Institute for Visualization and Interactive Systems

University of Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Bachelorarbeit

GPT-4-based Visualization Reasoning Dataset

Jo Hollenbeck

Course of Study: Informatik

Examiner: Prof. Dr. Bulling

Supervisor: Yao Wang

Commenced: 1. August 2023

Completed: 1. February 2024

Abstract

Visual data, such as charts and tables, is a widely used method to summarise data clearly. With the ascent of Artificial Intelligence lots of models have been created that provide users with answers to their questions on visual data. An analysis of the latest research reveals that the model accuracy and the reasoning of the results are still insufficient. These two major aspects are tackled in this work. The AI model used is OpenAI's large language model GPT-4. Tables are presented as text-only input, and charts are uploaded as images to GPT-4. A modified prompt guarantees a step-by-step reasoning as output. With the collected data, quantitative analyses are conducted to evaluate the numerical data and its influence on the response. Moreover, a qualitative analysis is performed determining the quality of answer in terms of clarity, relevance and reasoning. Additionally, responses on tables and charts are compared to get deeper insights on the model's performance.

Notable results are GPT-4's outstanding performance on the accuracy of the input formats, except for line charts and charts containing dense information. The model consistently produces good-quality answers when provided with either text-only or image-text input. This work demonstrates that GPT-4 performs well on visual data methods, but especially for complex chart images it exhibits room for improvement.

Kurzfassung

Grafische Darstellungen von Daten, wie Diagramme und Tabellen, sind eine weit verbreitete Methode, um Daten strukturiert und verständlich zusammenzufassen. Mit dem Fortschritt in der Künstlichen Intelligenz wurden viele Modelle entwickelt, die Nutzern Antworten auf ihre Fragen bezüglich dieser visuellen Daten liefern. Die Analyse der Modelle zeigt, dass sowohl die Evaluation der Genauigkeit als auch die Transparenz hinsichtlich der Schritte zum Erlangen des Ergebnisses unzureichend sind. Diese beiden Schwachstellen wurden in dieser Arbeit adressiert. Die verwendete KI ist das Modell von OpenAI, GPT-4. Tabellen werden in reiner Textform übermittelt und Diagramme als Bilder hochgeladen. Eine erweiterte Eingabe garantiert eine Begründung der Antwort seitens der KI. Mit den gesammelten Daten werden quantitative Analysen durchgeführt, um die numerischen Daten und ihren Einfluss auf die Antwort zu untersuchen. Darüber hinaus wird eine qualitative Analyse durchgeführt, um die Qualität der Antwort in Bezug auf Klarheit, Relevanz und Begründung zu bestimmen. Zusätzlich werden Antworten auf Tabellen und Diagramme verglichen, um tiefere Einblicke in die Leistung des Modells zu erhalten und mögliche Schwachstellen zu identifizieren.
Die Ergebnisse zeigen, dass GPT-4 bei den meisten Diagrammarten beeindruckend genau arbeitet, mit Ausnahme von Liniendiagrammen und insbesondere Grafiken mit hoher Informationsdichte. Das Modell erreicht eine durchweg gute Antwortqualität, sowohl bei rein textbasierten als auch bei bildbasierten Eingaben. Diese Arbeit zeigt, dass GPT-4 bei Aufgaben auf visuellen Darstellungen gut abschneidet, aber bei komplexen Diagrammen Verbesserungspotenzial aufweist.

Contents

List of Figures

List of Tables

1 Introduction

Charts and tables serve in expositions for various purposes as fundamental tools for an efficient visualisation and summarisation of data. The process of extracting elements from these visual representations to answer questions can be quite challenging, especially for complex question that require for example comparisons or calculations. In an era where Artificial Intelligence (AI) is evolving rapidly, the potential to use these capabilities for effective data interpretation is increasingly being spotlighted [KMA+17; KPCK18]. A wide range of different papers have already been published to create Artificial Intelligences that can handle these tasks and test them on their datasets [MGKK20; MLT+22]. For instance, Figure 1.1 shows an example, where the AI model has to extract the answer for the questions Q1 and Q2 out of the pie chart and provide the correct answer. For Q1 the AI needs to distinguish a colour and for Q2 a calculation has to be performed.

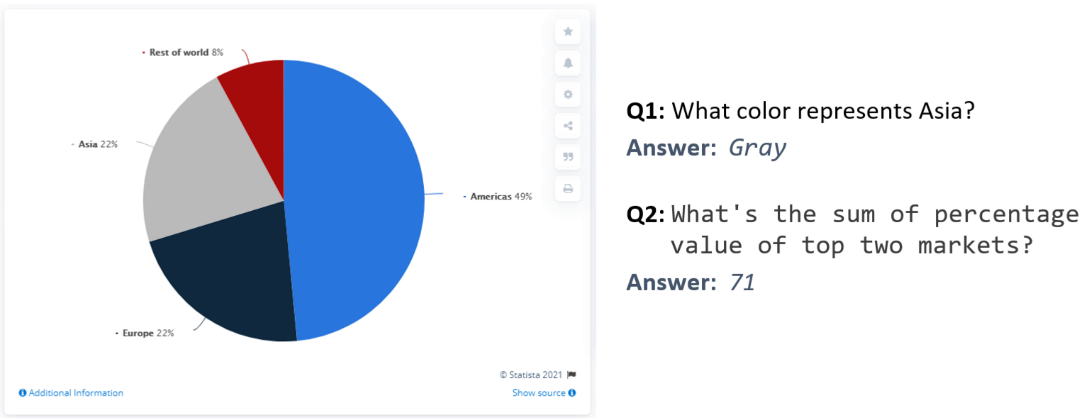

Figure 1.1: Example task for GPT-4

Promising approaches in Chart Question Answering (CQA) like PlotQA [MGKK20] and ChartQA [MLT+22] have focused on direct chart data extraction, dealing with deep neural networks to interpret visual data and generate relevant responses. Both approaches, i.e. the developed models, provided a decent accuracy, but the answers lacked in transparency and traceability. The researches mainly focused on the quantitative analysis of their model by only investigating the numerical data. This makes it challenging to understand the reasoning behind their answers. This limitation is especially distinct in data-rich visuals like charts and tables, as the rationale for a given answer on data visualisations is important to understand in order to reproduce the answer [ZTZ+22]. Furthermore, nowadays, there are a lot of models that can handle and work with complex charts and tables [GEL+23; OA+23; TAB+23].

This work introduces an approach to CQA, which extends to tables in the question answering. The model used for this task is OpenAI's GPT-4 [OA+23], an advanced Large Language Model (LLM), which interprets both chart images and text-only tables as input. The dataset from ChartQA serves as data source for this research. Its dataset is superior due to its diverse range of complex, human-authored questions and its inclusion of various chart types, sourced from multiple fields [MLT+22]. This dataset provides a range of questions, allowing an analysis of GPT-4's reasoning abilities across different visual formats. The method of this work emphasises on providing not only answers but also explaining the reasoning process, which is an important part in AI responses [GSC+19]. Therefore, it addresses the limitations of previous works, by dealing not only with the quantitative attributes but also determining and comparing the quality. The quantitative analysis extends to not only examining the accuracy but also searching for causing parameters. In addition, an analysis was conducted, to derive a correlation between the question length and the reasoned response. Lastly, the research also performed a user study, gathering the answer qualities of GPT-4's responses, that is the qualitative analysis.

Outline

The rest of this thesis is structured as follows:

Chapter 2 – Related Work: This Chapter presents and summarises related work. It focuses on other research on Chart and Table Question Answering and illustrates the improvements made.

Chapter 3 – Data Preparation & Collection: In Chapter 3, the modifications that had to be made to the ChartQA dataset are described to make it suitable for this work. Additionally, the process of the data collection is explained.

Chapter 4 – Parameters for Analysis: The fourth chapter gives a detailed explanation of the parameters, which are evaluated in the analysis.

Chapter 5 – Quantitative Analysis: This chapter presents the findings of the quantitative parameters, such as accuracy.

Chapter 6 – Qualitative Analysis: In Chapter 6, the conducted user study is evaluated, which provides the quality of GPT-4's response, and the results are presented.

Chapter 7 – Discussion & Future work: In this chapter, the findings of the previous two chapters are discussed, and feasible future work is presented.

Chapter 8 – Conclusion: The last chapter summarises the results and insights gained from this bachelor's thesis.

