Is GPT-4 a Good Data Analyst?

Liying Cheng[1] **Xingxuan Li** *[1,2] **Lidong Bing**[1]
[1]DAMO Academy, Alibaba Group [2]Nanyang Technological University, Singapore
{liying.cheng, xingxuan.li, l.bing}@alibaba-inc.com

Abstract

As large language models (LLMs) have demonstrated their powerful capabilities in plenty of domains and tasks, including context understanding, code generation, language generation, data storytelling, etc., many data analysts may raise concerns if their jobs will be replaced by artificial intelligence (AI). This controversial topic has drawn great attention in public. However, we are still at a stage of divergent opinions without any definitive conclusion. Motivated by this, we raise the research question of "*is GPT-4 a good data analyst?*" in this work and aim to answer it by conducting head-to-head comparative studies. In detail, we regard GPT-4 as a data analyst to perform end-to-end data analysis with databases from a wide range of domains. We propose a framework to tackle the problems by carefully designing the prompts for GPT-4 to conduct experiments. We also design several task-specific evaluation metrics to systematically compare the performance between several professional human data analysts and GPT-4. Experimental results show that GPT-4 can achieve comparable performance to humans. We also provide in-depth discussions about our results to shed light on further studies before we reach the conclusion that GPT-4 can replace data analysts. Our code, data and demo are available at: https://github.com/DAMO-NLP-SG/GPT4-as-DataAnalyst.

1 Introduction

LLMs such as GPT series have shown their strong abilities on various tasks in natural language processing (NLP) community, including data annotator (Ding et al., 2022), data evaluator (Chiang and Lee, 2023; Luo et al., 2023; Wang et al., 2023; Wu et al., 2023b; Shen et al., 2023), etc. Outside the NLP community, researchers also evaluate the LLM abilities in multiple domains, such as finance (Wu et al., 2023c), healthcare (Han et al., 2023; Li et al., 2023b), biology (Zheng et al., 2023), law (Sun, 2023), psychology (Li et al., 2023a), etc. Most of these researches demonstrate the effectiveness of LLMs when applying it to different tasks. However, the strong ability in understanding, reasoning, creativity causes some potential anxiety among certain groups of people.

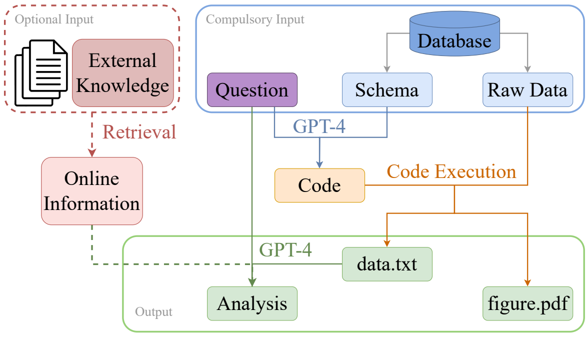


Figure 1: A figure showing the flow of our proposed framework that regarding GPT-4 as a data analyst. The compulsory input information containing both the business question and the database is illustrated in the blue box on the upper right. The optional input referring to the external knowledge source is circled in the red dotted box on the upper left. The outputs including the extracted data (i.e., *"data.txt"*), the data visualization (i.e., *"figure.pdf"*) and the analysis are circled in the green box at the bottom.

As LLMs are introduced and becoming popular not only in NLP community but also in many other areas, those people in and outside of the NLP community are considering or worrying whether AI can replace certain jobs (Noever and Ciolino, 2023; Wu et al., 2023a). One such job role that could be naturally and controversially "replaced" by AI is data analyst (Tsai et al., 2015; Ribeiro et al., 2015). The main and typical job scopes for a data analyst include extracting relevant data from several databases based on business partners' requirements, presenting data visualization in an easily understandable way, and also providing data

*Xingxuan Li is under the Joint Ph.D. Program between Alibaba and Nanyang Technological University.

analysis and insights for the audience. The overall process is shown in Figure 1, which a data analyst is often asked to fulfill during work. This job involves a relatively routine scope, which may become repetitive at times. It also requires several technical skills, including but not limited to SQL, Python, data visualization, and data analysis, making it relatively expensive. As this job scope may adhere to a relatively fixed pipeline, there is a heated public debate about the possibility of an AI replacing a data analyst, which attracts significant attention.

