency score and perceived importance.

Display Index (Figure 5.3b): Participants' ratings increased over the course of the experiment. We hypothesize that the participants report more conservative ratings at the beginning of the experiment to "leave enough room" for more extreme sentences and adapt their ratings to a more "calibrated" level over the course of the experiment. Interestingly, this trend does not seem to stop after our maximum number of 150 sentences. We leave the study of sufficient amount of training required for the effect to reach a peak to future work.



(a) Saliency (saturation)

(b) Temporal display index

(c) Word length

(d) Sentence length

(e) Word sentiment

(f) Saliency rank

Figure 5.3.: Partial effect plots for all significant smooth terms (note that y-axes are scaled per effect). Numbers in y-axis labels are estimated degrees of freedom (edf) of the respective smooth. The shaded area displays confidence intervals (plus and minus one standard error) including uncertainty about the overall mean.

Word Length (Figure 5.3c): With increasing word length, importance ratings rise up until a length of approximately eight characters and decrease again afterward. We hypothesize that the initial increase corresponds to an increase in the colored area that a longer word directly causes, as the saliency score is visualized within a box that is proportional to the number of characters. To interpret the subsequent decrease in perceived importance, we consider the interactions between word length and other factors. We find significant pairwise interactions of word length with (i) saliency, (ii) display index, and (iii) word frequency (Appendix A.5.2). For the interaction with display index, we observe that the decreasing effect of high word lengths grows with increasing display index up until around the 55th sentence. After this point, the effect decreases. While the latter decrease can be explained by the partial effect of increasing ratings with higher display indices (as shown in Figure 5.3b), the former decrease demands detailed investigation in future work.

Sentence Length (Figure 5.3d): Importance ratings decrease for words in longer sentences. A longer sentence leads to a higher number of color samples and therefore also to a larger expected color range. We argue that such an increased color range inhibits users to make very high importance ratings due to a missing "maximum color" anchor.

Sentiment Polarity (Figure 5.3e): The effect of a word's lemma's sentiment polarity on importance ratings. We observe a parabola-shaped curve with a minimum at slightly-positive sentiment. To the left, importance ratings increase with increasingly negative polarity, and to the right importance ratings increase with increasingly positive polarity. This suggests that users' ratings of "what was important to the model when classifying the sentence" are biased by their answer to "what is important to me when classifying the sentence myself". Such a substitution of a presumably complex-to-compute target attribute with a simpler heuristic attribute is a known cognitive bias and often referred to as *attribute substitution* or *substitution bias* (Kahneman and Frederick, 2002).

Saliency Rank (Figure 5.3f): The partial effect of a word's normalized saliency rank on participants' importance ratings. We normalize the rank by dividing by sentence length, as low ranks (i.e., larger numbers) would otherwise be strongly correlated to sentence length, and potentially cause stability issues within the model estimation. We observe that an increased rank (a value of one corresponds to the last rank, i.e., the lowest saliency score) corresponds to a decrease in rated importance. In contrast to the effect of saliency score shown in Figure 5.3a, the saliency rank is not only a property of a word but of a word in context of its sentence. A word's saliency score can remain unchanged while at the same time, its rank can be arbitrarily modified by changing the saliency scores of the other words in its sentence. We argue that the significant effect of saliency rank indicates that users interpret saliencies *in relation* to each other, i.e., their judgments are relative and lack a fixed anchoring point. This is supported by the qualitative analysis discussed in the following.

In addition to the statistical evaluation, we also evaluate the participants' voluntary free-text comments. Table 5.3 shows a selection of comments grouped into four categories:

Relative Judgment: Participants explicitly state that they make relative importance judgments. This supports our argumentation of relative judgments discussed for the effects of sentence length and saliency rank.

Own Opinion: Similarly, participant comments support our hypothesis that users' ratings are subject to the cognitive bias of attribute substitution as discussed for the effect of word sentiment polarity.

Type	Sentence & Saliency	Rating	Comment
relative	Best Electrician in Florence	2	"Best" highlighted in the light pink was not scored as high as the other words in deeper shades of red, so I assume the model didn't find it very important. (P11)
	Absolutely amazing job !	3	I see 4 different levels of highlights. Absolutely seems to be the third darkest so that's why I chose 3 (P20)
own opinion	Room was amazing .	3	I am uncertain why the period at the end of the sentence would be important, so I choose a 3, even though the AI coded it as red color. (P26)
	best	2	I would think that if it's one word then the word should be important. But I don't think it is important because it's such a light color (P20)
light color	Would do business with them again .	1	the symbol has no color code around it at all so I chose 1 (P26)
	David Bundren is the Tire GooRoo .	1	It probably didn't even notice the last name (P44)
other	Listened to my problem and took care of it .	7	Now I understand the range of red colors better. "it" outside of the phrase "care of it" is meaningless, but since blanks between words are NOT colored, I have to think that AI is judging "it" by itself. (P39)
	Great Place !	7	well you state that the redder the word is, the more influence it has...that's pretty red. (P44)

Table 5.3.: Comments of the participants of the English sentiment study. Participants were asked to rate the underlined word or symbol.

Light Color: Participants seem to make a categorical distinction between *very light color* and *seemingly no color* although this distinction does not exist in terms of the attribution score. This can be important when communicating very low influences and should be addressed in more detail in future work.

Other: Miscellaneous comments on, e.g., issues of word-level attribution and the resulting ambiguity in interpretation.

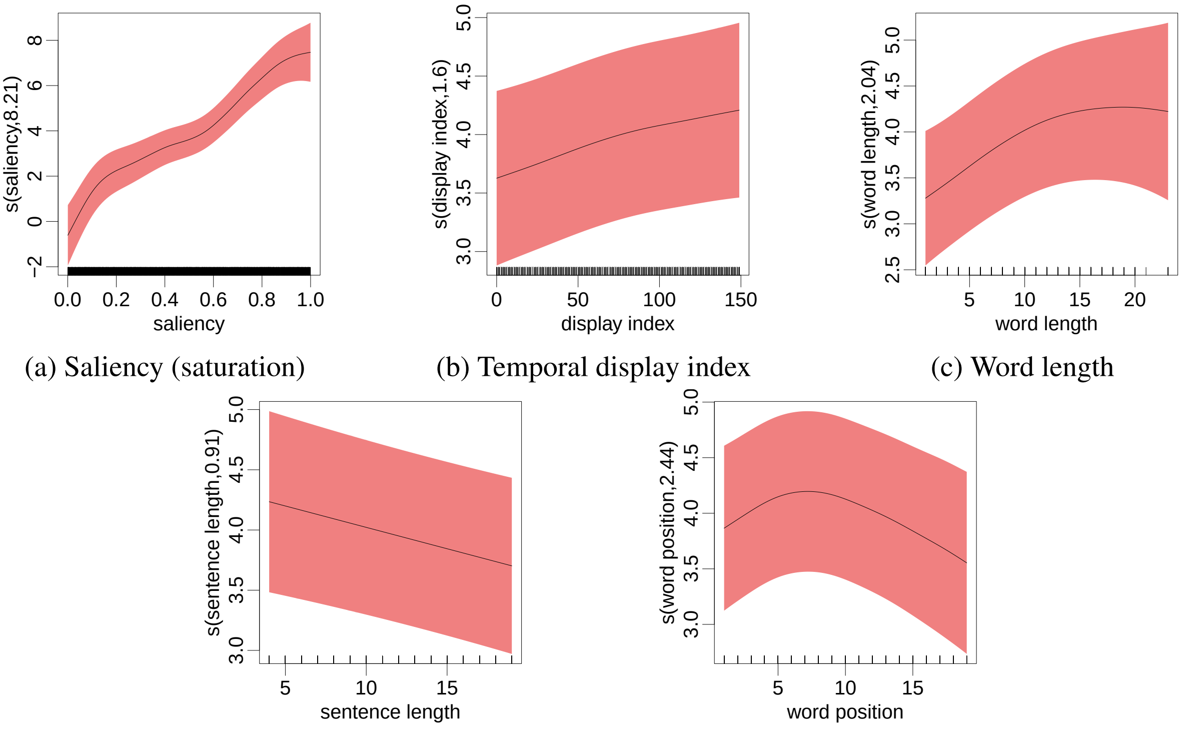
Generalization Across Tasks and Languages: Fact Checking in German. So far, we found indication that numerous factors (word length, saliency rank, etc.) significantly influence users' subjective importance ratings. Two important limitations are that (i) the findings are limited to English and (ii) they are limited to one AI task (sentiment classification). To assess whether the findings do generalize to another language and another task, we repeat the study identically with German sentences from the PUD Corpus[7] with a fact checking AI task. We collect responses from 25 German-speaking participants from a participant pool including Germany, Austria, and Switzerland. In total, this corresponds to 3750 ratings.

[7] https://universaldependencies.org/treebanks/de_pud/index.html.

Confirmed Effects: Our analysis confirms the significant effects of saliency, display index, word length, and sentence length. Figure 5.4 displays the respective partial effect plots. While the smooths for saliency (Figure 5.4a) and sentence length (Figure 5.4d) show high similarity to the respective smooths of the English study (see Figures 5.3a and 5.3d), we observe slight differences for display index (Figures 5.4b and 5.3b) and word length (Figures 5.4c and 5.3c). While the English display index smooth grows more or less linearly (edf=1.09), the respective German smooth reaches a plateau after around half the sentences (edf=1.60). We hypothesize that such a saturation effect will also be visible for English, but requires a larger number of sentences. We argue that this is caused by the fact that the sentences in the German study are longer than in the English study, which makes participants of the German study see more colored words and thereby "calibrates" their ratings faster in terms of number of sentences. Similarly, the German word length smooth saturates after around 15 characters, while the English smooth decreases after around eight characters. We hypothesize that this difference can be attributed to the overall longer words in German as well as the differences in compounding. The effect of saliency rank cannot be confirmed in the German experiment. Like in the English study, we find no indication that word frequency has a significant effect on importance ratings. We provide test statistics of parametric and smooth terms (univariate smooths and pairwise interactions) as well as coefficient estimates in Appendix A.5.3. As in the English study, we additionally qualitatively analyze the participants' free-text comments and observe (as in the English study) numerous instances for which participants mix their own estimate of importance with the communicated importance. We provide exemplary instances in Appendix A.5.3.

Additional Effects: In addition to the effects that we already observed in the English study, we also find that the word's position within the sentence (Figure 5.4e) as well as capitalization and dependency relation have significant effects on importance ratings. A full list of coefficient estimates along with further details is provided in Appendix A.5.3. The estimate for fully capitalized words is 1.91 (SE=0.96), the respective estimate for words with the first letter capitalized is 0.41 (SE=0.12).[8] This confirms the intuition that fully capitalized words receive the highest importance ratings, followed by first-letter-capitalized words. We argue that this effect — in particular for first-letter-capitalized words — is more visible in the German experiment as German uses more frequent capitalization (e.g., for all nouns). Regarding dependency relations, the highest estimate can be observed for temporal modifiers (obl:tmod, $\beta = 1.70$, SE=0.55) like "today" and numerical modifiers (nummod, $\beta = 1.39$, SE=0.36) like "one". The lowest estimate can be observed for clausal modifier of nouns (acl, $\beta = -1.22$, SE=0.64) like "sees"

[8] The estimate for lower-cased words is fixed to zero as the reference level. For dependency relations, we choose the (most frequent) punctuation relation.



(a) Saliency (saturation)

(b) Temporal display index

(c) Word length

(d) Sentence length

(e) Word position

Figure 5.4.: Partial effect plots for all significant smooth terms (note that y-axes are scaled per effect) for the German experiment. Numbers in y-axis labels are estimated degrees of freedom (edf) of the respective smooth. The shaded area displays confidence intervals (plus and minus one standard error) including uncertainty about the overall mean. (a) refers to color saturation.

in “the issues as he sees them” and indirect objects (iobj, $\beta = -0.48$, SE=0.52) like “me” in “she gave me the book”. We hypothesize that the grammatical function effect is larger here than in the previous experiment because, i.a., the use of temporality, numerals, and embedded clauses are more important for determining factuality than for determining sentiment.

Generalization to Model-based Saliencies. We want to assess whether our findings on the random saliency scores used in the previous two studies also hold for practically-used feature attribution scores. Therefore, we conduct an additional user study using Integrated Gradients (Sundararajan et al., 2017) instead of random saliencies.[9]

Study Modification: Within-Subject Design: We combine the evaluation of Integrated Gradient scores with a within-subject evaluation of three visualization methods which we detail in

[9] We make use of the Language Interpretability Toolkit (Tenney et al., 2020) to obtain normalized Integrated Gradient scores with respect to the SST2-base sentiment model and 30 interpolation steps.

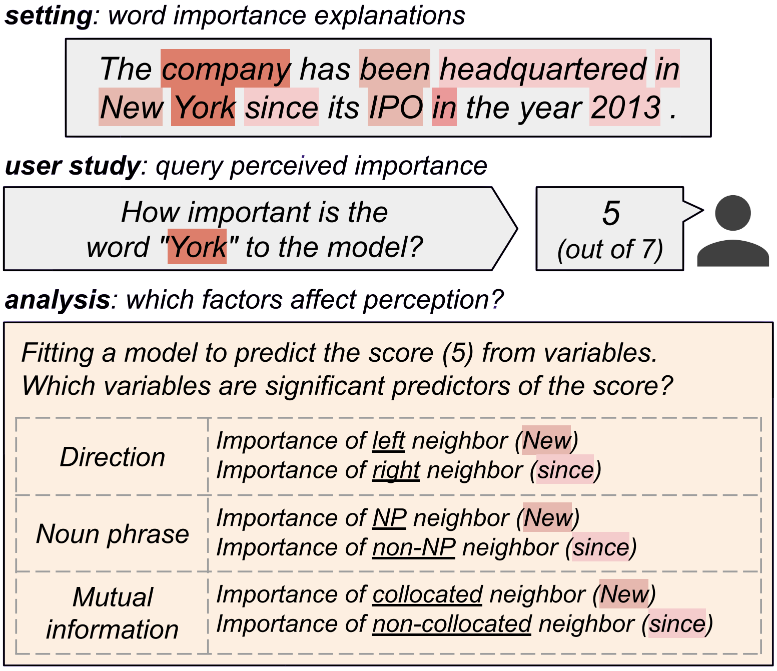
Section 5.1.3. In this section, we focus on the unmodified visualization as it is used in the two previously described studies. In the remainder of this section, this visualization method is referred to as *saliency*. We sample another 150 sentiment sentences from the sentence pool described in Section 5.1.1.5 and present them in the same sentiment classification context. Instead of using one saliency visualization method for all 150 sentences, we now use the three visualizations and show each participant 50 sentences per visualization.[10] We collect 9000 importance ratings from 60 participants and exclude participants from the previous study to avoid carry-over effects from previous exposures.

Model Modification Using Factor-Smooth Interactions: We again use an ordinal GAMM using the same covariates as in Section 5.1.1.5. We add a parametric term for the visualization condition to account for overall differences in rating intensities between the visualization conditions and include a random intercept to account for visualization order. We use factor-smooth interactions for each variable which leads to separate estimates for each variable per visualization (e.g., three smooths for word length, one per visualization). First, this yields smooths for the "original" saliency visualization, i.e., the heatmap visualization without corrections. In contrast to our first study, these smooths now correspond to effects on Integrated Gradient saliencies instead of random saliencies. First, comparing the smooths allows us to compare how factors influence importance ratings across visualizations, e.g., to assess whether the bar visualization did mitigate the biasing effect of word length. We discuss the respective results in Section 5.1.3.3. Second, analyzing the smooths relating to the original visualization allows us to evaluate which of the effects we observed in the first study do generalize to the Integrated Gradients attribution scores. We discuss the respective results in the following.

Results: We find significant effects of saliency score, word length, relative word frequency, and saliency rank. We provide details and test statistics on all parametric coefficients as well as smooth terms in Table A.17 in Appendix A.8. All of these variables except relative word frequency were also found to be significant in our first study and all of them except relative word frequency and saliency rank were confirmed in our German study. The significant influence of relative word frequency was observed for the first time.

Overall, three studies confirmed the presumably biasing influence of word length, (pairs of) two studies respectively confirmed the effect of sentence length, display index, and saliency rank, and one study (each) found significant effects of word position, sentiment polarity, word frequency, capitalization, and dependency relation. Together, these reflect the three sources of bias: anthropomorphism and belief bias, visual perception, and learning effects.

[10] The order of visualization methods is balanced across participants. Sentence order is fixed to ensure identical ordering effects for the three visualizations.


setting: word importance explanations
The company has been headquartered in New York since its IPO in the year 2013 .
user study: query perceived importance
How important is the word "York" to the model?
5 (out of 7)
analysis: which factors affect perception?
Fitting a model to predict the score (5) from variables. Which variables are significant predictors of the score?

Direction	Importance of left neighbor (New) Importance of right neighbor (since)
Noun phrase	Importance of NP neighbor (New) Importance of non-NP neighbor (since)
Mutual information	Importance of collocated neighbor (New) Importance of non-collocated neighbor (since)

Figure 5.5.: Illustration of the user study. We ask laypeople to rate the perceived importance of words following a word-importance explanation (*grey*). Then we analyze the effect of the importance of neighboring words on this interpretation, conditioned on the relationship between the words across various measures (*orange*).

5.1.2. Neighboring Words: Assimilation and Contrast

So far, we explored how word features (such as word length and capitalization) and sentence features (such as sentence length) affect human interpretation of importance scores visualized using heatmaps. We demonstrated that numerous factors have a distorting effect on human importance ratings. In the following, we extend our analysis to *phrase-level* features and their influence on the perceived importance of a particular word: Text is naturally constructed and comprehended in various levels of granularity that go beyond the word level (Chomsky, 1957; Xia, 2018). For example (Figure 5.5), the role of the word "*York*" is contextualized by the phrase "*New York*" that contains it. Given an explanation that attributes importance to "*New*" and "*York*" separately, what is the effect of the importance score of "*New*" on the explainee's understanding of the importance "*York*"? We investigate this question in the following.

As we demonstrated in Section 5.1.1, it is not trivial for an explanation of an AI system to successfully communicate the intended information to the explainee (Miller, 2019; Dinu et al., 2020; Fel et al., 2021a; Arora et al., 2021). In the case of *feature-attribution* explanations (Burkart and Huber, 2021; Tjoa and Guan, 2021), which commonly appear in NLP as explanations based on word importance (Madsen et al., 2023; Danilevsky et al., 2020), we must understand how the explainee interprets the role of the attributed inputs on model outputs (Nguyen et al., 2021; Zhou et al., 2022). Research shows that it is often an error to assume that explainees will interpret explanations "as intended" (Gonzalez et al., 2021; Ehsan et al., 2021).

As our approach discussed in Section 5.1.1, the following study involves two phases (Figure 5.5). First, we collect subjective self-reported ratings of importance by laypeople, in a setting of color-coded word importance explanations of a fact-checking NLP model (Section 5.1.2.1, Figure 5.6). Then, we fit a GAMM to map the importance of *neighboring words* to the word's rating, conditioned on various a priori measures of bigram constructs, such as the words' syntactic relation or the degree to which they collocate in a corpus Kolesnikova (2016).

We observe significant effects (Section 5.1.2.3) for (i) left-adjacency vs. right-adjacency, (ii) the difference in importance between the two words, and (iii) the phrase relationship between the words (common phrase vs. no relation). We then deduce likely causes for these effects from relevant literature (Section 5.1.2.4). We are also able to reproduce our findings from Section 5.1.1 in a different English language domain (Section 5.1.2.2). We release the collected data and analysis code.[11]

We conclude that laypeople's interpretation of word importance explanations in English *can be biased via neighboring words' importance*, likely moderated by reading direction and phrase units of language. Future work on feature-attribution should investigate more effective methods of communicating information (Mosca et al., 2022a; Ju et al., 2022), and implementations of such explanations should take care not to assume that human users interpret word-level importance objectively.

5.1.2.1. Study Specification

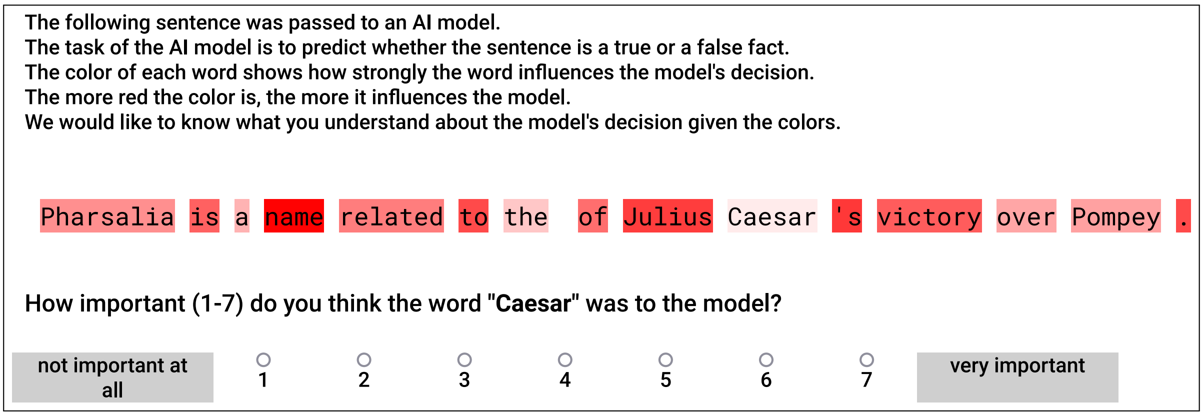
We first collect subjective interpretations of word-importances from laypeople, and then test for significant influence in various properties on the collected ratings — in particular, properties of *adjacent words* to the rated word.

Collecting Perceived Importance. As in Section 5.1.1, we ask laypeople to rate the importance of a word within a feature-importance explanation (Figure 5.6). We use the MTurk crowdsourcing platform to recruit a total of 100 participants.[12]

Explanations: As in Section 5.1.1, we use color-coding visualization of word importance explanations as the more common format in the literature (e.g., Arras et al., 2017; Wang et al., 2020;

[11] https://github.com/boschresearch/human-interpretation-saliency

[12] We select English-speaking raters from English-speaking countries and analyze responses from 64 participants for our first and 36 participants for our second experiment. Details are provided in Appendix A.6.1.



The following sentence was passed to an AI model.
The task of the AI model is to predict whether the sentence is a true or a false fact.
The color of each word shows how strongly the word influences the model's decision.
The more red the color is, the more it influences the model.
We would like to know what you understand about the model's decision given the colors.

Pharsalia is a name related to the of Julius Caesar 's victory over Pompey .

How important (1-7) do you think the word "**Caesar**" was to the model?

not important at all ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 very important

Figure 5.6.: Screenshot of the rating interface.

Tenney et al., 2020; Arora et al., 2021). We use importance values from two sources: Randomized, and SHAP-values[13] (Lundberg and Lee, 2017) for `facebook/bart-large-mnli`[14] (Yin et al., 2019; Lewis et al., 2020) as a fact-checking model.

Task: We communicate to the participants that the model is performing a plausible task of deciding whether the given sentence is fact or non-fact (Lazarski et al., 2021). The source texts are a sample of 150 Wikipedia sentences from the *Wikipedia Sentences* collection.[15] in order to select text in a domain that has a high natural rate of multi-word chunks.

Procedure: As in Section 5.1.1, we ask the explainee: "How important (1-7) do you think the word [...] was to the model?" and receive a point-scale answer with an optional comment field. This repeats for one randomly-sampled word in each of the 150 sentences.

Measuring Neighbor Effects. Ideally, the importance ratings of a word will be explained entirely by its color-coded saliency. However, as we showed in Section 5.1.1, this is not the case. Here, we are interested in whether and how much the participants' answers can be explained by properties of neighboring words, *beyond* what can be explained by the rated word's saliency.

Modeling: We analyze the collected ratings using an ordinal GAMM. We provide a brief introduction to GAMMs in Section 2.4.2.5. Its key properties are that it models the ordinal response variable (i.e., the importance ratings in our setting) on a continuous latent scale as a sum of smooth functions of covariates, while also accounting for random effects.[16]

[13] As the largest observed SHAP value in our data is 0.405, we normalize all SHAP values with 0.405^{-1} to cover the full color range.

[14] `https://huggingface.co/facebook/bart-large-mnli`

[15] `https://www.kaggle.com/datasets/mikeortman/wikipedia-sentences`

[16] Random effects allow to control for, e.g., systematic differences in individual participants' rating behavior, such as a specific participant with a tendency to give overall higher ratings than other participants.

Measure	Examples	Description
First-order constituent	highly developed, more than, such as	Smallest multi-word constituent subtrees in the constituency tree.
Noun phrase	tokyo marathon, ski racer, the UK	Multi-word noun phrase in the constituency tree.
Frequency	the United, the family, a species	Raw, unnormalized frequency.
Poisson Stirling	an American, such as a species	Poisson Stirling bigram score.
φ^2	Massar Egbari, ice hockey, Udo Dirkschneider	Square of the Pearson correlation coefficient.

Table 5.4.: Illustrative subset of our phrase measures.

Precedent model terms: We include all covariates tested in Section 5.1.1, including the rated word's saliency, word length, etc., in order to control for them when testing our new phrase-level variables. We follow our respective controls for all precedent main and random effects and exclude the pairwise interactions due to increased stability.

Novel neighbor terms: The following variables dictate our added model terms as the basis for the analysis: Left or right adjacency, rated word's saliency (color intensity), saliency difference between the two words, and whether the words hold a weak or strong relationship. We include four new bivariate smooth terms (Figure 5.7) based on the interactions of the above variables.

We refer to a bigram with a strong relationship as a chunk. To arrive at a reliable measure for chunks, we methodically test various measures of bigram relationships, in two different categories (Table 5.4): *syntactic*, via dependency parsing, and *statistical*, via word collocation in a corpus. Following Frantzi et al. (2000), we use both syntactic and statistical measures together, as first-order constituents among the 0.875 percentile for φ^2 collocations (our observations are robust to choices of statistical measure and percentile, see Appendix A.6.2).

5.1.2.2. Reproducing Prior Results

The described experiment largely overlaps with our experiments described in Section 5.1.1 in which we investigate the effects of word-level and sentence-level features. Thus, we investigate whether we can confirm our previous findings in a different language domain (medium-form Wikipedia texts vs. short-form restaurant reviews in Section 5.1.1), and SHAP-values vs. Integrated Gradients Sundararajan et al. (2017). The result is positive: We reproduce our previously reported significant effects of *word length*, *display index*, *capitalization*, and *dependency relation* for randomized explanations as well as SHAP-value explanations (details

	(e)df	**Ref.df**	**F**	**p**
s(saliency)	11.22	19.00	580.89	<**0.0001**
s(display index)	3.04	9.00	22.02	<**0.0001**
s(word length)	1.64	9.00	16.44	<**0.0001**
s(sentence length)	0.00	4.00	0.00	0.425
s(relative word frequency)	0.00	9.00	0.00	0.844
s(normalized saliency rank)	0.59	9.00	0.37	0.115
s(word position)	0.58	9.00	0.18	0.177
te(left diff.,saliency): no chunk	3.12	24.00	1.50	**0.002**
te(left diff.,saliency): chunk	2.24	24.00	0.51	**0.038**
te(right diff.,saliency): no chunk	2.43	24.00	0.47	**0.049**
te(right diff.,saliency): chunk	0.00	24.00	0.00	0.578
capitalization	2.00		3.15	**0.042**
dependency relation	35.00		2.92	<**0.0001**

Table 5.5.: (Effective) degrees of freedom, reference degrees of freedom and Wald test statistics for the univariate smooth terms (top) and parametric terms (bottom).

in Appendix A.6.1). This result reinforces prior observations that human users are at significant risk of biased perception of saliency explanations despite seemingly objective visualization.

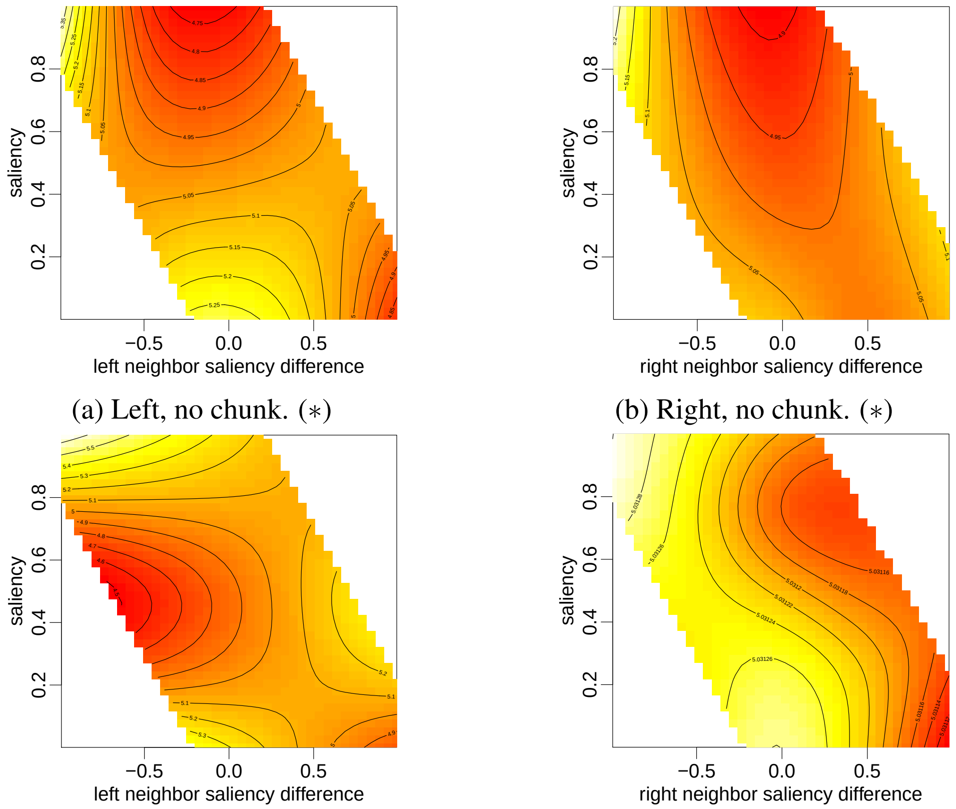
5.1.2.3. Neighbor Effects Analysis

In the following, we present our results for our two experiments using (a) random saliency values and (b) SHAP values.

Randomized Explanations. Regarding our additionally introduced neighbor terms, Figure 5.7 shows the estimates for the four described functions (left/right × chunk/no chunk). Table 5.5 lists all smooth and parametric terms along with Wald test results (Wood, 2013a,b). Appendix A.6.1 includes additional results.

Asymmetric influence: Figure 5.7a vs. Figure 5.7b and Figure 5.7c vs. Figure 5.7d reveal qualitative differences between left and right neighbor's influences. We quantitatively confirm these differences by calculating areas of significant differences (Fasiolo et al., 2020; Marra and Wood, 2012). Figures 5.8a and 5.8b show the respective plots of (significant) differences and probabilities for the chunk case. Overall, we conclude that the influence from left and right word neighbors is significantly different.

Chunk influence: We investigate the difference between neighbors that are within a chunk with the rated word vs. those that are not. We find qualitative differences in Figure 5.7 as well as statistically significant differences (Figures 5.8c and 5.8d).



(a) Left, no chunk. (∗)

(b) Right, no chunk. (∗)

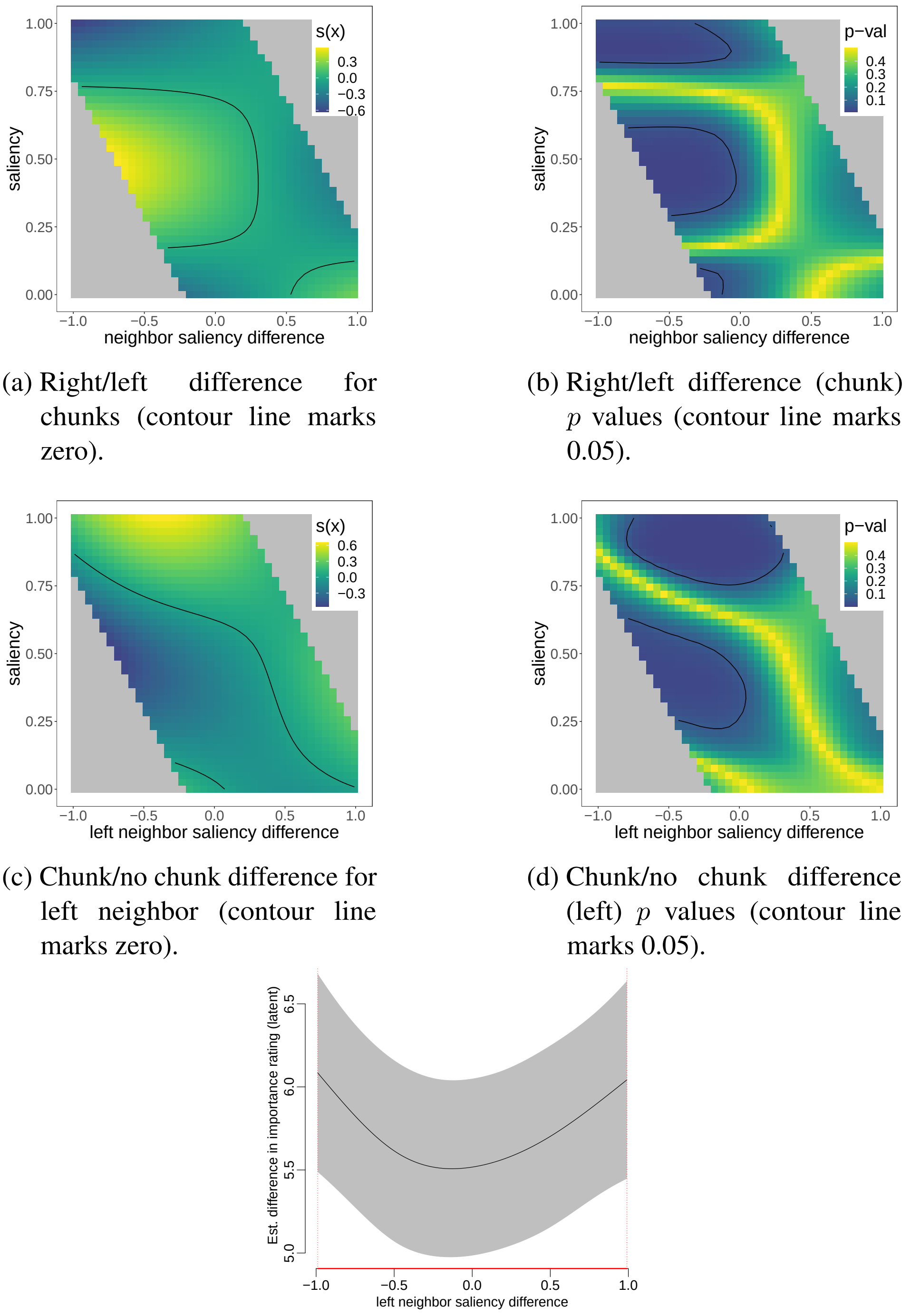
(c) Left, chunk. (∗)

(d) Right, chunk.

Figure 5.7.: Left and right neighbors. (∗) marks statistically significant smooths. Colors are normalized per figure.

Saliency moderates neighbor difference: Figure 5.7 shows that the effect of a neighbor's saliency difference (x-axis) is moderated by the rated word's saliency (y-axis). We confirm this observation statistically (Figure 5.8e) by comparing functions at a rated word saliency of 0.25 and 0.75, using unidimensional difference plots (Van Rij et al., 2015).

Combined effects: We identify two general opposing effects: assimilation and contrast. We borrow this terminology from psychology and will discuss links to related work in Section 5.1.2.4. We refer to *assimilation* as situations where a word's perceived saliency is perceived as more (or less) important based on whether its neighbor has a higher (or lower) saliency. We find assimilation effects from *left* neighbors that form a chunk with a moderate saliency (0.25–0.75) rated word. We refer to *contrast* as situations where a word's perceived saliency is perceived as less (or more) important based on whether its neighbor has a higher (or lower) saliency. We find contrast effects from left and right neighbors that do not form a chunk with the rated word. Note that although Figure 5.7d suggests a contrast effect, the color normalization inflates the minimal differences in this figure and the Wald tests did *not* signal a significant effect.



(a) Right/left difference for chunks (contour line marks zero).

(b) Right/left difference (chunk) p values (contour line marks 0.05).

(c) Chunk/no chunk difference for left neighbor (contour line marks zero).

(d) Chunk/no chunk difference (left) p values (contour line marks 0.05).

(e) Difference between rated saliency and left neighbor saliency.

Figure 5.8.: Difference plots. Red x-axis in (e) marks significant differences.

SHAP-value Explanations. For the SHAP-value explanations, we observe effects that (a) are shared with our observations for randomized saliencies and (b) differ from these.

Shared Results: Our SHAP-value experiment confirms our observation of (i) asymmetric influence of left/right neighbors (Figures A.22a and A.22b), (ii) chunk influence (Figures A.22c and A.22d), (iii) a moderating effect of saliency (Figure A.22e), and (iv) assimilation and contrast effects (Figure A.21d).

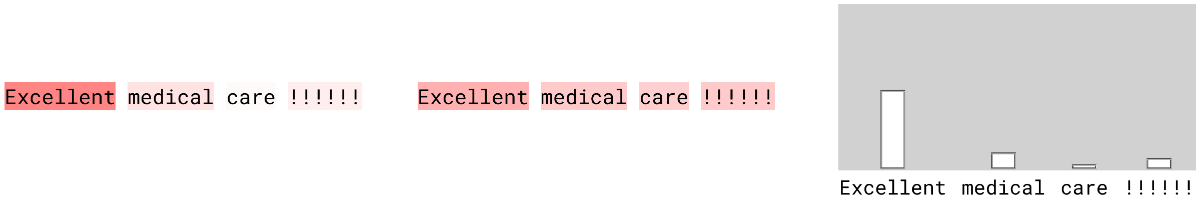
Variant Results: Notably, our SHAP-value results differ from our randomized saliency results with respect to the effects of left/right direction. For the randomized saliency experiment, we observe assimilation effects from left neighbors within a chunk (Figure 5.7c) and contrast effects from left and right neighbors outside a chunk (Figures 5.7a and 5.7b). For our SHAP-value experiment, we observe assimilation (low-rated word saliencies) and contrast effects (medium normalized rated word saliencies) from right neighbors within a chunk (Figure A.21d). We hypothesize that this difference can be attributed to the inter-dependencies of SHAP values as indicated in Figure A.23 in Appendix A.6.

Overall Results. Overall, we find that (a) left/right influences are not the same, (b) strong bigram relationships can invert contrasts into assimilation for left neighbors, (c) extreme saliencies can inhibit assimilation, and (d) biasing effects can be observed for randomized explanations as well as SHAP-value explanations.

5.1.2.4. Theoretical Grounds in Psychology

The assimilation effect is, of course, intuitive — it means that neighbor's importance "leaks" from neighbor to the rated word for strong bigram relationships. But is there precedence for the observed assimilation and contrast effects in the literature? How do they relate to each other? Psychology investigates how a prime (e.g., being exposed to a specific word) influences human judgment, as part of two categories: *assimilation* (the rating is "pulled" towards the prime) and *contrast* (the rating is "pushed" away from the prime) effects (i.a., Bless and Burger, 2016).

Förster et al. (2008) demonstrate how *global* processing (e.g. looking at the overall structure) vs. *local* processing (e.g., looking at details) leads to assimilation vs. contrast. We argue that some of our observations can be explained with their model: Multi-word phrase neighbors may induce global processing that leads to assimilation (for example, in the randomized explanation experiments, left neighbors) while other neighbors (in the randomized explanation experiments, right neighbors and unrelated left neighbors) induce local processing that leads to contrast. Future work may investigate the properties that induce global processing in specific contexts.



(a) Original saliency. (b) Corrected saliency. (c) Bars.

Figure 5.9.: The three different saliency visualization methods we compare.

5.1.3. Alternative Visualizations to Mitigate Biases

So far, we observed that various seemingly irrelevant factors influence human perception in unintended ways from the explicit and objective saliency information across different languages, tasks, and feature-attribution scores. Next, we explore two methods to decrease the bias in human perception (Figure 5.9): (i) controlling for the bias by modifying the color-coding to account for over-estimation and under-estimation of importance (over-estimated tokens will receive decreased color saturation, and vice versa) (Figure 5.9b), (ii) replacing the color-coding visualization with bar chart visualization (Figure 5.9c).

5.1.3.1. Model-Based Color Correction Technique

We compute an alternative color-coding visualization that a priori accounts for over-estimation and under-estimation of tokens based on the data collected in the previous experiments. Here we investigate whether it is possible to "correct" the explainees' saliency perception by superimposing the initial saliency values with a correction signal.

We require a procedure that increases the saliency scores for words that are predicted to be under-perceived (e.g., short words and words that appear in long sentences) and decrease the saliency scores for words that are predicted to be over-perceived (e.g., words with a high sentiment polarity or words that appear in short sentences). Briefly, the trained GAMM model from the English sentiment study (Section 5.1.1.6) allows us to map a combination of a saliency score together with word/sentence properties to a perceived importance score (on a continuous latent scale). By grounding this prediction of perceived importance to a prediction conditioned on a particularly chosen reference level, we can iteratively globally correct the explained scores over the sentence such that the (predicted) perception bias is decreased in each iteration. Table 5.6 displays examples of the application of this correction. In Appendix A.7, we discuss the full algorithm including its components and motivating details, and provide an extended list of example applications in Table A.16 as well as an example of the gradual correction over the course of 100 correction steps in Table A.15.

	Saliency	Bias	Removed Bias
original	Great people !	Great people !	94.9%
corrected	Great people !	Great people !	
original	Horrible service .	Horrible service .	100.0%
corrected	Horrible service .	Horrible service .	
original	I remain unhappy .	I remain unhappy .	84.3%
corrected	I remain unhappy .	I remain unhappy .	

Table 5.6.: Examples of the bias reduction procedure. The *saliency* column shows the saliency explanations (how users would see them) before and after the bias correction procedure. The *bias* column shows the color-coded bias estimates. Predicted over-estimations are colored in red whereas predicted under-estimations are colored in blue. More examples can be found in Table A.16 in Appendix A.7.

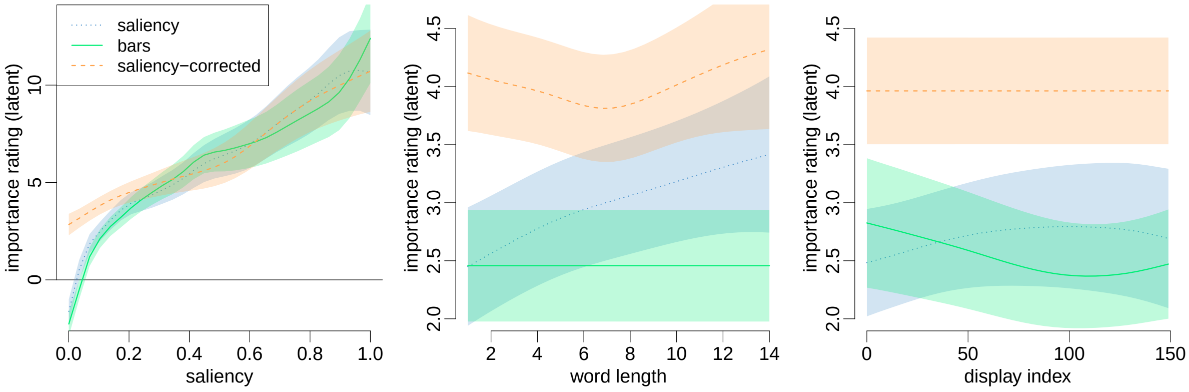
5.1.3.2. Bar Chart Visualization

As an alternative to color-coding visualization, we consider bar charts (Figure 5.9c): we investigate whether a sufficiently distinct visualization will result in different perception. We hypothesize that this is related to visual perception bias.

We note two visual qualities of bars that differentiate it from color-coding, and therefore make it a relevant alternative visualization candidate: (i) The bars are communicated with objective reference points of zero and one (the top and bottom of the draw area), while the results in, i.a., Section 5.1.1.6 indicate that participants perceive colored saliency in relation to each other, instead of in reference to zero and one (pure white and pure red, respectively) and (ii) the draw area for the bars is separate from the draw area for the input text, in contrast to color-coding, where they occupy the same space. This means that in color coding, for example, a word with more characters will receive a larger area of color, in comparison to a shorter word with the same color. As our studies in, i.a., Section 5.1.1.6 show, word length influenced explainee perception. In the bar chart visualization scheme, all words are treated identically within the draw area which communicates importance.

5.1.3.3. Results

We investigate how well the two proposed visualization alternatives counteract bias in user perception within the study described in Section 5.1.1.6. We find that visualization has a significant effect on importance ratings (df=2, F=35.45, p<0.0001) where the bar charts lead



(a) Saliency.

(b) Word length.

(c) Temporal display index.

Figure 5.10.: Selected summed-effects comparison plots of the visualization alternatives.

to lower importance ratings ($\beta = -0.5991$, SE=0.1579) and the correction method leads to higher ratings ($\beta = 1.1102$, SE=0.2515). Regarding the visualizations' effect on smooth terms, we focus on color saturation, word length, and display index in Figure 5.10.

Figure 5.10a shows that the saliency scores' effect on importance ratings is similar to the original saliency and the bar visualizations, while the corrected visualization leads to higher ratings in the lower end of the color saturation spectrum. These differences are neither "good" nor "bad" — we argue that the similarity between the original saliency visualization and the bar charts is remarkable as the two visualizations are fundamentally different.