2 Related Work

This chapter outlines related work and discusses how the contributions of this thesis expand on it. Research is presented, which covers parts of quantitative and qualitative analysis in regard to Visual Question Answering (VQA). Section 2.2 covers work which illustrate their findings of important criteria that constitutes good answer quality.

2.1 Visual Question Answering

Visual Question Answering has been a rapidly evolving field with several datasets contributing significantly to research and development [KMA+17; KPCK18; NHM+21; PL15]. Previous work already focused on testing and even creating models [CCS+21; YQZ+18] that are able to extract values out of visual data. This field of Visual Question Answering predominantly concentrates on creating datasets and conducting quantitative analysis on the results, by evaluating the model accuracy. For example, the work WikiTableQuestions [PL15], which used tables with a certain size from Wikipedia to investigate the model's performance. Beside the fact that only the accuracy is checked, this work did not consider other data visualisation formats. Papers that address charts as input are for example DVQA [KPCK18], FigureQA[KMA+17], PlotQA [MGKK20] and ChartQA [MLT+22]. The dataset of the latter one is superior in the field of Chart Question Answering due to the diversity in questions and chart types. This makes it more challenging for models to correctly answer the posed question but can give deeper insights in the analysis. ChartQA tested the accuracy of its dataset on various models, which are specialised in processing and understanding a combination of textual and visual data. The potential to identify possible factors was not exploited in their works, i.e. DVQA, FigureQA, PlotQA and ChartQA, despite low accuracy's of the tested models [CLTB21]. All of these only required the model to generate the answer, therefore the answer quality of the models could not be assessed. Furthermore, they solely focused on the accuracy in their evaluation, i.e. were only scratching the surface of the quantitative analysis.

Papers that require their model to provide a more reasoned response are TextVQA [WXL+21] and FeTaQA [NHM+21]. The first one included a quantitative and qualitative analysis. The dataset consists of real world images and the aim was it to let the model extract text out of them. Despite showing that qualitative analyses are important to get deeper insights in the performance and influencing factors, this work lacks in chart or table input. The work from Linyong Nan et al., FeTaQA [NHM+21], tackled the shortcomings for most parts. Here, the focus was not only on the accuracy but also on the model's ability to generate detailed and reasoned responses. Explaining the answers is especially important for complex questions, so the questioner can reconstruct it [GSC+19]. This thesis extends this by adding charts and working on an improved dataset, i.e the ChartQA dataset that provides the needed diversity of format types and includes more complex and differentiated questions. Furthermore, with integration of the two input formats, tables and charts, more discoveries can be made by comparing the quality of responses.

2.2 Answer Quality Evaluation

Due to limited prior work on answer quality in Visual Question Answering, the research of parameters which determine the answer quality was expanded. This section presents work, evaluating answer on questions in general and how they can be adopted in VQA.

Yiming Zhao et al. tackled in their paper "Evaluation of Google Question Answering Quality" the challenge of evaluating the complex answer boxes presented in Google's search results [ZZXL19]. A new evaluation criteria system was developed to assess Google's QA quality, where factors such as accuracy, clarity and relevance are important. The methodologies deployed for evaluation in this study can be adapted for chart Question Answering. Especially, the clarity and the relevance of the mentioned data in the answer are a significant factor for the answer quality.
Another work addressing the answer quality is "Evaluating and Predicting Answer Quality in Community QA" [SP10]. The data was taken from community-driven Question Answering platforms, and the answer quality was evaluated by Amazon Mechanical Turk workers. Due to the community-driven QA, the questions did not necessarily have a single correct solution. For that reason, the answer quality is determined with 13 different categories, such as 'easy to read', 'concise', 'relevant', or 'enough information', on a 5-point Likert scale [SP10]. The question answering applied in this thesis is only using questions with correct or incorrect answers, therefore some categories can be left out of consideration.
Building on the insights from these studies, the evaluation of answer quality in chart and table Question Answering was assisted. In this study, the quality is evaluated using three main criteria (see Chapter 4), which are based on the approaches addressed in the mentioned papers.

3 Data Preparation & Collection

This chapter is about the preparation and collection of the data needed for this work. This preparation stage not only sets the foundation for effective data analysis but also is significant for problem prevention and finding parameters for the analysis. Thereafter, the collection process is presented by highlighting the possible approaches and their constraints of this work. Furthermore, the defined conditions for GPT-4 are pointed out.

3.1 Preparation

In this thesis, the existing ChartQA dataset was used to analyse reasoned responses from GPT-4. This dataset comprises two types of questions: human-authored and automatically generated ones. Automatically generated questions were created by using an AI model. Since these questions are not as divers and complex as the human-authored queries, the analysis was exclusively based on the human-authored questions, ensuring a high complexity and diversity in questions throughout the thesis. It was significant to first test a smaller subset of the data to prevent problems in the data collection.

3.1.1 Subset

The composition of the test subset from the dataset was performed based on a specific criterion: the diversity of answer types. This set consisted of 20 data-question pairs, each requiring a different type of answer, such as a year, a string, a percentage, or a list. The method of evaluation was direct: providing the data-question pairs to GPT-4 and inspecting its responses. Providing the model merely with data-question pairs highlighted some problems that needed to be improved:

- **Query Attachment.** A significant aspect of the data collection strategy was forcing GPT-4 to provide a reasoned response on how the answer is retrieved from the data. This is accomplished by incorporating a query attachment, namely *"provide a step-by-step reasoning before providing your answer."*. This feature is the key for ensuring that GPT-4 not just provides the raw answer but also explains the steps that have to be taken to arrive at each result. Without any modifications to the prompt, the input to AI, i.e. only providing the data-question pair, GPT-4 answers correctly, but does not add the important reasoning how it can be conceived (see Figure 3.1). By adding the specified attachment, as shown in Figure 3.2, the response from GPT-4 becomes significantly more comprehensive. First, the model lists up every step that has to be taken to extract the correct value, and then in a final sentence the AI presents the answer to the user.

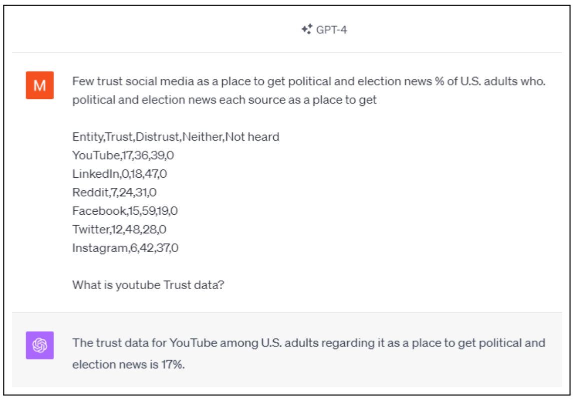

Figure 3.1: GPT-4 interface without query attachment

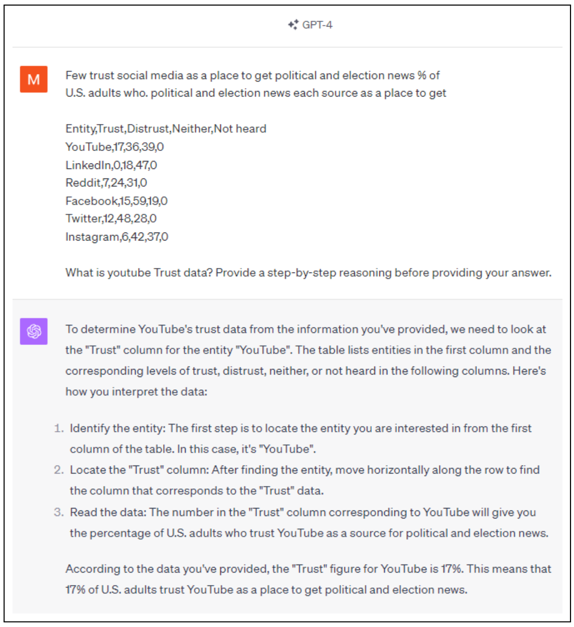

Figure 3.2: GPT-4 interface with query attachment

- **Addressing Missing Title.** In the ChartQA dataset, the text-only input files lacked essential information that GPT-4 needed to accurately answer the related questions. Frequently, these questions depended on specific details such as a country, year, or company referenced in the data. Without this information, GPT-4 could not recognise the context of the data. As a result, it often generated lengthy responses explaining that the question could not be answered due to these missing elements. This was solved by adding the corresponding title, which was provided in a separate file. As a result, GPT-4 had the needed information to correctly answer the question.
- **Addressing Files with Missing Data.** Certain text-only files had to be removed from the dataset, because the data had too many 'NaN' (not a number) values. In these instances, the model was not able to answer the question correctly, due to missing values.