In this paper, we aim to answer the following research question: *Is GPT-4 a good data analyst?* To answer this question, we conduct preliminary studies on GPT-4 to demonstrate its potential capabilities as a data analyst. We quantitatively evaluate the pros and cons of LLM as a data analyst mainly from the following metrics: performance, time, and cost. In specific, we treat GPT-4 as a data analyst to conduct several end-to-end data analysis problems. Since there is no existing dataset for such data analysis problems, we choose one of the most related dataset NvBench, and add the data analysis part on top. We design several automatic and human evaluation metrics to comprehensively evaluate the quality of the data extracted, charts plotted and data analysis generated. Experimental results show that GPT-4 can beat an entry level data analyst in terms of performance and have comparable performance to a senior level data analyst. In terms of cost and time, GPT-4 is much cheaper and faster than hiring a data analyst.

However, since it is a preliminary study on whether GPT-4 is a good data analyst, we provide fruitful discussions on whether the conclusions from our experiments are reliable in real life business from several perspectives, such as whether the questions are practical, whether the human data analysts we choose are representative, etc. To summarize, our contributions include:

- We for the first time raise the research question of whether GPT-4 is a good data analyst, and quantitatively evaluate the pros and cons.
- For such a typical data analyst job scope, we propose an end-to-end automatic framework to conduct data collection, visualization, and analysis.
- We conduct systematic and professional human evaluation on GPT-4's outputs. The data analysis and insights with good quality can be considered as the first benchmark for data analysis in NLP community.

2 Related Work

2.1 Related Tasks and Datasets

Since our task setting is new in NLP community, there is no existing dataset that is fully suitable for our task. We explore the most relevant tasks and datasets. First, the NvBench dataset (Luo et al., 2021) translates natural language (NL) queries to corresponding visualizations (VIS), which covers the first half of the main job scope of a data analyst. This dataset has 153 databases along with 780 tables in total and covers 105 domains, and this task (NL2VIS) has attracted significant attention in both commercial visualization vendors and academic researchers. Another popular subtask of the NL2VIS task is called text-to-sql, which converts natural language questions into SQL queries (Zhong et al., 2017; Guo et al., 2019; Qi et al., 2022; Gao et al., 2022). Spider (Yu et al., 2018), SParC (Yu et al., 2019b) and CoSQL (Yu et al., 2019a) are three main benchmark datasets for text-to-sql tasks. Since this work is more focused on imitating the overall process of the job scope of a data analyst, we adopt the NL2VIS task which has one more step forward than the text-to-sql task.

For the second part of data analysis, we also explore relevant tasks and datasets. Automatic chart summarization (Mittal et al., 1998; Ferres et al., 2013) is a task that aimed to explain a chart and summarize the key takeaways in the form of natural language. Indeed, generating natural language summaries from charts can be very helpful to infer key insights that would otherwise require a lot of cognitive and perceptual efforts. In terms of the dataset, the chart-to-text dataset (Kantharaj et al., 2022) aims to generate a short description of the given chart. This dataset also covers a wide range of topics and chart types. Another relevant NLP task is called data-to-text generation (Gardent et al., 2017; Dušek et al., 2020; Koncel-Kedziorski et al., 2019; Cheng et al., 2020). However, the output of all these existing works are descriptions or summaries in the form of one or a few sentences or a short paragraph. In the practical setting of data analytics work, one should highlight the analysis and insights in bullet points to make it clearer to the audience. Therefore, in this work, we aim to generate the data analysis in the form of bullet points instead of a short paragraph.