Figure 5.10b shows that the biasing effect of word length in the original visualization is successfully eliminated using the bar visualization as shown by the nearly constant smooth of the bar visualization (edf=0.0009). This confirms our hypothesis that bar charts evade word length bias. The correction visualization leads to a different effect than the original visualization, however, this effect indicates a different but equally distorting bias of word length.

Figure 5.10c indicates a successful application of our color correction technique. While the original visualization as well as the bar charts show a biasing effect regarding the model smooths, the saliency correction visualization leads to a nearly constant smooth (edf=0.0009). Regarding the original and the bar visualizations, the smooths indicate that, in contrast to the original visualization, the bar visualization leads to an initial overestimation of importances which decreases over time, while the original visualization lead to a respective underestimation. However, a difference plot between the two conditions (see Figure A.31c in Appendix A.8) shows no significant differences.

While these examples demonstrate indications for successful bias mitigation, we want to note that this mitigation cannot be observed for most of the other variables, in particular not for the effect of saliency rank, which we expected to be mitigated by the bar visualization. We provide summed-effect comparison plots for all effects under investigation in Figure A.30, difference plots between all conditions in Figures A.31 and A.32 as well as details and test statistics on all parametric coefficients as well as smooth terms in Table A.17 in Appendix A.8.

Tying back to our initial categorization of biases, we observe that our proposed visualization alternatives can successfully remove instances of visual bias (word length) and learning effect bias (display index). We hypothesize that belief biases (such as sentiment polarity) exhibit more distinct expression across individuals, which requires participant-adaptive correction methods and should be addressed by online estimation of individual participant slopes and intercepts within our GAMM model in future work.

5.1.4. Overall Discussion

Overall, our results show that *supposedly irrelevant factors, such as word length do affect how explainees perceive the influence of words in feature-attribution explanations, despite the explanations explicitly communicating this influence.* This is a surprising result, which raises important questions for explainability in NLP, and in general, about the ability of feature-attribution tools available today to convey the information that they intend to communicate: Even in the case of a relatively straightforward explanation, such as directly informing importance regions in the input, cognitive biases of explainees run deep, and may erroneously affect the understanding of the given information.

In a series of four studies, we demonstrate that (i) various word and sentence features distort users' explanation perception, (ii) distorting effects generalize across two languages and tasks, (iii) apply to experimental random as well as real attribution scores, and (iv) neighboring words affect each other's importance perception via assimilation and contrast effects depending on left/right neighborhood and whether they form a lexical chunk.

We explore two visualization alternatives to mitigate the effect of the observed biases and show that bar charts and color correction result in better-aligned human assessments in our setting on multiple bias factors. We urge researchers to not blindly trust that users perceive explanations as communicated, and to investigate if our findings transfer to their respective audience and context. We revisit and further refine our recommended bar chart visualization based on additional findings from our study of perceived predictability in Section 5.2.

Statements		**Agreement rating (1-7)**						
		strongly disagree	do not agree	somewhat disagree	neither agree nor disagree	generally agree	agree	strongly agree
EF-1	**The system behaves in a predictable manner.**				×			
EF-2	**I can tell which responses the system will likely give.**					×		
EP-1	**I observed enough system responses to predict how the system behaves.**			×				
EP-2	**Based on past system responses, I know the responses the system will likely give me.**			×				
AL-1	**I can tell the reasons for the system's decisions.**						×	
AL-2	**There is a consistent pattern in the system's behavior.**					×		

Example ratings result in an overall score of 4.3 (mean).

Table 5.7.: Final PSP scale with items, rating options, and a rating example. EF, EP, and AL refer to the effective, epistemic, and aleatory aspects of predictability covered in the scale. The example rating is indicated with "×" and illustrates how the six ratings are linked to numeric scores which are averaged to obtain the total system predictability score. The choice of the seven agreement anchors builds upon recent findings on optimal conceptual anchor distances (Casper et al., 2020).

5.2. Perceived System Predictability: Scale Development and Results

In this final section, we describe the development, validation, and usage of our novel PSP scale depicted in Table 5.7. The PSP scale is a highly-economic 6-item Likert scale to measure facets of perceived system predictability and is designed to be applicable to any system that takes decisions or makes predictions, such as automated decision making (ADM) systems, chatbots, robots, and many more.

In the following, we (i) outline why we need to measure PSP (Section 5.2.1), (ii) propose a theory of perceived predictability comprising three facets (epistemic, aleatory, and effective predictability) (Section 5.2.2.1), (iii) develop a novel 6-item scale to measure PSP (Section 5.2.2.2), (iv) evaluate our scale (Section 5.2.2.3), (v) use our scale to explore the effects of explanations and system stochasticity, and (vi) explore how PSP is related to, i.a., prediction correctness, trust, subjective information processing awareness (SIPA), and participants' NFC (Section 5.2.3).

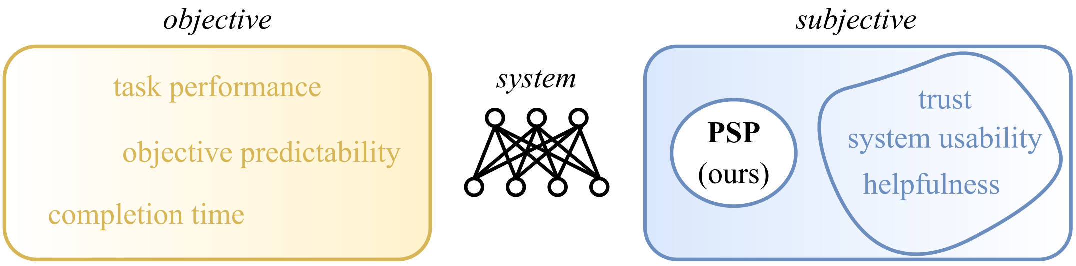


Figure 5.11.: We propose a novel scale to measure PSP. PSP belongs to subjective self-reported measures. We argue that it is necessary to assess subjective predictability in addition to objective predictability and that we cannot resort to existing subjective measures of related constructs.

5.2.1. The Need to Measure Perceived System Predictability

As we discussed in Section 4.2, the quality of explanations has numerous facets. Within this thesis, we measured how accurately users perform a task with the help of a system (Section 3.2), how much time users need to complete a task with a system (Sections 3.2 and 4.3), how usable users rate the system to be (Sections 3.2 and 4.3), or as how intelligent, human-like, or helpful users perceive the system to be (Section 3.2). As depicted in Figure 5.11, we can categorize these measures into objective scores and subjective ratings as we discussed in Section 4.4. While task performance and completion time are objective scores, ratings of usability and perceived system characteristics are subjective ratings. In addition to our work, there are numerous additional scores and ratings used in related work. We provide a non-exhaustive list of scores and ratings that are especially relevant to assess explanation quality in Table 4.9 on page 116 of this thesis.

A user's ability of being able to predict a system's behavior can be assessed on both, the objective as well as the subjective side of evaluation measures. While objective predictability, i.e., a user's demonstrated ability to predict how the system will behave in an unseen situation, has been assessed frequently (Goyal et al., 2019; Hase and Bansal, 2020; Wang and Vasconcelos, 2020; Wang and Yin, 2021), perceived predictability received less attention (Schulz et al., 2015; Schrills et al., 2022).

In the following, we will elaborate on (i) why measuring objective predictability is not enough, (ii) why we cannot resort to other subjective measures, and (iii) why we choose to develop a novel measurement instrument to score PSP.

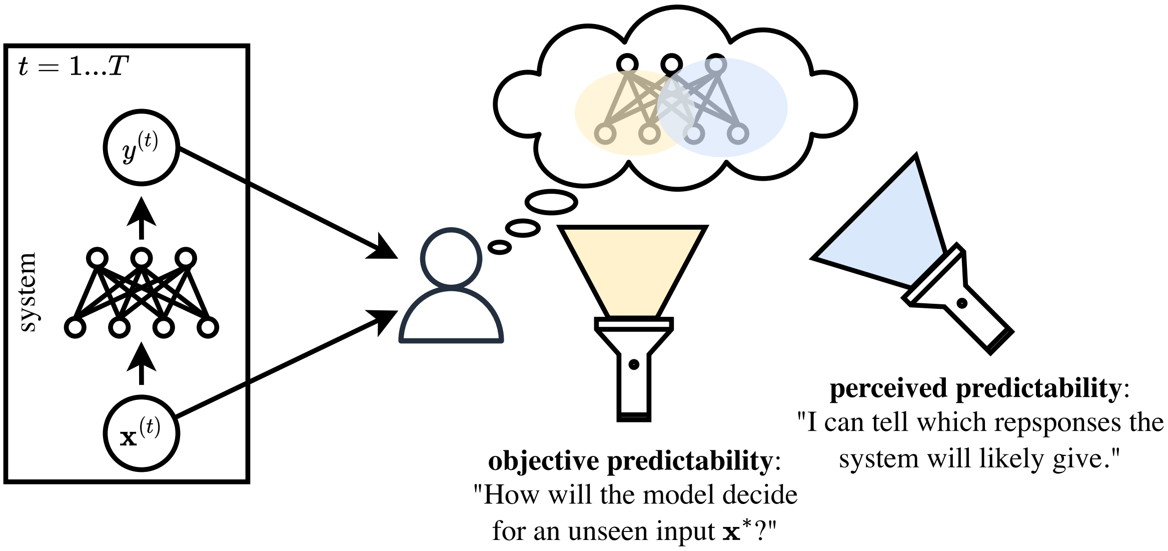

$t = 1...T$
system
$y^{(t)}$
$\mathbf{x}^{(t)}$
perceived predictability:
"I can tell which repsponses the system will likely give."
objective predictability:
"How will the model decide for an unseen input $\mathbf{x}^*$?"

Figure 5.12.: We argue that measuring *objective* and *perceived* predictability reflect distinct characteristics of the user's mental model of a system. We use the metaphor of two flashlights that illuminate separate aspects of the user's mental model that is, i.a., formed by the user's experience with the system indicated by the left box using plate notation.

5.2.1.1. Objective Measures of Predictability Are not Enough

It can be argued that measuring objective predictability is at least as informative or even more informative than measuring subjective predictability. Why should one investigate how well users *feel* to be able to predict a system's behavior when their ability to do so can be measured objectively? In the following, we motivate why we have to measure the subjective feeling in addition to the objective performance. For this, we draw parallels to two related evaluation settings: (i) measuring system usability in HCI and (ii) measuring the feeling of learning (FOL) in educational research. We illustrate our argument in Figure 5.12.

Subjective Versus Objective Usability. Measuring usability is a key aspect of quantifying user experience (Lewis, 2018). Similarly to our context, usability measures can be grouped into subjective measures (e.g., obtained from user satisfaction questionnaires) and objective measures (e.g., task completion time) (Nielsen and Levy, 1994; Hornbæk, 2006). Nielsen and Levy (1994) study how objective and subjective measures are related within a meta-analysis and found that — although there is an overall positive correlation — the two can yield contradictory conclusions. For example, users were observed to consistently prefer an interface with which they were slower than an alternative system with which they reached faster interactions before (Grudin and MacLean, 1985). Similarly, MacLean et al. (1985) observed that users preferred a slower input method as long as it was not more than 20% slower than the faster alternative. Hornbæk (2006) lists numerous arguments why both — subjective as well as objective — usability measures should be assessed. *Inter alia*, Hornbæk (2006) notes that objective and subjective measures can lead to different conclusions, for example, objective time measures and subjectively experienced time were shown to differ (Eisler, 1976; Tractinsky and Meyer, 2001; Czerwinski et al., 2001). Similarly, Hornbæk (2006) mentions the dissociation of objective performance and perceived workload discussed by Yeh and Wickens (1988).

Feeling of Learning Versus Actual Learning. In our second example, we consider the study of active learning classroom instructions (as opposed to passive lectures) of Deslauriers et al. (2019). Students were evaluated on (a) what they objectively learned during a class and (b) what they *felt* to have learned. While one may expect that the two measures are positively correlated, Deslauriers et al. (2019) find that students that participate in an active learning lecture (as opposed to a passive lecture) learn more but feel like they learn less. The authors argue that this observation is an effect of an increased cognitive effort leading to (i) an increased learning effect but at the same time causes (ii) a cognitive disfluency that students perceive as a decreased learning effect. This cognitive disfluency and the corresponding feeling of learning less can be major obstacles to the success and acceptance of active learning lectures as they pose a threat to students' motivation and engagement (Deslauriers et al., 2019). Measuring and studying FOL, allowed the authors to identify this disconnect and propose suitable interventions. If the authors would not have measured FOL, they would not have been able to uncover the described effect and would not have been able to address the respective problems. Similarly, we argue that measuring perceived predictability (in relation to objective predictability) can identify and help to eliminate obstacles to, i.a., explainability that remain hidden without access to an adequate measurement instrument.

5.2.1.2. Measuring Related Constructs Is not Enough

One might argue that, instead of measuring perceived predictability directly, it is sufficient to study it indirectly via related constructs, such as trust, usability, or perceived helpfulness (as depicted in Figure 5.11). While a strong relation between perceived predictability and related constructs, such as trust is plausible (as we demonstrate in Section 5.2.3), we argue that measuring perceived predictability enables insights beyond what we can explore with related constructs. As we will show in Section 5.2.3, the relation between trust and objective predictability differs from the relation between perceived and objective predictability.

Prior work compared objective and subjective measures of explanatory systems and indicates that subjective ratings are not predictive of objective measures of human interactions. Concretely, Buçinca et al. (2020) found that trust is not predictive of user performance with the system. Similarly, Hase and Bansal (2020) did not find that subjective explanation quality ratings ("Does this explanation show me why the system thought what it did?") are predictive of user correctness. One explanation of this phenomenon can be that (a) user-reported ratings of explanatory systems are dissociated from their objectively observed interaction with these systems in general. However, we want to raise attention that another explanation can be that (b) the previously used subjective measures capture constructs that are too “distant” from the user’s perceived system behavior to be predictive of their related objective measures. In Section 5.2.3, we find that (i) perceived predictability is a significant predictor of objective predictability, and (ii) that there is a significant negative association between trust and objective predictability.

5.2.1.3. The Need for a Novel Instrument

So far, we motivated why we need to measure perceived predictability. In the following, we discuss existing instruments and argue why we need to develop a new instrument.

Schulz et al. (2015) investigate how properties of a function (such as smoothness) affect the perceived predictability and ask participants to rate “how well could you predict this function?” with a 1-100 slider. While we agree that this scale is a suitable choice within the context studied by Schulz et al. (2015), we argue that we cannot apply it to general systems for two reasons. First, the wording of the scale is tailored to rate “this function” which is less suitable to refer to, e.g., a chatbot. While, of course, the wording can be adapted to “this system”, this highlights the second problem which is the lack of demonstrated psychometric validity. Concretely, we could only hope that “how well could you predict this system?” actually measures perceived predictability. This problem applies to all single-item scales (i.e., scales with a single question or statement) as a single item does not allow to estimate the item’s correlation to the latent variable (perceived predictability). When using multiple items to measure a construct, the items’ correlation to the latent variable can be indirectly quantified via the item-item correlation (DeVellis and Thorpe, 2021). Even when ultimately only a single item is retained, it is important to provide evidence that this item has a sufficiently strong relation to the construct that is intended to be measured. Subjective ratings of explanatory systems often rely on custom single-item scales that are applied without following a psychometric development and validation process. While the resulting scales *can* still be valid, without a dedicated analysis, there is no evidence that they actually *are* valid. An important exception is the SIPA scale by Schrills et al. (2022). This scale builds upon the theory of situation awareness

(Endsley, 1988) and collects ratings for transparency, understandability, and — importantly — predictability with two items each which are combined into a SIPA score. The development of the SIPA scale followed an established scale development process (which we will expand upon when describing the development of our own scale in Section 5.2.2) and demonstrated good statistical properties. While the SIPA scale can be used in explainability research[17], the construct of predictability only is one out of three facets of SIPA. This reduces the predictability-related measure to a two-item subscale which in turn models predictability as a unidimensional construct. In the following, we will provide (a) evidence from user interviews as well as (b) a theoretical link to uncertainty theory, that both support a more fine-grained measurement of perceived predictability. Concretely, our novel scale "zooms in" to predictability and measures predictability along three dimensions (effective, epistemic, and aleatory predictability) using two items each.

5.2.2. Scale Development and Validation

We define perceived predictability as *the degree to which a user feels to be able to predict how a system behaves*. In the following, we propose a theory of perceived predictability based on structured target population interviews (inductive) and uncertainty theory (deductive), which we combine into a multidimensional construct of perceived predictability.

5.2.2.1. A Theory of Perceived Predictability

Target Population Interviews. As our first step, we assess that the construct of system predictability exists in our target population's participants' notion of automated, in particular, AI systems. For this, we ask 20 crowdworkers on MTurk to state what "understanding an AI system" means to them.[18] We find that, besides transparency, technical details, and intended usage, perceived predictability is mentioned by the majority of participants and conclude that predictability is a natural aspect of users' system perception.

Already at this stage, we identify participant comments that suggest a distinction of predictability types: "*I would not trust a system that behaved randomly unless it was 'controlled random.' [...] Like if a system was programmed to randomly select from a set of pictures for example*". These comments indicate that users distinguish different dimensions of predictability.

In our second step, we ask another 20 participants what "having the feeling that you can predict how an AI system behaves" means to them to obtain a focused picture of the participants'

[17] We investigate the relation between the SIPA scale and our own scale in Section 5.2.3.

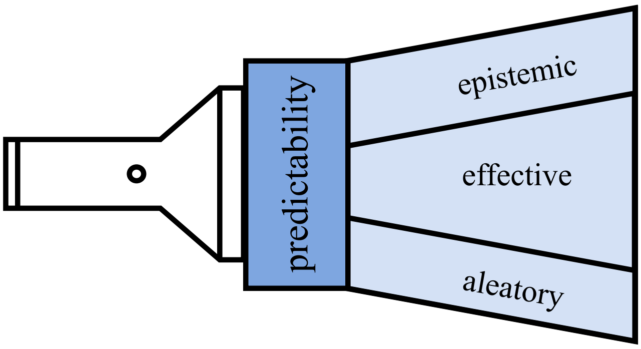
[18] We recruit crowdworkers from the US, Australia, and the UK within this and the following studies.

epistemic	– I feel like I can predict an AI system's behavior when I have interacted with it many times before. – If I use the AI frequently, I know what I can and cannot ask or do with the AI. – Being able to anticipate how it will respond after some moderate use [...]. – I feel like I can predict how an AI system behaves the more I interact with it. – I feel that as I gained more experience with the system it would become more and more predictable to me. – knowing that result it is likely to give me, either based on it's past responses or my own assumptions
aleatory	– [...] given a certain input it should always product the same output. – means to me that the AI normally behaves in a consistent manner – It may start following a certain pattern and I get an idea of how the algorithm works. – I would definitely want such a system to give consistent and reliable results. – I have a basic understand of the various rules or conditions that the AI system uses to make it's judgments
effective	– its more of like you can guess how it will react. – It means that I can guestimate how it comes to its conclusions. – I can know what to expect. – I can usually predict if the AI can answer the question I have in mind for it. – a good overall understanding of the AI

Table 5.8.: Target population comments on what "having the feeling that you can predict how an AI system behaves" means to individual participants. We identify three aspects of perceived predictability: epistemic, aleatory, and effective predictability.

notion of perceived predictability. Again, we observe responses that suggest that there exist multiple aspects of predictability: "*[...] it is following a set of rules [...], so given a certain input it should always produce the same output. I feel that as I gained more experience with the system it would become more and more predictable to me*" in which the first sentence can be interpreted as aleatory uncertainty and the second sentence as epistemic uncertainty which we will delineate in the following. Table 5.8 shows a list of target group participant comments.

An additional observation is that users relate a slight level of unpredictability with a preferable or more powerful system by comments, such as "*If it's too predictable then I wonder what the point of having it is. The same as if it is too unpredictable. Sometimes like on a racing game with really good AI you can have races that you really can't tell if it is a human or a bot. The same goes for chatbots that people test on MTurk. Some of them are uncanny at how real they are. I guess to sum it up AI needs to be just unpredictable enough and in the right way for me to like it and see a need for it.*" or "*with AI currently it seems you can usually within some reason, predict how the AI will answer or what actions it will give, i think its because its not a true AI as it stands. there is no conscious thought, just the data we gave it to act like its own self, which its just a shell of that. i do think in the somewhat future, not sure on distant or close, that AI will have its own conscious thought and then be slightly unpredictable in its answers/actions*".



"I observed enough system responses to predict how the system behaves."

"I can tell which responses the system will likely give."

"There is a consistent pattern in the system's behavior."

Figure 5.13.: We decompose perceived predictability into three (partially overlapping) facets. In contrast to uncertainty (in statistics), in which the total uncertainty is the sum of the epistemic and the aleatory uncertainty, our notion of "effective" predictability includes additional information beyond epistemic and aleatory predictability. Exemplary items from our final PSP scale are added to their respective facet on the right side of the figure.

We will revisit this effect in the light of our quantitative results in Section 5.2.3

Uncertainty Theory: Epistemic and Aleatory Uncertainty. We argue that our evaluation of target population participant comments indicates that perceived predictability has multiple facets. We argue that the observed categories can be related to epistemic and aleatory uncertainty. In uncertainty theory, Fox and Ülkümen (2011) distinguish two types of uncertainty: (a) *epistemic* uncertainty that relates to the uncertainty of not knowing something that could be known, e.g. due to a lack of observations of a phenomenon and (b) *aleatory* uncertainty that refers to a phenomenon's inherent stochasticity and that cannot be addressed with a higher number of observations. For example, we can have perfect epistemic certainty about how a die functions and yet be unable to predict the outcome of rolling dice due to the aleatory uncertainty rising from the dice's randomness.

Dimensions of Predictability: Epistemic, Aleatory, and Effective. We argue that the concepts of epistemic and aleatory uncertainty can be transferred to predictability (objective and perceived) as failing to predict a system's behavior (or the respective perceived ability) can be caused by (i) a lack of system behavior observations, and/or (ii) inconsistencies in the system's behavior. Regarding the former, in the extreme case, if users had no exposure to a system at all, their (perceived) prediction abilities are reduced to their general notion of an unknown system's predictability. On the contrary, having observed infinite system decisions removes this barrier to predictability. Regarding the latter, in the extreme case, the system

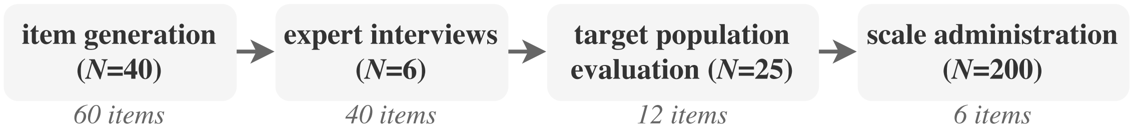

item generation (*N*=40)
60 items
expert interviews (*N*=6)
40 items
target population evaluation (*N*=25)
12 items
scale administration (*N*=200)
6 items

Figure 5.14.: Overview of our scale development process. For each step, we report the number of participants and the resulting number of items retained in our scale in *italics*.

takes completely random decisions. In this case, users will not be able to predict the system's behavior, even with access to infinite observations. In contrast, a completely deterministic system can (in theory) be fully predictable by observing all possible contexts before making the prediction. For uncertainty, epistemic and aleatory uncertainty can be summed to yield the overall uncertainty. We argue, that for perceived predictability, this does not hold and the *effective* predictability covers more than the sum of epistemic and aleatory predictability. We illustrate our theory in Figure 5.13.

5.2.2.2. Scale Development

In the following, we report the process and result of our scale development. We follow best practices as discussed by Boateng et al. (2018); Menold and Bogner (2016); DeVellis and Thorpe (2021). Figure 5.14 displays an overview of the separate development steps and the respective number of participants and retained scale items.

Initial Item Pool. We generate an initial item pool of 60 items based on (i) the target population interviews described above and (ii) our proposed theory of perceived predictability. We report the full item pool in Appendix A.9.1.1.

Expert Ratings. Following the typical scale development process, we ask experts to review each item within our initial item pool in order to ensure content validity, i.e., that our items capture the intended domain of perceived predictability. We had six experts rate our initial item pool in terms of the two dimensions relevance and clarity and collect additional textual feedback from each rater for each item. Based on the experts' ratings, we remove eight items experts found to be irrelevant and 13 items that experts found to be unclear. We additionally modify the wording of five items for which experts indicated a need for revision and add one new item based on their comments. Our revised item pool thus contains 40 items.

Target Population Evaluation. In order to assert the face validity of our scale, i.e., that our items are appropriately designed for their target population (e.g., regarding the items' wording), we conduct a crowdsourced adaption of cognitive interviews (Beatty and Willis, 2007) following probes from Willis (2004) with a total of 25 participants. In our first round of cognitive interviews, we present 20 MTurk crowdworkers (twelve identified as female, eight as male; mean age of 42.0 years (SD=11.1 years)) predictions of a hypothetical classification system. We ask participants to rate their agreement to each item on a numeric 1-7 Likert scale ranging from "strongly disagree" (1) to "strongly agree" (7). We discuss details of the classification system and the Likert items in Section 5.2.2.3. In addition, we ask to crowdworkers to (i) repeat each item statement in their own words and (ii) describe how they did get to their answer. The participants' responses allow us to detect ambiguous and unclear items. Based on the participants' feedback, we remove 15 items, modify five items, and add one new item. Next, we merge similar items, streamline item wording, and add additional items based on discussions of the revised item pool, resulting in a pool of 16 items.

Full Verbalization and Optimized Response Anchors: We noticed that for some items, participants explicitly mentioned that "*[they] will neither agree or disagree of this statement*" and rated 4 (as intended), so we keep the middle point. However, some participants also note that they gave a neutral rating, but in fact, rated a 5. We, therefore, choose to provide explicit response anchors for each possible rating (1-7). This is in line with Menold and Bogner (2016) who find in their review that fully verbalized scales are preferable to scales that only have endpoint labels. To choose an appropriate selection of response anchors, we build upon recent findings of Casper et al. (2020) who investigate how response anchors can be optimally chosen regarding the conceptual distance between anchors. They find that, for 7-point agreement ratings, the labels "strongly disagree", "do not agree", "somewhat disagree", "neither agree nor disagree", "generally agree", "agree" and "strongly agree" yield minimum overlap and approximate equal mean intervals. We thus update our scale to a fully-verbalized 7-point Likert scale using optimized response anchors and conduct another round of cognitive interviews with five crowdworkers (three identified as female, two as male; mean age of 43.6 years (SD=8.4 years)). We remove another four items and modify one item based on the participants' responses. Our updated item pool contains twelve items. We provide all items of this pool in Appendix A.9.1.2.

5.2.2.3. Scale Evaluation

In order to further reduce the number of items in our scale and to assess its psychometric properties, we collect scale ratings from 200 participants across five predictability scenarios of a fictional classification system.

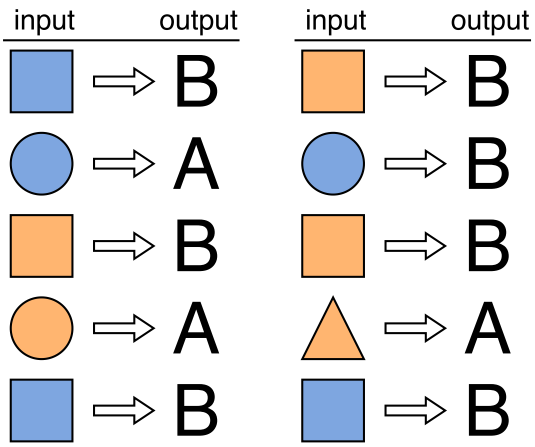
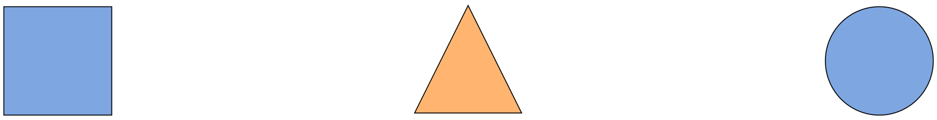


Figure 5.15.: One of the five fictional classification system prediction scenarios we show to participants. The class prediction for the blue circle is inconsistent, introducing aleatory uncertainty along the epistemic uncertainty caused by, i.a., the lack of blue triangle inputs. We refer to this scenario as MIXED.

Scenarios. To elicit different levels of perceived predictability, we design different versions of a fictional classification system. Our system maps colored shapes (red or orange circles, squares, or triangles) to one of two classes (A or B). We intentionally design a system that does not make predictions for which users might be biased by their own idea of what the correction is and will revisit scale ratings in the context of a real AI system in Section 5.2.3. We design five scenarios for which we vary (a) the number of shown predictions and (b) the randomness of the system's predictions. Figure 5.15 shows a scenario with mixed epistemic (few examples and non-exhaustive input examples) and aleatory (inconsistent predictions for the blue circle) uncertainty. We refer to this scenario as MIXED. We provide the shown examples for the additional four scenarios in Appendix A.9.2.1. Given the example predictions of the respective scenario, we ask participants to rate which class the model will predict for each of the three inputs shown in Figure 5.16. We use this task to motivate participants to analyze and reflect upon the provided system predictions in order to build a certain level of perceived predictability.

Scale Rating. After presenting the respective scenario and asking the users to rate which output they think the system will produce for the three symbols, we ask participants to rate each of the twelve remaining items. In this first evaluation, we randomize the order of items for each participant to reduce the confounding impact of potential carry-over effects or rating patterns on the estimate of, i.a., an item's discrimination strength, which we will discuss in the following. The order of items in the final version of our scale is not randomized. We review the psychometric properties of this fixed scale in Section 5.2.3.



(a) Blue square (predictable with high certainty).

(b) Orange triangle (predictable with lower certainty).

(c) Blue circle (unpredictable).

Figure 5.16.: The three symbols we ask users to predict the system's output for. Given the scenario shown in Figure 5.15, the system response for the first two shapes can be predicted (with different levels of certainty) and the system response to the third input is unpredictable.

Participants. We recruit 200 participants from the United States, Australia, and the United Kingdom on MTurk. Participants had a mean age of 40.8 (SD=11.2) years. 80 participants identified as female, 119 as male, and one participant as non-binary.

Item Reduction. We assess items along inter-item correlations, item-total correlations, item discrimination, and item difficulty and reduce the scale to six items following quantitative indicators as well as semantic overlap. We report the reasons for each of our removal decisions in Appendix A.9.1.2. Table 5.9 displays item difficulty and item discrimination values for the items retained in our final scale. As we intend our scale to be applicable in a broad spectrum of research contexts, we strive for an as short as possible scale while retaining enough items to obtain fine-grained measurements. In order to be able to estimate reliability coefficients for each of the three hypothesized facets, and thereby analyze their adequacy to be used as subscales, we choose to retain two items per dimension. In the following, we study our resulting 6-item PSP scale.

Reliability. The overall Cronbach's α (Cronbach, 1951) of our PSP scale equals 0.9606, coefficient ω (Raykov, 2001) equals 0.9607.[19] Following the evaluation of the SIPA scale Schrills et al. (2022), we use Spearman-Brown coefficients to quantify the reliability of the two-item subscales. Eisinga et al. (2013) recommend the Spearman-Brown coefficient as the reliability score of choice for two-item (sub)scales. We find R=0.908 for the epistemic subscale, R=0.889 for the aleatory subscale, and R=0.881 for the effective subscale. Overall, the observed reliability coefficients indicate a high internal consistency of our scale as a whole as well within its three subscales.

[19] We use a CFA-based calculation of coefficient ω as described by Furr (2022) (p. 515) using the R package semTools (Jorgensen et al., 2022). We calculate alternative McDonald's ω (Mcdonald, 1999) indices ω_h and ω_t based on hierarchical factor analysis equal using the R package psych (Revelle, 2022). The values equal 0.894 and 0.967 respectively.

	Item	Diff.	Discr.
EF-1	The system behaves in a predictable manner.	0.72	0.89
EF-2	I can tell which responses the system will likely give.	0.71	0.86
EP-1	I observed enough system responses to predict how the system behaves.	0.67	0.87
EP-2	Based on past system responses, I know the responses the system will likely give me.	0.70	0.90
AL-1	I can tell the reasons for the system's decisions.	0.65	0.84
AL-2	There is a consistent pattern in the system's behavior.	0.70	0.89

Table 5.9.: Item difficulty and item discrimination values of the items in our final scale (mean inter-item-correlation = 0.804, Cronbach's α=0.961).

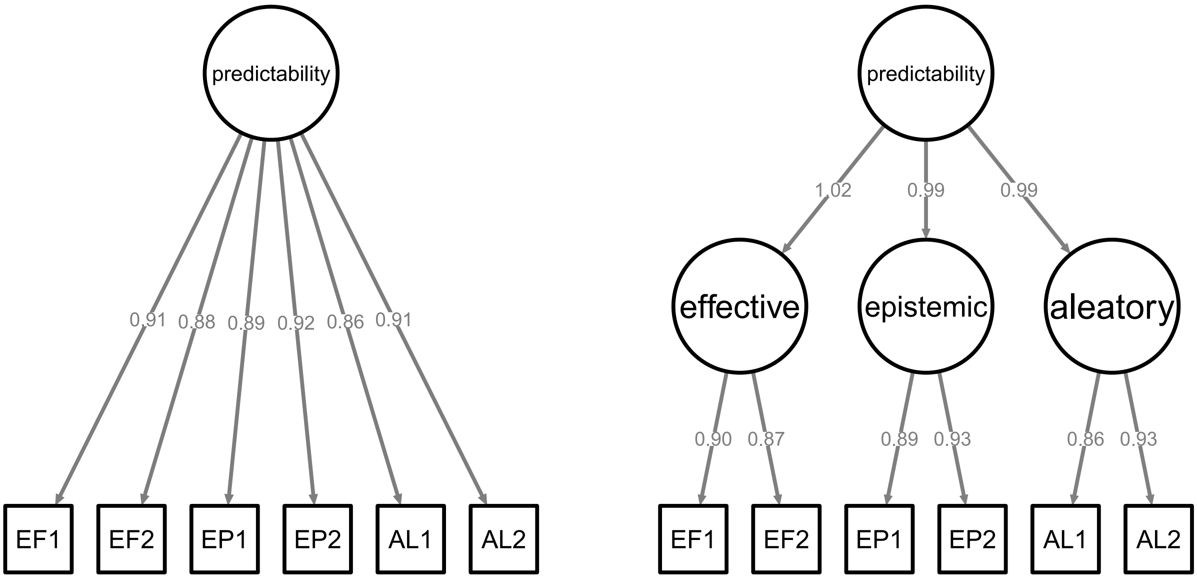
Confirmatory Factor Analysis. We conduct confirmatory factor analysis (CFA) to compare two models: a unidimensional model (Figure 5.17a) and a three-factor model (Figure 5.17b).[20] We report common model fit measures in Table 5.10 and detailed model fits in Appendix A.9.2.3. Hu and Bentler (1999) proposed — the now widespread — cutoff criteria for various fits, which are 0.06 for RMSAE, 0.08 for SRMR, and 0.95 for CFI and TLI. Both models fulfill the recommended criteria for RMSAE, CFI, and TLI, the three-factor model additionally fulfills the recommended SRMR criterion. While a comparison of the χ^2, RMSEA, SRMR, CFI, and TLI between the two models indicates a slight preference for the three-factor over the one-factor model, the differences between the two models are minor and both models can be considered adequate. Note that the three scale dimensions are strongly correlated with each other. Concretely, we observe Pearson correlations of 0.901 ($p < 0.001$) between effective and epistemic predictability, 0.889 ($p < 0.001$) between effective and aleatory predictability, and 0.876 ($p < 0.001$) between epistemic and aleatory predictability.[21] Overall, along with the identical differentiation patterns across the subscales and the total scale, which we detail in the following, our scale evaluation indicates that a unifactorial usage of our scale (i.e., measuring perceived predictability as the mean value over all six items) is warranted. However, the three-factor model showed slightly better model fit measures and the subscale scores should be evaluated in addition to the total scores for each application of our scale, especially when levels of epistemic and aleatory predictability can be expected to differ.[22]

Differentiation by Known Groups. One common way to assess a scale's construct validity is to evaluate how well the scale can differentiate between known groups that are expected to induce different scale scores. We assess whether our scale can differentiate between

[20] Note that the three-factor model is visualized as a higher-order model. With three first-order dimensions, a higher-order model is just-identified and equivalent to a correlated factors model with three dimensions.

[21] Pearson correlations are calculated using list-wise deletion as implemented in sjPlot (Lüdecke, 2023).

[22] For example, when users have a very high epistemic predictability (stemming from, e.g., a long usage period) but very low aleatory predictability.



(a) One-factor model.

(b) Three-factor model.

Figure 5.17.: The two models we compare in confirmatory factor analysis (CFA) along with standardized coefficients. The rectangular boxes correspond to the six items of our scale as denoted in Table 5.9. Dashed lines indicate estimates fixed to one (before standardization). Note that standardized coefficients are not bound to be lower than one as discussed by Deegan (1978) and Jöreskog (1999).

Fit Measure	**Model**	
	One-factor	Three-factor
χ^2 (↓)	12.67 (p=0.18)	7.140305 (p=0.31)
RMSEA (↓)	0.045 ([0.000, 0.098])	0.031 ([0.000, .101])
SRMR (↓)	0.010	0.008
CFI (↑)	0.997	0.999
TLI (↑)	0.995	0.998
AIC (↓)	3354.034	3354.502
BIC (↓)	3393.613	3403.977
Adj. BIC (↓)	3355.596	3356.455

Table 5.10.: Fit indices reported per model following Dunn and McCray (2020). Higher-is-better fit measures are marked with (↑) while lower-is-better fit measures are marked with (↓). We argue that, overall, fit measures indicate that the three-factor model is preferable over the one-factor model. The RMSEA value is reported along with a 90% confidence interval.

the different levels of predictability induced by the different classifier scenarios. Concretely, we conduct a one-way ANOVA to test for an effect of the independent variable "scenario" on total PSP scores. We find that the main effect of scenario is statistically significant and large ($F(4, 195) = 16.25$, $p < 0.001$; $\eta^2 = 0.25$, 95% CI [0.16, 1.00]). We conduct a Tukey HSD test to find pairs of scenarios with significantly different mean scores. Out of the ten model pairs, we find significant differences between all pairs except four which align to the four most similar pairs (MIXED vs. MIXED-LESS-ALEATORY, MIXED-MORE-EPISTEMIC vs. HIGH-BOTH, MIXED-LESS-ALEATORY vs. LOW-ALEATORY, and MIXED vs. MIXED-MORE-EPISTEMIC). We report detailed statistics in Appendix A.9.2.2.

Concurrent and Predictive Validity. We assess how PSP scores are associated to related measures that are expected to overlap in an additional study. While we discuss the relation of PSP to trust, objective predictability, and NFC in detail in Section 5.2.3, we already briefly discuss the main results regarding concurrent and predictive validity here. Regarding concurrent validity, we find that SIPA scores are strongly correlated with PSP scores ($r = 0.856$, $p < 0.001$) but not as strongly correlated as the PSP subscales among each other (average correlation of $r = 88.9$ as reported above). We expected this effect due to the two scales sharing an overlapping but not identical theoretical foundation. Regarding predictive validity, we find that, while PSP is a significant non-linear predictor of objective predictability, we find no significant association between SIPA and objective predictability. Again, this confirms our theoretical expectation: While PSP scores should be related to objective predictability, SIPA is a higher-level construct, which we expected to be related to objective predictability in a much weaker way or not at all.

5.2.3. Predictors, Objective Predictability, and Effects of Explanations

So far, we focused on developing and validating our PSP scale in the context of a fictional task and classification system. Now, we investigate how PSP is associated with objective predictability, trust, SIPA, and NFC in the context of a realistic sentiment classification system. In particular, we evaluate these constructs under different levels of true system predictability (via system stochasticity) as well as different explanation visualization methods including the saliency and bar chart explanations discussed in Section 5.1.

5.2.3.1. Experiment Design

We conduct another user study with 200 participants to further investigate PSP and its relation to (supposedly) related constructs and measures in the context of a simple, but real sentiment classification system.

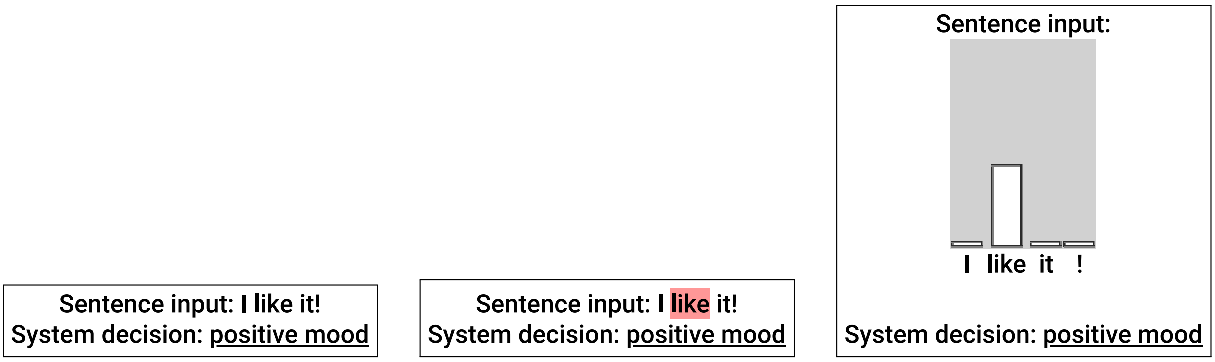
Sentiment Classifier. Our sentiment classifier makes use of word polarity scores provided in SentiWordNet3[23] (Baccianella et al., 2010) which are summed over all words in the sentence to obtain an overall sentiment score. The score corresponds to a positive sentiment prediction in case it is greater than zero and to a negative sentiment prediction when it is not. We choose to use this simple system because we want to use a system that (i) is easy to interpret (the word polarity scores can be considered to be technically faithful explanations), (ii) makes systematic errors that can be identified by humans (the system cannot handle negations or contractions by design), and (iii) can be executed on-demand as we — in contrast to our previous studies — also explore explanations in combination with an interactive user interface within this study. We additionally evaluate stochastic versions of the classifier. For this, we add independent, normally distributed noise to each word's polarity score.[24]

Explanation Modalities. Within our experiment, we compare six levels of explanations. First, we compare (i) no explanations, (ii) heatmap explanations, and (iii) bar chart explanations as discussed in Section 5.1.3. Figure 5.18 depicts the three explanation types. As in Section 5.1.3, the heatmap and bar chart explanations show absolute importance values, i.e., we do not communicate class-specific importance scores. In addition, we include interactive versions of these three explanation types. Concretely, we provide users an additional interface that allows them to enter arbitrary texts and receive the respective sentiment prediction, and — in the combinations involving heatmap or bar chart explanations — the respective explanation. Figure 5.19 shows these interactive interfaces for heatmap and bar chart explanations.

Procedure. We use a between-subject experiment design and assign 25 participants to each of the six explanation modalities, using the noiseless prediction system. To explore the effect of noisy systems, we additionally assign 25 participants to the medium noise system as well as the high noise system each, using system prediction examples without added explanations or system interaction interface. We ask each user to complete three phases of our experiment.

[23] `https://github.com/aesuli/SentiWordNet`

[24] We sample noise from $\mathcal{N}(0,1)$, scale this noise with 0.4 (0.8) for a medium (high) noise level, and clip the sum of the original score and the scaled noise to [-1,1].



Sentence input: I like it!
System decision: positive mood

Sentence input: I like it!
System decision: positive mood

Sentence input:

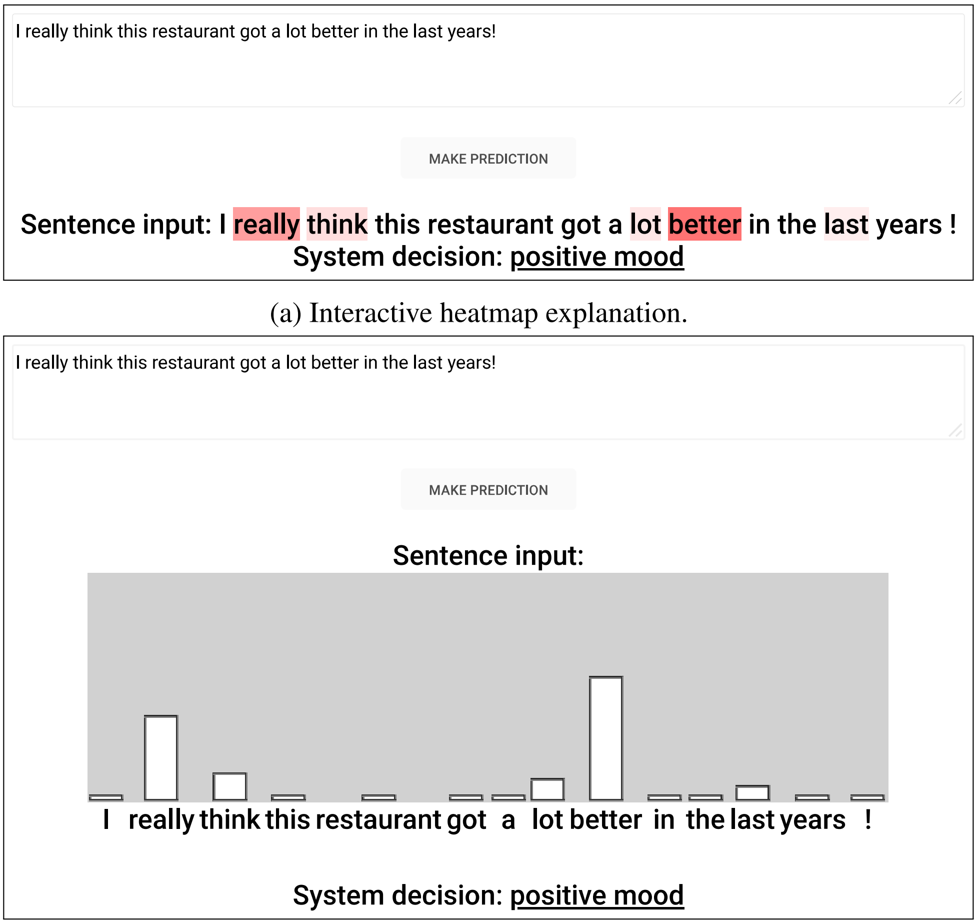
System decision: positive mood

(a) No explanation.