With these improvements, GPT-4 reasoned all its responses and provided a promising accuracy. This evaluation of the subset not only improved and optimised the prompt, but additionally presented first analysis parameters, which are covered in Chapter 4.

3.1.2 Data Composition

The valuable improvements from the subset could then be transferred to the entire dataset of this work. The goal was it to have a single file for the text-only data collection and a single file, besides the chart image itself, for the visual part, the chart-text input. This procedure led to an immense amount of time saved in the following data collection process.

Text-only input. The text-only table input needed more preparation, because more data from different files had to be merged. On top was the title, which prevented GPT-4 from having the previously mentioned problems with possible missing information. Followed by the corresponding data from the filtered dataset. At the end the question and the important query attachment, *"Provide a step-by-step reasoning before providing your answer."*, were added. This merge was handled automatically and done with a total of 591 text-only data-question pairs.

Chart-text input. In contrast to the text-only approach, the title was included in the chart image. The chart-text input required the collection of only two elements: the question and the query attachment.
This format did not have as many data-question pairs as the text-only input. The criterion for the chart-only subset selection was to ensure a balanced representation of different chart types, namely line, pie, horizontal bar (*h_bar*) and vertical bar (*v_bar*) charts. A total of 100 image-text input pairs with an equal distribution of 25 instances for each chart type was compiled.

3.2 Collection

This section discusses the process of the data collection. The first part presents the possible approaches and highlights the constraints the work faced. Thereafter, the data collection procedure of the two input formats and the conditions for GPT-4 are explained.

3.2.1 Limitations

During the research direct access to the GPT-4 Application Programming Interface (API) was not available to the public. With the API the ability for automated interactions with the model would be given, having the opportunity to collect more data easily.

The user interface (UI) of GPT-4 was utilised as alternative for the data collection. Initially, a script-based approach was considered to automate the process, but the script-operated browser failed the human verification required by the GPT-4 website.

3.2.2 Data Input & Output

Compensating the lack of API access and the unfeasible script-based approach, data was manually entered into the GPT-4 UI. Before starting with the collection, a few prerequisites were set, to ensure the same conditions for all inputs. Every input was only posed to GPT-4 once, i.e. GPT-4 has only one try to provide the correct answer. Furthermore, each data-question pair was inserted in a new GPT-4 chat to avoid the model being influenced by the previous data and questions.

For text-only inputs, the prepared data was pasted directly into the GPT-4 prompt field, as seen in Figure 3.3. The response from GPT-4 to these text-only inputs was then recorded.

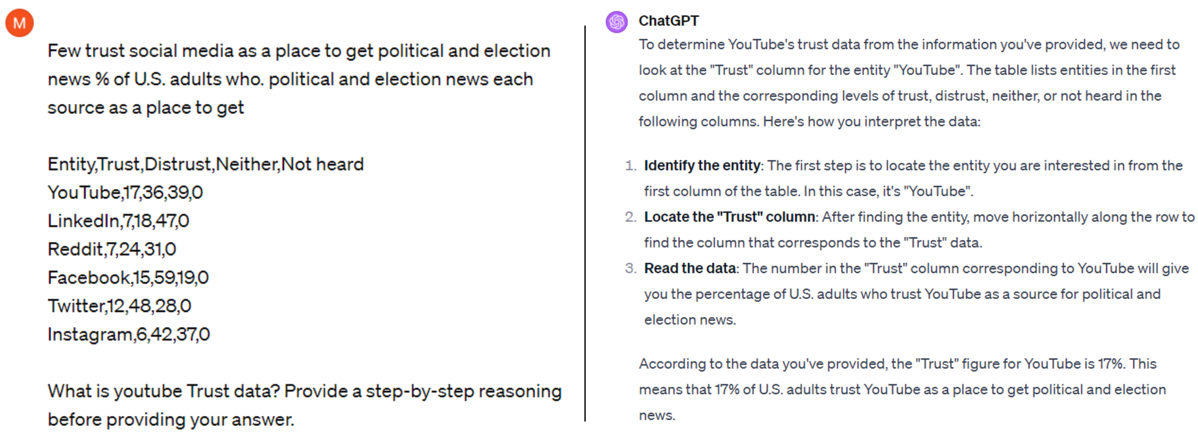

Figure 3.3: Text-only input & GPT-4 response

In the case of chart-text inputs, an image of the relevant chart was uploaded to the GPT-4 interface in a first step. Then the corresponding question with the attachment was added to the prompt (see Figure 3.4). The output generated to these chart-text inputs were appropriately recorded as well. This process was conducted with all the prepared 100 chart-text input pairs.

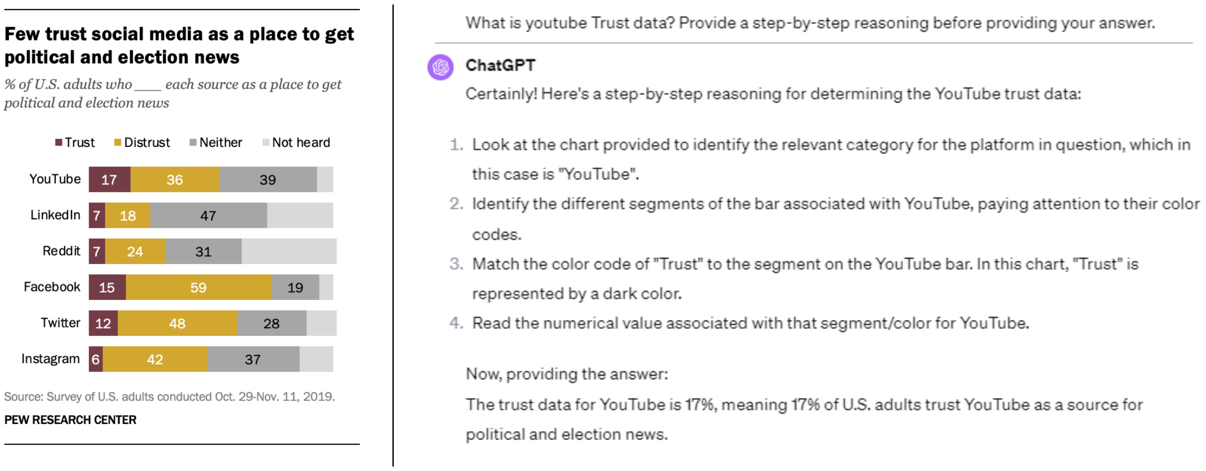

Figure 3.4: Chart-text input & GPT-4 response

4 Parameters for Analysis

In this chapter the parameters are introduced, which are evaluated in the upcoming analysis. The first set of attributes described in Section 4.1 are for the quantitative analysis. Thereafter, Section 4.2 covers the attributes for the qualitative analysis.

4.1 Quantitative Parameters

Quantitative parameters are data attributes that are numerically measurable, referring to the input or output data. Parameters corresponding to the output are the accuracy and the response length of GPT-4. On the other hand, parameters referring to the input are the question taxonomy and the data itself – size or type.

4.1.1 Accuracy & Length of Response

The corresponding answers to the questions never required opinion or interpretation. Therefore, the accuracy could exactly be determined and became an attribute of the quantitative analysis. The importance of correct answers is crucial [LLLZ19], as an inaccurate answer leaves the questioner's state of knowledge unchanged. Mistakes made in the value extraction were equally treated as erroneous calculations or comparisons made in the step-by-step reasoning. In this thesis the parameter accuracy indicates whether an answer is correct or not, at which, only the final answer given by GPT-4 was important.
Besides the accuracy of the answer, this work is also addressing the response length in the analysis. With the incorporation of the query attachment, the response lengthens. A question of research is the dependency of the response length from various factors. By investigating this in Chapter 5, the aim was to find the potential factors influencing it. The findings can then be used in the future to manipulate the length of answers by GPT-4.

4.1.2 Question Taxonomy

All questions in the dataset are human-authored, providing various formulations and types. A grouping of these into distinct categories enabled an analysis of the influence on the responses from GPT-4. It is possible that a question falls in more than one category. The categories are as follows:

1. **Finding Extrema:** This category contains questions that require the identification of extreme values within a dataset [WJBB22].