2.2 Abilities of GPT-3, ChatGPT and GPT-4

Researchers have demonstrated the effectiveness of GPT-3 and ChatGPT on several tasks (Ding et al., 2022; Chiang and Lee, 2023; Shen et al., 2023; Luo et al., 2023; Wang et al., 2023; Wu et al., 2023b; Li et al., 2023a; Han et al., 2023; Li et al., 2023b). Ding et al. (2022) evaluated the performance of GPT-3 as a data annotator. Their findings show that GPT-3 performs better on simpler tasks such as text classification than more complex tasks such as named entity recognition (NER). Wang et al. (2023) treated ChatGPT as an evaluator, used ChatGPT to evaluate the performance of natural language generation (NLG) and study its correlations with human evaluation. They found that the ChatGPT evaluator has a high correlation with humans in most cases, especially for creative NLG tasks.

GPT-4 is proven to be a significant upgrade over the existing models, as it is able to achieve more advanced natural language processing capabilities (OpenAI, 2023). For instance, GPT-4 is capable of generating more diverse, coherent, and natural language outputs. It is also speculated that GPT-4 may be more capable for providing answers to complex and detailed questions and performing tasks requiring deeper reasoning and inference. These advantages will have practical implications in various industries, such as customer service, finance, healthcare, and education, where AI-powered language processing can enhance communication and problem-solving. In this work, we regard GPT-4 as a data analyst to conduct our experiments.

3 Task Description

3.1 Background: Data Analyst Job Scope

The main job scope of a data analyst involves utilizing business data to identify meaningful patterns and trends from the data and provide stakeholders with valuable insights for making strategic decisions. To achieve their goal, they must possess a variety of skills, including SQL query writing, data cleaning and transformation, visualization generation, and data analysis.

To this end, the major job scope of a data analyst can be split into three steps based on the three main skills mentioned above: data collection, data visualization and data analysis. The initial step involves comprehending business requirement and deciding which data sources are pertinent to answering it. Once the relevant data tables have been identified, the analyst can extract the required data via SQL queries or other extraction tools. The second step is to create visual aids, such as graphs and charts, that effectively convey insights. Finally, in data analysis stage, the analyst may need to ascertain correlations between different data points, identify anomalies and outliers, and track trends over time. The insights derived from this process can then be communicated to stakeholders through written reports or presentations.

3.2 Our Task Setting

Following the main job scope of a data analyst, we describe our task setting below. Formally, as illustrated in Figure 1, given a business-related question (q) and one or more relevant database tables (d) and their schema (s), we aim to extract the required data (D), generate a graph (G) for visualization and provide some analysis and insights (A).

More specifically, according to the given question, one has to identify the relevant tables and schemes in the databases that contain the necessary data for the chart, and then extract the data from the databases and organize it in a way that is suitable for chart generation. One example question can be: "Show me about the correlation between Height and Weight in a scatter chart". As it can be seen, the question also includes the chart type information, thus one also has to choose an appropriate chart type based on the nature of the data and the question being asked, and to generate the chart using a suitable software or programming language. Lastly, it is required to analyze the data to identify trends, patterns, and insights that can help answer the initial question.

4 Our Framework

To tackle the above task setting, we design an end-to-end framework. With GPT-4's abilities on context understanding, code generation, data storytelling being demonstrated, we aim to use GPT-4 to automate the whole data analytics process, following the steps shown in Figure 1. Basically, there are three steps involved: (1) code generation (shown in blue arrows), (2) code execution (shown in orange arrows), and (3) analysis generation (shown in green arrows). The algorithm of our framework is shown in Algorithm 1.

4.1 Step 1: Code Generation

The input of the first step contains a question and database schema. The goal here is to generate

Algorithm 1 GPT-4 as a data analyst

Require: Question q; Database schema s; Database d; Online o
Require: Instruction prompts for code generation p_{code}, analysis generation $p_{analysis}$
Require: LLM $f(\cdot)$; LM decoding temperature τ
Require: An external knowledge retrieval model $g(\cdot)$
Require: Python compiler $h(\cdot)$

```
Q, C ← f(q, s, p_code, τ)                    ▷ Generate SQL query (Q) and Python code (C).
D, G ← h(Q, C, s, d)                         ▷ Execute code to get data (D) and graph (G).
if o is true then                            ▷ Only use online information when instructed.
    I ← g(q, D)                              ▷ Query information from external knowledge source.
    A ← f(q, p_analysis, D, I, τ)            ▷ Generate analysis (A) from data (D) and online information (I).
    return D, G, A
else if o is false then
    A ← f(q, p_analysis, D, τ)               ▷ Generate analysis (A) from data (D).
    return D, G, A
end if
```

Question: [question]

conn = sqlite3.connect([database file name])

[database schema]

Write python code to select relevant data and draw the chart. Please save the plot to "figure.pdf" and save the label and value shown in the graph to "data.txt".