(b) Heatmap explanation.

(c) Bar chart explanation.

Figure 5.18.: The three explanations forms underlying the six explanation modalities used in our experiment.



I really think this restaurant got a lot better in the last years!

Sentence input: I really think this restaurant got a lot better in the last years !
System decision: positive mood

(a) Interactive heatmap explanation.

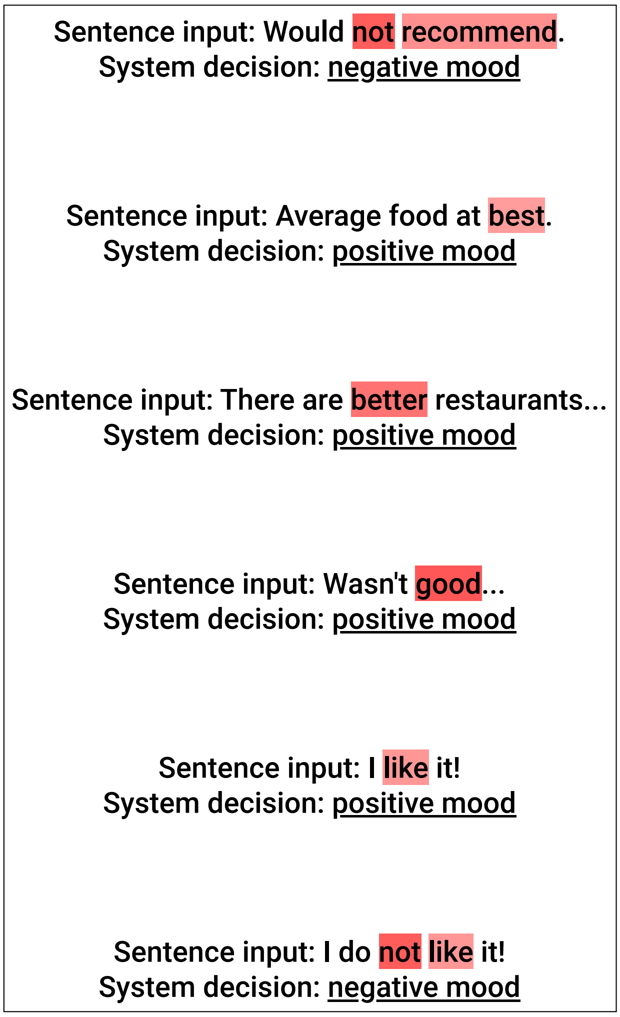
I really think this restaurant got a lot better in the last years!

Sentence input:

System decision: positive mood

(b) Interactive bar chart explanation.

Figure 5.19.: Two of the three additional interactive explanation modalities.



Sentence input: Would not recommend.
System decision: negative mood

Sentence input: Average food at best.
System decision: positive mood

Sentence input: There are better restaurants...
System decision: positive mood

Sentence input: Wasn't good...
System decision: positive mood

Sentence input: I like it!
System decision: positive mood

Sentence input: I do not like it!
System decision: negative mood

Figure 5.20.: Subset of the system predictions shown to users in the heatmap conditions.

Training Phase: In the first step, we show each user 20 system predictions using the respective explanation modalities. For the interactive modalities, we present users the same 20 examples of fictional restaurant reviews and *additionally* provide them with the interactive interface to give them the same context and ensure they receive the same “hints” regarding the system’s behavior. Figure 5.20 depicts six of the 20 example predictions we display to users in the heatmap conditions. We provide the full list of examples in Appendix A.9.3. Each user receives the same 20 prediction examples but in a randomized order to mitigate carry-over effects. We compose the 20 fictional review sentences in such a way that half of them correspond to a positive system decision, and half of them to a negative decision, and such that each of these halves contains five correct system decisions and five incorrect system decisions. Further, the examples are chosen in a way that demonstrates that the model treats contractions, such as “don’t” differently from “do not” and does not correctly resolve negations. In addition, the examples contain two sentences that each are repeated to make the users aware of (non-)deterministic model behavior which is relevant for the noisy systems described above.

Prediction Phase: In the second step, we ask users to predict which predictions the system will make for new, unseen sentence inputs (e.g., “I love the food at this place!” and “I expected it to be better.”). We provide the full list of sentences in Appendix A.9.3. We again randomized the order of sentences across participants.

Questionnaire Phase: In the third step, we ask participants for self-reports of PSP using our scale, SIPA using the respective scale proposed by Schrills et al. (2022), system trust using the facets of system trustworthiness (FOST) scale (Franke et al., 2015), and the user’s individual NFC using the NCS-6 scale (de Holanda Coelho et al., 2018). We choose to include a measure of user’s NFC based on Buçinca et al. (2021) who find that their intervention on explanation presentation to reduce over-reliance disproportionately affected users who reported a high NFC.

Participants. We recruit 200 participants from the United States, Australia, and the United Kingdom on MTurk. Participants had a mean age of 43.2 (SD=12.2) years. 89 participants identified as female, 110 as male, and one participant as non-binary. Across the entire experiment, the population’s mean PSP rating is 5.39 (SD=0.89).

5.2.3.2. Reliability Reproduction

Before analyzing the relation between constructs and objective performance measures, we first re-assess each scale’s internal reliability. Table 5.11 displays the respective values for Cronbach’s α (Cronbach, 1951), and McDonald’s ω (Mcdonald, 1999). As in our first relia-

Reliability measure	Scale			
	PSP	SIPA	FOST	NCS-6
α	0.874	0.852	0.780	0.766
ω_h	0.713	0.593	0.100	0.392
ω_t	0.896	0.886	0.830	0.834

Table 5.11.: Measures of internal reliability estimated for the four scales measured in our second experiment ($N = 200$) including our PSP scale. α refers to Cronbach's α (Cronbach, 1951), ω_h and ω_t refer to versions of McDonald's ω (Mcdonald, 1999).

bility assessment, we quantify the reliability of the PSP subscales using the Spearman-Brown coefficient and find $R = 0.718$ for the epistemic subscale, $R = 0.726$ for the aleatory subscale, and $R = 0.698$ for the effective subscale, indicating high reliability across subscales.

5.2.3.3. Predictors of Perceived Predictability

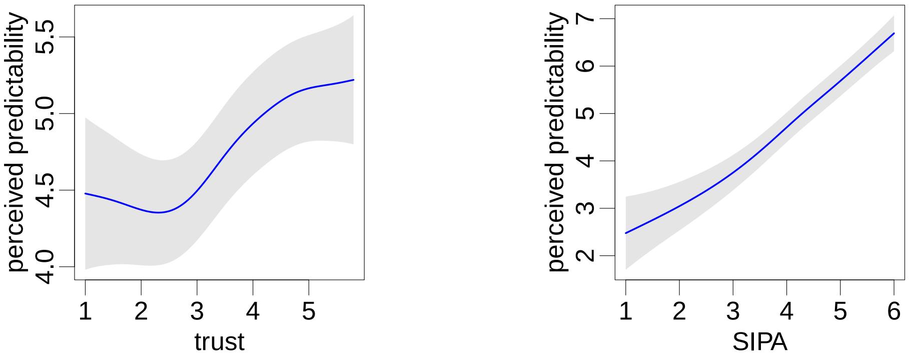
Next, we focus on which factors are predictors of PSP scores. In particular, we explore whether PSP scores can be predicted from objective measures, such as completion time or prediction correctness. For this, we model PSP scores using a GAM model.

Model. As discussed in Section 2.4, GAM models offer the advantage to model additive smooth non-linear effects of numeric covariates. We include smooth terms for prediction correctness, completion time of the prediction phase, FOST trust scores, SIPA scores, NSC-6 NFC scores, and participant age. We additionally include parametric terms for explanation form (none, saliency, and bar charts), interactivity[25], noise level, and the participant's identification to account for the experiment design and control for potentially confounding effects. We additionally add an interaction term between explanation form and interactivity as we expect that the different explanation forms induce different levels of interaction motivation.

Results. We report Wald tests for the parametric and smooth terms in Table 5.12 and Table 5.13 respectively.

Objective Scores Do not Predict PSP: Notably, we do not find significant effects of the objective scores (i.e., prediction correctness and completion time) on PSP scores, supporting our

[25] We find that 21.3% participants in the interactive conditions did not use the interactive prediction interface and thus define the "interactivity" factor to distinguish between participants that did interact with a prediction interface and participants who did not (including participants that could use the interface but did not do so).



(a) Trust

(b) SIPA

Figure 5.21.: Partial effect plots of factors with a significant effect on perceived predictability within our GAMM analysis. The plots show all significant smooth effects accounting for various additional parametric and smooth effects. Note that y-axes are scaled per plot. Remarkably, no objective score, such as prediction correctness or completion time is found to be a significant predictor of PSP.

assumption that PSP measures a distinct concept than objective predictability does. Separate Pearson correlations between PSP and prediction correctness ($r = 0.050$, $p = 0.478$) and completion time ($r = 0.109$, $p = 0.125$) yield the same conclusion. This result indicates that PSP scores capture additional information, that we cannot substitute with automatic scores and supports our hypothesis that we cannot cut corners and evaluate systems without subjective human evaluation as we argued in Section 4.4. In fact, we will demonstrate that, in the opposite direction (i.e., predicting prediction correctness from PSP scores), subjective ratings are significant predictors of prediction correctness in Section 5.2.3.4.

Strong Association Between PSP and SIPA: We find that increases in trust and SIPA scores can be associated with increases in PSP scores, which is supported by the Pearson correlations reported in Table 5.16. Figure 5.21 displays the respective partial effects of trust and SIPA ratings. Again, the strong association between PSP scores and SIPA scores is consistent with the theoretical background of PSP and SIPA: as SIPA models predictability as one facet of situational information processing awareness, PSP "zooms in" to the predictability facet and models predictability using the three proposed facets.

Effects of Explanation Modality and no Effect of Noise Level: For the parametric terms, we find a significant main effect of explanation form as well as a significant interaction effect between explanation form and interactivity as reported in Table 5.12. We report detailed parametric estimates in Appendix A.9.3. A post hoc Wald comparison of the contrasts for explanation formats revealed significant differences between saliency and bar chart explanation forms

	df	**F**	**p**
explanation format	2.00	3.44	**0.03**
interactivity	1.00	1.80	0.18
explanation format:interactivity	2.00	3.18	**0.04**
noise level	2.00	0.41	0.67
identification	2.00	0.37	0.69

Table 5.12.: Wald tests for the parametric terms in our model of PSP scores. Explanation format (none, saliency, or bar chart) was found to have a significant effect on PSP scores.

	edf	**Ref.df**	**F**	**p**
s(prediction correctness)	0.69	9.00	0.15	0.14
s(trust)	3.20	9.00	5.82	<**0.001**
s(completion time)	0.00	9.00	0.00	0.39
s(SIPA)	2.79	9.00	44.13	<**0.001**
s(NFC)	0.00	9.00	0.00	0.53
s(age)	0.00	9.00	0.00	0.66

Table 5.13.: Wald tests for the smooth terms in our model of PSP scores. Trust and SIPA ratings have significant effects on PSP scores. Notably, no objective score (correctness and completion time) is found to be a significant predictor of PSP scores.

($\chi^2(1) = 6.619$, $p = 0.010$) for which saliency explanations are associated with significantly higher PSP ratings. A respective joint post hoc test for explanation formats and interactivity revealed differences between non-interactive bar charts and non-interactive no-explanation sentences ($\chi^2(1) = 3.963$, $p = 0.047$), non-interactive bar charts and interactive bar charts ($\chi^2(1) = 4.759$, $p = 0.029$), and non-interactive bar charts and non-interactive saliencies ($\chi^2(1) = 6.619$, $p = 0.010$). Detailed estimates are reported in Appendix A.9.3. Interestingly, we do not find a significant effect of noise level on PSP scores. We revisit this observation in the context of the effect of noise level on prediction correctness in the following.

5.2.3.4. Perceived Predictability versus Prediction Correctness

In the previous analysis, we explored which factors are predictors of PSP. Now, we investigate if and how PSP scores and additional factors are related to objective prediction correctness.

Model. We analyze the relation between perceived predictability and prediction correctness within another GAM model. We consider smooth terms for PSP scores, FOST trust scores, SIPA scores, NCS-6 NFC scores, completion time, and participant age. In addition, we consider parametric terms for explanation format, interactivity, noise level, and participant identification. As for the GAM model discussed above, we also include an interaction term for explanation format and interactivity.

Results. We report Wald tests for the parametric and smooth terms in Table 5.14 and Table 5.15 respectively. Details are reported in Appendix A.9.3.

Subjective Ratings Are Predictive of Objective Correctness: While we found that objective prediction correctness did not have a significant effect on subjective PSP scores in Section 5.2.3.3, we find a significant effect of PSP scores on prediction correctness. As shown in Figure 5.22a, our model associates an increase in PSP scores with a moderate increase in objective prediction correctness. In addition, we find a strong contrary effect of trust. As shown in Figure 5.22b, increasing levels of participant's trust in the system correspond to a decrease in prediction correctness. Hase and Bansal (2020) found that subjective explanation quality ratings of "Does this explanation show me why the system thought what it did?" are not predictive of user correctness. Similarly, Buçinca et al. (2020) found that trust ratings are not predictive of performance. In contrast to Hase and Bansal (2020) and Buçinca et al. (2020), we find that PSP and trust ratings are predictive of prediction correctness. While the specific combinations of the particular scale and score as well as the usage context, do not allow to draw general conclusion in either direction, we note that in contrast to Hase and Bansal (2020) and Buçinca et al. (2020), we model non-linear effects of subjective ratings on prediction correctness using GAM models and also find the resulting function estimates to be non-linear (Figures 5.22a and 5.22b).

Better Predictions Need Time: Figure 5.22c shows that, within our model, an increase in prediction completion times is associated with a moderate increase in prediction correctness. We argue that this effect is to be explained by high differences in the participants' interest in the prediction task and their corresponding willingness to think about the system's behavior. While this finding is not surprising, it supports the use of time-based insufficient effort responding detection which aligns with the findings of Bowling et al. (2021).

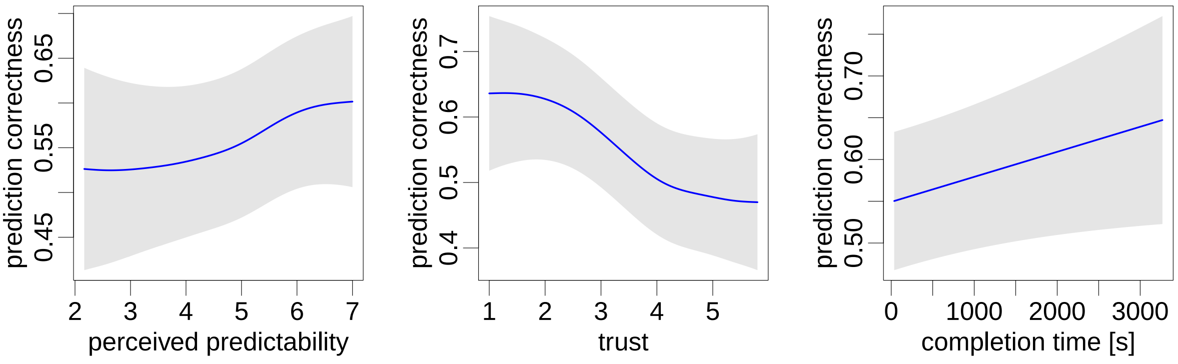
Noise Level Affects Prediction Correctness: Among the parametric terms, we find that only noise level has a significant effect on prediction correctness. In particular, explanation format did not have a significant effect. A post hoc Wald comparison of the contrasts for noise level revealed significant differences between a noise level of 0.0 and 0.4 ($\chi^2(1) = 28.196$, $p < 0.001$) and 0.0 and 0.8 ($\chi^2(1) = 34.539$, $p < 0.001$). The difference between 0.4 and 0.8 was not significant ($\chi^2(1) = 0.347$, $p = 0.556$). These findings are consistent with the box plots for prediction correctness (i.e., objective predictability) shown in Figure 5.23.

	df	F	p
explanation format	2.00	1.44	0.24
interactivity	1.00	0.55	0.46
explanation format:interactivity	2.00	0.85	0.43
noise level	2.00	22.13	<**0.01**
identification	2.00	0.66	0.52

Table 5.14.: Wald tests for the parametric terms in our model of prediction correctness scores. Noise level (i.e., the level of system stochasticity) is found to have a significant effect on prediction correctness.

	edf	Ref.df	F	p
s(PSP)	1.66	9.00	0.79	**0.01**
s(trust)	2.34	9.00	2.45	<**0.01**
s(completion time)	0.80	9.00	0.45	**0.02**
s(SIPA)	0.00	9.00	0.00	0.85
s(NFC)	0.86	9.00	0.20	0.11
s(age)	0.48	9.00	0.08	0.22

Table 5.15.: Wald tests for the smooth terms in our model of prediction correctness. We find PSP scores, trust scores, and completion time to have significant effects on prediction correctness.



(a) PSP

(b) Trust

(c) Prediction time

Figure 5.22.: Factors with a significant effect on objective rating correctness within our GAM analysis. The plots show all significant smooth effects accounting for various additional parametric and smooth effects. Note that y-axes are scaled per plot.

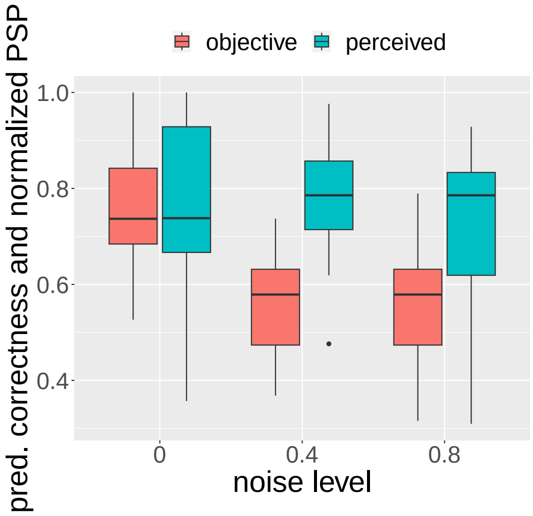


Figure 5.23.: Boxplot showing the distributions of prediction correctness and normalized PSP scores for different levels of system stochasticity and no additional explanations.

Subjective ≠ Objective: Comparing our findings on subjective PSP scores and objective prediction correctness, we observe that PSP scores are affected by explanation modality and prediction correctness is affected by noise level. While using saliency visualizations instead of bar charts results in a significant increase of PSP ratings, we find no effect of explanation form on prediction correctness. Similarly, adding noise to the system's word polarity estimates corresponds to a significant drop in prediction correctness without affecting PSP ratings. This observation supports our hypothesis that objective predictability and subjective predictability measure distinct characteristics of the user's mental model of a system (as illustrated in Figure 5.12). Our results further support our observations that supposedly minor visualization decisions can affect users' perception of explanations, which we explored in Section 5.1.

Hallucinated (Lack of) Predictability: While the objective information in the heatmaps and the bar charts is identical, we hypothesize that heatmaps have properties that induce a higher level of perceived information gain, than, e.g., bar charts. Note that the results of this experiment do not allow us to judge whether this increase in perceived predictability is beneficial or misleading. Similarly, our results do not show whether the observed difference is due to an increase induced by heatmaps or a decrease induced by bar charts. Related work on perceptual misinterpretation of bar charts found the "within-the-bar bias", users' tendency to perceive values contained within the bar (i.e., below the top edge) as more probable when, e.g., inspecting means depicted using bars and being asked about the probability of equidistant values above and below the bar's edge (Newman and Scholl, 2012). We hypothesize that (a part) of our observations can

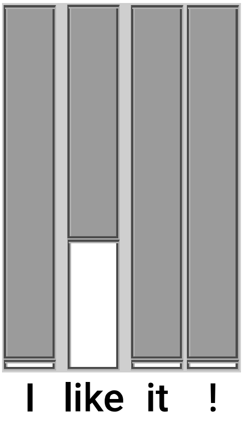


Figure 5.24.: Enhanced bar chart explanation visualization using cumulative bars as proposed by Kang et al. (2021) to mitigate "within-the-bar bias" (Newman and Scholl, 2012) and the supposedly associated underestimation of importance scores.

be explained by this effect and a corresponding underestimation of importance scores., which, in turn, lead to a less concise mental model of the system. Kang et al. (2021) investigate modifications of the bar chart visualization and find that cumulative bar charts (i.e., also filling the upper part using a different color) can reduce bias. Consequently, we adapt our bar chart visualization and recommend to use cumulative bar charts as shown in Figure 5.24.

Illusion of Explanatory Depth: The observed invariance of PSP scores under variation of noise levels along with the observed drop in prediction correctness and the demonstrated discrimination of known groups in Section 5.2.2.3 raises the question of why participants did not notice their reduced ability to predict the system's behavior while, at the same time, being sensitive to the choice of explanation visualization. We hypothesize that this phenomenon can potentially be explained by the illusion of explanatory depth explored by Rozenblit and Keil (2002). Concretely, we hypothesize that the participants' own judgments of the sentences' sentiment create a predisposition to form an illusion of explanatory depth while the fictional shape classification task does not allow them to fall back to their own judgments and alleged familiarity with the domain and forces them to reflect upon their knowledge. Similarly, Gonzalez et al. (2021) explore how explainees are affected by belief bias and argue that fictional domains might mitigate its distorting influence.

5.2.3.5. Relation to Trust, SIPA, and Need for Cognition

In Section 5.2.3.3 we explored to which extent PSP scores can be predicted in terms of a sum of smooth functions of other subjective ratings (such as trust ratings), objective performance measures (such as completion time), and additional variables related to, e.g., the explanation

	SIPA	**FOST**	**NCS-6**
PSP	**0.856** (< 0.001)	**0.558** (< 0.001)	**0.248** (< 0.001)
SIPA		**0.424** (< 0.001)	**0.272** (< 0.001)
FOST			0.118 (0.096)

Table 5.16.: Pearson correlation coefficients between PSP, SIPA, trust (FOST), and NFC (NCS-6). Numbers in parentheses correspond to Holm-adjusted p values. Significant correlation coefficients are highlighted in **bold** font.

format used. The focus of our GAM analysis was to *combine* all provided factors into a prediction of PSP scores. Among the three collected related scale scores for trust, SIPA, and NFC, we found that the smooth terms for trust and SIPA scores were significant. In turn, accounting for all other terms, we did not find a significant effect of NFC.

In the following analysis, we investigate a related, but slightly different question. Concretely, we ask how strongly ratings of a scale in *isolation* are related to ratings of another scale (without accounting for the respective relations to other scales). For this, we explore *pairwise* Pearson correlations between the (paired) responses collected with the four scales.

Results. Table 5.16 displays Pearson correlations and the respective Holm-adjusted p values.

Confirming the Association Between PSP and SIPA as Well as Trust: We observe that PSP and SIPA ratings show a strong (linear) correlation ($r = 0.856, p < 0.001$), which is consistent with our theory-driven expectations (see Section 5.2.3.3) as well as our empirical results from our GAM analysis (see Figure 5.21b). Similarly, our correlation analysis confirms the association between PSP and trust ($r = 0.558, p < 0.001$) which we observed in our GAM analysis.

Predictors of Trust: Schrills et al. (2022) also evaluate the correlations between SIPA scores and FOST trust scores as well and find that — across three different samples — correlations vary between 0.55 and 0.84. We confirm the finding of Schrills et al. (2022) that SIPA and trust have a significant (linear) association. We statistically compare the strength of the correlation of SIPA and trust to the strength of the correlation between PSP and trust using the cocor R package (Diedenhofen and Musch, 2015).[26] We find that, in our study, the correlation between PSP and trust is significantly higher than the correlation between SIPA and trust. We argue

[26] The cocor package provides numerous tests including Dunn and Clark's z (Dunn and Clark, 1969) and Zou's confidence interval (Zou, 2007). All implemented tests indicate a significant difference between the two correlation coefficients.

that this observation can be explained by the theoretical foundations of the PSP and the SIPA scales. While SIPA measures the facets transparency, understandability, and predictability, our PSP scale focuses on (three facets of) predictability. We hypothesize that trust has a stronger dependence on predictability and less on, potentially preceding, effects of e.g., perceived understandability. To test this hypothesis, we analyze the correlations of trust with the respective SIPA subscales. Following our hypothesis, the SIPA predictability subscale should have significantly higher correlations with trust compared to the transparency and understandability subscales. We find that the SIPA predictability subscale has a stronger correlation to trust ($r = 0.447, p < 0.001$) than the transparency subscale ($r = 0.318, p < 0.001$) and the understandability subscale ($r = 0.351$, $p < 0.001$). The respective statistical tests of differences in correlation strengths using cocor all indicate that the predictability subscale has higher correlations with trust than the remaining two subscales. This supports our hypothesis, that the stronger correlation between our PSP scale and trust (compared to the SIPA scale and trust) can be attributed to the focused construct of the PSP scale along with a stronger association between PSP and trust. In another experiment, Ford et al. (2020) find that an increased rate of system misclassifications was associated with a decrease in self-reported levels of trust. We fit another GAM model to assess whether we can replicate their observations.[27] In contrast to their results, we do not find a significant effect of noise level on trust. However, we find a significant effect of explanation modality on trust. A post hoc Wald comparison of the contrasts of explanation modality reveals significant pairwise differences between bar charts and saliency explanations ($\chi^2(1) = 8.007$, $p = 0.005$) as well as between no explanation and saliency explanations ($\chi^2(1) = 8.161, p = 0.004$). For both pairs, saliency explanations are associated with a significantly reduced trust level. We report detailed test statistics for the discussed and additional effects in Appendix A.9.3. This association opposes our observation for the effect of explanation format on PSP for which saliency explanations were associated with a significant increase in PSP compared to bar chart explanations (see Section 5.2.3.3). This conflict indicates that PSP ratings do not only offer an additional perspective in addition to objective scores but also yield information beyond measures of related subjective constructs whose association to PSP should be explored in-depth in future work.

Effect of High Need for Cognition: Our correlation analysis finds a significant (linear) association of PSP and NFC. As the results of Buçinca et al. (2021) indicate that users' NFC affects the effect that explanations have on their decision behavior, it is plausible to assume that NFC has an effect on perceived predictability as well. Interestingly, we do not find an effect of

[27] We include the same smooth and parametric terms as for our analysis of PSP ratings and swap PSP and trust.

NFC scores on prediction correctness in our scenario.[28] An analysis of the correlation between NFC and prediction correctness confirms this result ($r = 0.071$, $p = 0.318$). At the same time, we do not observe a significant correlation between NFC and trust. We thus hypothesize that participants' need for cognition affects their system perception in a way that neither is captured by prediction correctness nor trust but still positively affects PSP. Future work should build upon this initial observation and leverage our scale to further investigate the relation between PSP and further constructs, such as mental demand. Additional qualitative evaluation, such as think-aloud studies with participants with different NFC might shed light on the underlying relation between NFC and PSP.

5.2.4. Overall Discussion

In this section, we developed and evaluated a novel 6-item scale, and applied it to explore the effects of explanations and system stochasticity on (perceived) system predictability and the relation of PSP to, i.a., prediction correctness, trust, SIPA, and participants' NFC.

Overall, we collected opinions from 40 participants to guide our theory development based on uncertainty theory, incorporate feedback from six experts to improve our initial item pool, conduct written cognitive interviews with 25 participants to further enhance and filter our items, conduct a study on a functional shape classification study with 200 participants to distill and evaluate the final version of our scale, and conduct an additional study on different varieties of a sentiment classification system to confirm our evaluation and explore the relation of PSP to related scales and prediction correctness.

Our scale evaluations demonstrated that our PSP scale exhibits desirable psychometric properties, such as a consistently high internal reliability, and indicate that it can be used as (a) a unidimensional measure of perceived predictability and (b) a hierarchical measure of perceived predictability with three subscales for epistemic, aleatory, and effective predictability.

Our results suggest that (a) PSP cannot be predicted from automatic measures, such as prediction correctness or completion time, (b) vice versa, prediction correctness is significantly affected by subjective scores (higher PSP scores are associated with higher prediction correctness while higher trust scores are associated with notably lower prediction correctness), (c) the choice of explanation modality affects PSP but not prediction correctness, and (d) higher system stochasticity affects prediction correctness but not PSP.

Overall, we find that subjective PSP and objective prediction correctness measure distinct aspects of users' mental models of a system and that the two measures can diverge – highlighting

[28] We additionally evaluate a binarized NFC covariate following a partition of subjects above and under the median NFC as applied by Buçinca et al. (2021) and still do not find a significant effect of NFC.

the need to explore both, subjective as well as objective predictability. We link our observations to the known “within-the-bar bias” (Newman and Scholl, 2012) and the illusion of explanatory depth (Rozenblit and Keil, 2002) and refine our recommended bar chart visualization to include cumulative bars as shown in Figure 5.24.

5.3. Related Work

In the following, we (i) review related work on the effects explanations have on users (Section 5.3.1) including perceived system performance, usefulness, trust, and user behavior, (ii) discuss risks and misuses that arise when explanations are provided to users (Section 5.3.2), and (iii) provide a high-level overview of scale development and validation (Section 5.3.3).

5.3.1. Effects of Explanations on Users

In the following, we review related work on how presenting explanations to users affects their behavior and perception of the system that is explained.

5.3.1.1. Effects on User Perception

The effects of exposing users to explanations that accompany system predictions have been studied across a broad range of explanation methods, tasks, and user populations. In the following, we present a non-exhaustive summary that should equip the reader with a notion of the diverse dimensions of user perceptions that explanations can affect.

Perceived System Performance. Nourani et al. (2019) investigate local explanations for image classification and find that participants significantly underestimate system accuracy when implausible explanations (compared to plausible explanations) are provided. Biran and McKeown (2017) study explanations for a stock price prediction classifier and observe that providing explanations improves users’ ability to estimate whether classifier predictions are correct or not. The clinical decision support system study of Bussone et al. (2015) found that providing explanations can lead to over-reliance on the system. Similarly, we found both over- and under-estimation effects in our prior work on explainable QA (Schuff, 2020).

Trust. Related work came to different conclusions regarding the effect of explanations on user trust. For example, the study of Cramer et al. (2008) did not find an effect of explaining a recommender decision on user trust. Similarly, Ribes et al. (2021) investigated explanations for a news aggregator system and found no effect on user trust. In contrast, Khurana et al. (2021) explored explanations from chatbots and found an improvement in trust. We argue that the diversity in tasks, explanation methods, and experiment designs does not allow to draw general conclusions on the effect of explanations on user trust. However, as trust was found to be correlated with, e.g., acceptability (Nadarzynski et al., 2019), it can be assumed that it is (at least indirectly) affected by explanations.

Acceptability. Herlocker et al. (2000) investigate explanations in recommender systems and observe that providing explanations can foster system acceptability. Cramer et al. (2008) observe a similar effect in their study of art recommender systems.

Perceived Usefulness. The mentioned chatbot study of Khurana et al. (2021) found that explanations enhance perceived usefulness. Similarly, Bansal et al. (2021) observed that high-quality explanations increase usefulness ratings. However, a high perceived usefulness does not necessarily have to translate to an actual usefulness as we discuss in the following.

5.3.1.2. Effects on User Behavior

Besides studying the users' subjective experience of using an AI system (with explanations), the extent to which provided explanations affect the users' (decision) behavior has to be monitored.

Agreement and Human-AI Performance. A broad body of research reported improvements in user decisions when users receive explanations along the decisions of an AI system (Lundberg et al., 2018; Lai and Tan, 2019; Feng and Boyd-Graber, 2019; Green and Chen, 2019; Lai et al., 2020; Zhang et al., 2020b; Buçinca et al., 2020; Poursabzi-Sangdeh et al., 2021).[29] Bansal et al. (2021) call the generalizability of these improvements into question and find that providing explanations does generally increase the rate at which users accept a system's predictions. Importantly, this effect also holds for erroneous system decisions. Thus, improvements observed in earlier studies could largely be due to the used AI systems performing distinctly better than humans. When a AI system in fact is performing distinctly better than its users, this might still be a desirable effect. However, increasing the users' tendency to blindly agree to a given system decision clearly should not be a method's goal.

Likeability-Effectiveness Trade-off. In addition to the effect of explanations on user agreement and user performance, Buçinca et al. (2020) investigate whether an explanation that

[29] See Bansal et al. (2021) for an overview of these studies.

is *perceived to be helpful* by explainees also corresponds to an explanation that *actually is helpful* to them. The authors showed that between two decision support systems, users preferred one system (in terms of rating it as more helpful and trusted), although their actual performance was significantly better with the less-favored system. In their follow-up work, Buçinca et al. (2021) found a trade-off between subjective system quality ratings and effective human-AI performance for explainable AI systems.

Overall, the effects of providing explanations to users are still actively studied and — as with the effects on user perceptions — it is likely that the effects on user behavior do not generalize across tasks, AI systems, and explanation methods.

5.3.2. Risks and Misuses of Explanations

While explanation methods have a great potential to create added value for their users and empower individuals affected by decisions of AI systems used for automated decision making, providing explanations is also associated with certain risks, which we discuss in the following.

Fairwashing, Manipulation, and Remote Explainability. Aïvodji et al. (2019) warn against misusing explainability methods to create the false impression that a system adheres to some ethical values while it, in fact, does not. They demonstrate how an unfair black-box model can be *fairwashed* using their proposed LaundryML method.

Similarly, Lakkaraju and Bastani (2020) demonstrate how high-fidelity explanations can be constructed such that they still allow the model to discriminate without making this discrimination visible in the explanation. Concretely, their approach exploits input feature correlations that allow their method to cover the usage of obviously discriminating features, such as race by using correlated features, such as zip code. They demonstrate their approach's potential to manipulate user trust within a user study involving domain experts from criminal justice.

In a related argument, Merrer and Trédan (2019) introduce the *bouncer problem*. In analogy to a club bouncer that might cover discriminating customer reject with untruthful explanations (e.g., explaining a reject with non-matching cloth while the decision actually is based on the guest's age), providing remote explanations of decisions of a black-box model (e.g., via an API) allows malicious model providers to cover discriminating decisions with presumably plausible explanations using a "public relations attack" method which the authors propose.

Dark Patterns and Explainability Pitfalls. Besides directly forging an explanation or modifying it to make the underlying model appear more favorable than it is, explainability can also be misused when designing explanation presentation and control. Maclure (2021) discuss *dark patterns of explainability* building on the dark patterns in user experience design discussed by Gray et al. (2018). Dark patterns refer to deceptive design strategies that, e.g., artificially complicate an interaction process to make the user act in someone else's interest. Transferred to explainability, such a pattern would be, e.g., a strategically induced information overload that demotivates a user to understand a given explanation.

Apart from intentionally malicious strategies, explanation methods can also have unexpected negative effects that emerge even when designers and developers have the best intentions. Ehsan and Riedl (2021) introduce the term *explainability pitfalls* to refer to such unintended negative effects. In their case study, they, e.g., find that simply showing users unlabeled numbers without context of what these numbers refer to, increased users' trust and perceived intelligence of a reinforcement learning (RL) agent.

Accountability and the Responsibility Gap. Lima et al. (2022) argue that overemphasizing explanation methods can undermine accountability. When AI systems are perceived as blameworthy agents, explanations can help to misuse these systems as scapegoats to shift away responsibility from their designers. This problem highlights a specific instance of the *responsibility gap* (Matthias, 2004) (or *liability problem* (Asaro, 2016)), which refers to the problem of ascribing responsibility in the context of AI decisions which differ from the traditional responsibility and liability regarding a machine for which there is no “gap” between the manufacturer's responsibility and the operator's responsibility.

Overall, the risks that explanation methods can pose have to be considered carefully when researching and deploying explanations. This is not to say that research should stop investigating explainability. On the contrary, we have to understand which explanation methods affect users in which way to (i) spot and stop careless or malicious use of explanations, and (ii) develop explanation methods that serve users and society. To study the effect of explanations on users, we have to evaluate how they perceive and interact with explanations.

5.3.3. Scale Development

A scale is designed to quantify a construct, e.g., “system usability”, that may comprise multiple aspects, called dimensions, e.g., — for system usability — efficiency, effectiveness, and satisfaction (Brooke, 1996; Finstad, 2010). The most common type of scale is the Likert scale, containing (multiple) items, rated by the user on a discrete range. The overall score for a dimension or construct is calculated by combining the numbers related to the answer from each item. Depending on the exact scale, the procedure used may vary, e.g., items can be weighted. Note that the single questions are not scales themselves but rather are items and the group of items together constitutes the scale. Using multiple items instead of a single rating allows one to assess the scale's internal consistency, e.g., via Cronbach's alpha (DeVellis and Thorpe, 2021). Although one cannot directly assess how well an item is related to the latent variable of interest (e.g., perceived system predictability) because this is the construct to be captured via the items, one still can quantify these relationships indirectly via item-item correlations. If the items have a high correlation with the latent variable, they will have a high correlation with each other (DeVellis and Thorpe, 2021).

Designing a valid and reliable scale requires a precise development process, summarized by Boateng et al. (2018) and explained in detail by DeVellis and Thorpe (2021). For NLP, the fields of psychology, HCI, and robotics already offer a valuable range of scales. Validated questionnaires exist, for example, for evaluating trust (Körber, 2018), usability (Brooke, 1996; Finstad, 2010), cognitive load (Hart and Staveland, 1988), social attribution (Carpinella et al., 2017), or user interface language quality (Bargas-Avila and Brühlmann, 2016). However, to the best of our knowledge, there is no such scale available to measure (dimensions of) PSP. A potential pitfall in designing and applying (Likert) scales is to use scales that have not been validated (DeVellis and Thorpe, 2021). Although such unvalidated scales *can* yield valid measurements, the researcher does not know for certain that they will and runs the danger of not measuring the construct one intended to. In order to obtain a reliable and valid scale, we thus follow best practices of scale development (DeVellis and Thorpe, 2021; Boateng et al., 2018) in the development and validation of our PSP scale.

6. Conclusion and Future Work

In this thesis, we explored explainable NLP and provided contributions on (i) a technical machine learning level, (ii) a methodological evaluation level, as well as (iii) a human-centered level. This final chapter summarizes our contributions and discusses next steps for future work.

System Architectures and Explanations

In Chapter 3, we (a) addressed external knowledge integration for explanation generation and (b) explored model self-correction.

For external knowledge integration, we studied explainable NLI models and investigated how different knowledge sources (such as knowledge bases and language models) affect the models' classification performance as well as the generated free-text explanations' quality. We found that fine-tuned language models reach the highest performance on the explainable NLI dataset e-SNLI (Camburu et al., 2018) as well as the highest average accuracy within the NLI stress test evaluation Naik et al. (2018). However, their performance broke down on numerical reasoning and negations. In addition, we conducted a large-scale human crowdsourcing evaluation and found that, surprisingly, high differences in accuracy (up to 3.2%), BLEU (up to 10.17 points), or BLEURT scores do not reflect in significant differences in human ratings of explanation correctness, commonsense inclusion, grammar, or entailment prediction correctness.

While we explored how external knowledge affects explanations of NLI systems, one potential direction for future work is to investigate the effects of external knowledge on further explainable systems. Concretely, we consider studies of the impact of external knowledge on post hoc saliency attribution explanations as well as counterfactual explanations to be promising next steps. A further direction for future work is to leverage the information that our collected human ratings provide. Concretely, an extended psychometric analysis of how perceived explanation correctness, perceived commonsense inclusion, perceived grammatical correctness as well as perceived entailment prediction correctness are interrelated has the potential to extend our knowledge of user-perceived explanation quality. To facilitate such and similar studies, we make all collected human ratings available at `https://github.com/boschresearch/external-knowledge-explainable-nli`.

For model self-correction, we set out to enable classification models (regardless of the task), to re-consider and correct their own predictions. In contrast to humans who can solve complex problems by creating a sequence of ideas (involving an intuitive decision, reflection, error correction, etc.) in order to reach a conclusive decision, today's machine learning models are mostly trained to map an input to an output in one single step. To provide models with the capability of having a second, third, and k-th "thought", we took inspiration from philosophy and developed our thought flow method based on Hegel's dialectics. Our method comprises an architecture extension as well as adapted training and prediction algorithms. In particular, we train models to estimate the correctness of their own prediction and use this estimate's gradient to iteratively update the models' prediction towards higher self-estimated correctness. We applied our method to QA systems and demonstrate our method's ability to correct its own predictions and its potential to notably improve model performances (>9% absolute F_1-score). In addition, we performed a qualitative analysis of thought flow correction patterns and explored how thought flow predictions affect human users within a crowdsourcing study. We find that thought flows enable improved user performance and are perceived as more natural, correct, and intelligent as single and/or top-3 predictions.

As our oracle-stopping experiments demonstrated, halting the self-correction at the right time is crucial to avoid over-corrections. We encourage future work to explore automatic optimal stopping and expect geometric features of the correction trajectory (e.g., a very steep ascent) to yield an effective heuristic approach. Further, directly learning a policy to take correction steps (or stop) based on the correction module's gradients — similar to learning an optimizer (Li and Malik, 2016) — can be a promising advancement of our method. An additional extension of our work that could be addressed in future work is to apply our thought flow method to further classification as well as regression tasks. We expect that our method will yield promising results in any task for which predicting the correct answer is substantially harder than verifying whether a given answer is plausible. We thus hypothesize that our method will be particularly useful for (a) tasks that involve multiple reasoning steps, (b) structured predictions, and (c) generation tasks. Beyond the multi-step QA task we explored, multi-step reasoning NLP tasks include multi-step reading comprehension (Lin et al., 2019b) or multi-step numerical reasoning (Zhao et al., 2022). Structured prediction tasks in NLP include, i.a., dependency parsing or multi-label document classification. A concrete example of a generation task are dialog systems for which our flow method could prevent systems from generating inconsistent responses by re-assessing their answer generation with respect to the dialog history. As we demonstrated in our vision experiments, our method can also easily be applied beyond NLP, motivating further applications in more complex vision tasks, such as scene comprehension or image segmentation. Besides applying our method to further tasks, we also encourage a deeper study of thought flows' effects on users. While our user study revealed numerous advantages over, e.g., single-step predictions, it remains to be investigated which factors cause participants to perceive thought flow predictions to be, e.g., more natural and how thought flow predictions achieve an increased user performance without the sacrifice of significantly increased completion times that is required to increase user performance via top-3 predictions.

Evaluating and Quantifying Explainability

In Chapter 4, we (a) proposed new proxy scores to quantify prediction-explanation coupling in explainable QA systems, (b) discovered an alarming disconnect between automatic evaluation and human evaluation, and (c) raised awareness of the shortcomings of today's evaluation practices and proposed guidelines to overcome them.

We first extended our prior work (Schuff, 2020) and proposed two novel proxy scores to measure how strongly an explainable QA model's predicted answer is coupled to its prediction of supporting facts which should serve as evidence for the model's answer. Concretely, we proposed the FARM(k) score as well as the LOCA score. For FARM(k), we remove k facts that the model predicted to be relevant/irrelevant from the input context and relate the number of resulting answer changes caused by removing facts that the model predicted to be relevant (should be high) to the number of resulting answer changes caused by removing facts that the model predicted to be irrelevant (should be low). For LOCA, we consider the location of answers within the input context and evaluate how frequently model answers are located within facts that the model predicted to be relevant. We evaluated the two scores against the human ratings we obtained in our prior work (Schuff, 2020) and found that our scores reflect various human ratings better than standard metrics, such as F_1-score. The underlying human ratings are available at `https://github.com/boschresearch/f1-is-not-enough`.

While we developed our scores specifically for explainable QA, the challenge of measuring prediction-explanation coupling applies to all types of explanations. In the year we published our work (Schuff et al., 2020), DeYoung et al. (2020) proposed scores to measure *comprehensiveness* and *sufficiency* to evaluate saliency explanations. Similar to the answer change fractions within our FARM score, comprehensiveness, and sufficiency measure how class probabilities change when tokens with high/low saliency are removed. While various authors (DeYoung et al., 2020; Atanasova et al., 2020) consider these and similar scores to quantify faithfulness, we argue that — along the definition of socially aligned faithfulness by Jacovi and Goldberg (2021) — faithfulness evaluation has to consider how human explainees attribute a given explanation and future work should study which proxy scores respect social alignment.

The human evaluations we conducted for our proposed explainable NLI models as well as the analysis of our proposed scores indicate that today's *de facto* standard proxy scores, such as F_1-score, accuracy or BLEU, fail to capture human-rated explanation quality. We thus explored whether, and, if yes, how strongly, automatic evaluation is detached from human evaluation within explainable NLP. We evaluated ten real-world model submissions to the official HotpotQA (Yang et al., 2018) leaderboard which ranks explainable QA models and conducted an extensive crowdsourced user study. Assessing models from a real NLP leaderboard on the publicly unavailable test set allowed us to study a uniquely representative and realistic pool of systems. Additionally, our study exceeded previous studies in the diversity of human rating dimensions including, e.g., ratings of usability, mental effort, or explanation utility. Our results confirm the disconnect between automatic evaluation and human evaluation. We additionally found that optimizing for a single proxy score (as it is usually done within NLP) can decrease the score's expressiveness over time (i.e., Goodhart's law), meaning that scores can "wear off" and ultimately undermine a leaderboard's utility.