   List of keywords (FE_list): 'maximum', 'minimum', 'peak', 'highest', 'lowest', 'largest', 'greatest', 'smallest', 'shortest', 'longest', 'tallest', 'biggest', 'most', 'least'

2. **Filtering:** The 'Filtering' category encompasses questions, which locate or identify specific elements within the data [WJBB22].

   List of keywords (F_list): 'in which', 'which', 'represent', 'when', 'find'

3. **Retrieving Value:** In this category, the focus is on extracting specific values within the data [WJBB22].

   List of keywords (RV_list): 'what', 'percent', 'distribution', 'value', 'percentage'

4. **Comparison:** This category involves questions, which require GPT-4 to compare values in terms of size or quantity within the data [YQZ+18].

   List of keywords (CP_list): 'than', 'compare', 'equal', 'more', 'difference', 'between', 'less', 'greater', 'higher', 'same', 'drop', 'gap', 'above', 'below'

5. **Calculation:** The 'Calculation' category comprises terms related to mathematical operations and logical reasoning. The questions require arithmetic operations to derive answers. It is interesting to see how well GPT-4 handles these questions, because in some instances the LLM has problems doing this accurately [YDL+23].

   List of keywords (CL_list): 'how many', 'calculate', 'add', 'substract', 'deduct', 'subtract', 'time', 'ratio', 'average', 'median', 'mode', 'total', 'sum', 'adding', 'summation', 'combined'

The majority of questions are sorted into the 'Retrieving Value' category. In the remaining four, the number of questions in each category of the text-only input is equally distributed. The chart-text dataset has a bit more questions in the 'Calculation' and 'Comparison' categories compared to the 'Find Extrema' and 'Filtering' categories. As shown in Table 4.1 the distribution of questions in the categories between the two input types is pretty similar. Consequently, this allows a more precise evaluation of the comparison between the text-only and chart-text inputs.

Category	Percentage of Questions	
	text	chart+text
FE_list	38.2%	27%
F_list	32.3%	24%
RV_list	73.4%	71%
CP_list	34.1%	42%
CL_list	42.3%	54%

Table 4.1: Distribution of the categories

4.1.3 Input Characteristics

Besides the question taxonomy, other input characteristics could have a potential influence, that are relevant for the quantitative analysis. This section is divided into three subsections: Questions, text-only data and chart data.

Question lengths. In the dataset the lengths of the questions vary due to its diversity. In the quantitative analysis the influence of the question length on the response length is inspected. A longer request could require a more detailed and reasoned reply, especially with the query attachment asking for a step-by-step reasoning. Therefore, it is expected that longer questions lead to longer responses from GPT-4.

Text-only data. In this case GPT-4 processes data that is formatted in a tabular form, specifically as comma-separated values (CSV). This format contains parameters that can potentially influence the answers generated. On the one hand, the number of characters in the CSV-file offers a sense of amount of information, which might affect the model's response. Other possible factors considered were the total number of cells, rows, and columns in the table. They indicate the overall volume of information within the input file, representing its complexity, which are considered in the evaluation as well.

Chart data. The ChartQA dataset provides four different chart types: line, pie, horizontal bar and vertical bar charts (see Figures 4.1, 4.2, 4.3, 4.4). Within the context of chart-text input, these chart types present another variable that could potentially impact the accuracy. The objective was to examine whether GPT-4 demonstrates a differential skill in interpreting various chart types.

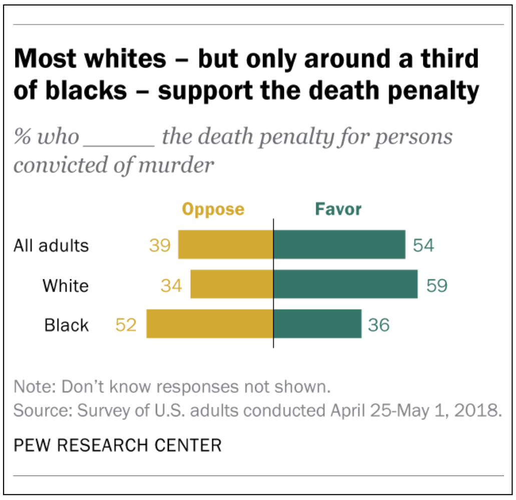

Figure 4.1: Horizontal bar chart

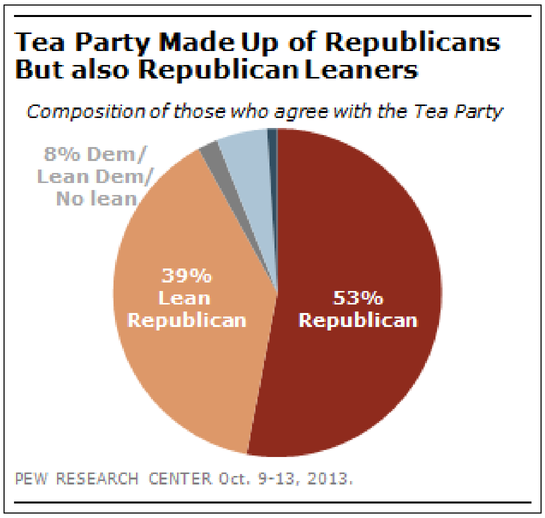

Figure 4.2: Pie chart

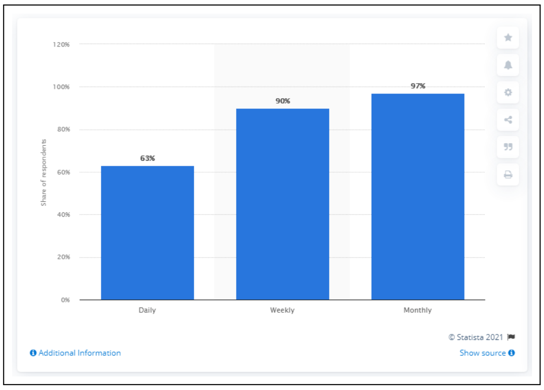

Figure 4.3: Vertical bar chart

Figure 4.4: Line chart

4.2 Qualitative Parameters

Besides the quantitative analysis, this work adds a qualitative analysis. A focal point is on the reasoning part from GPT-4. It is important to determine the quality of the model's answer to analyse how well the model works on charts and text-only tables. Therefore, the accuracy is not the only important criteria, the entire answer needs to be assessed, including the reasoning steps. To achieve a good answer quality GPT-4 should fulfil the following criteria:

1. **Clarity**: The clarity compromises a set of quality criteria, that were applied in the work of Chirag Shah and Jefferey Pomerantz [SP10]. Responses must be of appropriate length and avoiding unnecessary redundancy. Furthermore, the

sequence of steps leading to the answer should be logical to ensure good readability and comprehensibility.

2. **Relevance**: Only data that is directly relevant to the query should be included in the response. This criterion helps to maintain the focus on the primary objective of the question, ensuring that the response does not contain any potential confusion [ZZXL19].

3. **Reasoning**: Answers should not only solve the query but also include the steps taken to arrive at the solution, else the main goal is not fulfilled. Reasoning is a very important aspect in answers, especially for the more complex data, such as charts and tables, because the correctness can then be easily verified [Nua21]. Additionally, this builds more trust towards the Artificial Intelligence, leading to a higher perceived quality [Shi21].

Clarity, relevance, and reasoning are subjective criteria where every user has a different view on good answer quality. Therefore, a user study (see Chapter 6) was conducted, that let users determine these criteria of GPT-4's responses.

5 Quantitative Analysis

This chapter presents findings from examining numerical data collected during the study. In Section 5.1 the accuracy of GPT-4 is evaluated and the causes are presented. Thereafter, Section 5.2 showcases the correlation between the question length and output length.

5.1 GPT-4's Accuracy

The quantitative analysis starts with the general accuracy of GPT-4. In this regard, GPT-4 had demonstrated a noteworthy performance in response accuracy, as depicted in Figure 5.1. The model's proficiency is particularly evident in text-only input scenarios, where it achieves an impressive accuracy of 92.6%. This statistic is derived from 591 data-question pairs, with GPT-4 correctly responding to 547 of these requests. Conversely, the model's performance in chart-text input scenarios, while still respectable, is comparatively lower, with an accuracy of 81%. This suggests that the presence of charts does indeed influence the accuracy, because even when considering only the subset of data-question pairs that involve chart-text, the accuracy remains consistent at 93% (see Figure 5.1).