Table 1: Prompt for the first step in our framework: code generation. Text in blue: the specific question, database file name and database schema.

the code for extracting data and drawing the chart in later steps. We utilize GPT-4 to understand the questions and the relations among multiple database tables from the schema. Note that only the schema of the database tables is provided here due to the data security reasons. The massive raw data is still kept safe offline, which will be used in the later step. The designed prompt for this step is shown in Table 1. By following the instructions, we can get a piece of python code containing SQL queries. An example code snippet generated by GPT-4 is shown in Appendix A.

4.2 Step 2: Code Execution

As mentioned earlier in the previous step, to maintain the data safety, we execute the code generated by GPT-4 offline. The input in this step is the code generated from Step 1 and the raw data from the database, as shown in Figure 1. By locating the data directory using "conn = sqlite3.connect([database file name])" as shown in Table 1 in the code, the massive raw data is involved in this step. By executing the python code, we are able to get the chart in "figure.pdf" and the extracted data saved in "data.txt".

4.3 Step 3: Analysis Generation

Question: [question]

[extracted data]

Generate analysis and insights about the data in 5 bullet points.

Table 2: Prompt for the third step in our framework: analysis generation. Text in blue: the specific question and the extracted data as shown in "data.txt".

After we obtain the extracted data, we aim to generate the data analysis and insights. To make sure the data analysis is aligned with the original query, we use both the question and the extracted data as the input. Our designed prompt for GPT-4 of this step is shown in Table 2. Instead of generating a paragraph of description about the extracted data, we instruct GPT-4 to generate the analysis and insights in 5 bullet points to emphasize the key takeaways. Note that we have considered the alternative of using the generated chart as input as well, as the GPT-4 technical report (OpenAI, 2023) mentioned it could take images as input. However, this feature is still not open to public yet. Since the extracted data essentially contains at least the same amount of information as the generated figure, we only use the extracted data here as input for now. From our preliminary experiments, GPT-4 is able to understand the trend and the correlation from the data itself without seeing the figures.

In order to make our framework more practical such that it can potentially help human data analysts boost their daily performance, we add an option of utilizing external knowledge source, as shown in Algorithm 1. Since the actual data analyst role usually requires relevant business background knowledge, we design an external knowledge retrieval model $g(\cdot)$ to query real-time online information (I) from external knowledge source (e.g. Google). In such an option, GPT-4 takes both the data (D) and online information (I) as input to generate the analysis (A).

5 Experiments

5.1 Dataset

Since there is no exactly matched dataset, we choose the most relevant one, named the NvBench dataset. We randomly choose 100 questions from different domains with different chart types and different difficulty levels to conduct our main experiments. The chart types cover the bar chart, the stacked bar chart, the line chart, the scatter chart, the grouping scatter chart and the pie chart. The difficulty levels include: easy, medium, hard and extra hard. The domains include: sports, artists, transportation, apartment rentals, colleges, etc. On top of the existing NvBench dataset, we additionally use our framework to write insights drawn from data in 5 bullet points for each instance and evaluate the quality using our self-designed evaluation metrics.

5.2 Evaluation

To comprehensively investigate the performance, we carefully design several human evaluation metrics to evaluate the generated figures and analysis separately for each test instance.

Figure Evaluation We define 3 evaluation metrics for figures:

- information correctness: is the data and information shown in the figure correct?
- chart type correctness: does the chart type match the requirement in the question?
- aesthetics: is the figure aesthetic and clear without any format errors?