While challenges regarding the disconnect between proxy scores and human ratings within NLG are well known and extensively researched for various tasks (Callison-Burch et al., 2006; Liu et al., 2016; Novikova et al., 2017; Sulem et al., 2018; Reiter, 2018), this issue has received — apart from important exceptions, such as the work of Clinciu et al. (2021) — little attention in explainable NLP research yet. Exploring to which extent other explainable NLP tasks are affected by this disconnect is a pressing need that future work should address in order to ensure that we are not developing systems for the sake of improving scores on leaderboards, but to provide an actual benefit to the systems' users.

A key challenge in evaluating explanation quality is its often vague and implicit definition. We thus addressed the question of what makes a "good" explanation. For this, we proposed fundamental characteristics of explanation quality including insights from the behavioral sciences. Next, we demonstrated how today's evaluation practices violate them. As a remedy, we proposed guidelines to overcome some of the main challenges that explanation quality evaluation is facing. The core of our recommendations is to conduct human evaluation in order to validate automatic evaluation, combine the strengths of automatic and human evaluation, and evaluate explanations beyond quantitatively measurable dimensions. In order to support NLP researchers that want to get started with how to design, conduct, and evaluate user studies, we published a survey paper that reviews the aspects of human evaluation that we consider to be most important in NLP (Schuff et al., 2023b). We additionally developed Pareto front leaderboards as an alternative to the single-score rankings that are predominantly used in NLP today. Pareto front leaderboards construct model ranks based on multiple dimensions without weighting or averaging the single dimensions and provide an alternative approach to defining new state-of-the-art models and multi-criteria model selection.

Future work should seek to unify and standardize explanation quality evaluation. While an increasing amount of explainable NLP publications includes human evaluation in their method assessment, the choice of collected signals and ratings, experiment design, and statistical evaluation vary strongly. Although there is no one-fits-all solution to explanation quality evaluation, establishing a shared evaluation framework will allow to compare methods across publications using standardized norms. We are convinced that a well-developed evaluation foundation will accelerate research, strengthen reproducibility, and foster acceptance of human evaluation as an integral part of (explainable) NLP research.

Human Perception and Explanations

In Chapter 5, we (a) explored how cognitive biases distort human perception of explanations and (b) developed and validated a questionnaire to measure perceived system predictability.

For our study of cognitive biases, we considered saliency explanations over text. Saliency explanations overlay text with a heatmap of, typically red, color in which light shades of red correspond to less important parts of the input while strong shades of red mark more important parts. In a series of six experiments, we demonstrated that, although a word's importance information is communicated via its color only, the information that explainees understand is influenced by unrelated superficial factors, such as the word's length or the word's capitalization. We replicated these biasing effects across languages, tasks, domains, and saliency scores (including the commonly used attribution scores Integrated Gradients and SHAP). Besides a word's influence on the perception of its explanation, we also investigated how *other* words influence a word's perception. Concretely, we analyzed how left and right neighbors influence importance perception and discovered left-right asymmetries as well as significant differences between neighboring words that share a lexical chunk versus unrelated neighboring words. We linked the observed effects to known cognitive biases and connected the observed effects between neighboring words to a psychological theory of assimilation and contrast. We further explored how the distorting effects can be mitigated and proposed three visualization alternatives. We demonstrated that bar charts can mitigate visual bias, our novel model-based saliency correction method can mitigate bias from learning effects, and choosing an appropriate color range can prevent distorting influences from neighboring words. We release the collected human ratings as well as our GAMM analysis code at `https://github.com/boschresearch/human-interpretation-saliency`.

One important direction for future work is to study whether our conclusions can be replicated using alternative methodological approaches. Concretely, we envision a replication of our study using eye-tracking and electroencephalography . Further, future work can build upon our proposed visualization alternatives and derive improved visualization and general explanation methods based on our findings. Lastly, we hypothesize that the effects we observed can also be relevant to work in psychology, in which our results might either confirm existing theories or shed light on previously unexplored effects that are of interest beyond their relevance in explainable NLP.

In the last part of this thesis, we developed and validated a scale to measure perceived system predictability. Following best practices of scale development, we designed a 6-item Likert questionnaire that allows us to measure the degree to which users feel able to predict a system's behavior. We conducted extensive experiments that demonstrate the validity of our instrument and showed that it can be used to measure perceived predictability either as a unidimensional construct or as a three-dimensional construct in which we distinguish between effective, epistemic, and aleatory predictability inspired by uncertainty theory. We (a) explored how perceived predictability is related to, i.a., user trust, (b) analyzed how subjective perceived predictability and objective accuracy of users' system prediction estimates are related, and (c) investigated how different types of explanations and levels of system stochasticity affect perceived predictability. We found that measuring perceived system predictability yields information beyond measuring objective system predictability and that objective and perceived predictability can diverge and differ in their relation to explanations and system stochasticity.

Future work should further study the psychometric properties of our scale and assess its validity within new usage contexts as well as for additional user populations. Most importantly, we consider our scale together with our fundamental research on perceived predictability to be a valuable tool for researchers within NLP, broader explainability research, and various areas of HCI. We envision applications of our scale to range on a broad spectrum from the evaluation of chatbots to user experience evaluation in automated driving.

Essential Takeaways

The key argument that this thesis made is that developing effective explainable NLP systems requires an interdisciplinary approach. We showed that addressing explainable NLP from a purely technical perspective poses the danger of having "the inmates [NLP researchers] running the asylum [explainability research]" (Miller et al., 2017), i.e., developing explainability methods that suit their developers but not their users. Instead, we argued that adding a human-centered perspective allows us to question and improve our evaluation methods, diversifies model development, and ultimately drives explainable NLP research towards explanations that matter to their users. We envision an ideal explanation method development process as one that (i) proposes a new explanation method based on a principled technical foundation as well as knowledge of human cognition, (ii) is accelerated by automatic evaluation using meaningful proxy scores, and (iii) finally verifies the proposed explanation method using standardized human evaluation.

A. Appendix

A.1. External Knowledge for NLI

A.1.1. Knowledge Requirement Annotation

Table A.1 lists the annotation guidelines used to decide on low/high levels of required external knowledge as discussed in Section 3.1.3.3. Table A.2 shows example annotations.

Low Level	Pattern Matching	The entailment can be decided by matching identical parts in the premise and the hypothesis. Premise: *A water scene with a sunset in the background.* Hypothesis:*There is a water scene with the sunset in the back.*
	Unrelated Negation	The entailment can be decided by identifying an unrelated negation. Premise: *Children bathe in water from large drums.* Hypothesis: *The kids are not reading.*
	Rephrasing	The entailment can be decided by simple rephrasing (e.g. replacing a word with a synonym). Premise: *A boy dressed in an orange shirt and a helmet is riding a dirt bike in the woods.* Hypothesis: *A boy in orange rides his dirt bike.*
	Easily-Distinguishable Concepts	The entailment can be decided by identifying unrelated concepts that have no semantic relation. Premise: *Firefighters in full gear are walking up a ladder.* Hypothesis: *The firefighters are eating lunch.*
High Level	Complex Reasoning	The entailment can be decided by resolving more complex relations and reasoning using common sense knowledge. Premise: *Soccer players are playing a night game and the ball is in the air, while the two teams fight for it.* Hypothesis: *The sun was shining during the soccer match.*
	Abstract Concepts	The entailment can be decided using common sense knowledge about abstractions of concepts. Premise: *A girl reaches up to kiss a cat, which is sitting on the counter.* Hypothesis: *A girl is showing affection towards a cat.*

Table A.1.: Annotation guidelines used during the annotation of low/high levels of required external knowledge with examples.

Low Level	Premise: There is a group of children getting their picture taken with presents. Hypothesis: Two men carry a Christmas tree.
	Premise: A woman looks at a plate filled with steam. Hypothesis: The woman is out shopping at the mall.
	Premise: Man sitting on bench with a suitcase in front of PADDINGTON sign. Hypothesis: A man sitting with a sign.
	Premise: A man grilling a hamburger. Hypothesis: The man is swimming at the bottom of the ocean.
	Premise: The African American man protests against unlawful sex. Hypothesis: The man protests.
High Level	Premise: A boy in a red jacket and black hat sliding on his knees down a snowy hill Hypothesis: A child is playing outside.
	Premise: A man playing a piano. Hypothesis: The man's hands are on the keys of a piano.
	Premise: 3 girls chatting and laughing on the stairwell. Hypothesis: Girls are not having a good time.
	Premise: A man visiting a friend in the hospital. Hypothesis: A man and a patient in a hospital room.
	Premise: Two girls pose along a tree-lined path and blow kisses towards the camera. Hypothesis: Two girls are taking pictures outside.

Table A.2.: Pairs from the low/high external knowledge requirement annotations sampled from pairs for which annotators agreed.

A.2. Thought Flow Nets

A.2.1. Question Answering Experiments

A.2.1.1. Dataset Details.

We use the HOTPOTQA dataset (Yang et al., 2018), which is an English multi-hop QA data set. It covers 90,564 training instances, 7,405 test validation instances, and 7,405 test instances per setting (there are a distractor and a fullwiki setting). Training instances are grouped by difficulty and cover 18,089 easy, 56,814 medium, and 15,661 hard questions. We refer to (Yang et al., 2018) for more details.

A.2.2. User Study

A.2.2.1. Questionnaire Items

Per-System Questionnaires. We collected the following ratings per-system, i.e., after interaction with all instances.

Usability: The UMUX usability scale (Finstad, 2010, 2013) uses the following four 5-point Likert items:

- This system's capabilities meet my requirements.
- Using this system is a frustrating experience.
- This system is easy to use.
- I have to spend too much time correcting things with this system.

Mental Effort: The Pass mental effort scale (Paas, 1992) uses a single 9-point Likert item:

- Please rate the mental effort required to decide if the system's answer is correct. The nine points are labeled from "very, very low mental effort" to "very, very high mental effort".

Anthropomorphism: The Godspeed anthropomorphism subscale (Bartneck et al., 2009) uses five 5-point semantic differential scales that ask the user to rate the system in a spectrum of:

- fake – natural
- machinelike – humanlike
- unconscious – conscious
- artificial – lifelike
- (moving rigidly – moving elegantly) (We exclude this item as it is not applicable to question answering systems.)

Perceived Intelligence: The Godspeed perceived Intelligence subscale (Bartneck et al., 2009) uses five 5-point semantic differential scales that ask the user to rate the system in a spectrum of:

- incompetent – competent
- ignorant – knowledgeable
- irresponsible – responsible
- unintelligent – intelligent
- foolish – sensible

Per-Item Questionnaires. In addition to the per-system ratings, we also collect ratings on a fine-grained, per-instance level.

Perceived Answer Correctness: We use a single binary item to collect perceived answer correctness ratings:

- I think the system's answer is correct.

Perceived Helpfulness: We use a single 5-point Likert item to collect helpfulness ratings:

- I think the system's answer enables me to give the correct answer.

Perceived Understanding: We use a single 5-point Likert item to collect understanding ratings:

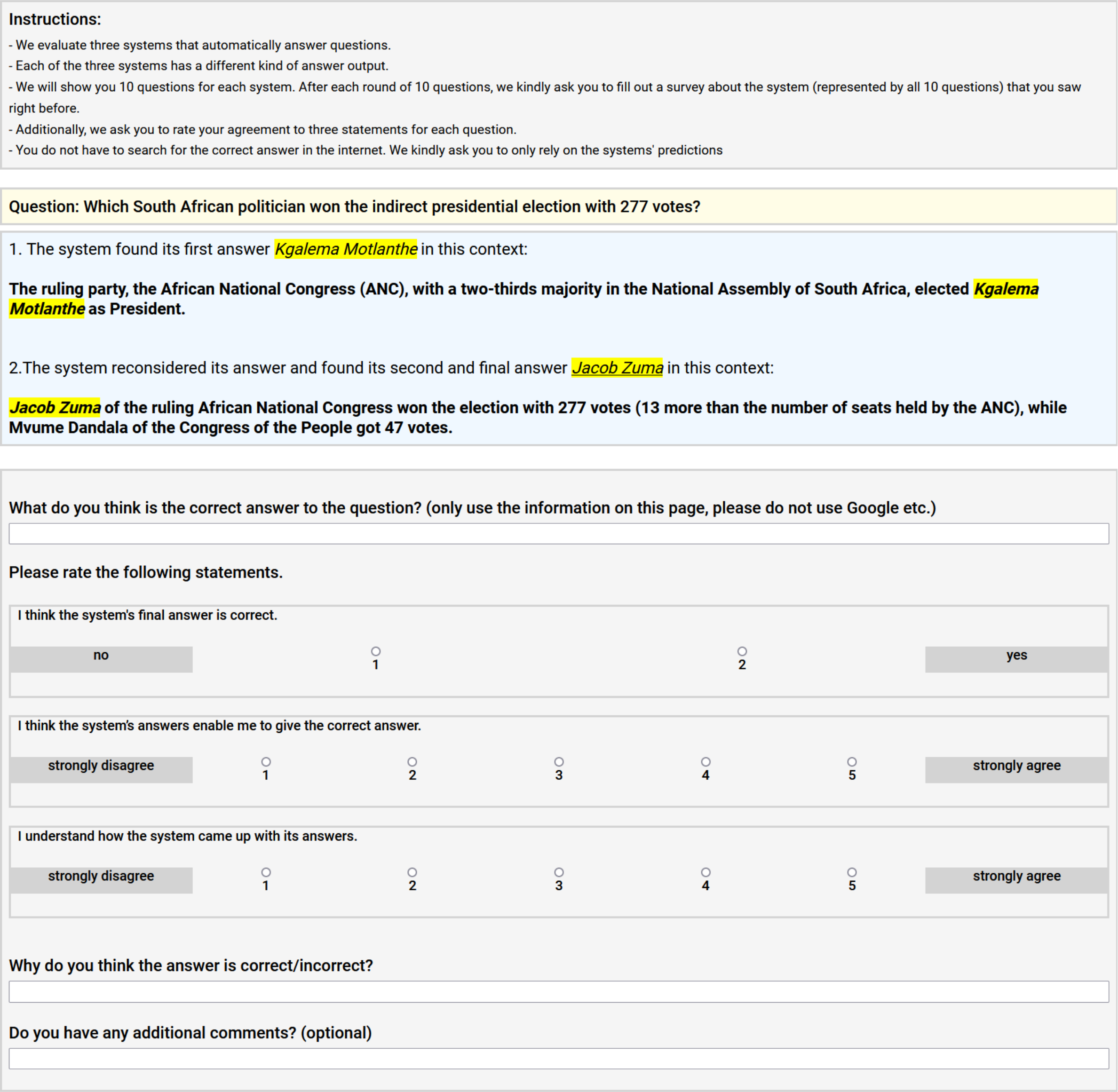
- I understand how the system came up with its answer.

A.2.2.2. Interface

Figures A.1 to A.3 show screenshots of our experiment interface for the three studied prediction conditions TF, TOP-3 and SINGLE. Figure A.4 depicts an attention check question.

A.3. Novel Proxy Scores

Figures A.5 and A.6 compare model scores and human measures grouped into F_1-scores and our proposed FARM(4) and LOCA scores. Rows alternate between F_1-scores and our scores.



Instructions:

- We evaluate three systems that automatically answer questions.
- Each of the three systems has a different kind of answer output.
- We will show you 10 questions for each system. After each round of 10 questions, we kindly ask you to fill out a survey about the system (represented by all 10 questions) that you saw right before.
- Additionally, we ask you to rate your agreement to three statements for each question.
- You do not have to search for the correct answer in the internet. We kindly ask you to only rely on the systems' predictions

Question: Which South African politician won the indirect presidential election with 277 votes?

1. The system found its first answer *Kgalema Motlanthe* in this context:

The ruling party, the African National Congress (ANC), with a two-thirds majority in the National Assembly of South Africa, elected *Kgalema Motlanthe* as President.

2.The system reconsidered its answer and found its second and final answer *Jacob Zuma* in this context:

***Jacob Zuma* of the ruling African National Congress won the election with 277 votes (13 more than the number of seats held by the ANC), while Mvume Dandala of the Congress of the People got 47 votes.**

What do you think is the correct answer to the question? (only use the information on this page, please do not use Google etc.)

Please rate the following statements.

I think the system's final answer is correct.

no	○ 1	○ 2	yes

I think the system's answers enable me to give the correct answer.

strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	strongly agree

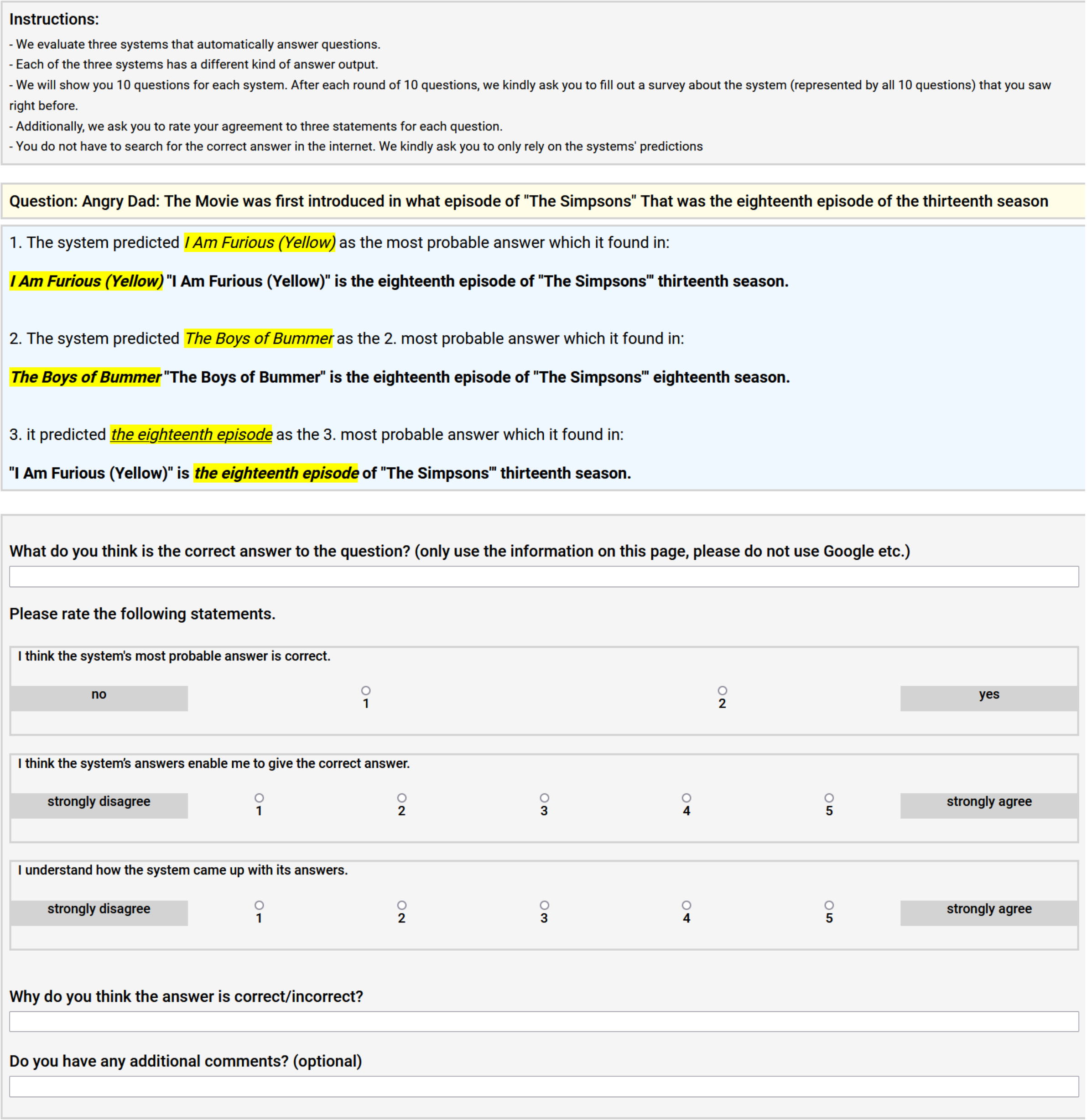
I understand how the system came up with its answers.

strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	strongly agree

Why do you think the answer is correct/incorrect?

Do you have any additional comments? (optional)

Figure A.1.: User study interface showing the TF condition (ours).



Instructions:

- We evaluate three systems that automatically answer questions.
- Each of the three systems has a different kind of answer output.
- We will show you 10 questions for each system. After each round of 10 questions, we kindly ask you to fill out a survey about the system (represented by all 10 questions) that you saw right before.
- Additionally, we ask you to rate your agreement to three statements for each question.
- You do not have to search for the correct answer in the internet. We kindly ask you to only rely on the systems' predictions

Question: Angry Dad: The Movie was first introduced in what episode of "The Simpsons" That was the eighteenth episode of the thirteenth season

1. The system predicted *I Am Furious (Yellow)* as the most probable answer which it found in:

I Am Furious (Yellow) **"I Am Furious (Yellow)" is the eighteenth episode of "The Simpsons"' thirteenth season.**

2. The system predicted *The Boys of Bummer* as the 2. most probable answer which it found in:

The Boys of Bummer **"The Boys of Bummer" is the eighteenth episode of "The Simpsons"' eighteenth season.**

3. it predicted *the eighteenth episode* as the 3. most probable answer which it found in:

"I Am Furious (Yellow)" is *the eighteenth episode* of "The Simpsons"' thirteenth season.

What do you think is the correct answer to the question? (only use the information on this page, please do not use Google etc.)

Please rate the following statements.

I think the system's most probable answer is correct.

no ○ 1 ○ 2 **yes**

I think the system's answers enable me to give the correct answer.

strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 **strongly agree**

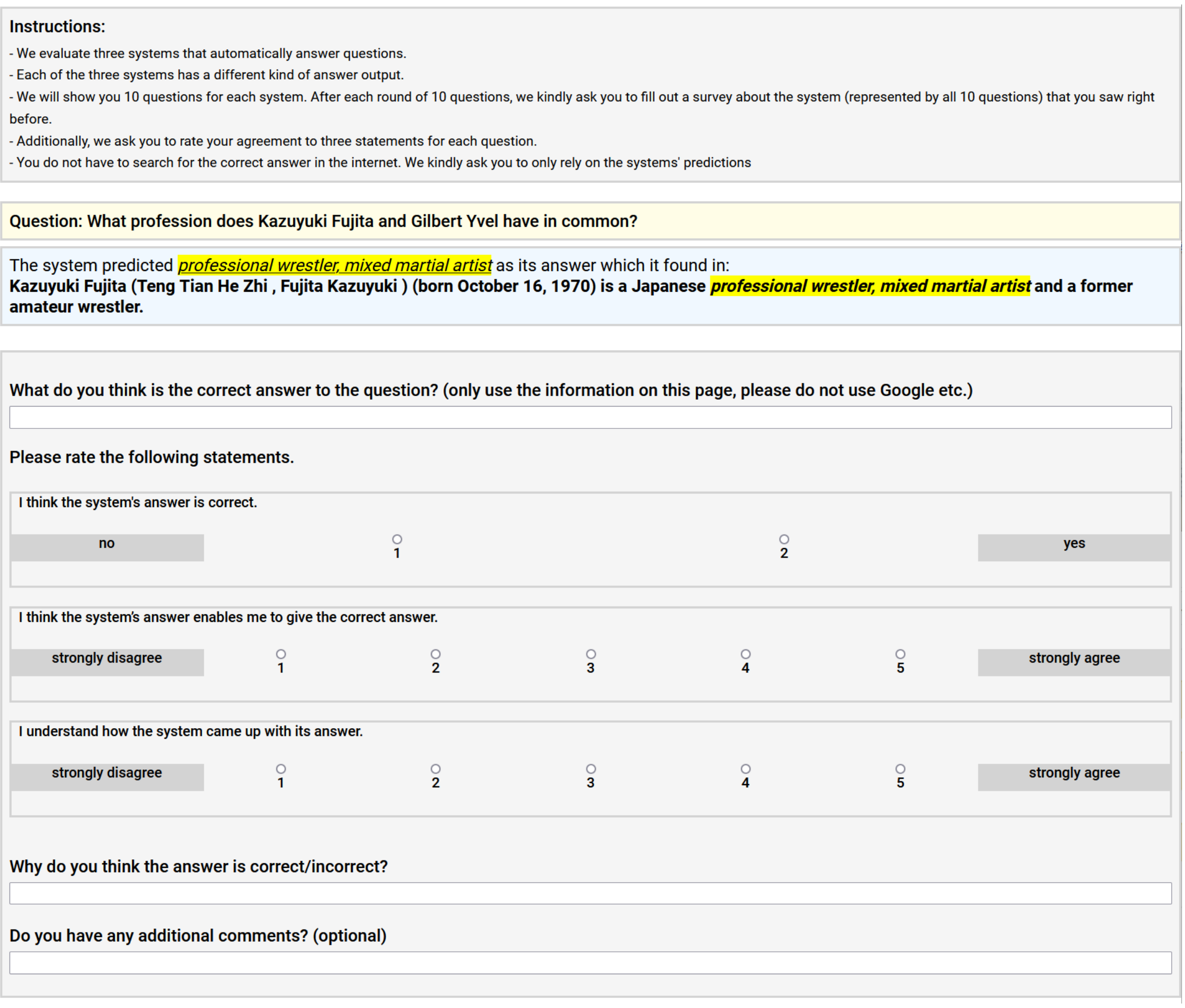
I understand how the system came up with its answers.

strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 **strongly agree**

Why do you think the answer is correct/incorrect?

Do you have any additional comments? (optional)

Figure A.2.: User study interface showing the TOP-3 condition.



Instructions:

- We evaluate three systems that automatically answer questions.
- Each of the three systems has a different kind of answer output.
- We will show you 10 questions for each system. After each round of 10 questions, we kindly ask you to fill out a survey about the system (represented by all 10 questions) that you saw right before.
- Additionally, we ask you to rate your agreement to three statements for each question.
- You do not have to search for the correct answer in the internet. We kindly ask you to only rely on the systems' predictions

Question: What profession does Kazuyuki Fujita and Gilbert Yvel have in common?

The system predicted *professional wrestler, mixed martial artist* as its answer which it found in:
Kazuyuki Fujita (Teng Tian He Zhi , Fujita Kazuyuki) (born October 16, 1970) is a Japanese *professional wrestler, mixed martial artist* and a former amateur wrestler.

What do you think is the correct answer to the question? (only use the information on this page, please do not use Google etc.)

Please rate the following statements.

I think the system's answer is correct.

no ○ 1 ○ 2 yes

I think the system's answer enables me to give the correct answer.

strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 strongly agree

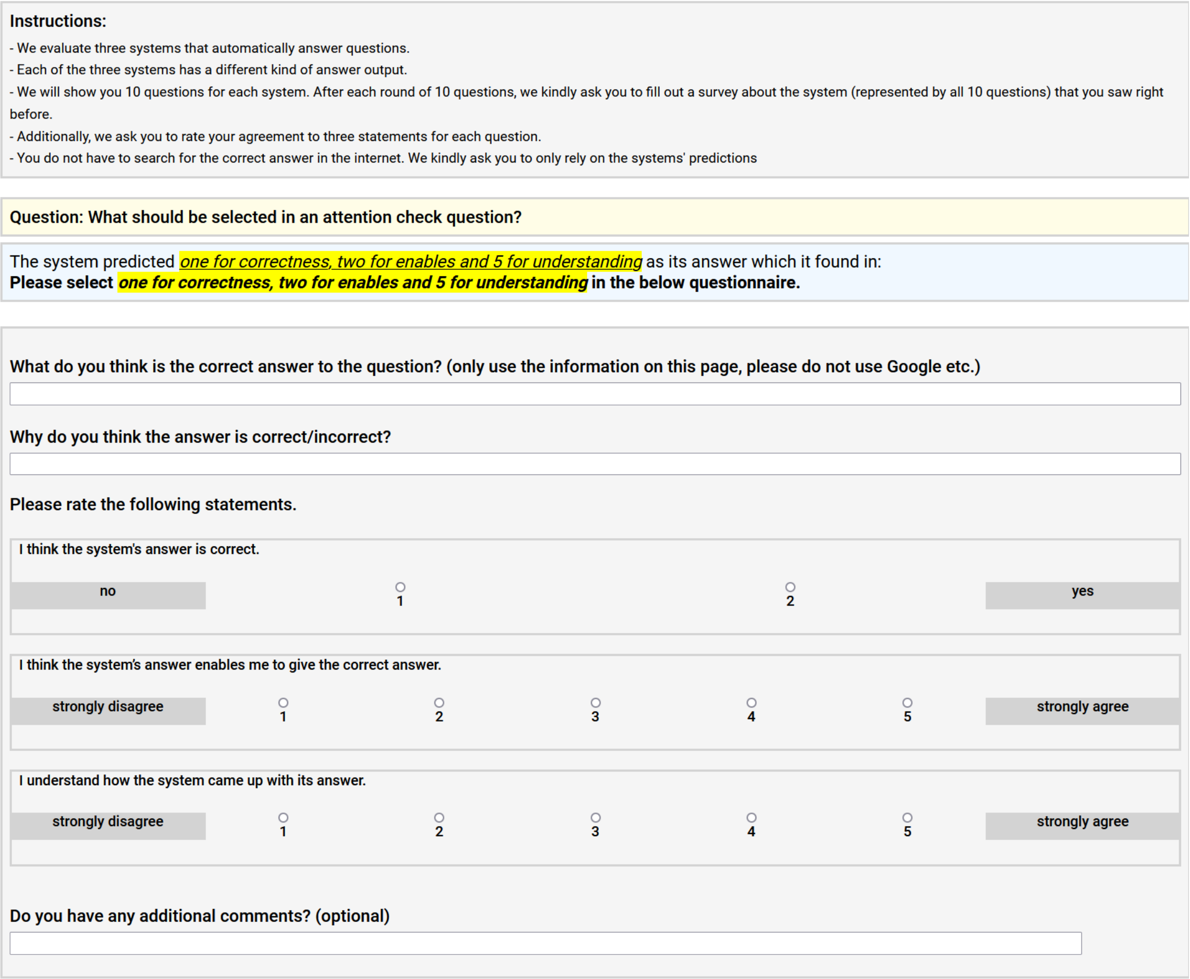
I understand how the system came up with its answer.

strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 strongly agree

Why do you think the answer is correct/incorrect?

Do you have any additional comments? (optional)

Figure A.3.: User study interface showing the SINGLE condition.



Instructions:

- We evaluate three systems that automatically answer questions.
- Each of the three systems has a different kind of answer output.
- We will show you 10 questions for each system. After each round of 10 questions, we kindly ask you to fill out a survey about the system (represented by all 10 questions) that you saw right before.
- Additionally, we ask you to rate your agreement to three statements for each question.
- You do not have to search for the correct answer in the internet. We kindly ask you to only rely on the systems' predictions

Question: What should be selected in an attention check question?

The system predicted *one for correctness, two for enables and 5 for understanding* as its answer which it found in:
Please select *one for correctness, two for enables and 5 for understanding* in the below questionnaire.

What do you think is the correct answer to the question? (only use the information on this page, please do not use Google etc.)

Why do you think the answer is correct/incorrect?

Please rate the following statements.

I think the system's answer is correct.

no	○ 1	○ 2	yes

I think the system's answer enables me to give the correct answer.

strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	strongly agree

I understand how the system came up with its answer.

strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	strongly agree

Do you have any additional comments? (optional)

Figure A.4.: User study interface showing an attention check.

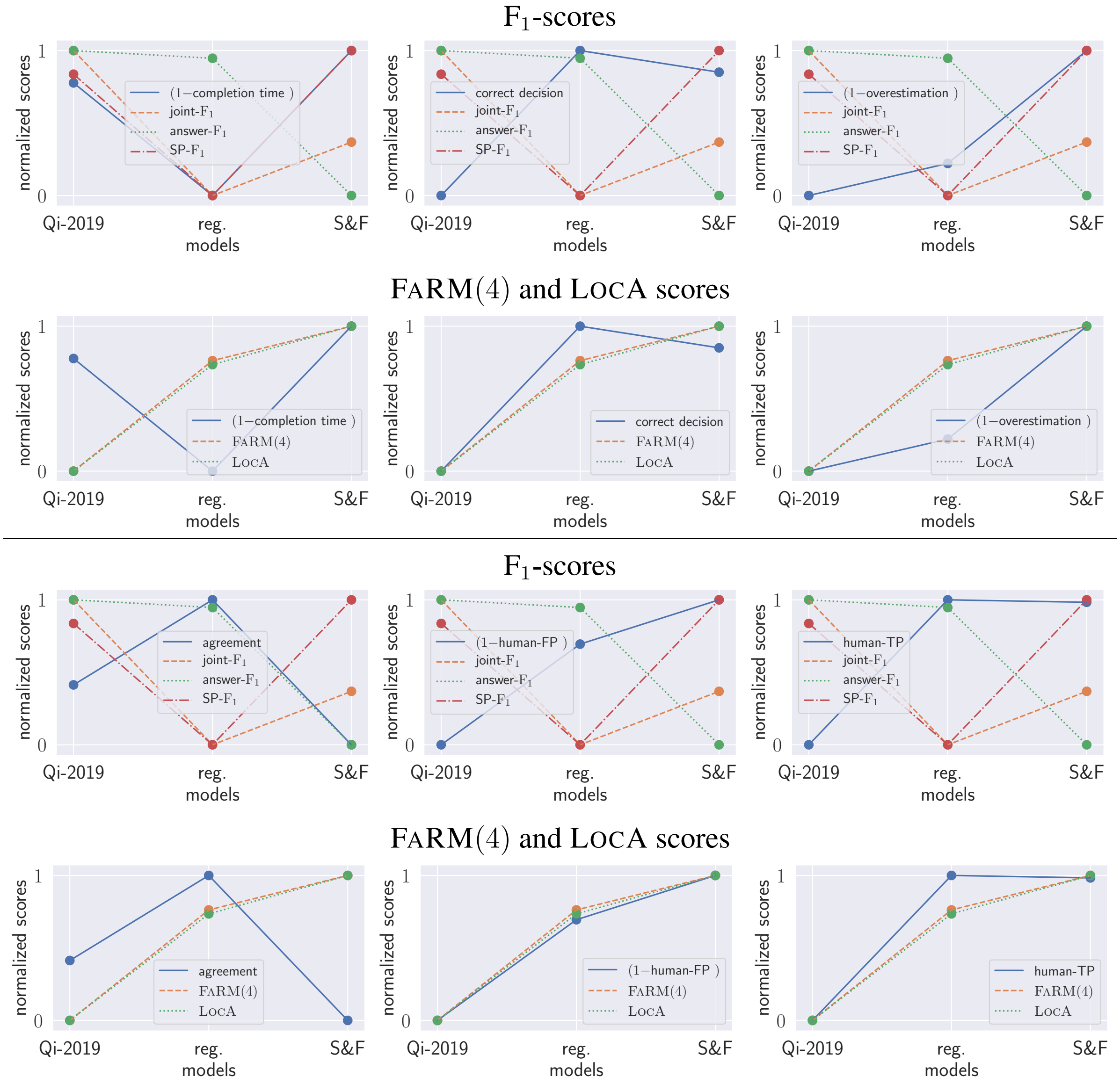


Figure A.5.: Comparisons between human measures and model scores. All scores are normalized before plotting by subtracting the minimum score and re-scaling the score span to $[0, 1]$. Human measures for which lower values correspond to better performance are plotted as $(1-\text{score})$ for the convenience of the reader. The figure shows scores for completion time, fraction of correct user decisions, overestimation, agreement, false positives, and true positives.

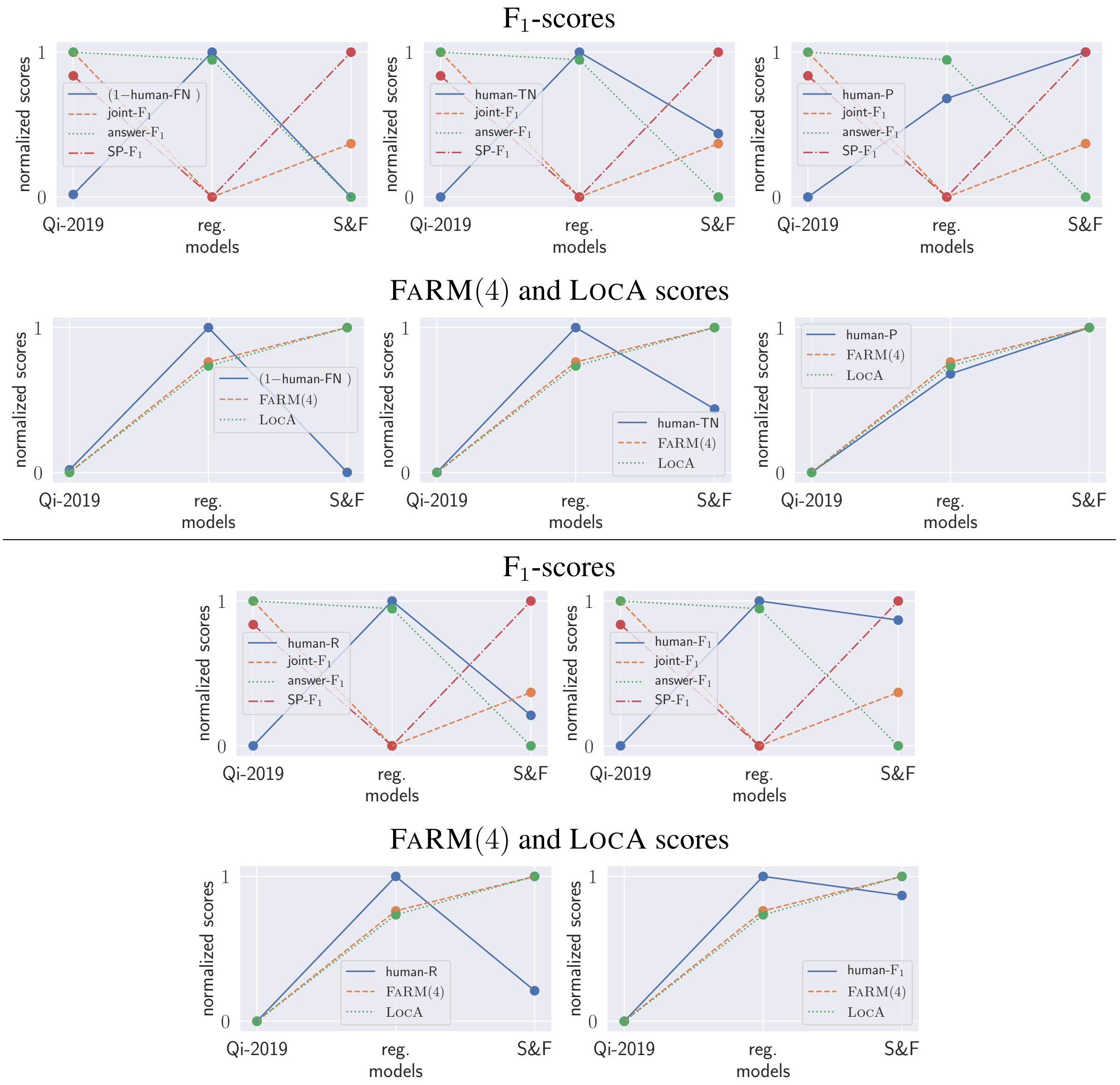


Figure A.6.: Comparisons between human measures and model scores. All scores are normalized before plotting by subtracting the minimum score and re-scaling the score span to $[0, 1]$. Human measures for which lower values correspond to better performance are plotted as $(1-\text{score})$. The figure shows scores for false negatives, true negatives, precision, recall, and user F$_1$-score.

A.4. HotpotQA Case Study

A.4.1. Detailed Proxy Scores

We provide values of all analyzed proxy scores across models in Table A.3 and Kendall's τ correlation coefficients between automatic scores in Figure A.9.

A.4.2. User Study Details

We show screenshots of our study interface in Section A.4.2.1 and report human ratings along all measured rating dimensions in Section A.4.2.2.

A.4.2.1. User Study Interface

We provide screenshots of the user study interface that we showed to participants. Figure A.7 displays the rating interface we showed *for each question*. Figure A.8 displays the post hoc questionnaire we asked participants to fill out at the end of the study.

A.4.2.2. Detailed Human Ratings

Table A.4 displays the human ratings and completion times we obtained within the user study for the ten leaderboard systems as well as our five synthetic systems.

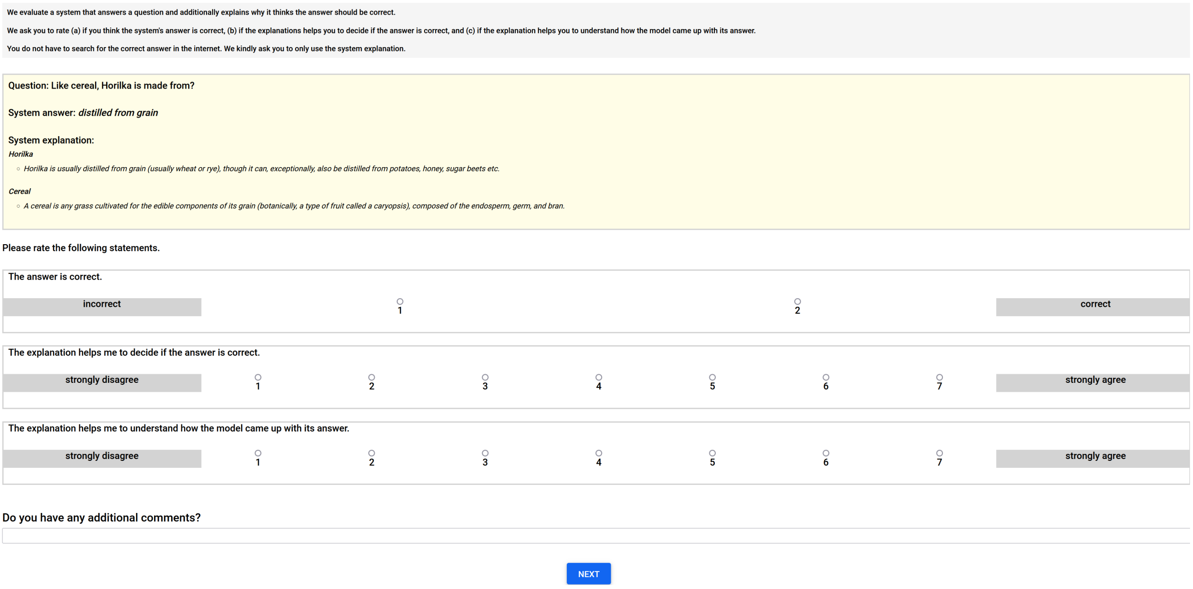
A.4.2.3. Proxy Scores and Human Ratings

Figure A.11 displays the Kendall's τ correlations between proxy scores and human ratings. We additionally provide Bonferroni-corrected significance levels. We further evaluate (i) grouped weighted κ inter annotator agreements (IAAs) Cohen (1968) as an appropriate IAA measure for ordinal responses and (ii) standard deviations to provide an additional perspective on the ratings' variances. We observe $\kappa = 0.42$ / SD= 0.43 for correctness, $\kappa = 0.3$ / SD= 1.88 for utility and $\kappa = 0.33$ / SD= 2.13 for consistency. These IAAs and standard deviations signal a low agreement / high variability which is commonly interpreted to correspond to low-quality *annotations*.[1] However, we want to emphasize that the purpose of our study is not (and should not be) to collect clean *annotations* of specific explanation instances but instead to capture the relation between automatic scores and intentionally and potentially noisy *subjective human ratings* as these are the exact ratings that constitute human assessment of explanation quality.

[1] We note that this interpretation can be challenged and low IAAs are not necessary to collect highly reliable data (Beigman Klebanov and Beigman, 2009).

	joint-F_1	F_1	SP-F_1	LOCA	# words	# facts	# excess facts
∗ *gold*	99.99	99.99	100.00	1.00	58.43	2.43	0.00
FE2H on ALBERT	76.54	84.44	89.14	0.98	56.05	2.30	-0.13
AMGN	74.20	82.80	88.12	0.95	54.05	2.22	-0.22
Longformer	73.17	81.26	88.34	0.72	56.50	2.33	-0.10
S2G-large	72.26	80.24	87.61	0.12	55.90	2.31	-0.13
HGN	71.04	79.37	87.33	0.97	57.36	2.36	-0.07
Text-CAN	65.96	73.99	85.76	0.92	56.65	2.33	-0.10
SAE	62.92	72.77	82.82	0.86	57.50	2.38	-0.05
IRC	59.22	72.53	79.36	0.77	70.34	2.94	0.51
GRN	58.48	66.72	84.11	0.89	57.35	2.37	-0.05
∗ *gold-answers-all-facts*	11.79	99.99	11.80	1.00	923.96	41.26	38.83
∗ *random-answers-gold-facts*	1.93	1.93	100.00	0.12	58.43	2.43	0.00
∗ *random-answers-random-facts*	0.00	1.89	0.00	0.11	55.95	2.43	-0.01
∗ *gold-answers-random-facts*	0.00	99.99	0.00	0.03	55.83	2.43	-0.01

Table A.3.: Extended HotpotQA leaderboard including synthetic systems derived from the gold test set (marked with "∗" and *italics*). DecompRC only reports answer metrics.



We evaluate a system that answers a question and additionally explains why it thinks the answer should be correct.

We ask you to rate (a) if you think the system's answer is correct, (b) if the explanations helps you to decide if the answer is correct, and (c) if the explanation helps you to understand how the model came up with its answer.

You do not have to search for the correct answer in the internet. We kindly ask you to only use the system explanation.

Question: Like cereal, Horilka is made from?