	Text-only	**Chart-text**	**Text-only subset**
Total questions	591	100	100
Correct answers	547	81	93
Accuracy	92.6%	81%	93%

Table 5.1: Accuracy of inputs

Comparing the results of the chart-text input with the accuracy of the models tested in the ChartQA study, GPT-4 significantly outperforms the models in interpreting chart-text queries [MLT+22]. A model called VL-T5 [CLTB21] exhibited superior performance, with a 40.08% success rate on the ChartQA dataset [MLT+22]. Notwithstanding the fact that the volume of data is significantly smaller, it can be asserted that GPT-4, with an accuracy of 81%, answered correctly at a rate twice that of the VL-T5.

5.1.1 Text-only Accuracy

Further analysis of the text-only input data revealed that the categories of the input questions do not significantly alter the accuracy of GPT's responses. The accuracy across different categories remains fairly consistent, ranging between 93% and 97%. Expected factors such as number of cells, rows, columns or characters of the input data do not influence the accuracy. The incorrect responses were due to reasons such as wrong extracted values, mixing up values in the response itself or computational errors. In some instances, GPT-4 provided verbal explanations, instead of performing the required calculations (see Figure 5.1), i.e. it described the differences between two attributes, focusing on their definitions instead of returning the mathematical difference.

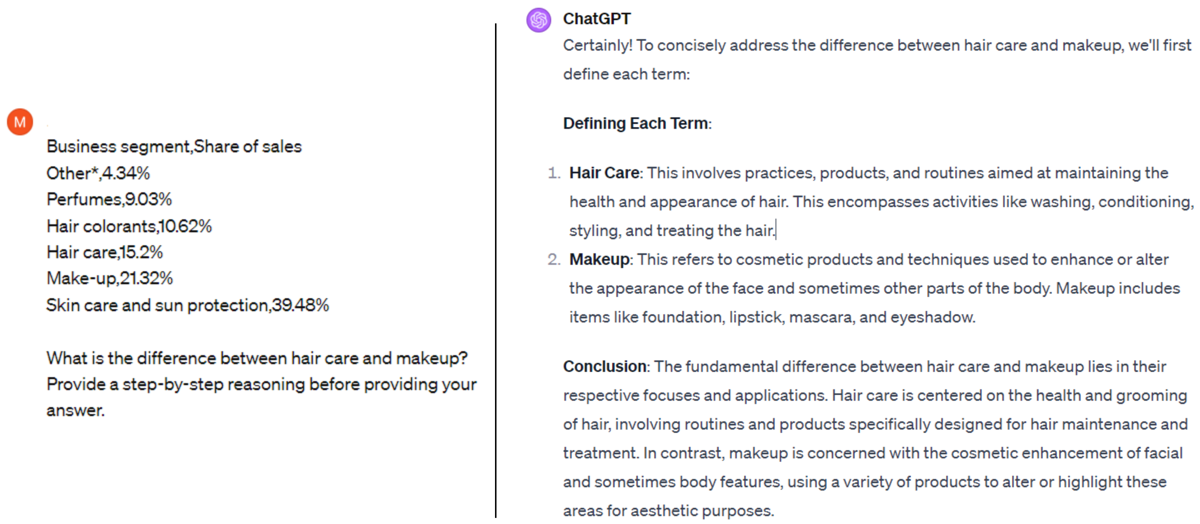

Figure 5.1: Verbal explanation by GPT-4

5.1.2 Chart-text Accuracy

The analysis of chart-text input data revealed more significant parameters influencing the accuracy. At first an analysis with a comparative assessment of the different chart types was performed. Particularly for questions related to line charts GPT-4 struggles to reliably provide correct answers. Its accuracy for these charts stands at just 60%, marking it the lowest among the four chart types included in the dataset (see Figure 5.2).

In comparison, GPT-4 demonstrates improved performance in responding to questions based on pie charts and vertical bar charts, achieving accuracy rates of 84% and 80%. GPT-4 stands out in interpreting horizontal bar charts, accurately answering every query posed in this category. The cause of the varying accuracy across the different chart types are different parameters.

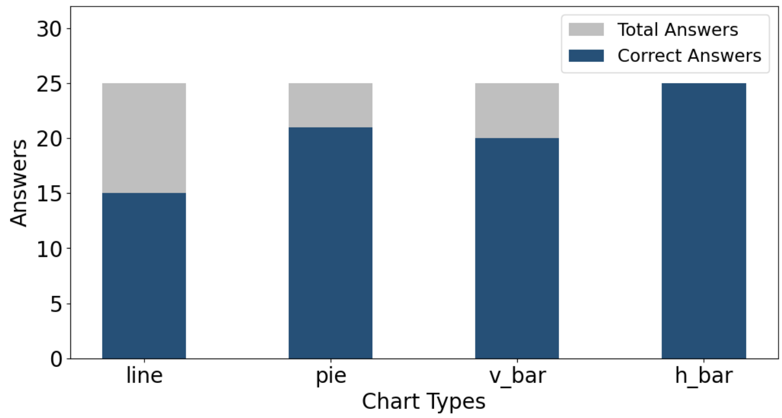

Figure 5.2: Accuracy distribution across chart-types

Line Charts. The main influence was discovered by investigating the different categories. The only category in which GPT-4 performs well is the 'Filtering' category. The four other categories – 'Find Exterma', 'Retrieve Value', 'Comparison', 'Calculation'– range from an accuracy from 46% to 62%. Deeper analysis revealed that the LLM encounters

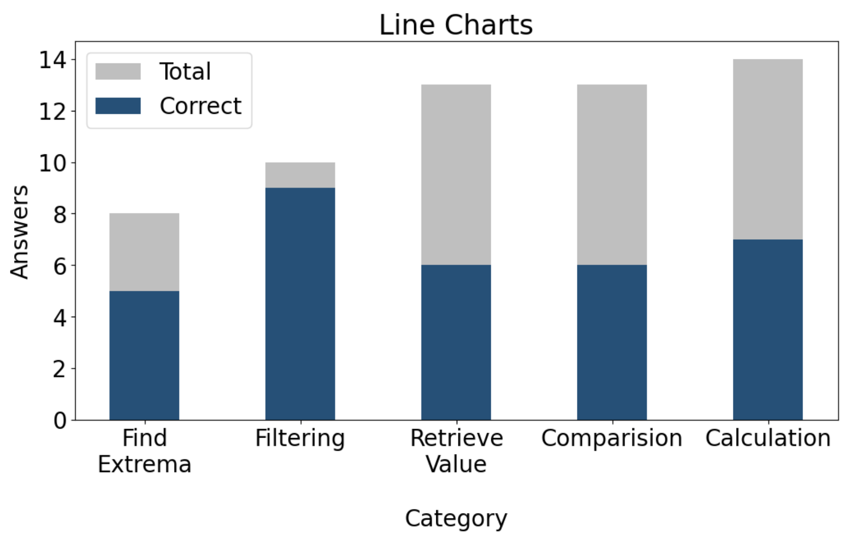

Figure 5.3: Accuracy for line charts

significant challenges in accurately extracting data from line charts. The model had to estimate values of different points on the line, because the lines rarely were labelled with numbers. Inaccurate estimations were leading to incorrect results. Therefore, GPT-4 failed most frequently on questions requiring the extraction of multiple data points, which increases the likelihood of extracting an incorrect value. This issue is particularly pronounced in queries requiring comparison, calculations and finding extrema, where the accuracy articulately decreases (see Figure 5.3). The fact that the retrieving value category performs that bad, is mainly because the queries are a subset from the other three poorly performing categories.

Pie Charts. The accuracy of GPT-4's responses to pie charts is invariant across different categories. Pie charts are mostly clear in the used dataset, therefore GPT-4 does not have difficulties extracting values. The instances of incorrect responses are due to GPT-4 explaining differences verbally, misidentifying colours or omitting calculations. No definitive cause for these errors was identified.