The information correctness and chart type correctness are calculated from 0 to 1, while the aesthetics is on a scale of 0 to 3.

Analysis Evaluation For each bullet point generated in the analysis and insight, we define 4 evaluation metrics as below:

- correctness: does the analysis contain wrong data or information?
- alignment: does the analysis align with the question?
- complexity: how complex and in-depth is the analysis?
- fluency: is the generated analysis fluent, grammatically sound and without unnecessary repetitions?

We grade the correctness and alignment on a scale of 0 to 1, and grade complexity and fluency in a range between 0 to 3.

To conduct human evaluation, 6 professional data annotators are hired from a data annotation company to annotate each figure and analysis bullet point following the evaluation metrics detailed above. The annotators are fully compensated for their work. Each data point is independently labeled by two different annotators.

5.3 Main Results

GPT-4 Performance Table 3 shows the performance of GPT-4 as an data analyst on 200 samples. We show the results of each individual evaluator group and the average scores between these two groups. For chart type correctness evaluation, both evaluator groups give almost full scores. This indicates that for such a simple and clear instruction such as "draw a bar chart", "show a pie chart", etc., GPT-4 can easily understand its meaning and has background knowledge about what the chart type means, so that it can plot the chart in the correct type accordingly. In terms of aesthetics score, it can get 2.73 out of 3 on average, which shows most of the figures generated are clear to audience without any format errors. However, for the information correctness of the plotted charts, the scores are not so satisfied. We manually check those charts and find most of them can roughly get the correct figures despite some small errors. Our evaluation criteria is very strict that as long as any data or any label of x-axis or y-axis is wrong, the score has to be deducted. Nevertheless, it has room for further improvement.

Evaluator	Chart			Data Analysis			
	Info. Correctness	Chart Type	Aesthetics	Correctness	Complexity	Alignment	Fluency
Evaluator Group 1	0.76	0.99	2.68	0.92	2.14	0.99	3.00
Evaluator Group 2	0.77	0.99	2.78	0.95	2.18	1.00	3.00
Average	0.78	0.99	2.73	0.94	2.16	1.00	3.00

Table 3: Performance of GPT-4 as a data analyst.

Annotator	# of Samples	Chart				Data Analysis				
		Info. Correctness	Chart Type	Aesthetics	Time (s)	Correctness	Complexity	Alignment	Fluency	Time (s)
Senior Data Analyst 1	10	0.80	1.00	2.90	645	0.94	1.70	0.94	2.98	225
GPT-4		0.80	0.90	2.90	49	0.88	2.08	1.00	3.00	31
Senior Data Analyst 2	8	0.75	1.00	3.00	420	1.00	2.35	1.00	2.95	428
GPT-4		0.63	1.00	2.00	50	0.75	2.05	0.98	3.00	34
Junior Data Analyst 1	9	0.72	1.00	3.00	633	0.88	1.68	1.00	3.00	293
GPT-4		0.78	1.00	2.89	45	0.91	2.20	1.00	3.00	34

Table 4: Overall Comparison between several senior/junior data analysts and GPT-4 on a few (8 to 10) random examples. Time spent is shown in seconds (s).

Source		Median/Average Annual Salary (USD)	Cost per instance (USD)
levels.fyi	Entry Level DA	37,661	4.81
	Senior DA	90,421	10.71
Glassdoor	Junior DA	50,000	6.38
	Senior DA	86,300	10.22
Our Annotation	Junior DA	-	7
	Senior DA	-	11
GPT-4		-	**0.05**

Table 5: Cost Comparison from different sources.

For analysis evaluation, both alignment and fluency get full marks on average. It again verifies generating fluent and grammatically correct sentences is definitely not a problem for GPT-4. We notice the average correctness score for analysis is much higher than the information correctness score for figures. This is interesting because despite the wrong figure generated, the analysis could be correct. It again verifies our explanation earlier for the information correctness scores of figures. As mentioned, since the figures generated are mostly consistent with the gold figures, thus some of the bullet points can be generated correctly. Only a few bullet points related to the error parts from the figures are considered wrong. In terms of the complexity scores, 2.16 out of 3 on average is reasonable and satisfying.