System answer: *distilled from grain*

System explanation:

Horilka

- *Horilka is usually distilled from grain (usually wheat or rye), though it can, exceptionally, also be distilled from potatoes, honey, sugar beets etc.*

Cereal

- *A cereal is any grass cultivated for the edible components of its grain (botanically, a type of fruit called a caryopsis), composed of the endosperm, germ, and bran.*

Please rate the following statements.

The answer is correct.

incorrect ○ 1 ○ 2 correct

The explanation helps me to decide if the answer is correct.

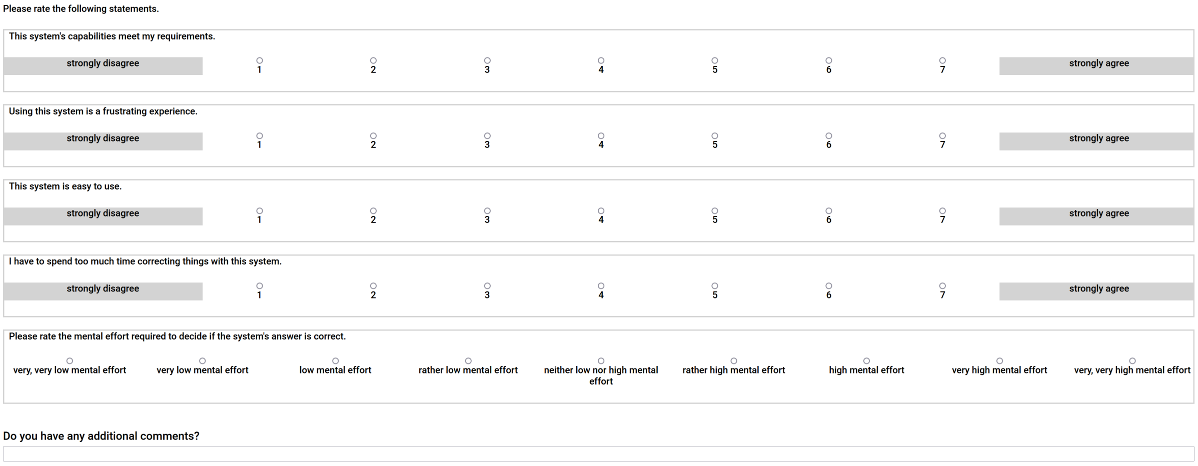
strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 strongly agree

The explanation helps me to understand how the model came up with its answer.

strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 strongly agree

Do you have any additional comments?

Figure A.7.: MTurk interface to rate a system prediction.



Please rate the following statements.

This system's capabilities meet my requirements.

strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	strongly agree

Using this system is a frustrating experience.

strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	strongly agree

This system is easy to use.

strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	strongly agree

I have to spend too much time correcting things with this system.

strongly disagree	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	strongly agree

Please rate the mental effort required to decide if the system's answer is correct.

○ very, very low mental effort	○ very low mental effort	○ low mental effort	○ rather low mental effort	○ neither low nor high mental effort	○ rather high mental effort	○ high mental effort	○ very high mental effort	○ very, very high mental effort

Do you have any additional comments?

Figure A.8.: Post questionnaire of the MTurk interface.

	Usability (UMUX)	Consistency	Utility	Answer Correctness	Mental Effort	Completion Time (seconds)
AMGN	86.7	5.8	5.6	0.9	5.8	80.2
DecompRC	78.3	5.0	4.8	0.9	5.8	43.2
FE2H on ALBERT	**97.5**	**6.3**	6.2	1.9	**4.0**	81.8
∗ *gold*	83.3	6.1	6.2	**2.0**	5.6	**41.4**
∗ *gold-answers-all-facts*	85.8	5.0	5.6	1.8	5.8	75.4
∗ *gold-answers-random-facts*	15.8	2.3	2.4	1.7	7.8	43.8
GRN	68.3	5.4	5.8	1.7	4.8	75.1
HGN	90.0	**6.3**	**6.3**	1.9	4.2	64.4
IRC	83.3	6.0	**6.3**	1.8	5.8	118.0
Longformer	86.7	5.9	**6.3**	1.9	5.0	42.0
∗ *random-answers-gold-facts*	20.8	2.1	5.4	1.0	4.6	44.4
∗ *random-answers-random-facts*	23.3	2.4	2.9	1.0	5.2	48.7
S2G-large	88.3	6.1	6.1	1.8	**4.0**	50.9
SAE	86.7	5.9	**6.3**	1.8	4.2	86.6
Text-CAN	86.7	6.0	**6.3**	1.9	4.6	94.2

Table A.4.: Human ratings of the systems we assessed within our human evaluation (synthetic systems are marked with "$*$" and *italics*). Best values are marked **bold**. Answer correctness ratings are scaled to $[0, 2]$ to allow a finer-grained differentiation between systems.

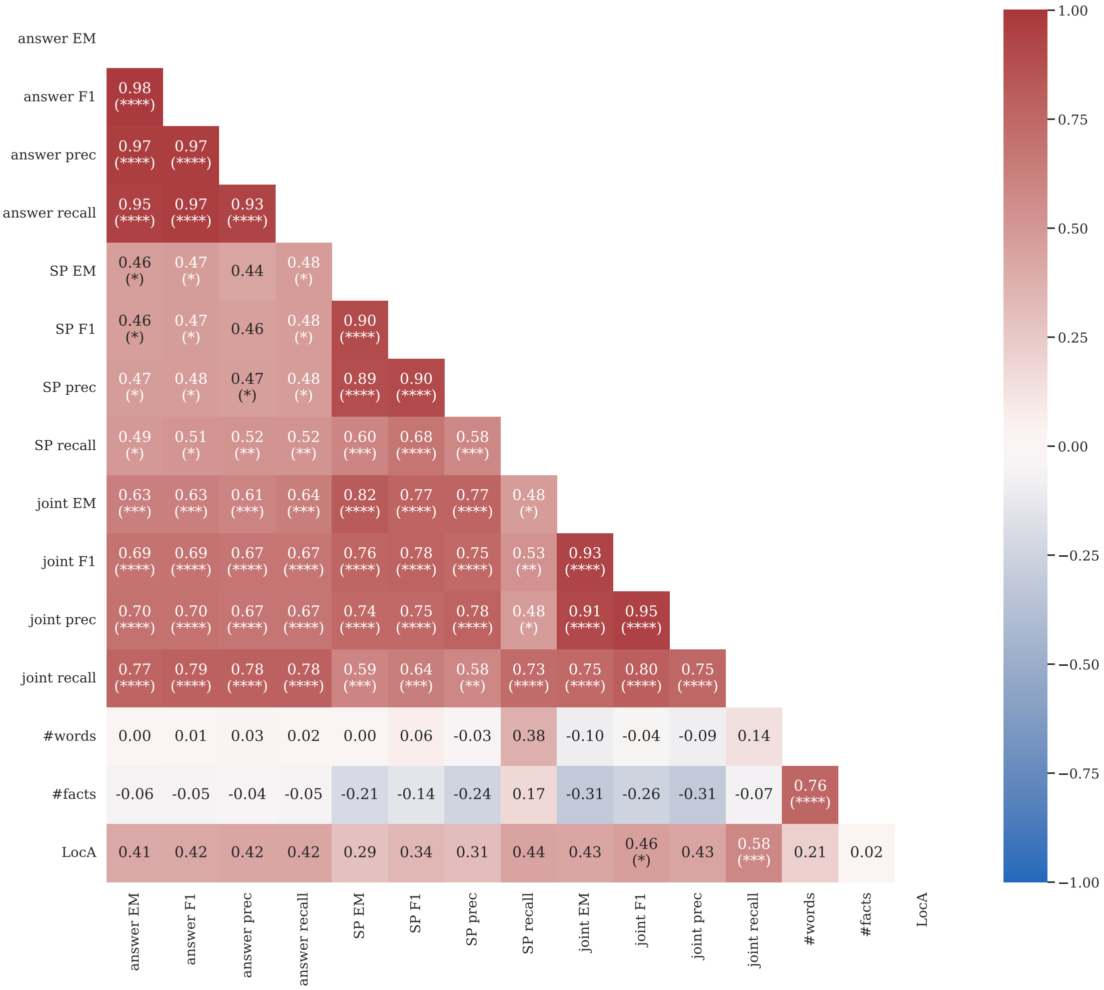


Figure A.9.: Kendall's τ correlation coefficients between automatic scores to quantify model behavior related to explanation quality on the HotpotQA dataset. Significance levels are corrected using Bonferroni correction. ($*$: $p \leq 0.05$, $**$: $p \leq 0.01$, $***$: $p \leq 0.001$ and $****$: $p \leq 0.0001$).

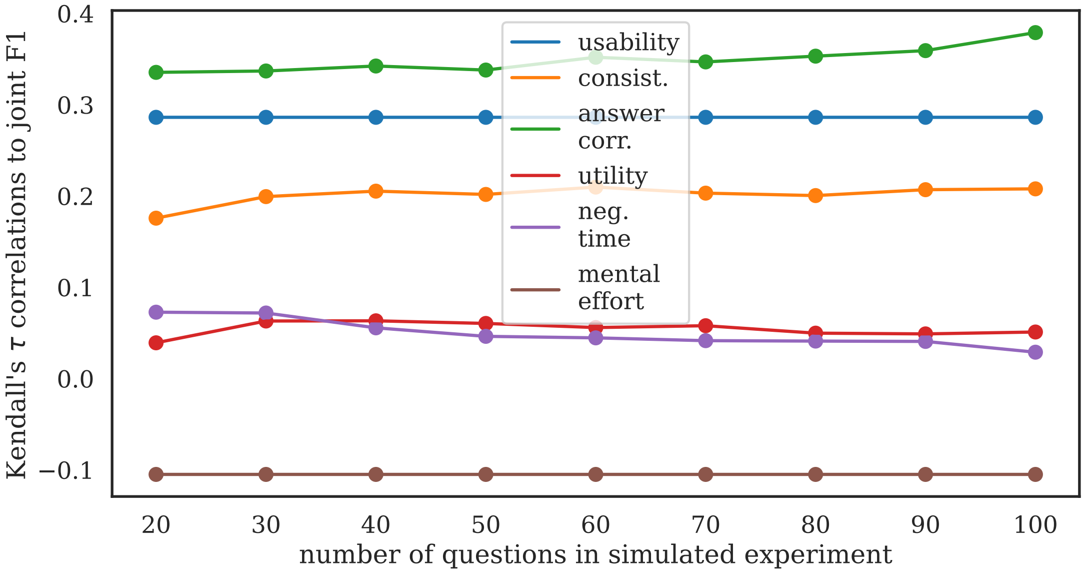


Figure A.10.: Kendall’s τ correlation coefficients between human ratings and joint-F1.

A.4.2.4. Question Pool Size Simulations.

In order to support our assumption that our pool of 100 questions is sufficiently representative, we simulate experiments with various question subsets. Figure A.10 shows that correlations already stabilize for 20 questions and that there are no qualitative or quantitative differences to using 100 (all τ differences<=0.04).

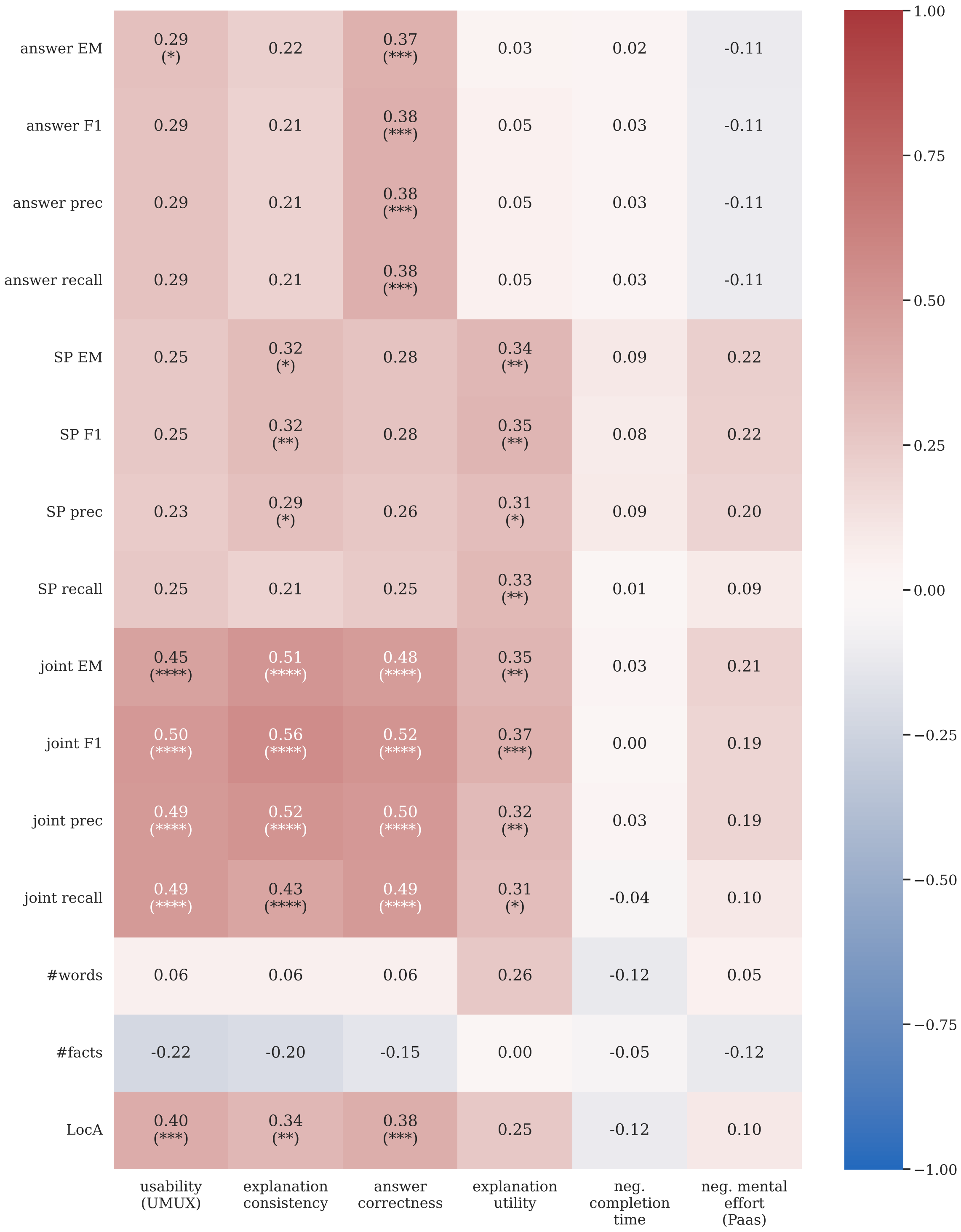
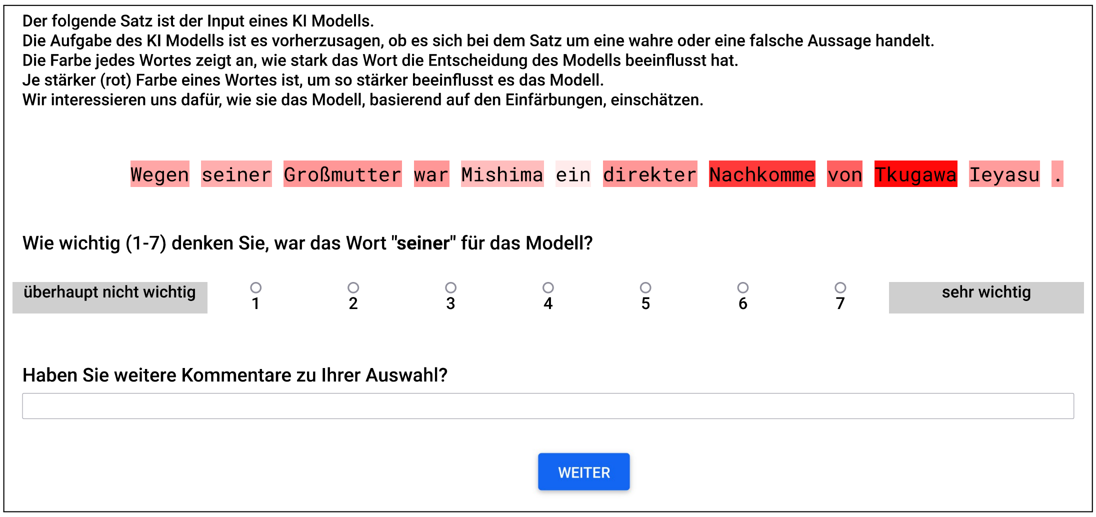


Figure A.11.: Kendall's τ correlations (per HIT). Significance levels are corrected using Bonferroni correction. ($*$: $p \leq 0.05$, $**$: $p \leq 0.01$, $***$: $p \leq 0.001$ and $****$: $p \leq 0.0001$).

A.5. Human Interpretation

A.5.1. Study Interfaces

In addition to the screenshot shown in Figure 5.2, Figure A.12 shows the interface of the German study and Figure A.13 shows an interface that uses the alternative bar chart visualization. Figure A.14 displays one of the three trap questions we use to detect participants that do not pay attention to the task.



Der folgende Satz ist der Input eines KI Modells.
Die Aufgabe des KI Modells ist es vorherzusagen, ob es sich bei dem Satz um eine wahre oder eine falsche Aussage handelt.
Die Farbe jedes Wortes zeigt an, wie stark das Wort die Entscheidung des Modells beeinflusst hat.
Je stärker (rot) Farbe eines Wortes ist, um so stärker beeinflusst es das Modell.
Wir interessieren uns dafür, wie sie das Modell, basierend auf den Einfärbungen, einschätzen.

`Wegen seiner Großmutter war Mishima ein direkter Nachkomme von Tkugawa Ieyasu .`

Wie wichtig (1-7) denken Sie, war das Wort "**seiner**" für das Modell?

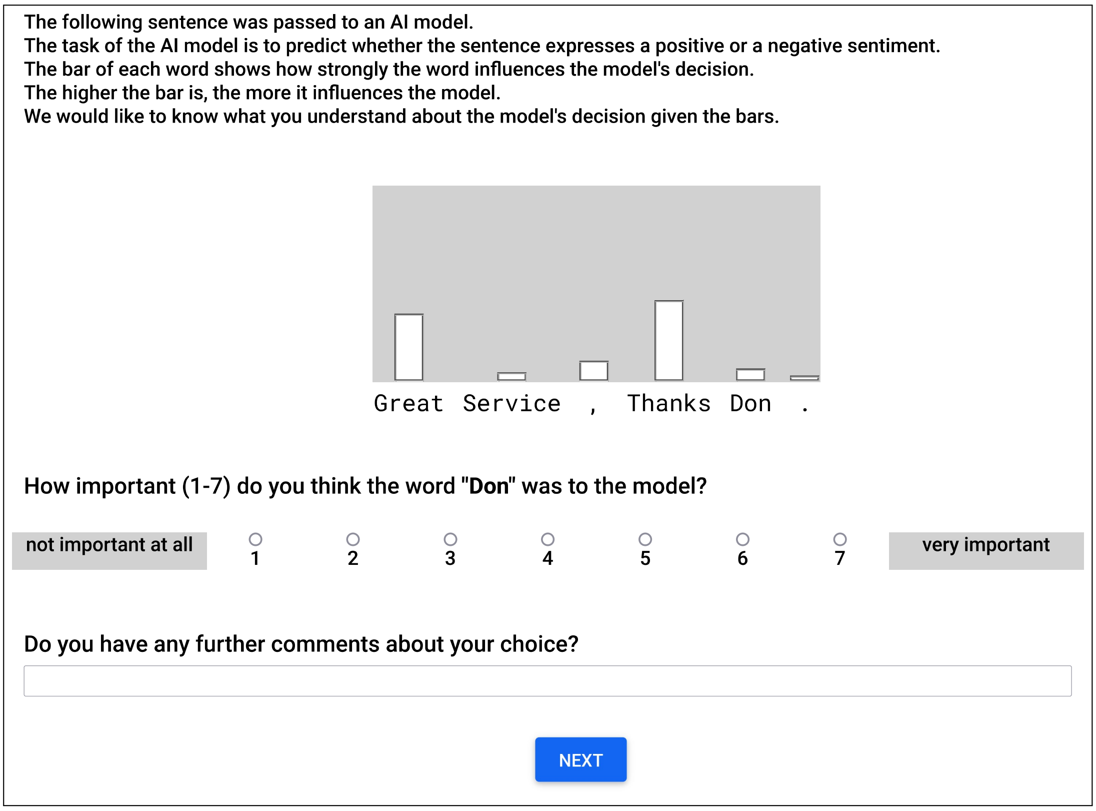
überhaupt nicht wichtig ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 sehr wichtig

Haben Sie weitere Kommentare zu Ihrer Auswahl?

Figure A.12.: Screenshot of the importance rating interface for German fact checking sentences using saliency visualization.

A.5.2. English Study Details

Table A.5 displays test statistics for all smooth pairwise interactions. We make use of tensor interaction smooths following a functional ANOVA decomposition. Figure A.15 shows summed-effect plots for the respective significant interactions. Ordered categorical cut points are located at -1, 1.31, 3.29, 5.15, 7.1, and 9.22.



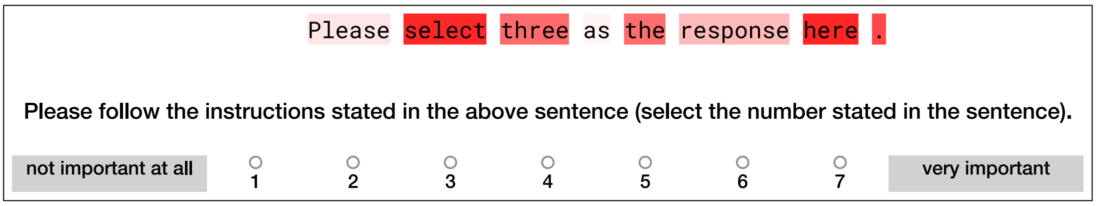
The following sentence was passed to an AI model.
The task of the AI model is to predict whether the sentence expresses a positive or a negative sentiment.
The bar of each word shows how strongly the word influences the model's decision.
The higher the bar is, the more it influences the model.
We would like to know what you understand about the model's decision given the bars.

How important (1-7) do you think the word "**Don**" was to the model?

not important at all ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 very important

Do you have any further comments about your choice?

Figure A.13.: Screenshot of the importance rating interface for English sentiment sentences using bar visualization.



Please select three as the response here .

Please follow the instructions stated in the above sentence (select the number stated in the sentence).

not important at all ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 very important

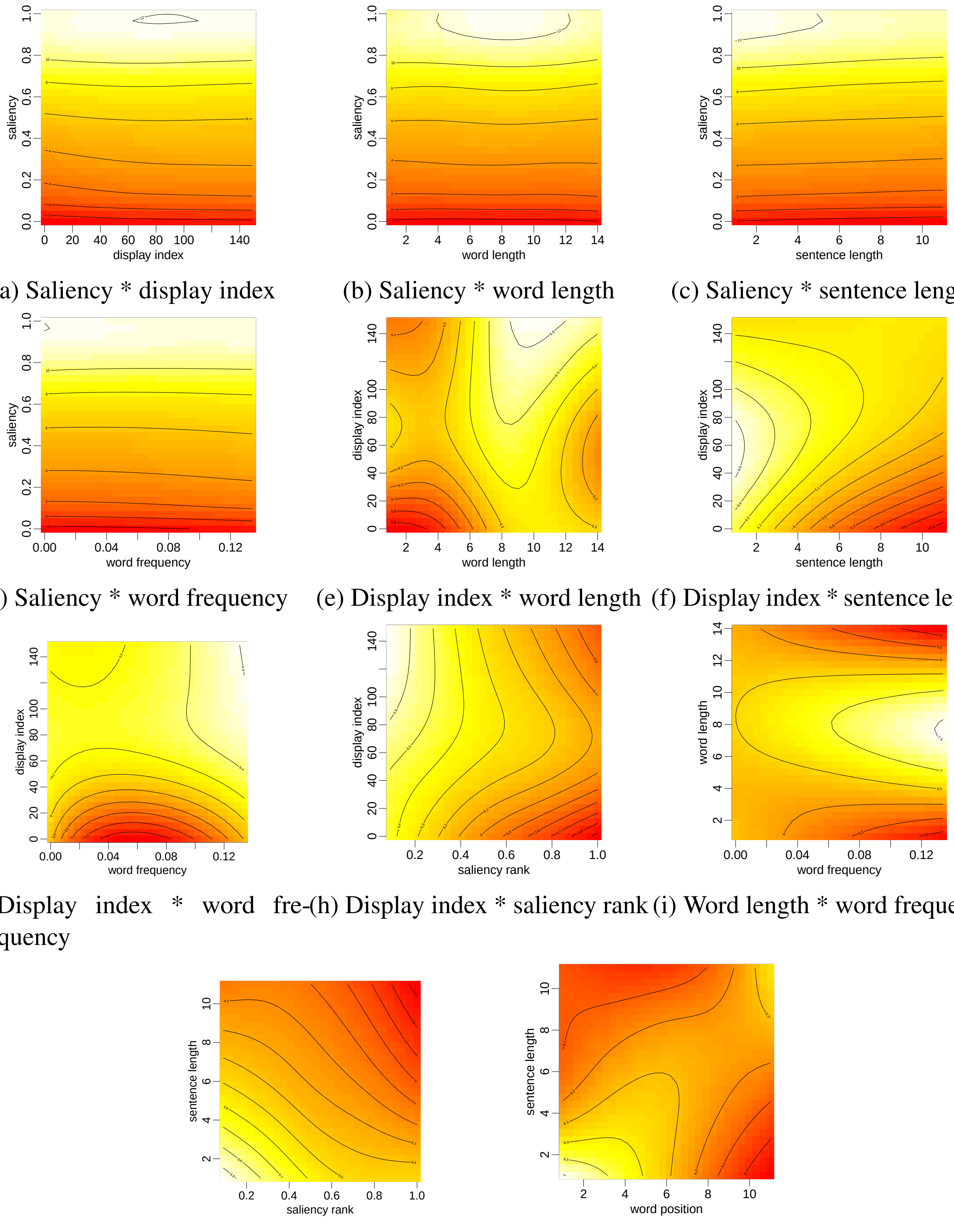
Figure A.14.: Screenshot of one of three trap sentences used to validate that the participant pays attention to the task.

	edf	ref. df	F	p
ti(saliency,display index)	2.5102	16	2.1075	0.0001
ti(saliency,word length)	6.0566	16	2.2698	0.0001
ti(saliency,sentence length)	3.1609	16	1.1203	0.0020
ti(saliency,word frequency)	0.9176	12	1.8325	0.0004
ti(saliency,sentiment polarity)	2.9357	16	0.5553	0.0814
ti(saliency,saliency rank)	0.0004	16	0.0000	0.5864
ti(saliency,word position)	0.6254	16	0.1144	0.1276
ti(display index,word length)	1.5112	16	0.6637	0.0026
ti(display index,sentence length)	1.2776	16	1.0159	0.0010
ti(display index,word frequency)	2.6938	16	1.7810	0.0001
ti(display index,sentiment polarity)	0.5386	16	0.0853	0.1678
ti(display index,saliency rank)	1.3966	16	0.5272	0.0174
ti(display index,word position)	3.3649	16	0.6625	0.0520
ti(word length,sentence length)	0.0004	16	0.0000	0.9236
ti(word length,word frequency)	2.1540	16	6.5510	< 0.0001
ti(word length,sentiment polarity)	0.0014	16	0.0000	0.6790
ti(word length,saliency rank)	2.2175	16	0.3503	0.0573
ti(word length,word position)	1.0296	16	0.1270	0.1222
ti(sentence length,word frequency)	0.0005	16	0.0000	0.8608
ti(sentence length,sentiment polarity)	0.0013	16	0.0001	0.5113
ti(sentence length,saliency rank)	1.3045	16	0.2651	0.0453
ti(sentence length,word position)	3.1995	16	0.8487	0.0067
ti(word frequency,sentiment polarity)	0.0015	16	0.0001	0.1969
ti(word frequency,saliency rank)	0.0022	15	0.0001	0.3230
ti(word frequency,word position)	2.0375	16	0.3168	0.0924
ti(sentiment polarity,saliency rank)	0.0006	16	0.0000	0.8407
ti(sentiment polarity,word position)	0.0005	16	0.0000	0.9558
ti(saliency rank,word position)	0.0006	16	0.0000	0.6542
s(sentence_id)	0.0006	150	0.0000	0.9276
s(saliency,sentence_id)	9.1441	150	0.0676	0.2305
s(worker_id)	48.1065	49	10640.8475	< 0.0001
s(saliency,worker_id)	48.0654	50	6593.7769	< 0.0001

Table A.5.: Wald tests for the pairwise interactions (tensor interactions) (upper) and random effects (lower) of the English user study.

	df	F	p
capitalization	2	7.62	0.0005
dependency relation	33	2.57	< 0.0001

Table A.6.: Wald tests for the parametric terms of the German user study.



(a) Saliency * display index (b) Saliency * word length (c) Saliency * sentence length

(d) Saliency * word frequency (e) Display index * word length (f) Display index * sentence length

(g) Display index * word frequency (h) Display index * saliency rank (i) Word length * word frequency

(j) Sentence length * saliency rank (k) Sentence length * word position

Figure A.15.: Summed-effect plots of all significant pairwise interactions.

	edf	**ref. df**	**F**	**p**
s(saliency)	8.2052	19	148.1115	< 0.0001
s(display index)	1.5999	9	2.4742	< 0.0001
s(word length)	2.0440	9	3.9174	< 0.0001
s(sentence length)	0.9073	9	1.7657	0.0003
s(word frequency)	0.0017	9	0.0002	0.2816
s(saliency rank)	0.0004	9	0.0000	0.7016
s(word position)	2.4429	9	2.8142	0.0002
ti(saliency,display index)	0.0007	16	0.0000	0.5846
ti(saliency,word length)	2.4114	16	1.1662	0.0013
ti(saliency,sentence length)	1.8496	16	0.7410	0.0125
ti(saliency,word frequency)	0.6953	11	0.3084	0.0549
ti(saliency,saliency rank)	1.4340	16	0.4958	0.0142
ti(saliency,word position)	0.0765	16	0.0053	0.2970
ti(display index,word length)	0.3529	16	0.0477	0.1968
ti(display index,sentence length)	0.1902	16	0.0171	0.2622
ti(display index,word frequency)	0.0005	15	0.0000	0.7096
ti(display index,saliency rank)	0.2967	16	0.0332	0.2325
ti(display index,word position)	1.1440	16	0.4244	0.0168
ti(word length,sentence length)	0.9858	16	0.3138	0.0290
ti(word length,word frequency)	0.9622	11	1.0293	0.0050
ti(word length,saliency rank)	0.0005	16	0.0000	0.8581
ti(word length,word position)	0.8285	16	0.5132	0.0091
ti(sentence length,word frequency)	0.0009	15	0.0001	0.3536
ti(sentence length,saliency rank)	0.0005	16	0.0000	0.9945
ti(sentence length,word position)	0.0005	16	0.0000	0.6862
ti(word frequency,saliency rank)	0.0003	16	0.0000	0.9438
ti(word frequency,word position)	0.0005	15	0.0000	0.6085
ti(saliency rank,word position)	0.0004	16	0.0000	0.8379
s(sentence ID)	0.0004	149	0.0000	0.9007
s(saliency,sentence ID)	36.6567	150	0.3534	0.0087
s(worker ID)	23.5324	24	8128.6327	< 0.0001
s(saliency,worker ID)	23.6122	25	5645.0812	< 0.0001

Table A.7.: Wald tests for the smooth terms of the German user study.

Coefficients	β	**SE**	**t**	**p**
capitalization: all capital	1.9051	0.9638	1.9767	0.0481
capitalization: first capital	0.4074	0.1151	3.5390	0.0004
dependency relation: acl	-1.2155	0.6428	-1.8910	0.0587
dependency relation: acl:relcl	1.3605	0.5947	2.2878	0.0222
dependency relation: advcl	0.8647	0.7154	1.2087	0.2269
dependency relation: advmod	0.3741	0.2369	1.5790	0.1144
dependency relation: amod	0.4794	0.2653	1.8072	0.0708
dependency relation: appos	0.2823	0.4119	0.6852	0.4932
dependency relation: aux	0.6395	0.3138	2.0379	0.0416
dependency relation: aux:pass	-0.0679	0.3798	-0.1789	0.8581
dependency relation: case	0.1169	0.2082	0.5613	0.5746
dependency relation: cc	0.1126	0.2571	0.4379	0.6615
dependency relation: cc:preconj	0.8039	1.1491	0.6996	0.4842
dependency relation: ccomp	1.1850	0.5206	2.2763	0.0229
dependency relation: compound	0.8738	0.4488	1.9470	0.0516
dependency relation: compound:prt	0.4114	0.4577	0.8989	0.3688
dependency relation: conj	0.1673	0.2900	0.5769	0.5640
dependency relation: cop	0.4169	0.2598	1.6043	0.1087
dependency relation: csubj	1.0154	0.7533	1.3480	0.1777
dependency relation: det	-0.1604	0.2088	-0.7682	0.4424
dependency relation: expl	-1.0130	0.4605	-2.1998	0.0279
dependency relation: flat:name	0.3786	0.5401	0.7010	0.4833
dependency relation: iobj	-0.4807	0.5162	-0.9312	0.3518
dependency relation: mark	0.1537	0.3646	0.4216	0.6734
dependency relation: nmod	0.4656	0.2787	1.6707	0.0949
dependency relation: nmod:poss	-0.0658	0.3251	-0.2025	0.8395
dependency relation: nsubj	0.4443	0.2369	1.8755	0.0608
dependency relation: nsubj:pass	0.4296	0.4305	0.9979	0.3184
dependency relation: nummod	1.3866	0.3609	3.8419	0.0001
dependency relation: obj	0.2406	0.2649	0.9082	0.3638
dependency relation: obl	0.3126	0.2679	1.1668	0.2434
dependency relation: obl:tmod	1.7042	0.5544	3.0739	0.0021
dependency relation: parataxis	-0.2780	0.8595	-0.3234	0.7464
dependency relation: root	0.5463	0.2432	2.2460	0.0248
dependency relation: xcomp	0.6718	0.4494	1.4948	0.1351

Table A.8.: Capitalization and dependency relation coefficients for the German user study.

Sentence with Saliency Explanation	Rating	Comment
Durch den Deal zwischen Aoun und Hariri kommen sich die beiden verfeindeten Bündnisse (vorerst) näher .	4	Wenn dann müssen beide Klammern weg (P4)
Jedes Gedicht erzählt nur von einem Teil des Krieges :	2	Das Symbol wird von der KI zu hoch bewertet. (P10)
Gewitterstürme sind selten , die Stadt berichtet nur an sieben Tagen pro Jahr von Gewittern .	7	Auch hier: "Gewitterstürme" viel zu gering gewichtet, "Jahr" zu hoch bewertet (P10)
Die Geschichte von Doss hat auch etwas Unglaubhaftes an sich , das sie nur umso attraktiver macht .	3	Der Artikel ist sicher wichtig, jedoch nicht zwingend für den Sinn verantwortlich. (P16)
Frau Hopley fügte hinzu : „ Der starke Anstieg des politischen Risikos sollte nicht unbeachtet bleiben. “	1	Es ist nur eine grammatische Kennzeichnung. Diese ist für KI meines Erachtens wenig bis garnicht relevant. (P16)
Wasser aus den Flüssen wird in über 500 Wasserkraftwerken genutzt , wobei 2900 Kilowatt Elekrizität generiert werden .	3	Die KI sollte schon den Wert einer Aussage kennen, die erst in der Zukunft eintritt und diese gegenüber aktuell bereits eingetretenen Ereignissen bewerten können. (P16)
Der Kunde kann die Forderung nach Veränderung verstärken .	5	Das Verb gibt dem Satz seinen Sinn. (P16)
Ich glaube , darum haben sie sich mit so vielen Mustern und Farben umgeben .	5	Das Adjektiv beschreibt eine wichtige Eigenschaft und ist für die Satzbewertung relevant. (P16)

Table A.9.: Comments of the participants of the German study. Participants were asked to rate the underlined word or symbol.

A.5.3. German Study Details

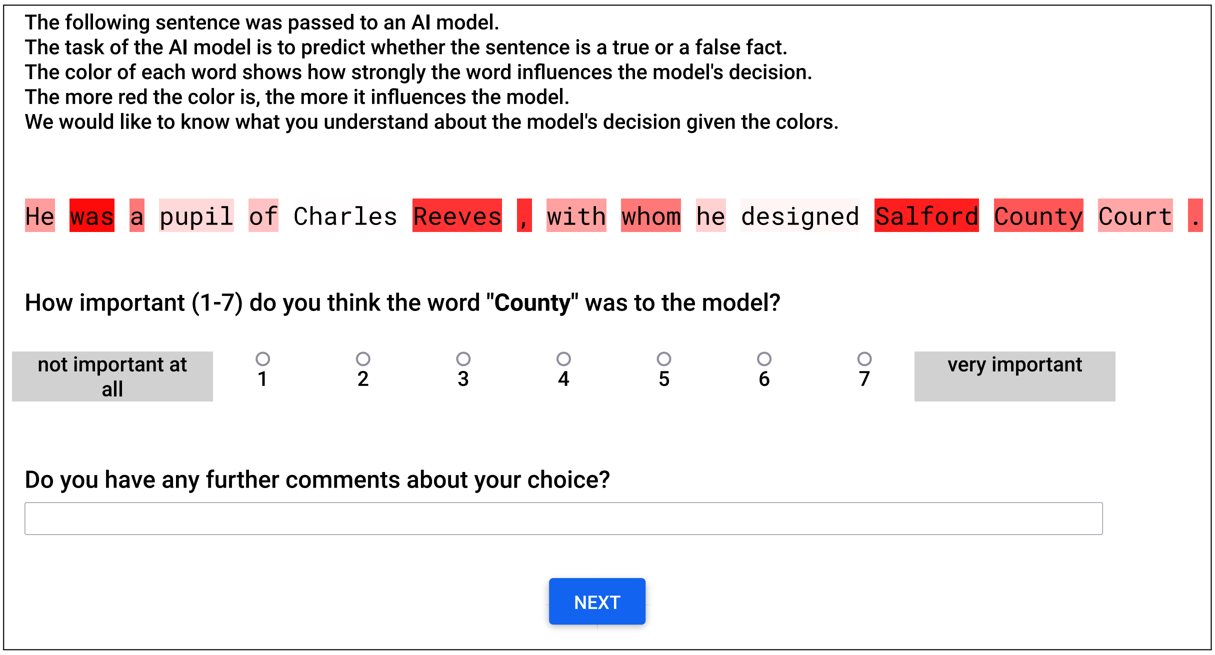
In this section, we provide details on the analysis of the German experiment. Table A.7 and Table A.6 display test statistics for the smooth and parametric terms of the fitted GAMM model. Table A.8 shows statistics regarding parametric coefficient estimates. Cut points are located at -1, 0.86, 2.42, 3.75, 5.53, and 7.67. Table A.9 lists exemplary participant comments.

A.6. Neighboring Words

A.6.1. User Study Details

A.6.1.1. Interface

Figure A.16 shows a screenshot of our rating interface. Figure A.17 shows a screenshot of an attention check.



The following sentence was passed to an AI model.
The task of the AI model is to predict whether the sentence is a true or a false fact.
The color of each word shows how strongly the word influences the model's decision.
The more red the color is, the more it influences the model.
We would like to know what you understand about the model's decision given the colors.

`He was a pupil of Charles Reeves , with whom he designed Salford County Court .`

How important (1-7) do you think the word "**County**" was to the model?

not important at all ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 very important

Do you have any further comments about your choice?

Figure A.16.: Screenshot of the rating interface.

A.6.1.2. Attention Checks

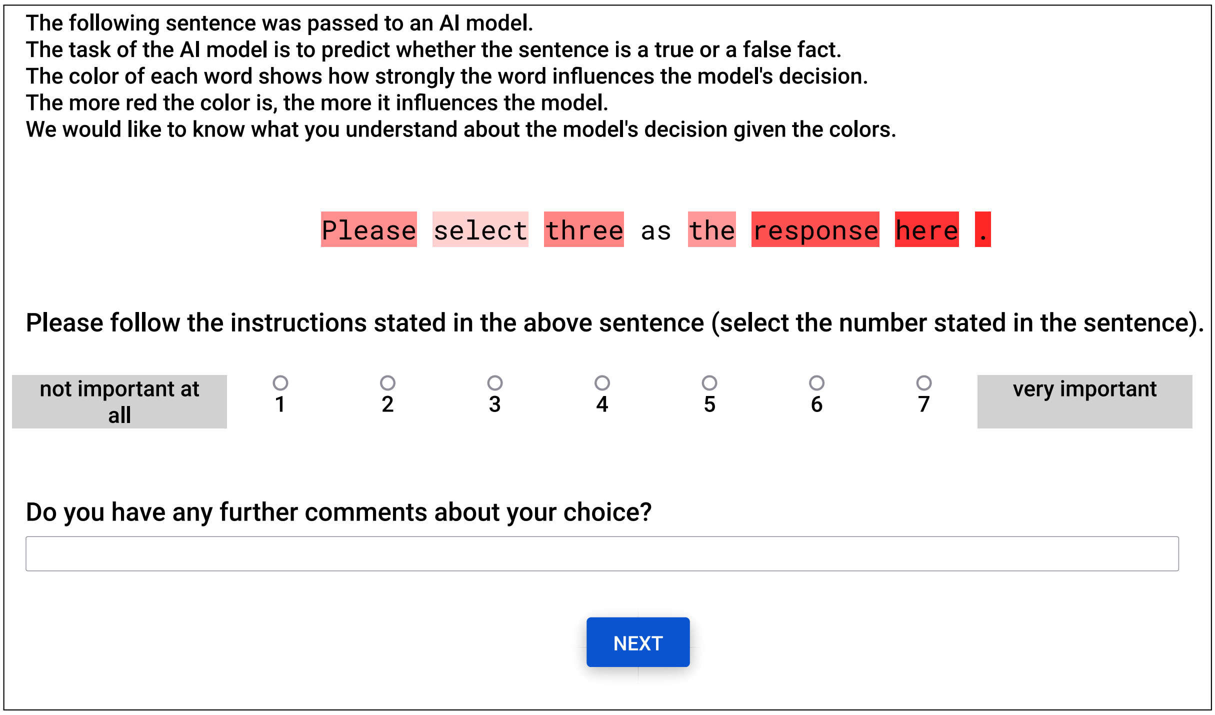
As in the previous studies, we include three attention checks per participant which we randomly place within the last two thirds of the study.

A.6.1.3. Participants

In total, we recruit 76 crowdworkers from English-speaking countries via MTurkfor our randomized explanation study and 36 crowdworkers for our SHAP-value explanation study. We require workers to have at least 5,000 approved HITs and a 95% approval rate. Raters are screened with three hidden attention checks that they must answer correctly to be included (but are paid fully regardless). Of the 76 workers, 64 workers passed the screening, i.e., we excluded 15.8% of responses on a participant level. From the 36 workers, all workers passed the screening. On average, participants were compensated with an hourly wage of US$ 8.95. We do not collect any personally identifiable data from participants.

A.6.1.4. Model Details in Our Analysis

We control for all main effects (word length, sentence length, etc.) as well as all random effects used in the previous study. We exclude the pairwise interactions due to model instability when including the interactions.



The following sentence was passed to an AI model.
The task of the AI model is to predict whether the sentence is a true or a false fact.
The color of each word shows how strongly the word influences the model's decision.
The more red the color is, the more it influences the model.
We would like to know what you understand about the model's decision given the colors.

Please select three as the response here .

Please follow the instructions stated in the above sentence (select the number stated in the sentence).

not important at all ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 very important

Do you have any further comments about your choice?

Figure A.17.: Screenshot of the rating interface for an attention check.

We additionally include four new novel bivariate smooth terms. Each of these terms models a tensor product of saliency (i.e. the rated word's color intensity) and the neighboring (left or right) word's saliency difference to the rated word. For each side (left and right), we model the smooths for neighbors that (i) are within a lexical chunk of the rated word and (ii) are not. Figure 5.7 shows the estimated four (bivariate) functions.

A.6.1.5. Data Preprocessing

We exclude ratings with a completion time of less than a minute (implausibly fast completion) and exclude words with a length of over 20 characters. We effectively exclude 1.8% of ratings.

In order to analyze left as well as right neighbors, we additionally have to ensure that we only include ratings for which both — left and right — neighbors exist. Therefore, we additionally exclude ratings for which the leftmost or rightmost word in the sentence was rated. This excludes 11.7% of ratings. In total, we thus use 9489 ratings to fit our model.

A.6.1.6. Chunk Measures

We explore and combine two approaches to identifying multi-word phrases (or "chunks)".

Syntactic Measures: Constituents. We first apply binary chunk measures based on the sentences' parse trees. We use Stanza (Qi et al., 2020) (version 1.4.2) to generate a parse tree for each sentence. We assess whether the rated word and its neighbor (left/right) share a constituent at the lowest possible level. Concretely, we (a) start at the rated word and move up one level in the parse tree and (b) start at the neighboring word and move up one level in the parse tree. If we now arrived at the same node in the parse tree, we consider the rated word and its neighbor to share a first-order constituent. If we arrived at different nodes, we consider them to do not. Restricting the type of first-level shared constituents to noun phrases yields a further category. We provide respective examples for shared first-level constituents and the respective noun phrase constituents extracted from our data in Table A.10 (upper part).

Statistical Measures: Co-occurrence Scores. We additionally explore numeric association measures and calculate all available bigram collocation measures available in NLTK's *BigramAssocMeasures* module[2]. The calculation is based on the seven million Wikipedia-2018 sentences in *Wikipedia Sentences*.[3] A description of each metric as well as top-scored examples on our data is provided in Table A.10 (lower part). We separate examples into examples that form a constituent vs. do not form a constituent to highlight the necessity to apply a constituent filter in order to get meaningful categorization into chunks vs. no chunks.