Vertical & Horizontal Bar Charts. The accuracy concerning vertical bar charts differs a lot from horizontal bar charts. This difference in accuracy is the most interesting, because both types are bar charts just rotated. The question categories do not influence the correctness, therefore the charts itself have to be the cause. Comparing the two types of bar charts show that the accuracy is linked to the complexity of the charts. In the dataset are a couple of vertical bar charts that are more complex than the horizontal bar charts. The errors occurred in these scenarios. The complexity is described by the segmentation and the information density. For example, in Figure 5.4 the individual

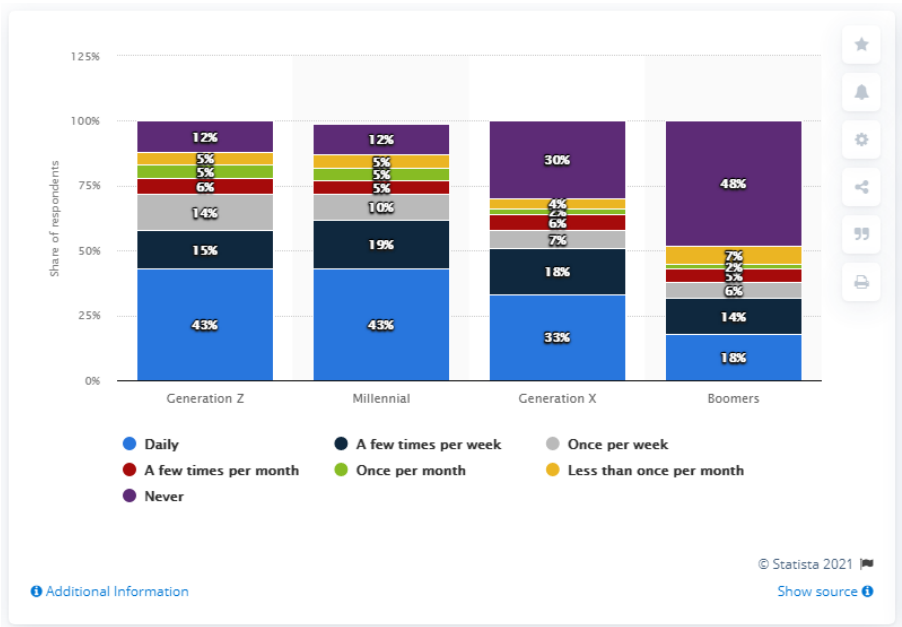

Figure 5.4: Vertical bar chart with multiple segments

bars in the bar chart are split into multiple segments. Additionally, each segment refers to a label by a colour, which GPT-4 could not consistently assign correctly and mixing up values leading to the incorrect answer. Lastly, the numbers are very close to each other, making in harder to extract the correct value. The horizontal bar charts do not have this level of complexity, that results in GPT-4 extracting incorrect values, resulting in incorrect answers.

5.2 Question and Response Lengths

A second conducted analysis was to explore the relationship between the length of questions posed to GPT-4 and the length of its corresponding reasoned responses. The investigation of the correlation between question and response length is interesting, due to the query attachment. The attachment prolongs the response in general and with more details asked this should lead to a longer response. This analysis was only on text-only input because the chart images are too diverse and have various potential factors that could influence the response length of GPT-4. In addition to that, the chart-text dataset size was not large enough to divide it into various subsets for this analysis.

5.2.1 Hypothesis

The hypothesis formulated for this analysis was that there is a correlation between the length of questions and the length of responses by GPT-4. It is expected that longer questions lead to longer responses. A more extensive request requires a more detailed and reasoned response.

5.2.2 Methodology

In this analysis, the Pearson Correlation Coefficient (Pearson's CC) [WJS+21] was used to assess the relationship. The Pearson's CC quantifies the correlation strength, with values ranging from -1, indicating a perfect negative correlation, to +1, indicating a perfect positive correlation. In addition, the statistical significance of the correlation is determined by the p-value [And19]. A p-value less than or equal to 0.05 is regarded as indicative of a statistically significant correlation, suggesting that the observed relationship is not a product of random chance. This analysis was conducted using Python. The lengths of the questions and responses were calculated based on the number of characters, considering only numeric and alphabetical characters in the count.

5.2.3 Results

Utilising the Pearson Correlation Coefficient and the p-value to analyse the relationship between question length and the response length by GPT-4, the study yielded the following results:

1. **Pearson Correlation Coefficient:** 0.19
2. **P-value:** 0.0

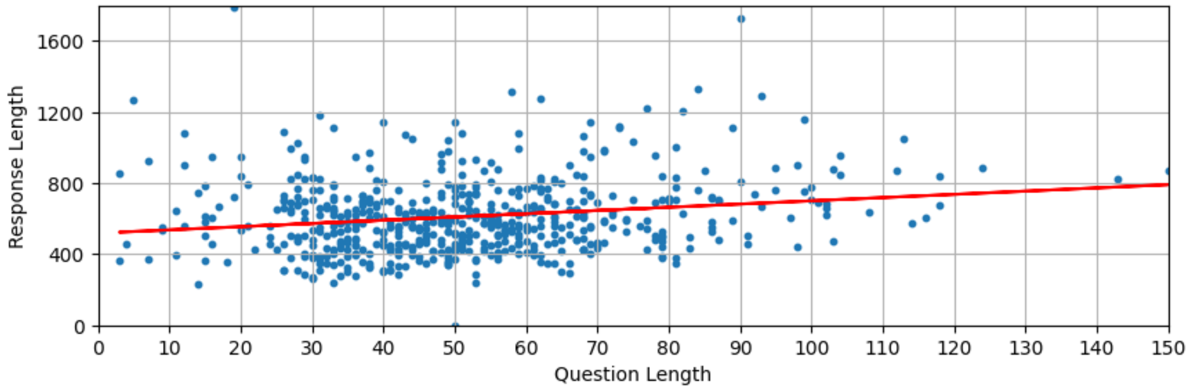

Figure 5.5: Pearson Correlation plot for text-only input

The weak positive Pearson's CC value of 0.19 suggests that the length of the question does have an influence on the length of GPT-4's responses, even though not a strong one (see Figure 5.5). With the p-value being 0.0, the observed correlation is statistically significant, and the relationship is not by random chance. It indicates that other factors might also influence the response length. A factor influencing the Pearson's CC and the p-value are the question categories. By calculating the Pearson correlation coefficient and p-value in the different categories, the Pearson's CC is increasing for the categories that require the extraction of multiple values (Table 5.2). In the 'Filtering' category, this relationship is not even statistically significant.

Even though the Pearson's CC value for the other categories is higher, it remains, stating that the correlation is not particularly strong. Nevertheless, the statistical significance confirms that the correlation between the question length and the response length is not by random chance.

Category	Results	
	Pearson's CC	P-value
FE_list	0.30	0.0
F_list	0.10	0.18
RV_list	0.31	0.0
CP_list	0.41	0.0
CL_list	0.28	0.0

Table 5.2: Pearson Correlation Coefficient in the categories

6 Qualitative Analysis

This chapter discusses the non-numerical attributes. Therefore, a study was conducted, comparing the answer quality of the text-only input with the chart-text input. First, the study questions are presented, followed by a clarification of the study design and finishing off with the results of this study.

To give a better understanding of the answer quality from GPT-4, the subjective perspectives needed to be collected. This study focuses on three criteria of answer quality: clarity, relevance, and reasoning. In addition, the study compares the quality of the text-only and chart-text input types.

6.1 Study Questions

The aim of this research is to understand how GPT-4 processes images, i.e. charts, and text-only inputs and whether there are any notable differences in its quality. Each of the three criteria in this study – clarity, relevance, and reasoning – is addressed by a specific research question (RQ), alternative hypothesis (H) and null hypothesis (H_0). RQ_1 pertains to clarity, RQ_2 to relevance, and RQ_3 to reasoning (see Table 6.1).

$\mathbf{RQ}_{\{1,2,3\}}$	How does the **{clarity, relevance, reasoning}** of GPT-4's responses compare between text-only and chart-text inputs?
$\mathbf{H}_{\{1,2,3\}}$	**{Clarity, Relevance, Reasoning}** is perceived to be higher in text-only inputs than in chart-text inputs.
$\mathbf{H}_{0_{\{1,2,3\}}}$	There is no significant difference in the perceived **{clarity, relevance, reasoning}** between chart-text and text-only inputs.