Comparison between Human Data Analysts and GPT-4 To further answer our research question, we hire professional data analysts to do these tasks and compare with GPT-4 comprehensively. We fully compensate them for their annotation. Table 4 shows the performance of several data analysts of different expert levels from different backgrounds compared to GPT-4. Overall speaking, GPT-4's performance is comparable to human data analysts, while the superiority varies among different metrics and human data analysts.

The first block shows 10-sample performance of a senior data analyst (i.e., Senior Data Analyst 1) who has more than 6 years' data analysis working experience in finance industry. We can see from the table that GPT-4 performance is comparable to the expert data analyst on most of the metrics. Though the correctness score of GPT-4 is lower than the human data analyst, the complexity score and the alignment score are higher.

The second block shows another 8-sample performance comparison between GPT-4 and another senior data analyst (i.e., Senior Data Analyst 2) who works in internet industry as a data analyst for over 5 years. Since the sample size is relatively smaller, the results shows larger variance between human and AI data analysts. The human data analyst surpasses GPT-4 on information correctness and aesthetics of figures, correctness and complexity of insights, indicating that GPT-4 still still has potential for improvement.

The third block compares another random 9-sample performance between GPT-4 and a junior data analyst who has data analysis working experience in a consulting firm within 2 years. GPT-4 not only performs better on the correctness of figures and analysis, but also tends to generate more complex analysis than the human data analyst.

Apart from the comparable performance between all data analysts and GPT-4, we can notice the time spent by GPT-4 is much shorter than human data analysts. Table 5 shows the cost comparison from different sources. We obtain the median annual salary of data analysts in Singapore from level.fyi and the average annual salary of data analysts in Singapore from Glassdoor. We assume there are around 21 working days per month and the daily working hour is around 8 hours, and calculate the cost per instance in USD based on the average time spent by data analysts from each level. For our annotation, we pay the data analysts based on the market rate accordingly. The cost of GPT-4 is approximately 0.71% of the cost of a junior data analyst and 0.45% of the cost of a senior data analyst.

6 Case Study

In this section, we show a few cases done by GPT-4 and our data analysts.

In the first case as shown in Table 6, GPT-4 is able to generate a python code containing the correct SQL query to extract the required data, and to draw a proper and correct pie chart according to the given question. In terms of the analysis, GPT-4 is capable of understanding the data by conducting proper comparisons (e.g., "most successful", "less successful", "diverse range"). In addition, GPT-4 can provide some insights from the data, such as: "indicating their dominance in the competition". These aforementioned abilities of GPT-4 including context understanding, code generation and data storytelling are also demonstrated in many other cases. Furthermore, in this case, GPT-4 can also make some reasonable guess from the data and its background knowledge, such as: "potentially due to their design, performance, or other factors".

The second case (Table 7) shows another question addressed by GPT-4. Again, GPT-4 is able to extract the correct data, draw the correct scatter plot and generate reasonable analysis. Although most of the bullet points are generated faithfully, if we read and check carefully, we can notice the numbers of the average height and weight are wrong. Apart from the well-known hallucination issue, we suspect that GPT-4's calculation ability is not strong, especially for those complex calculation. We also notice this issue in several other cases. Although GPT-4 generates the analysis bullet point in a very confident tone, but the calculation is sometimes inaccurate.