A.6.1.7. Detailed Results

As described in Section 5.1.2.3, we observe different influences of left/right neighbors, chunk/no chunk neighbors as well as rated word saliency levels for randomized explanation experiment. We report the detailed Wald test statistics for our randomized explanation experiment in Table A.11.

Left vs. Right Neighbors. Figure A.18 shows difference plots (and respective p values) between left and right neighbors for chunk neighbors (Figures A.18a and A.18b) and no chunk neighbors (Figures A.18c and A.18d).

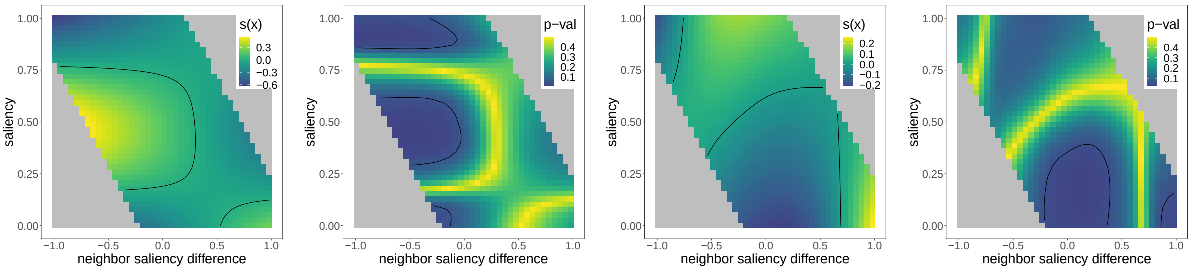
Chunk vs. No Chunk. Respectively, Figure A.19 shows difference plots (and respective p values) between chunk and no chunk neighbors for left neighbors (Figures A.19a and A.19b) and right neighbors (Figures A.19c and A.19d).

[2] https://www.nltk.org/_modules/nltk/metrics/association.html

[3] https://www.kaggle.com/datasets/mikeortman/wikipedia-sentences

Measure	Constituent Examples	No Constituent Examples	Description
First-order constituent	highly developed, more than, such as, DVD combo, 4 million	—	Smallest multi-word constituent subtrees in the constituency tree.
Noun phrase	Tokyo Marathon, ski racer, the UK, a retired, the city	—	Multi-word first-order noun phrase in the constituency tree.
Mutual information	as well, more than, ice hockey, United Kingdom, a species	is a, of the, in the, is an, it was	Bigram mutual information variant (per NLTK implementation).
Frequency	the United, the family, a species, an American, such as	of the, in the, is a, to the, on the	Raw, unnormalized frequency.
Poisson Stirling	an American, such as, a species, as well, the family	is a, of the, in the, is an, it was, has been	Poisson Stirling bigram score.
Jaccard	Massar Egbari, ice hockey, Air Force, more than, Udo Dirkschneider	teachers/students teaching/studying, is a, has been, it was, of the	Bigram Jaccard index.
φ^2	Massar Egbari, ice hockey, Udo Dirkschneider, Air Force, New Zealand	teachers/students teaching/studying, is a, has been, footballer who, is an	Square of the Pearson correlation coefficient.

Table A.10.: The list of phrase measures we tested for. Examples for numeric measures are chosen based on highest co-occurrence scores whereas the (boolean) noun phrase and constituent examples are chosen arbitrarily. For the numeric measures, we provide examples that (a) form a constituent with their neighbor and (b) do not. The examples underline the need to combine numeric scores with constituents.



(a) Difference between right - left (chunk). The contour line marks zero.

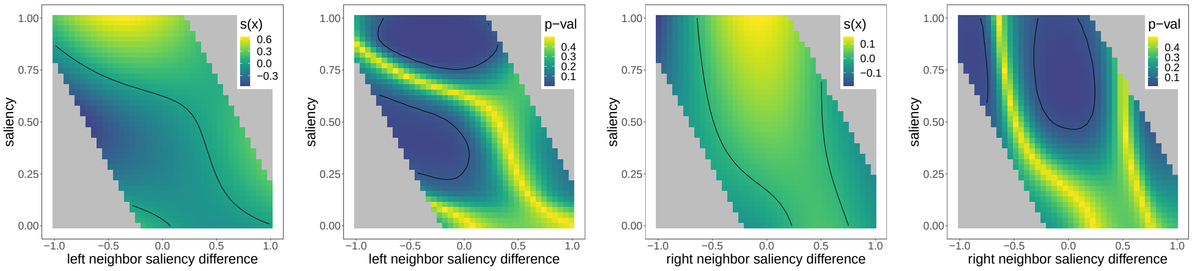
(b) Difference (chunk) p values. The contour line marks 0.05.

(c) Difference between right - left (no chunk). The contour line marks zero.

(d) Difference (no chunk) p values. The contour line marks 0.05.

Figure A.18.: Differences and p values for (no) lexical chunk neighbors for our randomized explanation experiment.

Differences Across Saliency Levels. Figure A.20 shows that the effects of saliency difference are significantly different between different levels of the rated word's saliency (0.25 and 0.75) for left neighbors (Figure A.20a) as well as right neighbors (Figure A.20b).



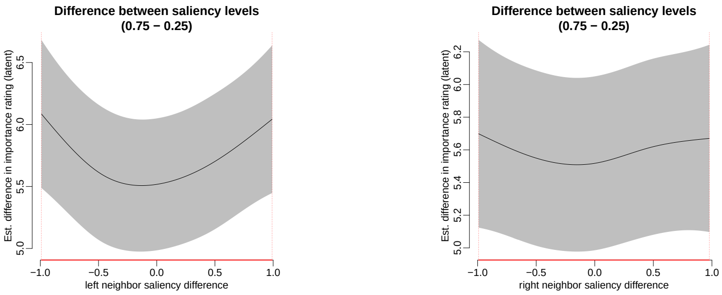
(a) Difference between (left neighbors) chunk - no chunk. The contour line marks zero.

(b) Difference (left neighbors) p values. The contour line marks 0.05.

(c) Difference between (right neighbors) chunk - no chunk. The contour line marks zero.

(d) Difference (right neighbors) p values. The contour line marks 0.05.

Figure A.19.: Differences and p values for left and right neighbors for our randomized explanation experiment.



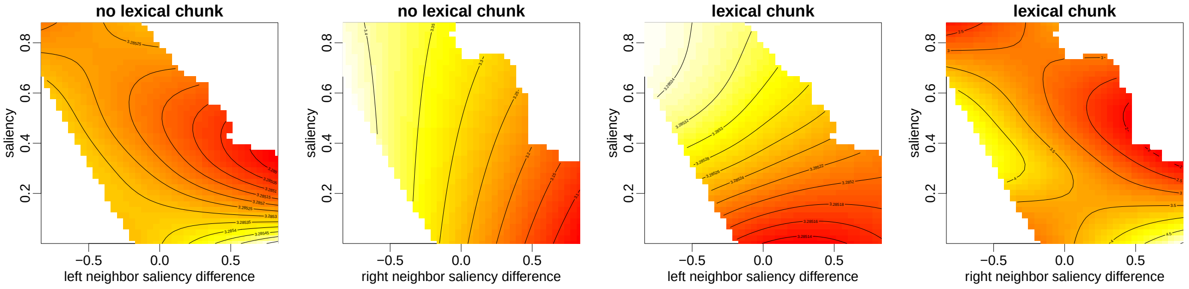
(a) Left.

(b) Right.

Figure A.20.: Difference plots between the influence of saliency differences between exemplary high (0.75) and low (0.25) rated word saliency levels. Red x-axis areas indicate significant differences.

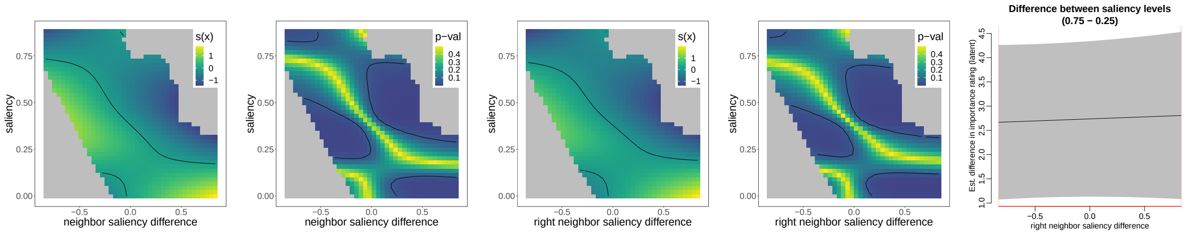
A.6.1.8. SHAP-value Results

We additionally report details regarding our SHAP-value experiment results. Figure A.21 shows the respective summed-effects plots. Figure A.22 displays left/right, chunk/no chunk, and rated word saliency level difference plots. We report the detailed Wald test statistics for our SHAP-value explanation experiment in Table A.12. Figure A.23 illustrates how the distribution of saliency scores is uniformly random for our randomized explanations in contrast to the distributions of SHAP values.



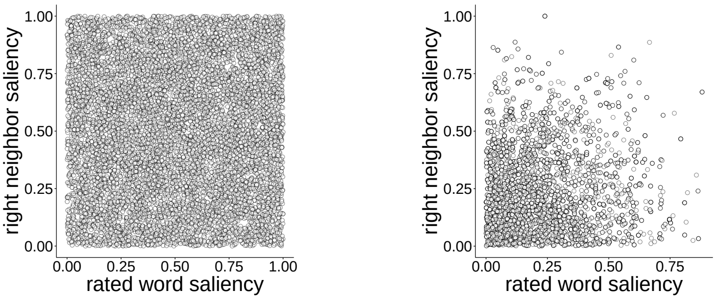
(a) Left, no chunk. (b) Right, no chunk. (c) Left, chunk. (d) Right, chunk. (∗)

Figure A.21.: Left and right neighbors in our SHAP-value experiment. (∗) marks statistically significant smooths. Colors are normalized per figure. Note that the first three plots correspond to non-significant effects and their respective color mappings cover a small value range.



(a) Right/left difference for chunks (contour line marks 0). (b) Right/left difference (chunk) p values (contour line marks 0.05). (c) Chunk/no chunk difference for right neighbor (contour line marks 0). (d) Chunk/no chunk difference (right) p values (contour line marks 0.05). (e) Difference between rated saliency and right neighbor saliency.

Figure A.22.: Difference plots of our SHAP-value experiment results. Red x-axis in (e) marks significant differences.



(a) Randomized saliency. (b) SHAP-value saliency.

Figure A.23.: Comparison of the distributions of rated word saliency and right neighbor saliency across our randomized explanations and our SHAP-value experiments.

	(e)df	**Ref.df**	**F**	**p**
s(saliency)	11.22	19.00	580.89	<**0.0001**
s(display index)	3.04	9.00	22.02	<**0.0001**
s(word length)	1.64	9.00	16.44	<**0.0001**
s(sentence length)	0.00	4.00	0.00	0.425
s(relative word frequency)	0.00	9.00	0.00	0.844
s(normalized saliency rank)	0.59	9.00	0.37	0.115
s(word position)	0.58	9.00	0.18	0.177
te(left diff.,saliency): no chunk	3.12	24.00	1.50	**0.002**
te(left diff.,saliency): chunk	2.24	24.00	0.51	**0.038**
te(right diff.,saliency): no chunk	2.43	24.00	0.47	**0.049**
te(right diff.,saliency): chunk	0.00	24.00	0.00	0.578
s(sentence ID)	0.00	149.00	0.00	0.616
s(saliency,sentence ID)	16.13	150.00	0.14	0.191
s(worker ID)	62.19	63.00	30911.89	<**0.0001**
s(saliency,worker ID)	62.11	64.00	16760.88	<**0.0001**
capitalization	2.00		3.15	**0.042**
dependency_relation	35.00		2.92	<**0.0001**

Table A.11.: (Effective) degrees of freedom, reference degrees of freedom and Wald test statistics for the univariate smooth terms (top), random effects terms (middle) and parametric fixed terms (bottom) using $t = 87.5\%$ and φ^2 measure.

A.6.1.9. Reproduction Analysis

We confirm our previous results from Section 5.1.1 and find significant effects of word length, display index, capitalization, and dependency relation. We report detailed statistics of our randomized saliency experiment in Table A.11 and our SHAP experiment in Table A.12.

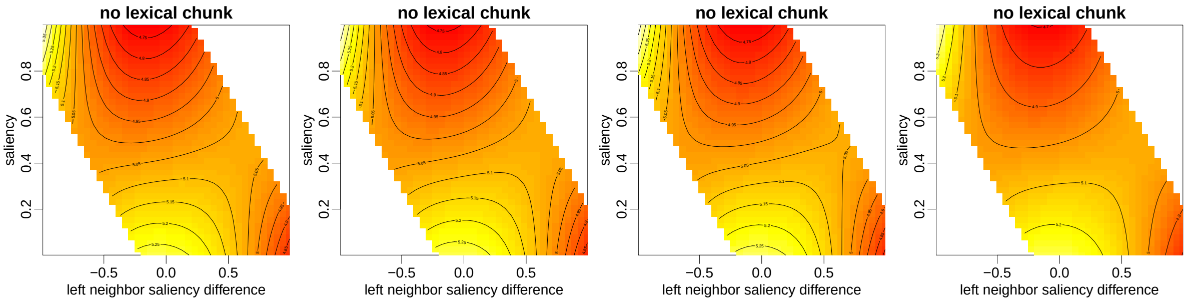
A.6.2. Robustness to Evaluation Parameters.

To ensure our results are not an artifact of the particular combination of threshold and co-occurrence measure, we investigate how our results change if we (i) vary the threshold within $\{0.5, 0.75, 0.875\}$ and (ii) vary the co-occurrence measure within {Jaccard, MI-like, φ^2, Poisson-Stirling}. We find significant interactions and observe similar interaction patterns as well as areas of significant differences (left/right, chunk/no chink as well as saliency levels) across all settings. We provide a representative selection of plots in Figures A.24 to A.29. Additionally, Tables A.13 and A.14 demonstrate that changing the threshold or co-occurrence measure leads to model statistics that are largely consistent with the results reported in Table A.11. We choose the φ^2 and a 87.5% threshold as no other model reaches higher deviance

	(e)df	**Ref.df**	**F**	**p**
s(saliency)	6.71	19.00	18.85	<**0.0001**
s(display index)	1.88	9.00	6.45	<**0.0001**
s(word length)	2.04	9.00	4.43	<**0.0001**
s(sentence length)	0.00	4.00	0.00	0.98
s(relative word frequency)	0.00	9.00	0.00	0.64
s(normalized saliency rank)	0.89	9.00	1.99	**0.002**
s(word position)	0.42	9.00	0.12	0.19
te(left diff.,saliency): no chunk	0.00	24.00	0.00	0.37
te(left diff.,saliency): chunk	0.00	24.00	0.00	0.49
te(right diff.,saliency): no chunk	0.99	24.00	0.20	0.06
te(right diff.,saliency): chunk	3.24	24.00	1.09	0.01
s(sentence ID)	0.00	149.00	0.00	0.52
s(saliency,sentence ID)	11.31	150.00	0.10	0.14
s(worker ID)	34.77	35.00	14185.28	<**0.0001**
s(saliency,worker ID)	62.11	64.00	16760.88	<**0.0001**
capitalization	2.00	0.35	0.71	
dependency relation	34.59	36.00	8468.22	<**0.0001**

Table A.12.: SHAP experiment results details. (Effective) degrees of freedom, reference degrees of freedom and Wald test statistics for the univariate smooth terms (top), random effects terms (middle) and parametric fixed terms (bottom) using $t = 87.5\%$ and φ^2 measure.

explained and a comparison of randomly-sampled chunk/no chunk examples across measures and thresholds yields the best results for this setting.

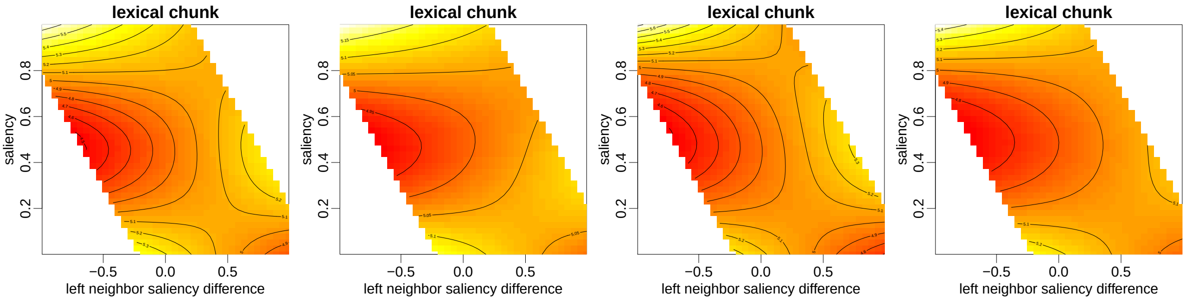


(a) φ^2. (b) Jaccard. (c) MI-like. (d) Poisson-Stirling.

Figure A.24.: Tensor product interactions for left saliency difference in the outside chunk setting across different choices of co-occurrence measures. We find similar patterns across all settings for our randomized explanation experiment. $t = 87.5$ is consistent for all plots.

	(e)df	**Ref.df**	**F**	**p**
s(saliency)	11.23	19.00	547.16	< **0.0001**
s(display_index)	3.10	9.00	20.93	< **0.0001**
s(word_length)	1.61	9.00	16.47	< **0.0001**
s(sentence_length)	0.00	4.00	0.00	0.436
s(relative_word_frequency)	0.00	9.00	0.00	0.814
s(normalized_saliency_rank)	0.58	9.00	0.36	0.120
s(word_position)	0.59	9.00	0.18	0.173
te(left diff.,saliency): no chunk	2.90	24.00	1.21	**0.003**
te(left diff.,saliency): chunk	3.34	24.00	0.92	**0.015**
te(right diff.,saliency): no chunk	2.50	24.00	0.67	**0.021**
te(right diff.,saliency): chunk	0.00	24.00	0.00	0.836
s(sentence_id)	0.00	149.00	0.00	0.601
s(saliency,sentence_id)	17.35	150.00	0.15	0.178
s(worker_id)	62.19	63.00	30421.05	< **0.0001**
s(saliency,worker_id)	62.11	64.00	17591.01	< **0.0001**
capitalization	2.00		3.01	**0.049**
dependency_relation	35.00		2.93	< **0.0001**

Table A.13.: (Effective) degrees of freedom, reference degrees of freedom and Wald test statistics for the univariate smooth terms (top), random effects terms (middle) and parametric fixed terms (bottom) using $t = 25\%$ and φ^2 measure.

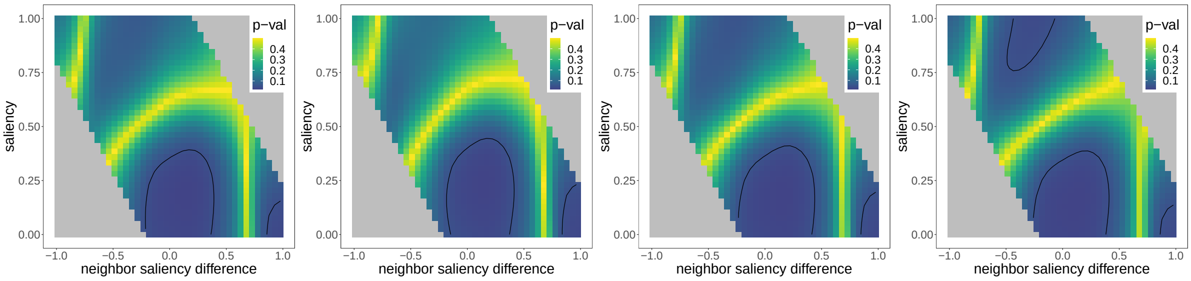


(a) φ^2. (b) Jaccard. (c) MI-like. (d) Poisson-Stirling.

Figure A.25.: Tensor product interactions for left saliency difference in the within chunk setting across different choices of co-occurrence measures. We find similar patterns across all settings for our randomized explanation experiment. $t = 87.5$ is consistent for all plots.

	(e)df	**Ref.df**	**F**	**p**
s(saliency)	11.21	19.00	584.57	< **0.0001**
s(display_index)	3.04	9.00	21.63	< **0.0001**
s(word_length)	1.63	9.00	16.66	< **0.0001**
s(sentence_length)	0.00	4.00	0.00	0.407
s(relative_word_frequency)	0.00	9.00	0.00	0.813
s(normalized_saliency_rank)	0.56	9.00	0.32	0.130
s(word_position)	0.65	9.00	0.22	0.159
te(left diff.,saliency): no chunk	3.10	24.00	1.57	**0.0010**
te(left diff.,saliency): chunk	1.79	24.00	0.34	0.082
te(right diff.,saliency): no chunk	2.37	24.00	0.47	**0.048**
te(right diff.,saliency): chunk	0.64	24.00	0.05	0.249
s(sentence ID)	0.00	149.00	0.00	0.638
s(saliency,sentence ID)	17.14	150.00	0.15	0.164
s(worker ID)	62.19	63.00	30521.95	< **0.0001**
s(saliency,worker ID)	62.11	64.00	16749.25	< **0.0001**
capitalization	2.00		3.23	**0.039**
dependency relation	35.00		2.94	< **0.0001**

Table A.14.: (Effective) degrees of freedom, reference degrees of freedom and Wald test statistics for the univariate smooth terms (top), random effects terms (middle) and parametric fixed terms (bottom) using $t = 87.5\%$ and MI-like measure.



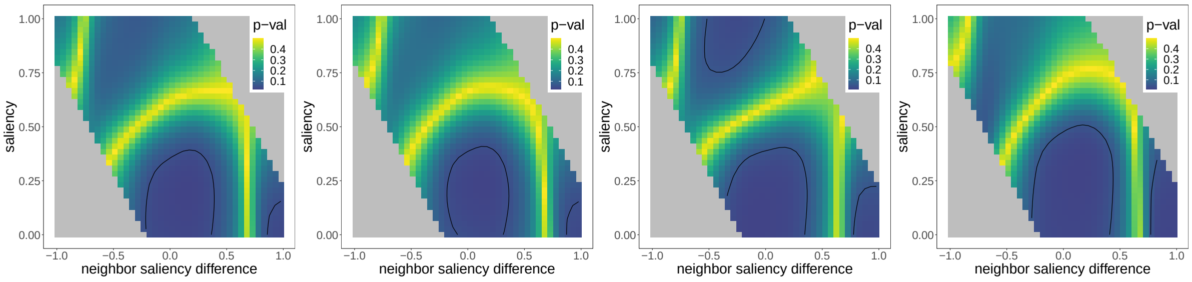
(a) φ^2.

(b) Jaccard.

(c) MI-like.

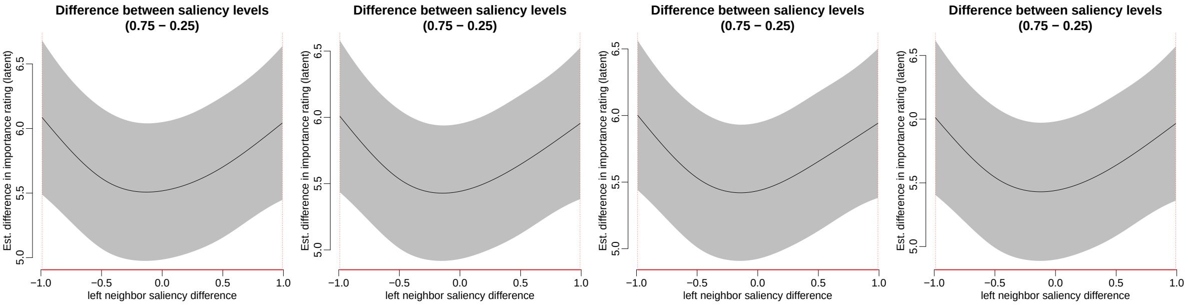
(d) Poisson-Stirling.

Figure A.26.: p values for differences between right - left for no lexical chunk neighbors across different choices of co-occurrence measures. We find similar patterns across all settings for our randomized explanation experiment. $t = 87.5$ is consistent for all plots.



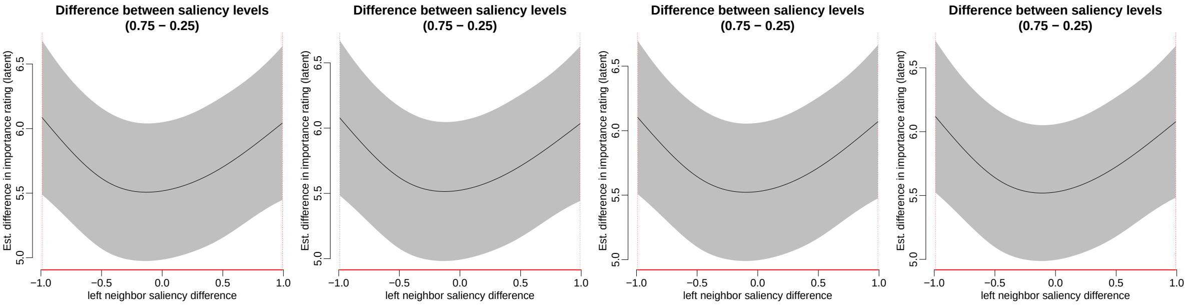
(a) $t = 87.5\%$ (b) $t = 75\%$ (c) $t = 50\%$ (d) $t = 25\%$

Figure A.27.: p values for differences between right - left for no lexical chunk neighbors across different choices of thresholds. We find similar patterns across all settings for our randomized explanation experiment. The φ^2 measure is used across all plots.



(a) $t = 87.5\%$ (b) $t = 75\%$ (c) $t = 50\%$ (d) $t = 25\%$

Figure A.28.: Difference plots between the influence of left saliency differences between exemplary high (0.75) and low (0.25) rated word saliency levels across different choices of thresholds for our randomized explanation experiment. We find similar patterns across all settings. The φ^2 measure is used across all plots.



(a) φ^2. (b) Jaccard. (c) MI-like. (d) Poisson-Stirling.

Figure A.29.: Difference plots between the influence of left saliency differences between exemplary high (0.75) and low (0.25) rated word saliency levels across different choices of co-occurrence measures for our randomized explanation experiment. We find similar patterns across all settings. $t = 87.5$ is consistent for all plots.

A.7. Model-based Bias Correction

Our second approach to bias mitigation is to leverage the previously described GAMM model of human saliency perception and to *correct* saliency perception by superimposing the initial saliency values with a correction signal.

Concretely, we want to increase the saliency scores for words that are predicted to be under-perceived (e.g., short words and words that appear in long sentences) and decrease the saliency scores for words that are predicted to be over-perceived (e.g., word's with a high polarity score or words that appear in very short sentences).

When we want to *correct* a user perception via the saliency scores, we cannot say whether a subjective user rating of importance is right or wrong. However, the previously described GAMM model allows us to map a combination of a saliency score together with word/sentence properties to a perceived importance score (on a continuous latent scale). In the following, we denote this mapping as

$$u(s, \mathbf{x}) : [0, 1] \times \mathbb{R}^d \to \mathbb{R}, \tag{A.1}$$

where s is a saliency score and $\mathbf{x}$ is a d-dimensional feature vector representing the word/sentence properties. This function allows us to take a fixed saliency score s (e.g., 0.7) and predict its perceived importance given word and sentence features $\hat{\mathbf{x}}$ (corresponding to, e.g., a word length of five characters and a sentence length of four). We define this predicted importance score as

$$p := u(s, \hat{\mathbf{x}}) . \tag{A.2}$$

Additionally, it allows us to predict the perceived importance of that same saliency (0.7) in a hypothetical reference context $\mathbf{x}_{\text{ref}}$ (corresponding to, e.g., a word length of three and a sentence length of six). We define this second predicted importance score as

$$p_{\text{ref}} := u(s, \mathbf{x}_{\text{ref}}) . \tag{A.3}$$

We can now define a *bias score* $b \in \mathbb{R}$ as the difference between the importance score for the saliency in the observed context and the importance score for the same saliency in the reference context

$$b := p - p_{\text{ref}}. \tag{A.4}$$

The predicted bias score b is positive if the saliency in the observed context is *over-perceived* with respect to the reference level and negative if it is *under-perceived* with respect to the reference level. A bias score of zero corresponds to an *unbiased* predicted perception. Intuitively, this formalization allows us to answer the question "*In which direction do I have to change the saliency such that the predicted bias with respect to the reference context is decreased?*".

To gain an executable process for bias mitigation we still lack (a) a way to handle the random effects in the model, i.e., participant IDs and sentence IDs, (b) a definition of the reference context, and (c) a procedure to minimize the absolute value of the bias score. We detail these three aspects in the following.

A.7.1. Including Random Effects

So far, our definition of the model function u ignores the random effects of the GAMM model, i.e., we did not specify which worker ID and which sentence ID should be used in predicting the importance score. However, the choice of the respective levels directly influences the model predictions not only via the random intercepts but also via the random slopes for each worker and sentence ID. We see two options to address this. While a first, intuitive remedy is to use an arbitrary worker ID and an arbitrary sentence ID for all predictions, this approach has the disadvantage of introducing an arbitrary bias. Therefore, we choose to make each model prediction not only for one participant ID and one sentence ID but instead for all combinations of participant IDs and sentence IDs ($50 \times 150 = 7500$ combinations). Thereby, we consider each combination of a participant and a sentence as equally relevant for the prediction of unseen participants and sentences and smooth-out extreme influences of single participants or sentence IDs. Formally, we thus update our definition of Equation A.1 to

$$u(s, \mathbf{x}, w, v) : [0, 1] \times \mathbb{R}^d \times W \times V \to \mathbb{R}, \tag{A.5}$$

where W is the set of participant (or crowdworker) IDs ($|W| = 50$) and V is the set of sentence IDs ($|V| = 150$). Consequently, a single evaluation of $u\,(s, \mathbf{x})$ is now replaced with

$$\frac{1}{|W||V|} \sum_{w \in W} \sum_{v \in V} u\,(s, \mathbf{x}, w, v)\,. \tag{A.6}$$

A.7.2. Choosing the Reference Context

So far, our definitions in Equations A.2 to A.4 do not impose any constraints on the choice of reference context. Why can we not just use an arbitrary reference context with, e.g., a word length of eight and a sentence length of one (and respective choices for all remaining covariates, such as sentiment polarity, etc.)? The problem that arises for that concrete context is that the model assigns a very high importance prediction to words with eight characters within a sentence with length one. Consequently, p_{ref} will be larger than p for most words and the bias score b would get negative, indicating an under-perception for most words. If we then increase all these words' saliency scores in order to minimize the absolute bias score, we, overall, have to make large changes to the saliencies. In other words, this specific reference context corresponds to an, overall, raised level of saliency intensities. While this is not bad per se, we favor a reference context that is as neutral as possible regarding its impact on predicted importance ratings.

In order to find such a reference context, we sample 10001 random points from the space of possible contexts defined as the cross product of intervals of observed values (e.g., 1-37 characters word length) per variable if the variable is numeric (e.g., word length) and the set of possible values if the variable is categorical (e.g., dependency relation). Each point is a candidate context. We evaluate the term in Equation A.6 for a saliency score of 0.5 and each candidate context. Among all predicted importance scores, we select the median score and choose the corresponding candidate context as our reference context $\mathbf{x}_{\text{ref}}$.[4]

A.7.3. Iterative Bias Minimization

In order to minimize the absolute predicted bias score, we have to modify each word's original saliency score $s^{(i)}_{\text{orig}} \in [0, 1]$ into a corrected saliency score $s^{(i)}_{\text{corr}} \in [0, 1]$. While this seems to be a straight-forward minimization at first glance there is one covariate in the model that complicates optimization. The value of the saliency rank variable depends on the saliencies of all words in the sentence. Thus, changing one word's saliency can impact all other words' saliency ranks. We, therefore, propose an iterative minimization that (i) sequentially picks a token in the sentence (one after the other, round-robin) and (ii) updates this token's saliency score into the direction of a decreased absolute bias score while keeping all other tokens' saliencies fixed. Algorithm 1 shows the complete correction procedure, Table A.15 shows the procedure's impact on an example sentence over the course of 100 optimization steps. Besides the examples shown in Table 5.6, we provide additional examples in Table A.16.

[4] The concrete $\mathbf{x}_{\text{ref}}$ corresponds to a "flat" dependency relation, a "first letter capitalized" capitalization, a display index of 129.7, a word length of 24.6, a sentence length of 4.1, a relative word frequency of 0.04, a sentiment polarity of -0.78, a normalized saliency rank of 0.11 and a word position index of 1.08. While non-integer values for, e.g. word length cannot occur in any prediction, this does not limit the utility of $\mathbf{x}_{\text{ref}}$ as the reference context as it only serves as an arbitrary, but neutral reference point.

Algorithm 1: Saliency color correction procedure.

```
Input: $s^{(i)}_{\text{orig}}$: Original saliency scores for each word of the sentence with length $l$.
Input: $\mathbf{x}_{\text{ref}}$: Feature representation of the reference input.
Output: $s^{(i)}_{\text{corr}}$: Corrected saliency scores for each word of the sentence.
$s^{(i)}_{\text{corr}} \leftarrow s^{(i)}_{\text{orig}}$ for all $i$. // Initialization
// Iterate for a fixed number of steps
for $k \leftarrow 1$ to $n_{steps}$ do
    // Each iteration goes over all tokens in the sentence
    for $i \leftarrow 1$ to $l$ do
        $\hat{\mathbf{x}} \leftarrow$ feature representation of the $i$-th word (also depends on all other $s^{(i)}_{\text{corr}}$ via the saliency rank feature)
        $p \leftarrow \frac{1}{|W||V|} \sum_{w \in W} \sum_{v \in V} u\left(s^{(i)}_{\text{corr}}, \hat{\mathbf{x}}, w, v\right)$ // Model-predicted perceived importance (on the latent continuous scale) averaged over participant IDs $W$ and sentence IDs $V$.
        $p_{\text{ref}} \leftarrow \frac{1}{|W||V|} \sum_{w \in W} \sum_{v \in V} u\left(s^{(i)}_{\text{orig}}, \mathbf{x}_{\text{ref}}, w, v\right)$ // Model-predicted perceived importance if the word would be the reference level word (in the reference level sentence).
        $b \leftarrow p - p_{\text{ref}}$ // Define bias.
        $s^{(i)}_{\text{corr}} \leftarrow s^{(i)}_{\text{corr}} - \alpha \cdot \left(1 - \frac{k-1}{n_{\text{steps}}}\right)^2 \cdot \text{sgn}(b)$ // Update saliency with quadratically-decaying step size (starting from $\alpha$) into the direction of reduced predicted bias.
        $s^{(i)}_{\text{corr}} \leftarrow \max\left(0, \min\left(s^{(i)}_{\text{corr}}, 1\right)\right)$ // Make sure we stay within $[0, 1]$.
    end
end
return $s^{(i)}_{\text{corr}}$ for all $i$.
```

Step	Saliency				Bias			
1	many	thanks	2scompany	...	many	thanks	2scompany	...
10	many	thanks	2scompany	...	many	thanks	2scompany	...
21	many	thanks	2scompany	...	many	thanks	2scompany	...
41	many	thanks	2scompany	...	many	thanks	2scompany	...
61	many	thanks	2scompany	...	many	thanks	2scompany	...
81	many	thanks	2scompany	...	many	thanks	2scompany	...
100	many	thanks	2scompany	...	many	thanks	2scompany	...

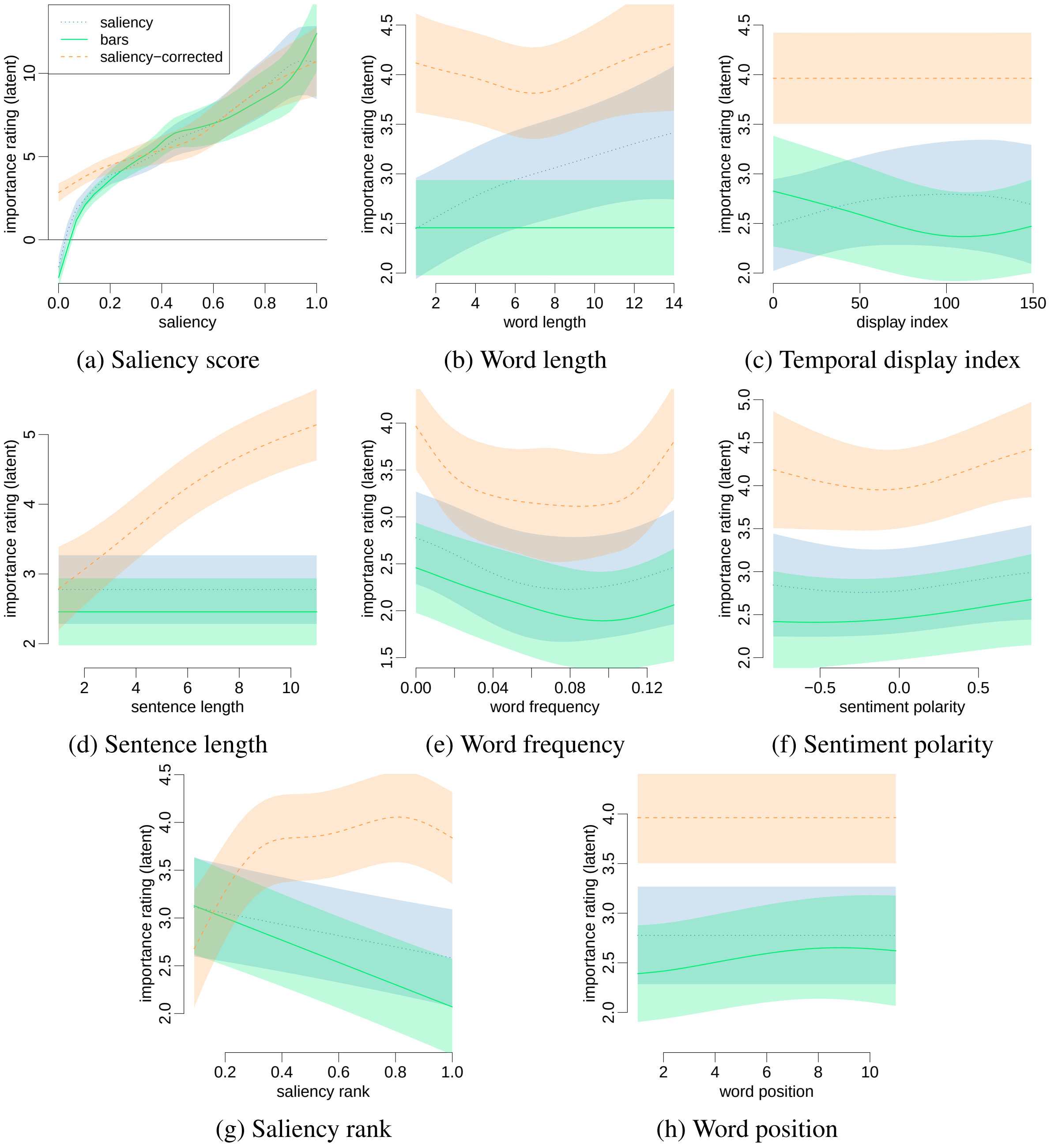
Table A.15.: Evolution of saliency scores and corresponding bias estimates across 100 optimization steps of our bias correction procedure. The first row corresponds to the initial saliency scores. The first row of the right column shows that our method predicts that the word “thanks” is perceived as overly important, while the other parts of the sentence (especially “...”) are under-perceived. After 100 steps, the saliencies of “many”, “2scompany” and “...” have been increased while the ones of “thanks” is decreased resulting in a removal of nearly all predicted bias.

A.8. Integrated Gradients and Correction Study

We report detailed estimates and test statistics regarding our third user study in Table A.17. Figure A.30 shows comparison plots for each smooth term and Figure A.31 as well as Figure A.32, visualize the respective difference functions between visualizations along with highlighted regions of significant differences. Cut points are located at -1, 0.95, 2.37, 3.67, 5.06, and 6.83.

	Saliency	Bias	Removed Bias
original	Wonderful Atmosphere	Wonderful Atmosphere	100.0%
corrected	Wonderful Atmosphere	Wonderful Atmosphere	
original	Craig and Nate are wonderful .	Craig and Nate are wonderful .	95.3%
corrected	Craig and Nate are wonderful .	Craig and Nate are wonderful .	
original	Love this place !!	Love this place !!	91.6%
corrected	Love this place !!	Love this place !!	
original	But not so .	But not so .	98.5%
corrected	But not so .	But not so .	
original	Usually very quick and timely .	Usually very quick and timely .	92.7%
corrected	Usually very quick and timely .	Usually very quick and timely .	
original	Just ask American Express	Just ask American Express	100.0%
corrected	Just ask American Express	Just ask American Express	
original	Rubbish	Rubbish	76.3%
corrected	Rubbish	Rubbish	
original	Great Manicure	Great Manicure	100.0%
corrected	Great Manicure	Great Manicure	
original	Fantastic couple of days .	Fantastic couple of days .	86.6%
corrected	Fantastic couple of days .	Fantastic couple of days .	
original	They are especially rude to women .	They are especially rude to women .	80.7%
corrected	They are especially rude to women .	They are especially rude to women .	
original	Not enough seating .	Not enough seating .	89.4%
corrected	Not enough seating .	Not enough seating .	
original	Not impressed .	Not impressed .	100.0%
corrected	Not impressed .	Not impressed .	
original	The food was incredibly bland .	The food was incredibly bland .	86.8%
corrected	The food was incredibly bland .	The food was incredibly bland .	
original	Dessert was good .	Dessert was good .	92.9%
corrected	Dessert was good .	Dessert was good .	
original	Horrible !	Horrible !	100.0%
corrected	Horrible !	Horrible !	

Table A.16.: Examples of our proposed bias reduction method. The table shows sentences along with their initial saliency scores and the respective corrected saliency scores. The *bias* column shows the color-coded bias estimates. Predicted overestimations are colored in red whereas predicted underestimations are colored in blue. For each example, we scale the range of biases to use the full color spectrum in one direction. The column *removed bias* lists how much of the predicted bias was removed in the corrected saliencies.



(a) Saliency score

(b) Word length

(c) Temporal display index

(d) Sentence length

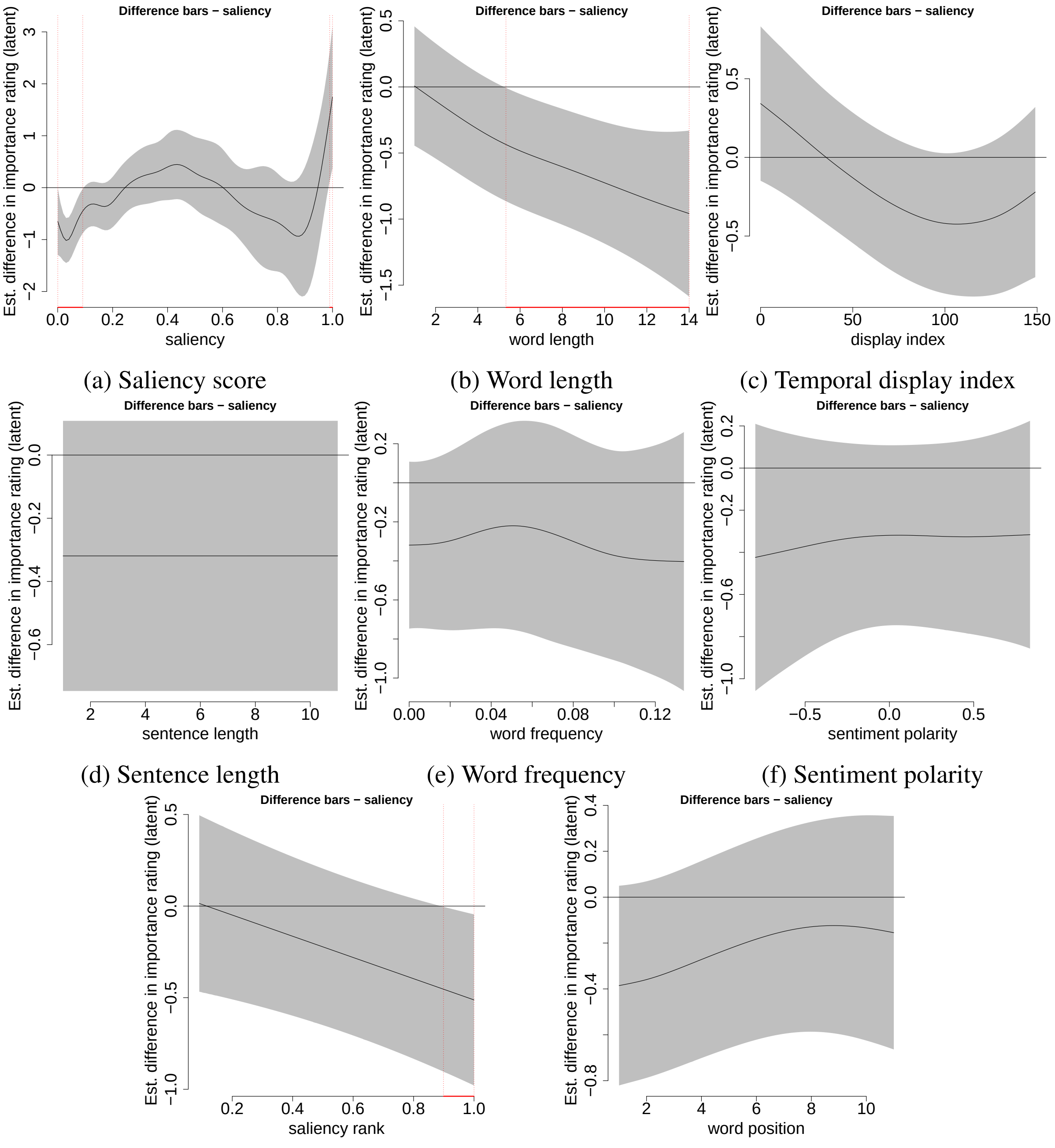
(e) Word frequency

(f) Sentiment polarity

(g) Saliency rank

(h) Word position

Figure A.30.: Summed-effects comparison plots of the correction methods.