Table 6.1: Research questions and hypotheses

6.2 Study Design

This study consists of two surveys with ten question blocks each. The blocks were designed to reflect GPT-4's user interface. Participants were first shown a question followed by the corresponding data. After that, they were presented with GPT-4's response to that data-question pair. At the end, there was an evaluation section for the participants to provide their feedback (see Figure 6.1). This evaluation section featured a five-point Likert scale, where the participator were asked to rank the three criteria. In this evaluation section each criterion was explained in a few words.

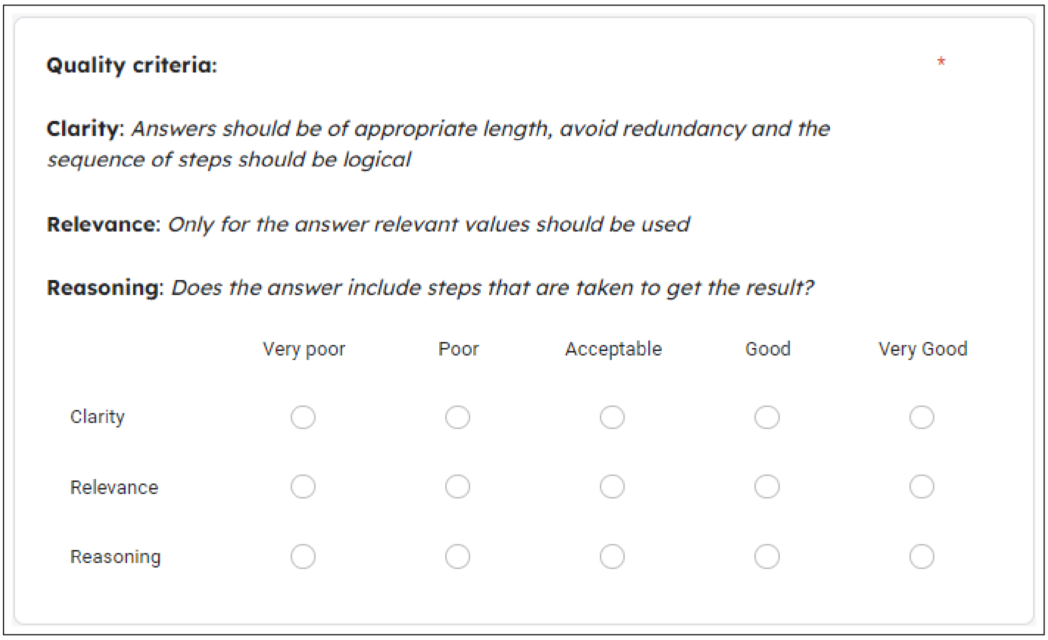

Figure 6.1: Evaluation section

The composition of the study was alternating text-only and chart-text inputs across the two surveys. For instance, if in survey 1 a question was presented with text-only data (Figure 6.2), in survey 2 the same question was paired with chart data (Figure 6.3) and vice versa. The participation of the same 14 individuals in both surveys provided the needed subjective rankings of the different criteria. The volume of each survey was 10 data-question pairs, comprising 5 chart-text and 5 text-only inputs, having a total of 20 different responses by GPT-4.

Figure 6.2: Survey 1

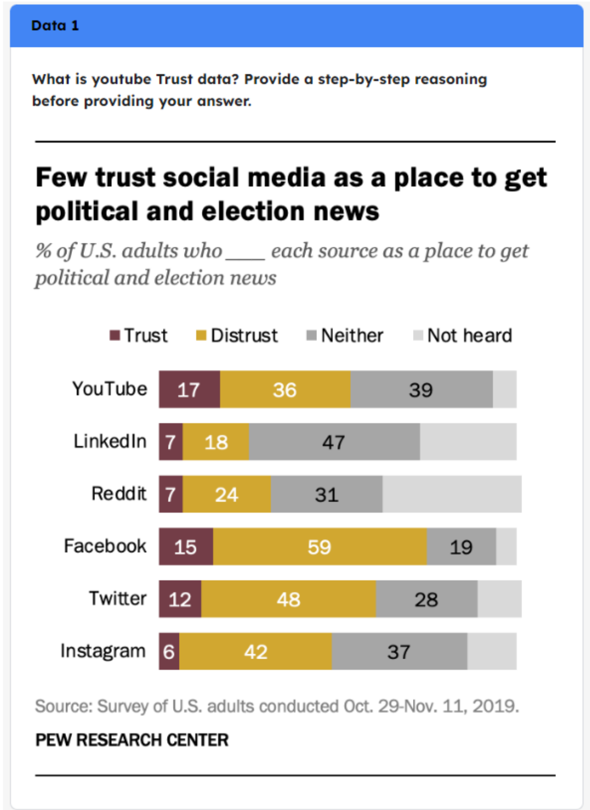

Figure 6.3: Survey 2

6.3 Results

A pairwise T-test [CGHH21] was used to evaluate the differences of the response quality between the two input formats. The degrees of freedom (df) for the T-test was calculated based on the number of pairs of inputs. The 10 chart-text and 10 text-only inputs result in 10 pairs of responses. Therefore, the degrees of freedom is 9. The T-Statistic is a numerical value used to assess whether there is a significant difference between the means of two input formats. A higher absolute T-statistic indicates a larger difference if it is statistical significant. The following table capitalises the T-test results:

Criterion	T-Statistic	P-Value
Clarity	0.94	0.36
Relevance	0.85	0.41
Reasoning	0.17	0.87

Table 6.2: Pairwise T-test results for GPT-4's response quality

Clarity. The hypothesis H_1 posited that text-only inputs would be perceived as clearer than chart-text inputs. However, the T-test yielded a p-value of 0.36, which does not support H_1. This indicates that there is no statistically significant difference in the clarity of responses between the two input formats, leading to the confirmation of the null hypothesis H_{0_1}. The participants rated the clarity of the responses of both input formats in average as good.

Relevance. Similarly are the results regarding the relevance. The Hypothesis H_2 is not supported, due to a p-value of 0.41. Hence, the null hypothesis H_{0_2} is accepted again. Inspecting the quality assessed by the participants, the overall relevance was ranked as good as well.

Reasoning. Lastly, the reasoning, due to the less complex text-only input a better performance of GPT-4 on this input is anticipated. However, a p-value of 0.87 implies no significant difference in reasoning quality between the two types of inputs. The quality of reasoning was rated slightly better than the other criteria, being between good and very good.

The outcomes of the T-test analysis indicate that the input formats do not statistically influence the perceived clarity, relevance, and reasoning. Moreover, the quality is ranked, throughout all criteria, as good.

7 Discussion & Future work

This chapter discusses the findings in Chapter 5 and Chapter 6. Furthermore, possible projects for future work are proposed. The chapter is divided into three sections: GPT-4's accuracy, correlation of input & output lengths and the perceived response quality.

7.1 GPT-4's Accuracy

The analysis indicates that GPT-4 has a high accuracy rate in text-only input scenarios (92.6%), but shows a reduced accuracy in chart-text scenarios (81%). This exhibits that the model's ability to understand and process information from visual elements like charts still requires improvement, while it is proficient in text interpretation. The significant variance in accuracy across different types of charts, showing particularly low values for line charts and high in horizontal bar charts, indicates that GPT-4's performance is heavily dependent on the type of the chart. This is due to the different complexities of the charts and the model's current capabilities in visual data processing [OA+23].

The following aspects should be considered for future work. First, this work limits its analyses to the first response of GPT-4 only, but in some instances, the LLM answered the questions correctly on the second try. The results might improve by providing GPT-4 the input more often and use the majority whether the response is correct or not. Secondly, on chart-text inputs a number of research questions are still open. For example, future work could analyse the model's performance between monochrome and coloured charts. This would give more insights, how much the colour influences the output. Another notable factor is the rapid improvement of GPT-4. The outputs collected over the last few months could now show a lot higher accuracy, due to the numerous updates released after.

7.2 Question & Response Lengths

The relationship between the length of questions posed to GPT-4 and the length of its corresponding reasoned responses was not the main focus of this work, but particularly in text-only input scenarios there is a correlation. The expectation that longer questions would naturally lead to longer, more detailed responses from GPT-4 was not proven. The test performed shows a statistical significance, but a pretty weak on, which indicates that other factors than just the length of the question and the categories are influencing the response length. Therefore, possibly the complexity and the depth of information are influences. Future work can focus on finding these factors to get a deeper understanding of the model's response length. This can be then used to determine how efficiently GPT-4 answers certain questions.