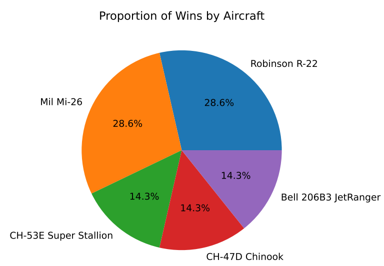
Question	Please list the proportion number of each winning aircraft.
SQL Query	SELECT a.Aircraft, COUNT(m.Winning_Aircraft) as wins FROM aircraft a JOIN match m ON a.Aircraft_ID = m.Winning_Aircraft GROUP BY a.Aircraft ORDER BY wins DESC
Figure	
Analysis	1. The Robinson R-22 and Mil Mi-26 are the most successful aircraft in the dataset, with each having 2 wins. 2. The CH-53E Super Stallion, CH-47D Chinook, and Bell 206B3 JetRanger each have 1 win, making them less successful compared to the Robinson R-22 and Mil Mi-26. 3. The Robinson R-22 and Mil Mi-26 together account for 50% of the total wins, indicating their dominance in the competition. 4. The dataset includes a diverse range of aircraft, from helicopters like the CH-53E Super Stallion to smaller aircraft like the Robinson R-22, showcasing the variety of winning aircraft. 5. Despite the variety of aircraft, the data suggests that certain models, such as the Robinson R-22 and Mil Mi-26, have a higher likelihood of winning, potentially due to their design, performance, or other factors.

Table 6: Case Study 1 by GPT-4.

Table 8 shows an example done by Senior Analyst 2. We can notice that this expert human data analyst can also understand the requirement, write the code to draw the correct bar chart, and analyze the extracted data in bullet points. Apart from this, we can summarize three main differences with GPT-4. First, different from GPT-4, the human data analyst can express the analysis with some personal thoughts and emotions. For example, the data analyst mentions "This is a bit surprising ...". In real-life business, personal emotions are important sometimes. With the emotional phrases, the audience can easily understand whether the data is as expected or abnormal. Second, the human data

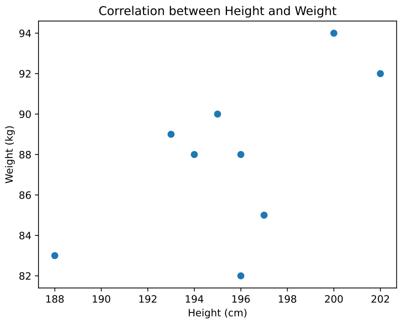
Question	Show me about the correlation between Height and Weight in a scatter chart.
Figure	
Analysis	1. The data provided consists of 9 data points, each representing the height and weight of an individual. 2. The height range in the dataset is from 188 cm to 202 cm, while the weight range is from 82 kg to 94 kg. 3. The average height of the individuals in the dataset is approximately 195.3 cm, and the average weight is approximately 88.1 kg. 4. There seems to be a positive correlation between height and weight, as taller individuals generally have higher weights. 5. A scatter chart would help visualize this correlation, with height on the x-axis and weight on the y-axis. The data points would likely form an upward trend, indicating that as height increases, weight also tends to increase.

Table 7: Case Study 2 by GPT-4.

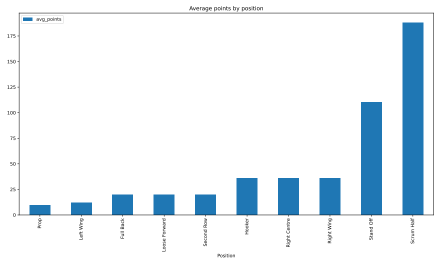
Question	List the position of players and the average number of points of players of each position. Visualize by bar chart, and could you sort by the total number in ascending?
Figure	
Analysis	1. There are 10 positions. Based on the names, this dataset is about Rugby. Rugby is a group sport that is commonly seen in the US, Canada and Commonwealth countries. 2. Stand Off and Scrum Half are having more than 100 average points while the rest are below 40. This is a bit surprising as usually the Right Wing and Left Wing are the star scorers. 3. Prop has the lowest average points of only 9.33, as Prop's main role is to be in the first row and push against the opposition's Prop. 4. Full Bck, Loose Forward and Second Row has the same points (20), while Hooker, Right Centre and Right Wing has the same points (36). 5. Overall the dataset is slightly different from the general understanding. If there's no data issue, it might worth to reevaluate the players and think about a better position assignment.