(a) Saliency score

(b) Word length

(c) Temporal display index

(d) Sentence length

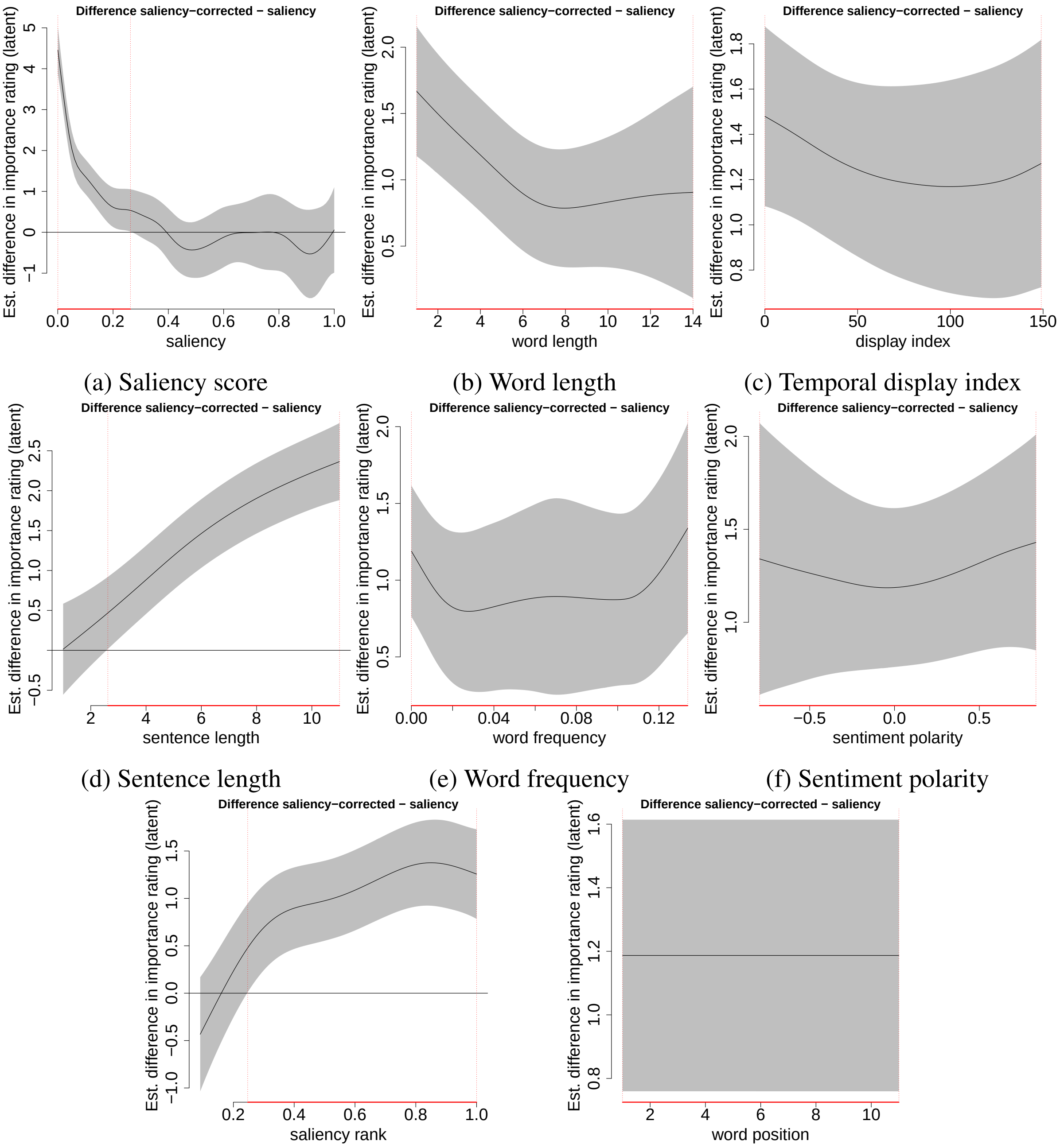
(e) Word frequency

(f) Sentiment polarity

(g) Saliency rank

(h) Word position

Figure A.31.: Difference plots between the bar visualization and the original visualization. Areas of significant differences are marked in red.



(a) Saliency score

(b) Word length

(c) Temporal display index

(d) Sentence length

(e) Word frequency

(f) Sentiment polarity

(g) Saliency rank

(h) Word position

Figure A.32.: Difference plots between the model-corrected saliencies and original saliencies. Areas of significant differences are marked in red.

Parametric Terms	β	**SE**	**t**	**p**
(Intercept)	2.1119	0.1994	10.5909	< 0.0001
bars	-0.5991	0.1578	-3.7974	0.0001
saliency-corrected	1.1102	0.2515	4.4135	< 0.0001
Smooth Terms	**edf**	**ref. df**	**F**	**p**
s(saliency):saliency	11.4304	19	283.3393	< 0.0001
s(saliency):bars	11.0767	19	321.0314	< 0.0001
s(saliency):saliency-corrected	5.5202	19	113.9321	< 0.0001
s(display index):saliency	1.4830	9	7.2492	0.2575
s(display index):bars	1.7044	9	15.3135	0.0254
s(display index):saliency-corrected	0.0009	9	0.0001	0.6438
s(word length):saliency	1.7724	9	4.1550	< 0.0001
s(word length):bars	0.0009	9	0.0001	0.3775
s(word length):saliency-corrected	2.3645	9	1.3936	0.0213
s(sentence length):saliency	0.0005	9	0.0001	0.2313
s(sentence length):bars	0.0004	9	0.0000	0.8967
s(sentence length):saliency-corrected	2.4024	9	22.4406	< 0.0001
s(word frequency):saliency	1.8086	9	2.3192	< 0.0001
s(word frequency):bars	1.7381	9	2.7043	< 0.0001
s(word frequency):saliency-corrected	2.8913	9	7.2153	< 0.0001
s(sentiment polarity):saliency	1.0751	9	0.4727	0.0633
s(sentiment polarity):bars	1.0022	9	0.5076	0.0507
s(sentiment polarity):saliency-corrected	1.6991	9	2.2243	0.0020
s(saliency rank):saliency	0.9279	9	2.0901	0.0002
s(saliency rank):bars	0.9764	9	6.5779	< 0.0001
s(saliency rank):saliency-corrected	4.1893	9	6.8094	< 0.0001
s(word position):saliency	0.0004	9	0.0000	0.9754
s(word position):bars	1.2970	9	0.7165	0.0167
s(word position):saliency-corrected	0.0005	9	0.0000	0.9615
s(capitalization):saliency	0.0009	2	0.0003	0.4268
s(capitalization):bars	0.0003	2	0.0001	0.4525
s(capitalization):saliency-corrected	1.0644	2	3.2665	0.0245
s(dependency relation):saliency	0.0057	29	0.0002	0.3443
s(dependency relation):bars	0.0010	28	0.0000	0.5819
s(dependency relation):saliency-corrected	1.4715	28	0.0731	0.1955
s(condition order):saliency	3.7653	6	30.7306	0.0044
s(condition order):bars	0.0007	6	0.0001	0.5619
s(condition order):saliency-corrected	4.4665	6	150.1092	< 0.0001
s(sentence ID)	12.7259	150	0.1028	0.2236
s(saliency,sentence ID)	68.0861	150	1.7605	< 0.0001
s(worker ID)	55.7637	59	313.9570	< 0.0001
s(saliency,worker ID)	53.3619	60	230.3436	< 0.0001

Table A.17.: Parametric and smooth coefficients of the GAMM corresponding to the third user study comparing the three visualizations.

A.9. Perceived Predictability Scale

A.9.1. Item Generation

A.9.1.1. Initial Item Pool

Our initial item pool contains 60 items. Concretely, these items are:

- "My knowledge about the system behavior is complete."
- "I know a lot about the system's behavior."
- "I do not need to learn more about the system's behavior."
- "I understand how the system functions."
- "The system behaves as expected (including 'controlled random')."
- "I can explain how the system functions."
- "I have a lot of experience with this system."
- "I observed the system's behavior in many different situations."
- "I know enough about the system to predict how it behaves."
- "Seeing more of the system's behavior will not surprise me."
- "Based on past responses, I know the responses the system will likely give me."
- "I have interacted with the system many times."
- "I am able to anticipate how the system will respond after having used it."
- "I have an understanding of the system based on its responses to the given input."
- "I have a comfortable feeling of knowing."
- "I engaged with the system a lot."
- "I have personal experiences with the system."
- "I have an educated guess on how the system will behave."
- "I have experience with the system."

- "I can identify rules and patterns in the system's decisions.
- "The system does not take random decisions."
- "The system takes consistent decisions."
- "I feel a sense of order and direction."
- "There is a consistent pattern in the system's decisions."
- "Given a fixed input, the system always takes the same decision"
- "I know the reasons for the system's decisions."
- "I know what brought the system to its decisions."
- "The system's logic is similar to mine."
- "The system's knowledge is similar to mine."
- "The system is following a certain pattern."
- "I get an idea of how the algorithm works."
- "I know how the system is likely to interpret input."
- "I feel like the results the system gives are reliable."
- "The system behaves in a predictable manner."
- "The system gives consistent results."
- "The system gives reliable results."
- "The system normally behaves in a consistent manner."
- "I know how the system will respond in a given situation."
- "I am able to predict how the system will react."
- "The system's decisions are predictable."
- "I have an understanding how the system makes its judgements."
- "The system's decision process is straightforward."

- "I know how the system ’thinks’."
- "I have a good overall understanding of the system."
- "I know the responses the system will likely give."
- "I can guess how the system will behave."
- "I know if the system is biased."
- "I know how the system was created."
- "I know the system’s quirks."
- "I have an understanding of why the systems responded in the way it did."
- "I can guess how the system will react."
- "I feel like I can predict how the system will behave."
- "I have knowledge about what the system is supposed to do."
- "I know what I can expect from the system."
- "I am certain about the system’s behavior."
- "I can guess how the system comes to its conclusions."
- "I can estimate how the system comes to its conclusions."
- "I can usually predict if the system can answer the question I have in mind for it."
- "I know what I can and cannot do with the system."
- "I know how to use the system efficiently."

A.9.1.2. Intermediate Item Pool

After the expert ratings and two rounds of cognitive interviews with target population participants, we filtered and refined the item pool used in our first large-scale evaluation to the following items. We mark items included in our final scale version with “$*$” and report reasons for our removal decisions in parentheses:

- $*$"I can tell which responses the system will likely give."
- "I can predict how the system will behave most of the time." (> 0.85 inter-item correlation to previous item)
- $*$"The system behaves in a predictable manner."
- "The system's responses are predictable for me." (strong content similarity to previous item, removed for brevity)
- "I can identify rules and patterns in the system's responses." (lower item discrimination value than all other items)
- "I understand the system well enough to predict how it behaves." (least categorizable item regarding our three-facet theory)
- $*$"There is a consistent pattern in the system's behavior."
- $*$"I can tell the reasons for the system's decisions."
- "I have an understanding of why the systems responded in the way it did." (strong similarity to previous item but more complex wording, removed for brevity)
- $*$"I observed enough system responses to predict how the system behaves."
- $*$"Based on past system responses, I know the responses the system will likely give me."
- "The number of system responses I have seen is large enough to predict the system's behavior." (> 0.85 inter-item correlation to previous item)

A.9.2. Scale Evaluation

A.9.2.1. Colored Shapes Experiment Interface

In addition to the scenario depicted in Figure 5.15, we included the scenarios depicted in Figures A.33 to A.36 as described in Section 5.2.2.

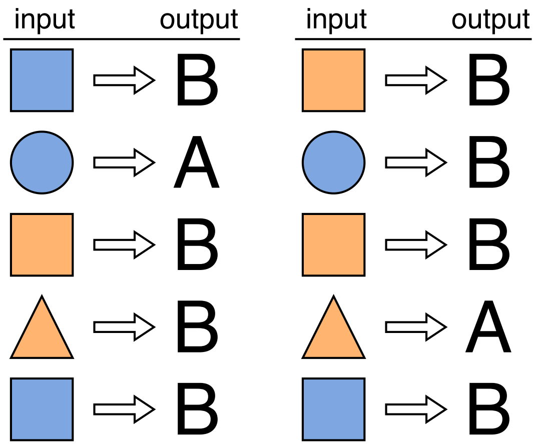


Figure A.33.: A scenario with mixed uncertainty, but slightly more aleatory uncertainty than the scenario shown in Figure 5.15 (i.e., less predictability). We refer to this scenario as MIXED-LESS-ALEATORY.

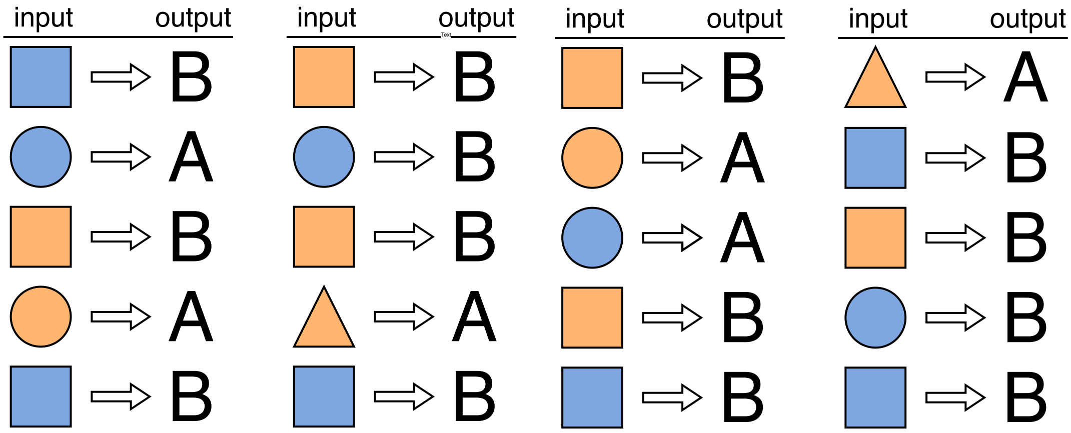


Figure A.34.: A scenario with mixed uncertainty, but twice the number of examples of the scenario shown in Figure 5.15 (i.e., more epistemic predictability). We refer to this scenario as MIXED-MORE-EPISTEMIC.

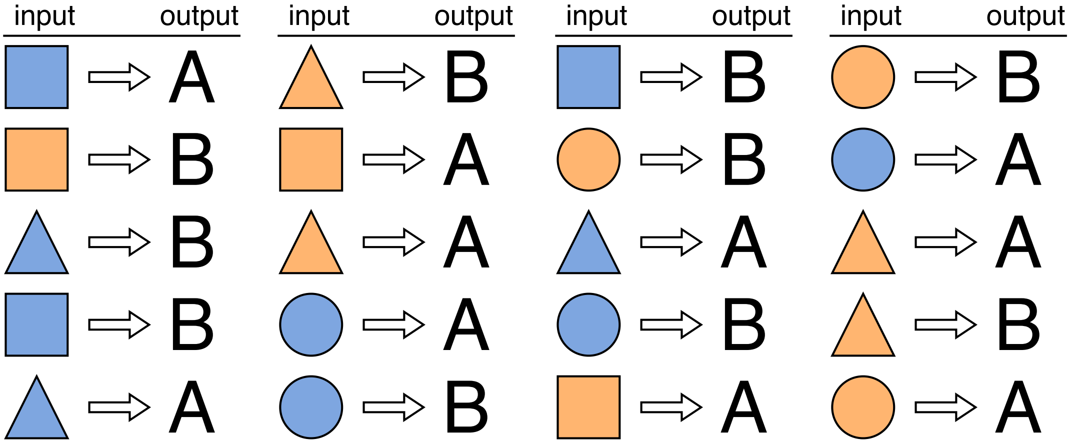


Figure A.35.: A scenario with strong aleatory uncertainty (i.e., low predictability). We refer to this scenario as LOW-ALEATORY.

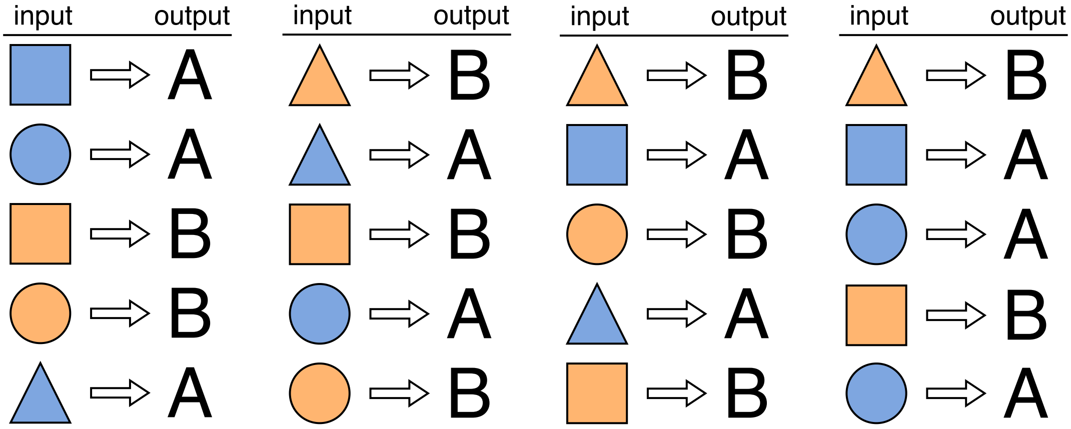


Figure A.36.: A scenario with a high degree of epistemic and aleatory certainty. We refer to this scenario as HIGH-BOTH.

A.9.2.2. Differentiation by Known Groups

Table A.18 shows details of the Tukey HSD post hoc test to determine significant differences between our scenarios.

Scenario pair	difference	CI-upper	CI-lower	p (adjusted)
HIGH-BOTH VS. LOW-ALEATORY	2.025	1.177	2.873	<**0.0001**
MIXED VS. LOW-ALEATORY	1.088	0.240	1.935	**0.005**
MIXED-LESS-ALEATORY VS. LOW-ALEATORY	0.250	-0.598	1.098	0.927
MIXED-MORE-EPISTEMIC VS. LOW-ALEATORY	1.588	0.740	2.435	<**0.0001**
MIXED VS. HIGH-BOTH	-0.938	-1.785	-0.090	**0.022**
MIXED-LESS-ALEATORY VS. HIGH-BOTH	-1.775	-2.623	-0.927	<**0.0001**
MIXED-MORE-EPISTEMIC VS. HIGH-BOTH	-0.438	-1.285	0.410	0.615
MIXED-LESS-ALEATORY VS. MIXED	-0.8375	-1.685	0.010	0.0546
MIXED-MORE-EPISTEMIC VS. MIXED	0.500	-0.348	1.348	0.484
MIXED-MORE-EPISTEMIC VS. MIXED-LESS-ALEATORY	1.338	0.490	2.185	<**0.001**

Table A.18.: Tukey HSD test result details for our PSP scores between known groups, i.e. scenarios. Pairs with significant differences are highlighted in **bold** font.

A.9.2.3. Confirmatory Factor Analysis

Table A.19 and Table A.20 show detailed parameter estimates of the one-factor and three-factor models.

LHS	op	RHS	estimate	SE	z	p	CI-lower	CI-upper
predictability	=∼	EF1	1.00	0.00			1.00	1.00
predictability	=∼	EF2	0.96	0.05	19.62	0.00	0.86	1.06
predictability	=∼	EP1	1.04	0.05	19.95	0.00	0.94	1.14
predictability	=∼	EP2	1.01	0.05	22.28	0.00	0.92	1.10
predictability	=∼	AL1	0.98	0.05	18.41	0.00	0.88	1.09
predictability	=∼	AL2	1.00	0.05	21.58	0.00	0.91	1.09
EF1	∼∼	EF1	0.48	0.06	8.11	0.00	0.36	0.59
EF2	∼∼	EF2	0.60	0.07	8.62	0.00	0.46	0.73
EP1	∼∼	EP1	0.66	0.08	8.53	0.00	0.51	0.81
EP2	∼∼	EP2	0.40	0.05	7.67	0.00	0.29	0.50
AL1	∼∼	AL1	0.77	0.09	8.89	0.00	0.60	0.95
AL2	∼∼	AL2	0.44	0.06	7.99	0.00	0.33	0.55
predictability	∼∼	predictability	2.28	0.27	8.34	0.00	1.75	2.82

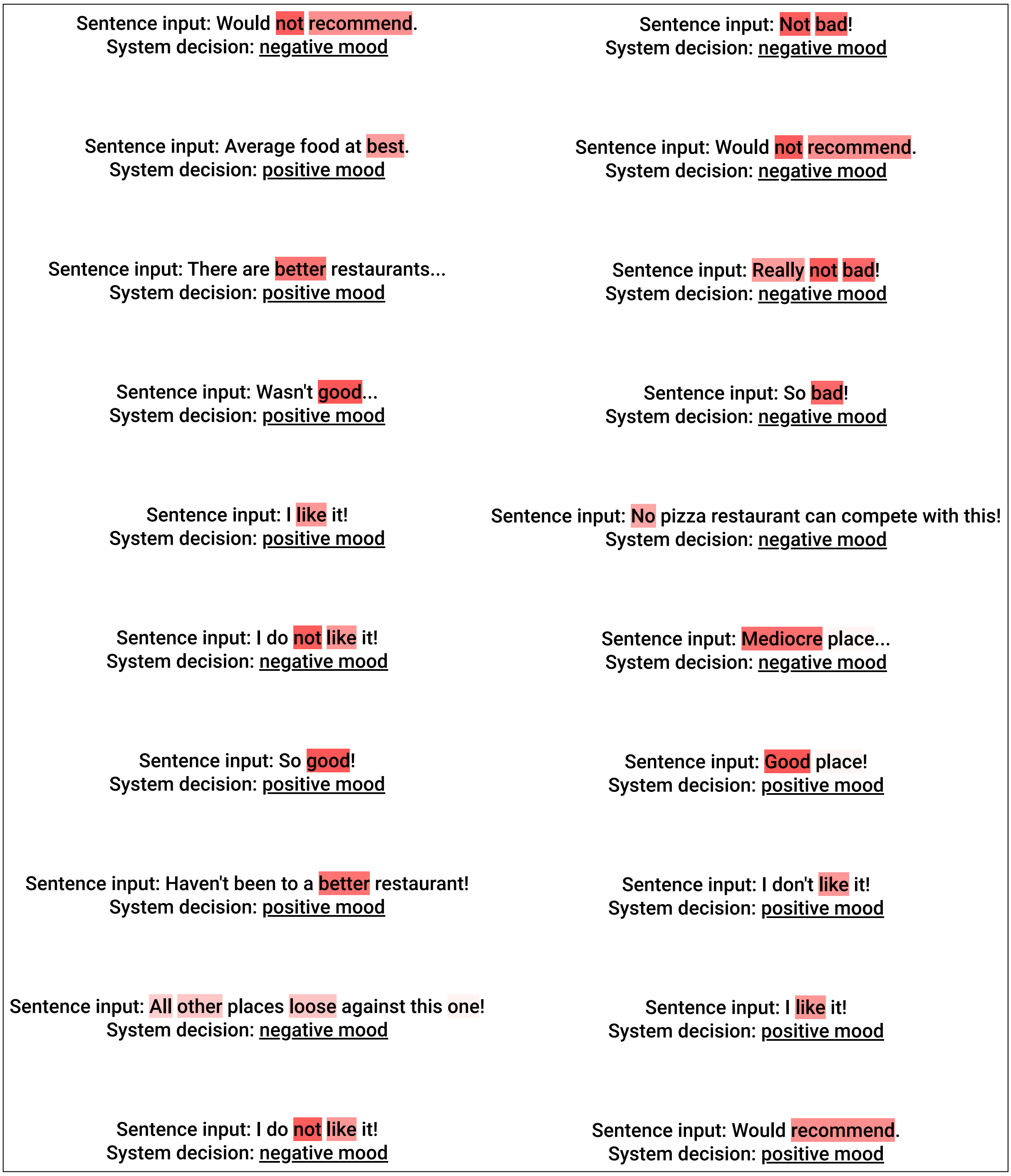
Table A.19.: Detailed parameter estimates of the one-factor model.

LHS	op	RHS	estimate	SE	z	p	CI-lower	CI-upper
effective	=∼	EF1	1.00	0.00			1.00	1.00
effective	=∼	EF2	0.96	0.05	18.99	0.00	0.86	1.06
epistemic	=∼	EP1	1.00	0.00			1.00	1.00
epistemic	=∼	EP2	0.97	0.05	21.34	0.00	0.88	1.06
aleatory	=∼	AL1	1.00	0.00			1.00	1.00
aleatory	=∼	AL2	1.02	0.05	19.11	0.00	0.92	1.13
predictability	=∼	epistemic	1.00	0.00			1.00	1.00
predictability	=∼	aleatory	0.94	0.06	16.83	0.00	0.83	1.05
predictability	=∼	effective	0.98	0.05	19.70	0.00	0.88	1.08
EF1	∼∼	EF1	0.52	0.07	7.49	0.00	0.39	0.66
EF2	∼∼	EF2	0.64	0.08	8.31	0.00	0.49	0.79
EP1	∼∼	EP1	0.63	0.08	7.96	0.00	0.47	0.78
EP2	∼∼	EP2	0.36	0.06	6.28	0.00	0.25	0.48
AL1	∼∼	AL1	0.75	0.09	8.37	0.00	0.58	0.93
AL2	∼∼	AL2	0.39	0.06	6.05	0.00	0.26	0.52
effective	∼∼	effective	-0.09	0.05	-1.89	0.06	-0.19	0.00
epistemic	∼∼	epistemic	0.07	0.05	1.28	0.20	-0.04	0.17
aleatory	∼∼	aleatory	0.06	0.05	1.14	0.26	-0.04	0.16
predictability	∼∼	predictability	2.43	0.31	7.88	0.00	1.83	3.03

Table A.20.: Detailed parameter estimates of the three-factor model.

A.9.3. Sentiment Classifier Experiments

Figure A.37 shows the full list of system prediction examples we showed to users. While Figure A.37 shows examples of the heatmap conditions, we respectively use bar charts or no explanations in the other conditions.



Sentence input: Would not recommend.
System decision: negative mood

Sentence input: Average food at best.
System decision: positive mood

Sentence input: There are better restaurants...
System decision: positive mood

Sentence input: Wasn't good...
System decision: positive mood

Sentence input: I like it!
System decision: positive mood

Sentence input: I do not like it!
System decision: negative mood

Sentence input: So good!
System decision: positive mood

Sentence input: Haven't been to a better restaurant!
System decision: positive mood

Sentence input: All other places loose against this one!
System decision: negative mood

Sentence input: I do not like it!
System decision: negative mood

Sentence input: Not bad!
System decision: negative mood

Sentence input: Would not recommend.
System decision: negative mood

Sentence input: Really not bad!
System decision: negative mood

Sentence input: So bad!
System decision: negative mood

Sentence input: No pizza restaurant can compete with this!
System decision: negative mood

Sentence input: Mediocre place...
System decision: negative mood

Sentence input: Good place!
System decision: positive mood

Sentence input: I don't like it!
System decision: positive mood

Sentence input: I like it!
System decision: positive mood

Sentence input: Would recommend.
System decision: positive mood

Figure A.37.: System predictions shown to users. Figure showing examples in the heatmap conditions, sentences were equal across conditions. Sentence order is randomized across participants.

The sentences we asked users to predict the system's decision for are (we provide the system predictions in parentheses and use *italics* to highlight wrong model decisions):

- I love the food at this place! (pos)
- Absolutely sensational! (pos)
- Quite nice! (pos)
- Tasty food! (pos)
- Super good place! (pos)
- Would not go there again. (neg)
- Pretty bad place. (neg)
- The food made me sick... (neg)
- Quite bad. (neg)
- Do not eat there! (neg)
- I don't like this restaurant very much! (*pos*)
- Wouldn't recommend. (*pos*)
- The water was the best part of the meal... (*pos*)
- Nice ads but didn't hold up the high expectations. (*pos*)
- I expected it to be better. (*pos*)
- Have not expected such a good place! (*neg*)
- I was so sad when I heard that they will close! (*neg*)
- I have not eaten at a better restaurant! (*neg*)
- Not too bad at all! (*neg*)
- Have not expected such a good place! (*neg*)

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	4.78	0.42	11.41	0.00
bar charts	-0.08	0.10	-0.78	0.44
saliency	0.22	0.11	1.90	0.06
interactivity	0.16	0.12	1.34	0.18
female	-0.18	0.42	-0.43	0.67
male	-0.13	0.42	-0.32	0.75
noise level L	0.06	0.08	0.77	0.44
noise level Q	-0.04	0.08	-0.44	0.66
bar charts : interactivity	0.09	0.17	0.55	0.58
saliency : interactivity	-0.32	0.18	-1.84	0.07

Table A.21.: Parametric terms details for our model of PSP scores.

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	0.56	0.11	5.12	0.00
bar charts	0.04	0.03	1.32	0.19
saliency	0.05	0.03	1.53	0.13
interactivity	0.02	0.03	0.74	0.46
female	0.10	0.11	0.88	0.38
male	0.11	0.11	0.97	0.33
noise level L	-0.12	0.02	-5.88	0.00
noise level Q	0.05	0.02	2.57	0.01
bar charts : interactivity	-0.04	0.04	-0.96	0.34
saliency : interactivity	0.01	0.05	0.29	0.77

Table A.22.: Parametric terms details for our model of prediction correctness scores.

	df	F	p
explanation format	2.00	5.15	**0.01**
interactivity	1.00	0.12	0.73
explanation format:interactivity	2.00	7.44	**0.00**
noise level	2.00	2.66	0.07
identification	2.00	0.11	0.90

Table A.23.: Wald tests for the parametric terms in our model of FOST trust scores.

	edf	Ref.df	F	p
s(prediction correctness)	2.58	9.00	3.38	**0.00**
s(PSP)	0.98	9.00	5.41	**0.00**
s(completion time)	0.84	9.00	0.57	**0.01**
s(SIPA)	1.01	9.00	0.18	0.18
s(NFC)	2.32	9.00	1.38	**0.00**
s(age)	0.50	9.00	0.11	0.16

Table A.24.: Wald tests for the smooth terms in our model of FOST trust scores.

Bibliography

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I. J., Hardt, M., and Kim, B. (2018). Sanity Checks for Saliency Maps. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9525–9536.

Aïvodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., and Tapp, A. (2019). Fairwashing: the risk of rationalization. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 161–170. PMLR.

Arora, S., Pruthi, D., Sadeh, N. M., Cohen, W. W., Lipton, Z. C., and Neubig, G. (2021). Explain, edit, and understand: Rethinking user study design for evaluating model explanations. *CoRR*, abs/2112.09669.

Arras, L., Horn, F., Montavon, G., Müller, K.-R., and Samek, W. (2017). " What is relevant in a text document?": An interpretable machine learning approach. *PloS one*, 12(8):e0181142. Publisher: Public Library of Science San Francisco, CA USA.

Arrieta, A. B., Rodríguez, N. D., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*, 58:82–115.

Asaro, P. M. (2016). The liability problem for autonomous artificial agents. In *2016 AAAI Spring Symposia, Stanford University, Palo Alto, California, USA, March 21-23, 2016*. AAAI Press.

Atanasova, P., Simonsen, J. G., Lioma, C., and Augenstein, I. (2020). A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Banino, A., Balaguer, J., and Blundell, C. (2021). Pondernet: Learning to ponder. *CoRR*, abs/2107.05407.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. S. (2021). Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In Kitamura, Y., Quigley, A., Isbister, K., Igarashi, T., Bjørn, P., and Drucker, S. M., editors, *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, pages 81:1–81:16. ACM.

Bargas-Avila, J. A. and Brühlmann, F. (2016). Measuring user rated language quality: development and validation of the user interface language quality survey (lqs). *International Journal of Human-Computer Studies*, 86:1–10.

Barra, A., Beccaria, M., and Fachechi, A. (2018). A new mechanical approach to handle generalized Hopfield neural networks. *Neural Networks*, 106:205–222.

Bartneck, C., Kulic, D., Croft, E. A., and Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robotics*, 1(1):71–81.

Bastings, J. and Filippova, K. (2020). The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online. Association for Computational Linguistics.

Bauer, L., Deng, L., and Bansal, M. (2021). ERNIE-NLI: Analyzing the Impact of Domain-Specific External Knowledge on Enhanced Representations for NLI. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 58–69, Online. Association for Computational Linguistics.

Bausell, R. B. and Li, Y.-F. (2002). *Power analysis for experimental research: a practical guide for the biological, medical and social sciences*. Cambridge University Press.

Beatty, P. C. and Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public opinion quarterly*, 71(2):287–311.

Beigman Klebanov, B. and Beigman, E. (2009). Squibs: From annotator agreement to noise models. *Computational Linguistics*, 35(4):495–503.

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The Long-Document Transformer. *CoRR*, abs/2004.05150. arXiv: 2004.05150.

Belz, A., Mille, S., and Howcroft, D. M. (2020). Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.

Belz, A. and Reiter, E. (2006). Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320, Trento, Italy. Association for Computational Linguistics.

Bender, R. and Lange, S. (2001). Adjusting for multiple testing—when and how? *Journal of clinical epidemiology*, 54(4):343–349.

Bevan, G. and Hood, C. (2006). What's measured is what matters: Targets and gaming in the English public health care system. *Public Administration*, 84(3):517–538. Publisher: John Wiley & Sons, Ltd.

Bhagavatula, C., Bras, R. L., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., Downey, D., Yih, W.-t., and Choi, Y. (2020). Abductive Commonsense Reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., and Eckersley, P. (2020). Explainable machine learning in deployment. In Hildebrandt, M., Castillo, C., Celis, L. E., Ruggieri, S., Taylor, L., and Zanfir-Fortuna, G., editors, *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 648–657. ACM.

Biran, O. and Cotton, C. V. (2017). Explanation and justification in machine learning : A survey or.

Biran, O. and McKeown, K. R. (2017). Human-centric justification of machine learning predictions. In Sierra, C., editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 1461–1467. ijcai.org.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. (2004). Confidence Estimation for Machine Translation. In *COLING 2004:*

Proceedings of the 20th International Conference on Computational Linguistics, pages 315–321, Geneva, Switzerland. COLING.

Bless, H. and Burger, A. M. (2016). Assimilation and contrast in social priming. *Current Opinion in Psychology*, 12:26–31. Social priming.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., and Young, S. L. (2018). Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Frontiers in Public Health*, 6:149.

Bojar, O., Federmann, C., Haddow, B., Koehn, P., Post, M., and Specia, L. (2016). Ten years of wmt evaluation campaigns: Lessons learnt. In *Proceedings of the LREC 2016 Workshop "Translation Evaluation–From Fragmented Tools and Data Sets to an Integrated Ecosystem*, pages 27–34.

Boonstra, E. A. and Slagter, H. A. (2019). The dialectics of free energy minimization. *Frontiers in systems neuroscience*, 13:42. Publisher: Frontiers.

Borgatta, E. F. and Bohrnstedt, G. W. (1980). Level of measurement: Once over again. *Sociological Methods & Research*, 9(2):147–160.

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., and Choi, Y. (2019). COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Bowling, N. A., Huang, J. L., Brower, C. K., and Bragg, C. B. (2021). The quick and the careless: The construct validity of page time as a measure of insufficient effort responding to surveys. *Organizational Research Methods*, 26:323 – 352.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs. *Biometrika*.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101.

Brooke, J. (1996). Sus: a "quick and dirty'usability. *Usability evaluation in industry*, page 189.

Buçinca, Z., Lin, P., Gajos, K. Z., and Glassman, E. L. (2020). Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In Paternò, F., Oliver, N., Conati, C., Spano, L. D., and Tintarev, N., editors, *IUI '20: 25th International Conference on Intelligent User Interfaces, Cagliari, Italy, March 17-20, 2020*, pages 454–464. ACM.

Buçinca, Z., Malaya, M. B., and Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in ai-assisted decision-making. *Proc. ACM Hum. Comput. Interact.*, 5(CSCW1):188:1–188:21.

Buckley, R. C. (2006). Choosing and using statistics: A biologist's guide, 2nd edition. *Austral Ecology*, 31:425–425.

Burkart, N. and Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *J. Artif. Intell. Res.*, 70:245–317.

Busemeyer, J. R. and Diederich, A. (2010). *Cognitive modeling*. Sage.

Bussone, A., Stumpf, S., and O'Sullivan, D. (2015). The role of explanations on trust and reliance in clinical decision support systems. In Balakrishnan, P., Srivatsava, J., Fu, W., Harabagiu, S. M., and Wang, F., editors, *2015 International Conference on Healthcare Informatics, ICHI 2015, Dallas, TX, USA, October 21-23, 2015*, pages 160–169. IEEE Computer Society.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Cambazoglu, B. B., Sanderson, M., Scholer, F., and Croft, W. B. (2020). A review of public datasets in question answering research. *SIGIR Forum*, 54(2):5:1–5:23.

Camburu, O., Rocktäschel, T., Lukasiewicz, T., and Blunsom, P. (2018). e-snli: Natural language inference with natural language explanations. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9560–9572.

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and program planning*, 2(1):67–90. Publisher: Elsevier.

Cardoso, W., Smith, G., and Garcia Fuentes, C. (2015). Evaluating text-to-speech synthesizers. In *Critical CALL–Proceedings of the 2015 EUROCALL Conference, Padova, Italy*, pages 108–113. Research-publishing. net.

Carifio, J. and Perla, R. (2008). Resolving the 50-year debate around using and misusing likert scales. *Medical education*, 42(12):1150–1152.

Carpinella, C. M., Wyman, A. B., Perez, M. A., and Stroessner, S. J. (2017). The Robotic

Social Attributes Scale (RoSAS): Development and Validation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 254–262, Vienna Austria. ACM.

Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832.

Casper, W., Edwards, B. D., Wallace, J. C., Landis, R. S., Fife, D. A., et al. (2020). Selecting response anchors with equal intervals for summated rating scales. *Journal of Applied Psychology*, 105(4):390.

Chaganty, A., Paranjape, A., Liang, P., and Manning, C. D. (2017). Importance sampling for unbiased on-demand evaluation of knowledge base population. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1038–1048, Copenhagen, Denmark. Association for Computational Linguistics.

Charness, G., Gneezy, U., and Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior and Organization*, 81:1–8.

Chen, A., Stanovsky, G., Singh, S., and Gardner, M. (2019). Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124.

Chen, Q., Zhu, X., Ling, Z.-H., Wei, S., Jiang, H., and Inkpen, D. (2017). Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Krishnapuram, B., Shah, M., Smola, A. J., Aggarwal, C. C., Shen, D., and Rastogi, R., editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 785–794. ACM.

Chen, Y., Huang, S., Wang, F., Cao, J., Sun, W., and Wan, X. (2018). Neural maximum subgraph parsing for cross-domain semantic dependency analysis. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 562–572, Brussels, Belgium. Association for Computational Linguistics.

Cheng, H. F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., and Zhu, H. (2019). Explaining decision-making algorithms through UI: strategies to help non-expert stakeholders. In Brewster, S. A., Fitzpatrick, G., Cox, A. L., and Kostakos, V., editors, *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 559. ACM.

Cheng, J., Dong, L., and Lapata, M. (2016). Long short-term memory-networks for machine reading. In Su, J., Carreras, X., and Duh, K., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 551–561. The Association for Computational Linguistics.

Chicco, D. and Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21.

Choi, J. W., Hecht, G. W., and Tayler, W. B. (2012). Lost in translation: The effects of incentive compensation on strategy surrogation. *The Accounting Review*, 87(4):1135–1163.

Choi, J. W., Hecht, G. W., and Tayler, W. B. (2013). Strategy selection, surrogation, and strategic performance measurement systems. *Journal of Accounting Research*, 51(1):105–133.

Chomsky, N. (1957). *Syntactic Structures*. De Gruyter Mouton, Berlin, Boston.

Chromik, M. and Schuessler, M. (2020). A taxonomy for human subject evaluation of black-box explanations in XAI. In Smith-Renner, A., Kleanthous, S., Lim, B. Y., Kuflik, T., Stumpf, S., Otterbacher, J., Sarkar, A., Dugan, C., and Tal, A. S., editors, *Proceedings of the Workshop on Explainable Smart Systems for Algorithmic Transparency in Emerging Technologies co-located with 25th International Conference on Intelligent User Interfaces (IUI 2020), Cagliari, Italy, March 17, 2020*, volume 2582 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Chu, E., Roy, D., and Andreas, J. (2020). Are visual explanations useful? A case study in model-in-the-loop prediction. *CoRR*, abs/2007.12248.

Clark, R., Silén, H., Kenter, T., and Leith, R. (2019). Evaluating long-form text-to-speech: Comparing the ratings of sentences and paragraphs. *CoRR*, abs/1909.03965.

Clinciu, M.-A., Eshghi, A., and Hastie, H. (2021). A study of automatic metrics for the evaluation of natural language explanations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2376–2387, Online. Association for Computational Linguistics.

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Cohen, J. (1988). The effect size index: d. *Statistical power analysis for the behavioral sciences*, 2(1).

Colin, J., Fel, T., Cadene, R., and Serre, T. (2021). What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods.

Colquhoun, D. (1971). *Lectures on biostatistics: an introduction to statistics with applications in biology and medicine*. David Colquhoun.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the*

2018 Conference on Empirical Methods in Natural Language Processing, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Corbière, C., Thome, N., Bar-Hen, A., Cord, M., and Pérez, P. (2019). Addressing Failure Prediction by Learning Model Confidence. In Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F. d., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Corder, G. W. and Foreman, D. I. (2014). *Nonparametric statistics: A step-by-step approach.* John Wiley & Sons.

Courtois, C. and Timmermans, E. (2018). Cracking the Tinder Code: An Experience Sampling Approach to the Dynamics and Impact of Platform Governing Algorithms. *Journal of Computer-Mediated Communication*, 23(1):1–16.

Covington, P., Adams, J., and Sargin, E. (2016). Deep neural networks for youtube recommendations. In Sen, S., Geyer, W., Freyne, J., and Castells, P., editors, *Proceedings of the 10th ACM Conference on Recommender Systems, Boston, MA, USA, September 15-19, 2016*, pages 191–198. ACM.

Cramer, H. S. M., Evers, V., Ramlal, S., van Someren, M., Rutledge, L., Stash, N., Aroyo, L., and Wielinga, B. J. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Model. User Adapt. Interact.*, 18(5):455–496.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16:297–334.

Czerwinski, M., Horvitz, E., and Cutrell, E. (2001). Subjective duration assessment: An implicit probe for software usability. In *Proceedings of IHM-HCI 2001 conference*, volume 2, pages 167–170.

Dacey, M. (2017). Anthropomorphism as cognitive bias. *Philosophy of Science*, 84(5):1152–1164.

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., and Sen, P. (2020). A survey of the state of explainable AI for natural language processing. In Wong, K., Knight, K., and Wu, H., editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020*, pages 447–459. Association for Computational Linguistics.

Darling, K. (2015). 'who's johnny?' anthropomorphic framing in human-robot interaction, integration, and policy. *SSRN Electronic Journal*.

de Holanda Coelho, G. L., Hanel, P. H. P., and Wolf, L. J. (2018). The very efficient assessment of need for cognition: Developing a six-item version*. *Assessment*, 27:1870 – 1885.

De Winter, J. and Dodou, D. (2010). Five-point likert items: t test versus mann-whitney-wilcoxon (addendum added october 2012). *Practical Assessment, Research, and Evaluation*, 15(1):11.

Dean, A., Voss, D., Draguljić, D., et al. (1999). *Design and analysis of experiments*, volume 1. Springer.

Deegan, J. P. (1978). On the occurrence of standardized regression coefficients greater than one. *Educational and Psychological Measurement*, 38:873 – 888.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society.

Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., and Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences*, 116(39):19251–19257.

DeVellis, R. F. and Thorpe, C. T. (2021). *Scale development: Theory and applications*. Sage publications.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

DeVries, T. and Taylor, G. W. (2018). Learning Confidence for Out-of-Distribution Detection in Neural Networks. *CoRR*, abs/1802.04865. _eprint: 1802.04865.

DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., and Wallace, B. C. (2020). ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Diedenhofen, B. and Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE*, 10.

Dinu, J., Bigham, J. P., and Kolter, J. Z. (2020). Challenging common interpretability assumptions in feature attribution explanations. *CoRR*, abs/2012.02748. arXiv: 2012.02748.

Divjak, D. and Baayen, H. (2017). Ordinal GAMMs: a new window on human ratings. In *Each venture, a new beginning: Studies in Honor of Laura A. Janda*, pages 39–56. Slavica Publishers.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR*, abs/2010.11929. _eprint: 2010.11929.

Dras, M. (2015). Squibs: Evaluating human pairwise preference judgments. *Computational Linguistics*, 41(2):309–317.

Dunn, K. and McCray, G. (2020). The place of the bifactor model in confirmatory factor analysis investigations into construct dimensionality in language testing. *Frontiers in Psychology*, 11.

Dunn, O. J. and Clark, V. A. (1969). Correlation coefficients measured on the same individuals. *Journal of the American Statistical Association*, 64:366–377.

Dušek, O., Sevegnani, K., Konstas, I., and Rieser, V. (2019). Automatic quality estimation for natural language generation: Ranting (jointly rating and ranking). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 369–376, Tokyo, Japan. Association for Computational Linguistics.

Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I., Muller, M. J., and Riedl, M. O. (2021). The who in explainable AI: how AI background shapes perceptions of AI explanations. *CoRR*, abs/2107.13509.

Ehsan, U. and Riedl, M. O. (2021). Explainability pitfalls: Beyond dark patterns in explainable AI. *CoRR*, abs/2109.12480.

Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., and Riedl, M. O. (2019). Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In Fu, W., Pan, S., Brdiczka, O., Chau, P., and Calvary, G., editors, *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI 2019, Marina del Ray, CA, USA, March 17-20, 2019*, pages 263–274. ACM.

Eisinga, R., te Grotenhuis, M., and Pelzer, B. (2013). The reliability of a two-item scale: Pearson, cronbach, or spearman-brown? *International Journal of Public Health*, 58:637–642.

Eisler, H. (1976). Experiments on subjective duration 1868-1975: A collection of power function exponents. *Psychological Bulletin*, 83(6):1154.

Endsley, M. R. (1988). Situation awareness global assessment technique (sagat). In *Proceedings of the IEEE 1988 national aerospace and electronics conference*, pages 789–795. IEEE.

Epley, N., Waytz, A., and Cacioppo, J. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4):864–886.

Ethayarajh, K. and Jurafsky, D. (2020). Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational

Linguistics.

Evans, J. S. B., Barston, J. L., and Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & cognition*, 11(3):295–306. Publisher: Springer.

Faldu, K., Sheth, A. P., Kikani, P., and Akabari, H. (2021). KI-BERT: infusing knowledge context for better language and domain understanding. *CoRR*, abs/2104.08145.

Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J., and Auli, M. (2019). ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Fang, Y., Sun, S., Gan, Z., Pillai, R., Wang, S., and Liu, J. (2020). Hierarchical graph network for multi-hop question answering. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8823–8838. Association for Computational Linguistics.

Fasiolo, M., Nedellec, R., Goude, Y., and Wood, S. N. (2020). Scalable visualization methods for modern generalized additive models. *Journal of computational and Graphical Statistics*, 29(1):78–86.

Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (2009). Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4):1149–1160.

Fel, T., Cadène, R., Chalvidal, M., Cord, M., Vigouroux, D., and Serre, T. (2021a). Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 26005–26014.

Fel, T., Colin, J., Cadène, R., and Serre, T. (2021b). What I cannot predict, I do not understand: A human-centered evaluation framework for explainability methods. *CoRR*, abs/2112.04417.

Feng, S. and Boyd-Graber, J. L. (2019). What can AI do for me?: evaluating machine learning interpretations in cooperative play. In Fu, W., Pan, S., Brdiczka, O., Chau, P., and Calvary, G., editors, *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI 2019, Marina del Ray, CA, USA, March 17-20, 2019*, pages 229–239. ACM.

Field, A. and Hole, G. (2002). *How to design and report experiments*. Sage.

Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, 22(5):323–327.

Finstad, K. (2013). Response to commentaries on 'The Usability Metric for User Experience'.

Interact. Comput., 25(4):327–330.

Ford, C., Kenny, E. M., and Keane, M. T. (2020). Play MNIST for me! user studies on the effects of post-hoc, example-based explanations & error rates on debugging a deep learning, black-box classifier. *CoRR*, abs/2009.06349.

Förster, J., Liberman, N., and Kuschel, S. (2008). The effect of global versus local processing styles on assimilation versus contrast in social judgment. *Journal of personality and social psychology*, 94(4):579.

Fox, C. R. and Ülkümen, G. (2011). Distinguishing two dimensions of uncertainty. *Fox, Craig R. and Gülden Ülkümen (2011),"Distinguishing Two Dimensions of Uncertainty," in Essays in Judgment and Decision Making, Brun, W., Kirkebøen, G. and Montgomery, H., eds. Oslo: Universitetsforlaget.*

Fox, J. (2003). Effect Displays in R for Generalised Linear Models. *Journal of Statistical Software*, 8(15).

Franke, T., Trantow, M., Günther, M., Krems, J. F., Zott, V., and Keinath, A. (2015). Advancing electric vehicle range displays for enhanced user experience: the relevance of trust and adaptability. In Burnett, G. E., Gabbard, J. L., Green, P. A., and Osswald, S., editors, *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI 2015, Nottingham, United Kingdom, September 1-3, 2015*, pages 249–256. ACM.

Frantzi, K. T., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms:. the c-value/nc-value method. *International Journal on Digital Libraries*, 3:115–130.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, 29(5):1189–1232.

Furr, R. M. (2022). *Psychometrics: an introduction*. SAGE publications.

Gage, P. (1994). A new algorithm for data compression. *The C Users Journal archive*, 12:23–38.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Balcan, M.-F. and Weinberger, K. Q., editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1050–1059. JMLR.org.

Gatti, L., Guerini, M., and Turchi, M. (2016). SentiWords: Deriving a High Precision and High Coverage Lexicon for Sentiment Analysis. *IEEE Trans. Affect. Comput.*, 7(4):409–421.

Gaudio, R., Burchardt, A., and Branco, A. (2016). Evaluating machine translation in a usage scenario. In *Proceedings of the Tenth International Conference on Language Resources* *and Evaluation (LREC'16)*, pages 1–8, Portorož, Slovenia. European Language Resources Association (ELRA).

Gharibshah, Z. and Zhu, X. (2022). User response prediction in online advertising. *ACM Comput. Surv.*, 54(3):64:1–64:43.

Gonzalez, A. V., Rogers, A., and Søgaard, A. (2021). On the interaction of belief bias and explanations. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2930–2942. Association for Computational Linguistics.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Goodhart, C. (1975). Problems of monetary management: The U.K. experience. *Papers in monetary economics 1975*, 1:1–20.

Goyal, P., Dollár, P., Girshick, R. B., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. (2017). Accurate, large minibatch SGD: training imagenet in 1 hour. *CoRR*, abs/1706.02677.

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., and Lee, S. (2019). Counterfactual visual explanations. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2376–2384. PMLR.

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2013). Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41.

Graves, A. (2016). Adaptive computation time for recurrent neural networks. *CoRR*, abs/1603.08983.

Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., and Toombs, A. L. (2018). The dark (patterns) side of UX design. In Mandryk, R. L., Hancock, M., Perry, M., and Cox, A. L., editors, *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*, page 534. ACM.

Green, B. and Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. *Proc. ACM Hum. Comput. Interact.*, 3(CSCW):50:1–50:24.

Greis, M., Karolus, J., Schuff, H., Wozniak, P. W., and Henze, N. (2017a). Detecting uncertain input using physiological sensing and behavioral measurements. In Henze, N., Wozniak, P. W., Väänänen, K., Williamson, J. R., and Schneegass, S., editors, *Proceedings of the*

16th International Conference on Mobile and Ubiquitous Multimedia, MUM 2017, Stuttgart, Germany, November 26 - 29, 2017, pages 299–304. ACM.

Greis, M., Schuff, H., Kleiner, M., Henze, N., and Schmidt, A. (2017b). Input Controls for Entering Uncertain Data: Probability Distribution Sliders. In *Proceedings of the ACM on Human-Computer Interaction*, volume 1, New York, NY, USA. Association for Computing Machinery. Issue: EICS.

Grimsley, C., Mayfield, E., and R.S. Bursten, J. (2020). Why attention is not explanation: Surgical intervention and causal reasoning about neural models. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1780–1790, Marseille, France. European Language Resources Association.

Grudin, J. and MacLean, A. (1985). Adapting a psychophysical method to measure performance and preference tradeoffs in human-computer interaction. In *Human-Computer Interaction - INTERACT '84*, pages 737–741. Elsevier Science Publishers B.V. (North-Holland).

Grundkiewicz, R., Junczys-Dowmunt, M., and Gillian, E. (2015). Human evaluation of grammatical error correction systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 461–470, Lisbon, Portugal. Association for Computational Linguistics.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5):93:1–93:42.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Hand, D. and Christen, P. (2018). A note on using the f-measure for evaluating record linkage algorithms. *Stat. Comput.*, 28(3):539–547.

Harpe, S. E. (2015). How to analyze likert and other rating scale data. *Currents in pharmacy teaching and learning*, 7(6):836–850.

Hart, S. G. and Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier.

Hartzog, W. (2015). Unfair and deceptive robots. *Maryland Law Review*, 74:785.

Hase, P. and Bansal, M. (2020). Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online. Association for Computational Linguistics.

Hashimoto, T. B., Zhang, H., and Liang, P. (2019). Unifying human and statistical evaluation for natural language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1689–1701, Minneapolis, Minnesota. Association for Computational Linguistics.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, volume 43. CRC press.

Hempel, C. G. and Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of science*, 15(2):135–175.

Herbrich, R., Minka, T., and Graepel, T. (2006). Trueskill™: A bayesian skill rating system. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems*, volume 19, Vancouver, Canada. MIT Press.

Herlocker, J. L., Konstan, J. A., and Riedl, J. (2000). Explaining collaborative filtering recommendations. In Kellogg, W. A. and Whittaker, S., editors, *CSCW 2000, Proceeding on the ACM 2000 Conference on Computer Supported Cooperative Work, Philadelphia, PA, USA, December 2-6, 2000*, pages 241–250. ACM.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, 107(1):65.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558. Publisher: National Acad Sciences.

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies*, 64(2):79–102.

Howcroft, D. M., Belz, A., Clinciu, M.-A., Gkatzia, D., Hasan, S. A., Mahamood, S., Mille, S., van Miltenburg, E., Santhanam, S., and Rieser, V. (2020). Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Howcroft, D. M. and Rieser, V. (2021). What happens if you treat ordinal ratings as interval data? human evaluations in NLP are even more under-powered than you think. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8932–8939, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hsieh, H.-F. and Shannon, S. E. (2005). Three approaches to qualitative content analysis.

Qualitative health research, 15(9):1277–1288.

Hu, L. and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis : Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6:1–55.

Huang, L., Le Bras, R., Bhagavatula, C., and Choi, Y. (2019). Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 329–338.

Iskender, N., Polzehl, T., and Möller, S. (2021). Reliability of human evaluation for text summarization: Lessons learned and challenges ahead. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 86–96, Online. Association for Computational Linguistics.

Iskender, N., Polzehl, T., and Möller, S. (2020). Best Practices for Crowd-based Evaluation of German Summarization: Comparing Crowd, Expert and Automatic Evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 164–175, Online. Association for Computational Linguistics.

Iyengar, S. (1993). *Explorations in political psychology*. Duke University Press.

Jacovi, A., Bastings, J., Gehrmann, S., Goldberg, Y., and Filippova, K. (2022). Diagnosing AI explanation methods with folk concepts of behavior. *CoRR*, abs/2201.11239.

Jacovi, A. and Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Jacovi, A. and Goldberg, Y. (2021). Aligning faithful interpretations with their social attribution. *Transactions of the Association for Computational Linguistics*, 9:294–310.

Jacovi, A., Schuff, H., Adel, H., Vu, N. T., and Goldberg, Y. (2023). Neighboring words affect human interpretation of saliency explanations. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics.

Jain, S. and Wallace, B. C. (2019). Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Jaki, S. and Smedt, T. D. (2019). Right-wing german hate speech on twitter: Analysis and automatic detection. *ArXiv*, abs/1910.07518.

Jamieson, S. (2004). Likert scales: How to (ab) use them? *Medical education*, 38(12):1217–1218.

Jannach, D. and Bauer, C. (2020). Escaping the McNamara fallacy: Towards more impactful recommender systems research. *AI Magazine*, 41(4):79–95. Section: Articles.

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*. Springer Science & Business Media.

Jijkoun, V. and de Rijke, M. (2005). Recognizing textual entailment using lexical similarity. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, pages 73–76.

Johnson, D. K. (2018). *Anthropomorphic Bias*, chapter 69, pages 305–307. John Wiley & Sons, Ltd.

Jöreskog, K. G. (1999). How large can a standardized coefficient be.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., and Rosseel, Y. (2022). *`semTools`: Useful tools for structural equation modeling*. R package version 0.5-6.

Joshi, M., Choi, E., Weld, D., and Zettlemoyer, L. (2017). TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Ju, Y., Zhang, Y., Liu, K., and Zhao, J. (2022). Generating hierarchical explanations on text classification without connecting rules. *CoRR*, abs/2210.13270.

Jurafsky, D. and Martin, J. H. (2023). Speech and language processing, 3rd edition.

Kadioglu, S. and Sellmann, M. (2009). Dialectic search. In *International Conference on Principles and Practice of Constraint Programming*, pages 486–500. Springer.

Kahneman, D. and Frederick, S. (2002). Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In Gilovich, T., Griffin, D., and Kahneman, D., editors, *Heuristics and Biases: The Psychology of Intuitive Judgment*, pages 49–81. Cambridge University Press.

Kang, H. S., Ji, J., Yun, Y., and Han, K. H. (2021). Estimating bar graph averages: Overcoming within-the-bar bias. *i-Perception*, 12.

Kaplan, A. (1964). The Conduct of inquiry: Methodology for behavioral science; Chandler Pub. *Co.: San Francisco, CA, USA*.

Katupitiya, J. and Gock, K. (2005). Neural network based iterative prediction of multivariable processes. In *IEEE International Conference Mechatronics and Automation, 2005*, volume 4, pages 2043–2048. IEEE.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., and Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA. Association for Computing Machinery.

Kaya, Y., Hong, S., and Dumitras, T. (2019). Shallow-Deep Networks: Understanding and Mitigating Network Overthinking. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 3301–3310. PMLR.

Kayser, M., Camburu, O., Salewski, L., Emde, C., Do, V., Akata, Z., and Lukasiewicz, T. (2021). e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1224–1234. IEEE.

Kennedy-Shaffer, L. (2019). Before $p < 0.05$ to beyond $p < 0.05$: Using history to contextualize p-values and significance testing. *The American Statistician*, 73:82 – 90.

Khurana, A., Alamzadeh, P., and Chilana, P. K. (2021). Chatrex: Designing explainable chatbot interfaces for enhancing usefulness, transparency, and trust. In Harms, K. J., Cunha, J., Oney, S., and Kelleher, C., editors, *IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC 2021, St Louis, MO, USA, October 10-13, 2021*, pages 1–11. IEEE.

Kim, J. and André, E. (2008). Emotion recognition based on physiological changes in music listening. *IEEE transactions on pattern analysis and machine intelligence*, 30(12):2067–2083.

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. (2019). The (Un)reliability of Saliency Methods. In Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R., editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*, pages 267–280. Springer.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S. (2017). Self-Normalizing Neural Networks. In Guyon, I., Luxburg, U. v., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December*

4-9, 2017, Long Beach, CA, USA, pages 971–980.

Kočiský, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K. M., Melis, G., and Grefenstette, E. (2018). The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Kolesnikova, O. (2016). Survey of word co-occurrence measures for collocation detection. *Computacion y Sistemas*, 20:327–344.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Körber, M. (2018). Theoretical considerations and development of a questionnaire to measure trust in automation. In *Congress of the International Ergonomics Association*, pages 13–30. Springer.

Krizhevsky, A. (2009). Learning Multiple Layers of Features from Tiny Images.

Kulesza, T., Stumpf, S., Burnett, M., and Kwan, I. (2012). Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 1–10, New York, NY, USA. Association for Computing Machinery.

Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., and Wong, W.-K. (2013). Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, pages 3–10. ISSN: 1943-6106.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86.

Kumar, S. and Talukdar, P. (2020). NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. (2019). Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S. J., and Doshi-Velez, F. (2019). Human evaluation of models built for interpretability. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 59–67.

Lai, G., Xie, Q., Liu, H., Yang, Y., and Hovy, E. (2017). RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Lai, V., Liu, H., and Tan, C. (2020). "why is 'chicago' deceptive?" towards building model-driven tutorials for humans. In Bernhaupt, R., Mueller, F. F., Verweij, D., Andres, J., McGrenere, J., Cockburn, A., Avellino, I., Goguey, A., Bjøn, P., Zhao, S., Samson, B. P., and Kocielnik, R., editors, *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–13. ACM.

Lai, V. and Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In danah boyd and Morgenstern, J. H., editors, *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 29–38. ACM.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, 40. Publisher: Cambridge University Press.

Lakkaraju, H. and Bastani, O. (2020). "how do I fool you?": Manipulating user trust via misleading black box explanations. In Markham, A. N., Powles, J., Walsh, T., and Washington, A. L., editors, *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7-8, 2020*, pages 79–85. ACM.

Langer, M. and König, C. J. (2018). Introducing and Testing the Creepiness of Situation Scale (CRoSS). *Frontiers in Psychology*, 9:2220.

Lazarski, E., Al-Khassaweneh, M., and Howard, C. (2021). Using nlp for fact checking: A survey. *Designs*, 5(3).

Lei, T., Barzilay, R., and Jaakkola, T. (2016). Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Leo, G. D. and Sardanelli, F. (2020). Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach. *European Radiology Experimental*, 4.

Levine, D. S. (2018). *Introduction to neural and cognitive modeling*. Routledge.

Lewis, J. R. (2018). Measuring perceived usability: The csuq, sus, and umux. *International Journal of Human–Computer Interaction*, 34:1148 – 1156.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

Li, A. H. and Sethy, A. (2019). Knowledge enhanced attention for robust natural language inference. *CoRR*, abs/1909.00102.

Li, K. and Malik, J. (2016). Learning to optimize. *CoRR*, abs/1606.01885.

Li, T. and Srikumar, V. (2019). Augmenting neural networks with first-order logic. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 292–302, Florence, Italy. Association for Computational Linguistics.

Li, T., Zhu, X., Liu, Q., Chen, Q., Chen, Z., and Wei, S. (2019). Several experiments on investigating pretraining and knowledge-enhanced models for natural language inference. *CoRR*, abs/1904.12104.

Li, X., Lei, W., and Yang, Y. (2022). From easy to hard: Two-stage selector and reader for multi-hop question answering. *CoRR*, abs/2205.11729.

Li, Z. and Wood, S. N. (2020). Faster model matrix crossproducts for large generalized linear models with discretized covariates. *Stat. Comput.*, 30(1):19–25.

Liao, Q. V., Zhang, Y., Luss, R., Doshi-Velez, F., and Dhurandhar, A. (2022). Connecting algorithmic research and usage contexts: A perspective of contextualized evaluation for explainable ai. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 147–159.

Lim, B. Y., Dey, A. K., and Avrahami, D. (2009). *Why and why not* explanations improve the intelligibility of context-aware intelligent systems. In Jr., D. R. O., Arthur, R. B., Hinckley, K., Morris, M. R., Hudson, S. E., and Greenberg, S., editors, *Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009, Boston, MA, USA, April 4-9, 2009*, pages 2119–2128. ACM.

Lima, G., Grgic-Hlaca, N., Jeong, J. K., and Cha, M. (2022). The conflict between explainable and accountable decision-making algorithms. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pages 2103–2113. ACM.

Lin, B. Y., Chen, X., Chen, J., and Ren, X. (2019a). KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2829–2839, Hong Kong, China. Association for Computational Linguistics.

Lin, K., Tafjord, O., Clark, P., and Gardner, M. (2019b). Reasoning over paragraph effects in situations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 58–62, Hong Kong, China. Association for Computational Linguistics.

Linzen, T. (2020). How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Lipsey, M. W. and Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American psychologist*, 48(12):1181.

Liu, C.-W., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., and Pineau, J. (2016). How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Liu, J., Liu, A., Lu, X., Welleck, S., West, P., Le Bras, R., Choi, Y., and Hajishirzi, H. (2022a). Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.

Liu, X., Sun, T., He, J., Wu, J., Wu, L., Zhang, X., Jiang, H., Cao, Z., Huang, X., and Qiu, X. (2022b). Towards efficient NLP: A standard evaluation and a strong baseline. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3288–3303, Seattle, United States. Association for Computational Linguistics.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, 10:464–470.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Lundberg, S. M. and Lee, S. (2017). A unified approach to interpreting model predictions. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774.

Lundberg, S. M., Nair, B. G., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D., Low, D. K.-W., Newman, S.-F., Kim, J. H., and Lee, S.-I. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2:749 – 760.

Lüdecke, D. (2023). *sjPlot: Data Visualization for Statistics in Social Science*. R package version 2.8.14.

MacCartney, B., Galley, M., and Manning, C. D. (2008). A phrase-based alignment model for natural language inference. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 802–811, Honolulu, Hawaii. Association for Computational Linguistics.

MacKenzie, I. S. (2013). *Human-Computer Interaction: An Empirical Research Perspective*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1st edition.

MacLean, A., Barnard, P., and Wilson, M. (1985). Evaluating the human interface of a data entry system: user choice and performance measures yield different tradeoff functions. *People and computers: Designing the interface*, 5(7):45–61.

Maclure, J. (2021). Correction to: Ai, explainability and public reason: The argument from the limitations of the human mind. *Minds Mach.*, 31(4):637.

Madsen, A., Meade, N., Adlakha, V., and Reddy, S. (2021). Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining. *CoRR*, abs/2110.08412. arXiv: 2110.08412.

Madsen, A., Reddy, S., and Chandar, S. (2023). Post-hoc interpretability for neural NLP: A survey. *ACM Comput. Surv.*, 55(8):155:1–155:42.

Malle, B. F. (2003). Folk theory of mind: Conceptual foundations of social cognition.

Manheim, D. (2018). Building less flawed metrics: Dodging Goodhart and Campbell's laws. *Munich Personal RePEc Archive*.

Manheim, D. and Garrabrant, S. (2018). Categorizing variants of goodhart's law. *CoRR*, abs/1803.04585.

Marra, G. and Wood, S. N. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, 39(1):53–74.

Marzouk, Z. (2018). Text marking: A metacognitive perspective.

Mathur, N., Baldwin, T., and Cohn, T. (2020). Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6:175–183.

Maybee, J. E. (2020). Hegel's Dialectics. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2020 edition.

McCulloch, C. E. and Neuhaus, J. M. (2005). Generalized linear mixed models. *Encyclopedia of biostatistics*, 4.

McDermid, J. A., Jia, Y., Porter, Z., and Habli, I. (2021). Artificial intelligence explainability: the technical and ethical dimensions. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 379.

Mcdonald, R. P. (1999). Test theory: A unified treatment.

McDuff, D. J., Hernandez, J., Gontarek, S., and Picard, R. W. (2016). *COGCAM: Contact-Free Measurement of Cognitive Stress During Computer Tasks with a Digital Camera*, page 4000–4004. Association for Computing Machinery, New York, NY, USA.

Menold, N. and Bogner, K. (2016). Design of rating scales in questionnaires. *GESIS survey guidelines*, 4.

Merrer, E. L. and Trédan, G. (2019). The bouncer problem: Challenges to remote explainability. *CoRR*, abs/1910.01432.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38.

Miller, T., Howe, P., and Sonenberg, L. (2017). Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. *CoRR*, abs/1712.00547.

Min, S., Zhong, V., Zettlemoyer, L., and Hajishirzi, H. (2019). Multi-hop reading comprehension through question decomposition and rescoring. In Korhonen, A., Traum, D. R., and Màrquez, L., editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6097–6109. Association for Computational Linguistics.

Montgomery, D. C. (2017). *Design and analysis of experiments*. John wiley & sons.

Mori, K.-i., Kidode, M., and Asada, H. (1973). An iterative prediction and correction method for automatic stereocomparison. *Comput. Graph. Image Process.*, 2(3-4):393–401.

Mosca, E., Demirtürk, D., Mülln, L., Raffagnato, F., and Groh, G. (2022a). GrammarSHAP: An efficient model-agnostic and structure-aware NLP explainer. In *Proceedings of the First Workshop on Learning with Natural Language Supervision*, pages 10–16, Dublin, Ireland. Association for Computational Linguistics.

Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D., and Groh, G. (2022b). SHAP-based explanation methods: A review for NLP interpretability. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4593–4603, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Moyé, L. A. (1998). P-value interpretation and alpha allocation in clinical trials. *Annals of Epidemiology*, 8(6):351–357.

Mueller, G. E. (1958). The Hegel Legend of" Thesis-Antithesis-Synthesis". *Journal of the*

History of Ideas, 19(3):411–414. Publisher: JSTOR.

Nadarzynski, T., Miles, O., Cowie, A., and Ridge, D. (2019). Acceptability of artificial intelligence (ai)-led chatbot services in healthcare: A mixed-methods study. *Digital health*, 5:2055207619871808.

Naik, A., Ravichander, A., Sadeh, N., Rose, C., and Neubig, G. (2018). Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Narang, S., Raffel, C., Lee, K., Roberts, A., Fiedel, N., and Malkan, K. (2020). Wt5?! training text-to-text models to explain their predictions. *CoRR*, abs/2004.14546.

Natarajan, M. and Gombolay, M. (2020). Effects of anthropomorphism and accountability on trust in human robot interaction. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, page 33–42, New York, NY, USA. Association for Computing Machinery.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., and Seifert, C. (2022). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384. Publisher: Wiley Online Library.

Newman, G. E. and Scholl, B. J. (2012). Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychonomic Bulletin & Review*, 19:601–607.

Nguyen, D. (2018). Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, New Orleans, Louisiana. Association for Computational Linguistics.

Nguyen, G., Kim, D., and Nguyen, A. (2021). The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 26422–26436.

Nielsen, J. and Levy, J. (1994). Measuring usability: Preference vs. performance. *Commun. ACM*, 37(4):66–75.

Nishida, K., Nishida, K., Saito, I., and Yoshida, S. (2021). Towards interpretable and reliable reading comprehension: A pipeline model with unanswerability prediction. In *International*

Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021, pages 1–8. IEEE.

Nourani, M., Kabir, S., Mohseni, S., and Ragan, E. D. (2019). The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7(1):97–105.

Novikova, J., Dušek, O., Cercas Curry, A., and Rieser, V. (2017). Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Novikova, J., Dusek, O., and Rieser, V. (2018). Rankme: Reliable human ratings for natural language generation. In Walker, M. A., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 72–78. Association for Computational Linguistics.

Oddone, E. Z., Blakeley, D., and Matchar, D. B. (1995). Noninvasive carotid artery testing. *Annals of Internal Medicine*, 123:634.

Ostertagova, E., Ostertag, O., and Kováč, J. (2014). Methodology and application of the kruskal-wallis test. In *Applied Mechanics and Materials*, volume 611, pages 115–120. Trans Tech Publ.

Ottenbacher, K. J. (1998). Quantitative evaluation of multiplicity in epidemiology and public health research. *American Journal of Epidemiology*, 147(7):615–619.

Overton, J. A. (2012). *Explanation in Science*. The University of Western Ontario (Canada).

Owczarzak, K., Conroy, J. M., Dang, H. T., and Nenkova, A. (2012). An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9, Montréal, Canada. Association for Computational Linguistics.

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. *Journal of educational psychology*, 84(4):429.

Pan, X., Sun, K., Yu, D., Chen, J., Ji, H., Cardie, C., and Yu, D. (2019). Improving question answering with external knowledge. In Fisch, A., Talmor, A., Jia, R., Seo, M., Choi, E., and Chen, D., editors, *Proceedings of the 2nd Workshop on Machine Reading for Question Answering, MRQA@EMNLP 2019, Hong Kong, China, November 4, 2019*, pages 27–37. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Paulus, R., Xiong, C., and Socher, R. (2018). A deep reinforced model for abstractive summarization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Pereira, D. G., Afonso, A., and Medeiros, F. M. (2015). Overview of friedman's test and post-hoc analysis. *Communications in Statistics-Simulation and Computation*, 44(10):2636–2653.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. (2019). Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Pezeshkpour, P., Jain, S., Singh, S., and Wallace, B. (2022). Combining feature and instance attribution to detect artifacts. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1934–1946, Dublin, Ireland. Association for Computational Linguistics.

Piepho, H.-P. (2004). An algorithm for a letter-based representation of all-pairwise comparisons. *Journal of Computational and Graphical Statistics*, 13(2):456–466.

Pimentel, T., Saphra, N., Williams, A., and Cotterell, R. (2020). Pareto probing: Trading off accuracy for complexity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3138–3153, Online. Association for Computational Linguistics.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., and Wallach, H. M. (2021). Manipulating and measuring model interpretability. In Kitamura, Y., Quigley, A., Isbister, K., Igarashi, T., Bjørn, P., and Drucker, S. M., editors, *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, pages 237:1–237:52. ACM.

Preece, A. D., Harborne, D., Braines, D., Tomsett, R., and Chakraborty, S. (2018). Stakeholders in explainable AI. *CoRR*, abs/1810.00184.

Pruthi, D., Bansal, R., Dhingra, B., Baldini Soares, L., Collins, M., Lipton, Z. C., Neubig, G., and Cohen, W. W. (2022). Evaluating explanations: How much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics*, 10:359–375.

Pruthi, D., Gupta, M., Dhingra, B., Neubig, G., and Lipton, Z. C. (2020). Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793, Online. Association for Computational Linguistics.

Qi, P., Lin, X., Mehr, L., Wang, Z., and Manning, C. D. (2019). Answering complex open-domain questions through iterative query generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2590–2602, Hong Kong, China. Association for Computational Linguistics.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Qian, P., Qiu, X., and Huang, X. (2016). A new psychometric-inspired evaluation metric for Chinese word segmentation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2185–2194, Berlin, Germany. Association for Computational Linguistics.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, (8):9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Rajani, N. F., McCann, B., Xiong, C., and Socher, R. (2019). Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Gruber, L., Holzleitner, M., Pavlovic, M., Sandve, G. K., Greiff, V., Kreil, D. P., Kopp, M., Klambauer, G., Brandstetter, J., and Hochreiter, S. (2020). Hopfield Networks is All You Need. *CoRR*, abs/2008.02217. _eprint: 2008.02217.

Ranney, M. and Thagard, P. (1988). Explanatory coherence and belief revision in naive physics. Technical report, Pittsburgh Univ PA Learning Research an Development Center.

Raykov, T. (2001). Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints. *The British journal of mathematical and statistical psychology*, 54 Pt 2:315–23.

Raza, S. and Ding, C. (2022). News recommender system: a review of recent progress, challenges, and opportunities. *Artif. Intell. Rev.*, 55(1):749–800.

Read, S. and Marcus-Newhall, A. (1993). Explanatory Coherence in Social Explanations: A Parallel Distributed Processing Account. *Journal of Personality and Social Psychology*, 65(3):429–447.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Reiter, E. (2018). A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Renshaw, T., Stevens, R., and Denton, P. D. (2009). Towards understanding engagement in games: an eye-tracking study. *On the Horizon*.

Revelle, W. (2022). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 2.2.9.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. In Krishnapuram, B., Shah, M., Smola, A. J., Aggarwal, C. C., Shen, D., and Rastogi, R., editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM.

Ribera, M. and Lapedriza, À. (2019). Can we do better explanations? A proposal of user-centered explainable AI. In Trattner, C., Parra, D., and Riche, N., editors, *Joint Proceedings of the ACM IUI 2019 Workshops co-located with the 24th ACM Conference on Intelligent User Interfaces (ACM IUI 2019), Los Angeles, USA, March 20, 2019*, volume 2327 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Ribes, D., Henchoz, N., Portier, H., Défayes, L., Phan, T., Gatica-Perez, D., and Sonderegger, A. (2021). Trust indicators and explainable AI: A study on user perceptions. In Ardito, C., Lanzilotti, R., Malizia, A., Petrie, H., Piccinno, A., Desolda, G., and Inkpen, K., editors, *Human-Computer Interaction - INTERACT 2021 - 18th IFIP TC 13 International Conference,* *Bari, Italy, August 30 - September 3, 2021, Proceedings, Part II*, volume 12933 of *Lecture Notes in Computer Science*, pages 662–671. Springer.

Riegel, K. F. (1973). Dialectic operations: The final period of cognitive development. *Human development*, 16(5):346–370. Publisher: Karger Publishers.

Ross, A. S., Hughes, M. C., and Doshi-Velez, F. (2017). Right for the right reasons: Training differentiable models by constraining their explanations. In Sierra, C., editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 2662–2670. ijcai.org.

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, pages 43–46.

Rozenblit, L. and Keil, F. C. (2002). The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive science*, 26 5:521–562.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215. Publisher: Nature Publishing Group.

Rupert, R. D. (2009). *Cognitive systems and the extended mind*. Oxford University Press.

Sakaguchi, K., Post, M., and Van Durme, B. (2014). Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 1–11, Baltimore, Maryland, USA. Association for Computational Linguistics.

Santhanam, S. and Shaikh, S. (2019). Towards best experiment design for evaluating dialogue system output. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 88–94, Tokyo, Japan. Association for Computational Linguistics.

Scharrer, L., Bromme, R., Britt, M. A., and Stadtler, M. (2012). The seduction of easiness: How science depictions influence laypeople's reliance on their own evaluation of scientific information. *Learning and Instruction*, 22(3):231–243.

Schedl, M., Zamani, H., Chen, C., Deldjoo, Y., and Elahi, M. (2018). Current challenges and visions in music recommender systems research. *Int. J. Multim. Inf. Retr.*, 7(2):95–116.

Schlegel, V., Guzman, E. M., and Batista-Navarro, R. (2022). Towards human-centred explainability benchmarks for text classification. In de Melo, P. O. S. V., Jeng, W., and Buntain, C., editors, *Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media, ICWSM 2022 Workshops, Atlanta, Georgia, USA [hybrid], June 6, 2022*.

Schrills, T. and Franke, T. (2020). How to answer why - evaluating the explanations of AI through mental model analysis. *CoRR*, abs/2002.02526.

Schrills, T. P. P., Kargl, S., Bickel, M., and Franke, T. (2022). Perceive, understand & predict -

empirical indication for facets in subjective information processing awareness.

Schuff, H. (2020). Explainable question answering beyond f1: metrics, models and human evaluation.

Schuff, H., Adel, H., Qi, P., and Vu, N. T. (2022a). Challenges in explanation quality evaluation. *CoRR*, abs/2210.07126.

Schuff, H., Adel, H., and Vu, N. T. (2020). F1 is Not Enough! Models and Evaluation Towards User-Centered Explainable Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7076–7095, Online. Association for Computational Linguistics.

Schuff, H., Adel, H., and Vu, N. T. (2021a). Thought flow nets: From single predictions to trains of model thought. *CoRR*, abs/2107.12220.

Schuff, H., Adel, H., and Vu, N. T. (2023a). Device and method for classifying a signal and/or for performing regression analysis on a signal. US Patent App. 17/831,750.

Schuff, H., Jacovi, A., Adel, H., Goldberg, Y., and Vu, N. T. (2022b). Human interpretation of saliency-based explanation over text. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 611–636, New York, NY, USA. Association for Computing Machinery.

Schuff, H., Vanderlyn, L., Adel, H., and Vu, N. T. (2023b). How to do human evaluation: A brief introduction to user studies in nlp. *Natural Language Engineering*, page 1–24.

Schuff, H., Yang, H.-Y., Adel, H., and Vu, N. T. (2021b). Does external knowledge help explainable natural language inference? automatic evaluation vs. human ratings. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 26–41, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Schulz, E., Tenenbaum, J. B., Reshef, D. N., Speekenbrink, M., and Gershman, S. (2015). Assessing the perceived predictability of functions. In Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock, T., Jennings, C. D., and Maglio, P. P., editors, *Proceedings of the 37th Annual Meeting of the Cognitive Science Society, CogSci 2015, Pasadena, California, USA, July 22-25, 2015*. cognitivesciencesociety.org.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2021). WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1351–1361. Association for Computational Linguistics.

Secară, A. (2005). Translation evaluation: A state of the art survey. In *Proceedings of the* *eCoLoRe/MeLLANGE workshop, Leeds*, volume 39, page 44. Citeseer.

Sedoc, J., Ippolito, D., Kirubarajan, A., Thirani, J., Ungar, L., and Callison-Burch, C. (2019). ChatEval: A tool for chatbot evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 60–65, Minneapolis, Minnesota. Association for Computational Linguistics.

Sedoc, J. and Ungar, L. (2020). Item Response Theory for Efficient Human Evaluation of Chatbots. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 21–33, Online. Association for Computational Linguistics.

Sellam, T., Das, D., and Parikh, A. (2020). BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Seyler, D., Dembelova, T., Del Corro, L., Hoffart, J., and Weikum, G. (2018). A study of the importance of external knowledge in the named entity recognition task. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 241–246, Melbourne, Australia. Association for Computational Linguistics.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317.

Shi, C., Liu, S., Ren, S., Feng, S., Li, M., Zhou, M., Sun, X., and Wang, H. (2016). Knowledge-based semantic embedding for machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2245–2254, Berlin, Germany. Association for Computational Linguistics.

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153. PMLR.

Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. (2016). Not just a black box: Learning important features through propagating activation differences. *CoRR*, abs/1605.01713.

Silveira, N., Dozat, T., de Marneffe, M.-C., Bowman, S., Connor, M., Bauer, J., and Manning, C. D. (2014). A gold standard dependency corpus for English. In *Proceedings of the Ninth*

International Conference on Language Resources and Evaluation (LREC-2014).

Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In Bengio, Y. and LeCun, Y., editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*.

Sivaraman, V., Bukowski, L. A., Levin, J., Kahn, J. M., and Perer, A. (2023). Ignore, trust, or negotiate: Understanding clinician acceptance of ai-based treatment recommendations in health care.

Smith, A. R. (1978). Color gamut transform pairs. *ACM Siggraph Computer Graphics*, 12(3):12–19. Publisher: ACM New York, NY, USA.

Smith, H. (2020). Algorithmic bias: should students pay the price? *AI Soc.*, 35(4):1077–1078.

Sokol, K. and Flach, P. A. (2020). Explainability fact sheets: a framework for systematic assessment of explainable approaches. In Hildebrandt, M., Castillo, C., Celis, L. E., Ruggieri, S., Taylor, L., and Zanfir-Fortuna, G., editors, *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 56–67. ACM.

Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). Beyond accuracy, f-score and ROC: A family of discriminant measures for performance evaluation. In Sattar, A. and Kang, B., editors, *AI 2006: Advances in Artificial Intelligence, 19th Australian Joint Conference on Artificial Intelligence*, volume 4304 of *Lecture Notes in Computer Science*, pages 1015–1021, Hobart, Australia. Springer.

Speer, R., Chin, J., and Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In Singh, S. P. and Markovitch, S., editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Speith, T. (2022). A review of taxonomies of explainable artificial intelligence (XAI) methods. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pages 2239–2250. ACM.

Sprent, P. (2012). *Applied nonparametric statistical methods*. Springer Science & Business Media.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103 2684:677–80.

Strathern, M. (1997). 'Improving ratings': audit in the British University system. *European review*, 5(3):305–321. Publisher: Cambridge University Press.

Streiner, D. L. and Norman, G. R. (2011). Correction for multiple testing: is there a resolution? *Chest*, 140(1):16–18.

Sulem, E., Abend, O., and Rappoport, A. (2018). BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Sullivan, G. M. and Feinn, R. (2012). Using effect size—or why the p value is not enough. *Journal of graduate medical education*, 4(3):279–282.

Sullivan Jr., J., Brackenbury, W., McNutt, A., Bryson, K., Byll, K., Chen, Y., Littman, M., Tan, C., and Ur, B. (2022). Explaining why: How instructions and user interfaces impact annotator rationales when labeling text data. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 521–531, Seattle, United States. Association for Computational Linguistics.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Suresh, H., Gomez, S. R., Nam, K. K., and Satyanarayan, A. (2021). Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., and Kroeker, K. I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digital Medicine*, 3.

Swanson, K., Yu, L., and Lei, T. (2020). Rationalizing text matching: Learning sparse alignments via optimal transport. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5609–5626, Online. Association for Computational Linguistics.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. (2019). CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Tenney, I., Wexler, J., Bastings, J., Bolukbasi, T., Coenen, A., Gehrmann, S., Jiang, E., Pushkarna, M., Radebaugh, C., Reif, E., and Yuan, A. (2020). The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, 12(3):435–467.

Thomas, R. L. and Uminsky, D. (2022). Reliance on metrics is a fundamental challenge for AI. *Patterns*, 3(5):100476. Publisher: Elsevier.

Tjoa, E. and Guan, C. (2021). A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32(11):4793—4813.

Tractinsky, N. and Meyer, J. (2001). Task structure and the apparent duration of hierarchical search. *International Journal of Human-Computer Studies*, 55(5):845–860.

Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., and Suleman, K. (2017). NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Tseng, P.-H., Carmi, R., Cameron, I. G., Munoz, D. P., and Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of vision*, 9(7):4–4. Publisher: The Association for Research in Vision and Ophthalmology.

Tu, M., Huang, K., Wang, G., Huang, J., He, X., and Zhou, B. (2020). Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9073–9080. AAAI Press.

Tuckey, D., Broda, K., and Russo, A. (2019). Saliency Maps Generation for Automatic Text Summarization. *CoRR*, abs/1907.05664. arXiv: 1907.05664.

van der Lee, C., Gatt, A., van Miltenburg, E., Wubben, S., and Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Van Rij, J., Wieling, M., Baayen, R. H., and van Rijn, D. (2015). itsadug: Interpreting time series and autocorrelated data using gamms.

VanVoorhis, C. W., Morgan, B. L., et al. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in quantitative methods for psychology*, 3(2):43–50.

Vasilyeva, N., Wilkenfeld, D. A., and Lombrozo, T. (2015). Goals Affect the Perceived Quality of Explanations. In *CogSci*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information*

Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008.

Vilar, D., Leusch, G., Ney, H., and Banchs, R. E. (2007). Human evaluation of machine translation through binary system comparisons. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 96–103.

Vilone, G. and Longo, L. (2021). Classification of explainable artificial intelligence methods through their output formats. *Mach. Learn. Knowl. Extr.*, 3(3):615–661.

Waldman, A. E. (2019). Power, process, and automated decision-making. *Fordham Law Review*, 88:613.

Wang, J., Tuyls, J., Wallace, E., and Singh, S. (2020). Gradient-based Analysis of NLP Models is Manipulable. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 247–258. Association for Computational Linguistics.

Wang, P. and Vasconcelos, N. (2020). SCOUT: self-aware discriminant counterfactual explanations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8978–8987. Computer Vision Foundation / IEEE.

Wang, X., Kapanipathi, P., Musa, R., Yu, M., Talamadupula, K., Abdelaziz, I., Chang, M., Fokoue, A., Makni, B., Mattei, N., and Witbrock, M. (2019). Improving Natural Language Inference Using External Knowledge in the Science Questions Domain. In *AAAI*, pages 7208–7215.

Wang, X. and Yin, M. (2021). Are explanations helpful? A comparative study of the effects of explanations in ai-assisted decision-making. In Hammond, T., Verbert, K., Parra, D., Knijnenburg, B. P., O'Donovan, J., and Teale, P., editors, *IUI '21: 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, April 13-17, 2021*, pages 318–328. ACM.

Watson, D. (2020). *The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence*, pages 45–65.

Watson, D. S. (2021). Interpretable machine learning for genomics. *Human Genetics*, 141:1499 – 1513.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E. H., Le, Q., and Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Wiegreffe, S. and Pinter, Y. (2019). Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design*. sage publications.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. *CoRR*, abs/1308.5499.

Wood, S., N., Pya, and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models (with discussion). *Journal of the American Statistical Association*, 111:1548–1575.

Wood, S. N. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society (B)*, 65(1):95–114.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467):673–686.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36.

Wood, S. N. (2013a). On p-values for smooth components of an extended generalized additive model. *Biometrika*, 100(1):221–228.

Wood, S. N. (2013b). A simple test for random effects in regression models. *Biometrika*, 100(4):1005–1010.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.

Wood, S. N., Li, Z., Shaddick, G., and Augustin, N. H. (2017). Generalized additive models for gigadata: modeling the UK black smoke network daily data. *Journal of the American Statistical Association*, 112(519):1199–1210. Publisher: Taylor & Francis.

Xia, X. (2018). An effective way to memorize new words—lexical chunk. *Theory and Practice in Language Studies*, 8:14941498.

Xiong, W., Wu, J., Wang, H., Kulkarni, V., Yu, M., Chang, S., Guo, X., and Wang, W. Y. (2019). TWEETQA: A social media focused question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5020–5031, Florence, Italy. Association for Computational Linguistics.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., and Manning, C. D. (2018). HotpotQA: A dataset for diverse, explainable multi-hop question answering. In

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Yeh, Y.-Y. and Wickens, C. D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors*, 30(1):111–120.

Yin, W., Hay, J., and Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3912–3921. Association for Computational Linguistics.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3.

Yu, R. and Shi, L. (2018). A user-based taxonomy for deep learning visualization. *Vis. Informatics*, 2(3):147–154.

Zaidan, O. and Eisner, J. (2008). Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 31–40, Honolulu, Hawaii. Association for Computational Linguistics.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine*, 15.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020a). Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhang, Y., Liao, Q. V., and Bellamy, R. K. E. (2020b). Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In Hildebrandt, M., Castillo, C., Celis, L. E., Ruggieri, S., Taylor, L., and Zanfir-Fortuna, G., editors, *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 295–305. ACM.

Zhang, Y., Tiño, P., Leonardis, A., and Tang, K. (2021). A survey on neural network interpretability. *IEEE Trans. Emerg. Top. Comput. Intell.*, 5(5):726–742.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., and Liu, Q. (2019). ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Zhao, Y., Li, Y., Li, C., and Zhang, R. (2022). MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6588–6600, Dublin, Ireland. Association for Computational Linguistics.

Zhou, X. and Zafarani, R. (2021). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5):109:1–109:40.

Zhou, Y., Booth, S., Ribeiro, M. T., and Shah, J. (2022). Do feature attribution methods correctly attribute features? In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 9623–9633. AAAI Press.

Zlotowski, J., Proudfoot, D., Yogeeswaran, K., and Bartneck, C. (2015). Anthropomorphism: Opportunities and challenges in human–robot interaction. *International Journal of Social Robotics*, 7:347–360.

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological methods*, 12 4:399–413.