7.3 Perceptions of Quality

The last part of the study is the qualitative analysis (see Chapter 6) of the survey results. This shows that there is no significant difference in clarity, relevance, and reasoning of GPT-4's responses between the two input formats. The hypothesis that responses of the chart-text input are ranked worse than the ones on the text-only input can not be proven. On the positive side this implies that the AI model maintains a consistent quality across different input formats, with an impressive quality.

The limitation of this study are the needed trustworthy participants, due to the challenging questions asked. With too many participants not answering thoughtfully, the results lead to incorrect conclusions. Furthermore, the study does not differentiate between the various chart types. Future research could extend this analysis by taking these factors into account. It is interesting to see if not only the accuracy suffers from the different types, but potentially also the three other criteria. Prolific[1] is a suitable tool for this study to increase the size of the participants pool and by that addressing the limitations.

[1] https://www.prolific.com/

8 Conclusion

In this work, Table and Chart Question Answering data were passed to GPT-4 and the responses were evaluated. A unique aspect of this study was the inclusion of an attachment, containing the instruction to provide a step by step explanation for arriving at the answers. By examining input formats, such as text-only and chart-text, various discoveries were gained into the capabilities and limitations of GPT-4. The model performs outstanding on text-only input. However, when it comes to interpreting and processing information from charts the accuracy is less impressive, especially with line charts and more complex charts, containing dense information. A qualitative analysis demonstrates, that the clarity, relevance, and reasoning is perceived as good, underscoring the high performance of GPT-4. Interestingly, there is no difference in quality between text input and chart text input, that is consistency maintained across both types. Additionally, the investigation of the relationship between the length of questions and responses uncovered a weak but significant correlation. This suggested that factors other than just the question length influenced the responses.
This work not only underscores GPT-4's strengths in text interpretation but also highlights potential areas for improvement when it comes to interacting with charts.

Bibliography

[And19] C. Andrade. “The P value and statistical significance: misunderstandings, explanations, challenges, and alternatives.” In: *Indian journal of psychological medicine* 41.3 (2019), pp. 210–215 (cit. on p. 33).

[CCS+21] W. Chen, M.-W. Chang, E. Schlinger, W. Wang, W. W. Cohen. *Open Question Answering over Tables and Text*. 2021. arXiv: 2010.10439 [cs.CL] (cit. on p. 15).

[CGHH21] K. Church, W. Gale, P. Hanks, D. Hindle. “Using statistics in lexical analysis.” In: *Lexical acquisition*. Psychology Press, 2021, pp. 115–164 (cit. on p. 39).

[CLTB21] J. Cho, J. Lei, H. Tan, M. Bansal. *Unifying Vision-and-Language Tasks via Text Generation*. 2021. arXiv: 2102.02779 [cs.CL] (cit. on pp. 15, 29).

[GEL+23] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, I. Misra. *ImageBind: One Embedding Space To Bind Them All*. 2023. arXiv: 2305.05665 [cs.CV] (cit. on p. 12).

[GSC+19] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z. Yang. “XAI—Explainable artificial intelligence.” In: *Science Robotics* 4 (2019). DOI: 10.1126/scirobotics.aay7120 (cit. on pp. 12, 16).

[KMA+17] S. E. Kahou, V. Michalski, A. Atkinson, Á. Kádár, A. Trischler, Y. Bengio. “Figureqa: An annotated figure dataset for visual reasoning.” In: *arXiv preprint arXiv:1710.07300* (2017) (cit. on pp. 11, 15).

[KPCK18] K. Kafle, B. Price, S. Cohen, C. Kanan. *DVQA: Understanding Data Visualizations via Question Answering*. 2018. arXiv: 1801.08163 [cs.CV] (cit. on pp. 11, 15).

[LLLZ19] M. Li, Y. Li, Y. Lu, Y. Zhang. “Evaluating indicators of answer quality in social Q&A websites.” In: (2019) (cit. on p. 23).

[MGKK20] N. Methani, P. Ganguly, M. M. Khapra, P. Kumar. “Plotqa: Reasoning over scientific plots.” In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2020, pp. 1527–1536 (cit. on pp. 11, 15).

[MLT+22] A. Masry, D. X. Long, J. Q. Tan, S. Joty, E. Hoque. "ChartQA: A benchmark for question answering about charts with visual and logical reasoning." In: *arXiv preprint arXiv:2203.10244* (2022) (cit. on pp. 11, 12, 15, 29).

[NHM+21] L. Nan, C. Hsieh, Z. Mao, X. V. Lin, N. Verma, R. Zhang, W. Kryściński, N. Schoelkopf, R. Kong, X. Tang, M. Mutuma, B. Rosand, I. Trindade, R. Bandaru, J. Cunningham, C. Xiong, D. Radev. *FeTaQA: Free-form Table Question Answering*. 2021. arXiv: 2104.00369 [cs.CL] (cit. on pp. 15, 16).

[Nua21] K. Nuamah. *Deep Algorithmic Question Answering: Towards a Compositionally Hybrid AI for Algorithmic Reasoning*. 2021. arXiv: 2109.08006 [cs.AI] (cit. on p. 27).

[OA+23] OpenAI et al. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL] (cit. on pp. 12, 41).

[PL15] P. Pasupat, P. Liang. *Compositional Semantic Parsing on Semi-Structured Tables*. 2015. arXiv: 1508.00305 [cs.CL] (cit. on p. 15).

[Shi21] D. Shin. "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI." In: *International Journal of Human-Computer Studies* 146 (2021), p. 102551 (cit. on p. 27).

[SP10] C. Shah, J. Pomerantz. "Evaluating and predicting answer quality in community QA." In: *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 2010, pp. 411–418 (cit. on pp. 16, 26).

[TAB+23] G. Team et al. *Gemini: A Family of Highly Capable Multimodal Models*. 2023. arXiv: 2312.11805 [cs.CL] (cit. on p. 12).

[WJBB22] Y. Wang, C. Jiao, M. Bace, A. Bulling. "VisRecall: Quantifying Information Visualisation Recallability via Question Answering." In: *IEEE Transactions on Visualization and Computer Graphics* 28.12 (Dec. 2022), pp. 4995–5005. ISSN: 2160-9306. DOI: 10.1109/tvcg.2022.3198163. URL: http://dx.doi.org/10.1109/TVCG.2022.3198163 (cit. on p. 24).

[WJS+21] A. Wooditch, N. J. Johnson, R. Solymosi, J. Medina Ariza, S. Langton. "Bivariate Correlation." In: *A Beginner's Guide to Statistics for Criminology and Criminal Justice Using R*. Cham: Springer International Publishing, 2021, pp. 227–244. ISBN: 978-3-030-50625-4. DOI: 10.1007/978-3-030-50625-4_14. URL: https://doi.org/10.1007/978-3-030-50625-4_14 (cit. on p. 33).

[WXL+21] Q. Wang, L. Xiao, Y. Lu, Y. Jin, H. He. "Towards reasoning ability in scene text visual question answering." In: *Proceedings of the 29th ACM International Conference on Multimedia*. 2021, pp. 2281–2289 (cit. on p. 15).

[YDL+23] Z. Yang, M. Ding, Q. Lv, Z. Jiang, Z. He, Y. Guo, J. Bai, J. Tang. “Gpt can solve mathematical problems without a calculator.” In: *arXiv preprint arXiv:2309.03241* (2023) (cit. on p. 24).

[YQZ+18] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, C. D. Manning. *HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering*. 2018. arXiv: 1809.09600 `[cs.CL]` (cit. on pp. 15, 24).

[ZTZ+22] B. Zhou, Z. Tan, Z. Zheng, D. Zhou, O. Savkovic, E. Kharlamov. “Towards A Visualisation Ontology for Reusable Visual Analytics.” In: *Proceedings of the 11th International Joint Conference on Knowledge Graphs* (2022). DOI: 10.1145/3579051.3579074 (cit. on p. 12).

[ZZXL19] Y. Zhao, J. Zhang, X. Xia, T. Le. “Evaluation of Google question-answering quality.” In: *Library Hi Tech* 37.2 (2019), pp. 308–324 (cit. on pp. 16, 27).

All links were last followed on Junuary 10, 2024.

Declaration

I hereby declare that the work presented in this thesis is entirely my own and that I did not use any other sources and references than the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted copies.

---

place, date, signature