Table 8: Case Study 3 by Senior Data Analyst 2.

analyst tends to apply some background knowledge. While GPT-4 usually only focuses on the extracted data itself, the human is easily linked with one's background knowledge. For example, as shown in Table 8, the data analyst mentions "... is commonly seen ...", which is more natural during a data analyst's actual job. Therefore, to mimic a human data analyst better, in our demo, we add an option of using Google search API to extract real-time online information when generating data analysis. Third, when providing insights or suggestions, a human data analyst tends to be conservative. For instance, in the 5th bullet point, the human data analyst mentions "If there's no data issue" before giving a suggestion. Unlike humans, GPT-4 will directly provide the suggestion in a confident tone without mentioning its assumptions.

7 Findings and Discussions

During our experiments, we notice several phenomena and conduct some thinking about them. In this section, we discuss our findings and hope researchers can address some of the issues in future work.

Generally speaking, GPT-4 can perform comparable to a data analyst from our preliminary experiments, while there are still several issues to be addressed before we can reach a conclusion that GPT-4 is a good data analyst. First, as illustrated in the case study section, GPT-4 still has hallucination problems, which is also mentioned in GPT-4 technical report (OpenAI, 2023). Data analysis jobs not only require those technical skills and analytics skills, but also requires high accuracy to be guaranteed. Therefore, a professional data analyst always tries to avoid those mistakes including calculation mistakes and any type of hallucination problems. Second, before providing insightful suggestions, a professional data analyst is usually confident about all the assumptions. Instead of directly giving any suggestion or making any guess from the data, GPT-4 should be careful about all the assumptions and make the claims rigorous.

The questions we choose are randomly selected from the NvBench dataset. Although the questions indeed cover a lot of domains, databases, difficulty levels and chart types, they are still somewhat too specific according to human data analysts' feedback. The questions usually contain the information such as: a specific correlation between two variables, a specific chart type. In a more practical setting, the original requirements are more general, which require a data analyst to formulate such a specific question from the general business requirement, and to determine what kind of chart would represent the data better. Our next step is to collect more practical and general questions to further test the problem formulation ability of GPT-4.

The quantity of human evaluation and data analyst annotation is relatively small due to budget limit. For human evaluation, we strictly select those professional evaluators in order to give better ratings. They have to pass our test annotations for several rounds before starting the human evaluation. For the selection of data analysts, we are even more strict. We verify if they really had data analysis working experience, and make sure they master those technical skills before starting the data annotation. However, since hiring a human data analyst (especially for those senior and expert human data analyst) is very expensive, we can only ask them to do a few samples.

8 Conclusions

The potential for large language models (LLMs) like GPT-4 to replace human data analysts has sparked a controversial discussion. However, there is no definitive conclusion on this topic yet. This study aims to answer the research question of whether GPT-4 can perform as a good data analyst by conducting several preliminary experiments. We design a framework to prompt GPT-4 to perform end-to-end data analysis with databases from various domains and compared its performance with several professional human data analysts using carefully-designed task-specific evaluation metrics. The results and analysis show that GPT-4 can achieve comparable performance to humans, but further studies are needed before concluding that GPT-4 can replace data analysts.

Acknowledgements

We would like to thank our data annotators and data evaluators for their hard work. Especially, we would like to thank Mingjie Lyu for the fruitful discussion and feedback.

References

Liying Cheng, Dekun Wu, Lidong Bing, Yan Zhang, Zhanming Jie, Wei Lu, and Luo Si. 2020. Ent-desc: Entity description generation by exploring knowledge graph. In *Conference on Empirical Methods in Natural Language Processing*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of ACL*.

Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq Joty, and Boyang Li. 2022. Is gpt-3 a good data annotator?

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge. *Computer Speech & Language*.

Leo Ferres, Gitte Lindgaard, Livia Sumegi, and Bruce Tsuji. 2013. Evaluating a tool for improving accessibility to charts and graphs. *ACM Trans. Comput. Hum. Interact.*, 20:28:1–28:32.

Chang Gao, Bowen Li, Wenxuan Zhang, Wai Lam, Binhua Li, Fei Huang, Luo Si, and Yongbin Li. 2022. Towards generalizable and robust text-to-sql parsing. In *Conference on Empirical Methods in Natural Language Processing*.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The webnlg challenge: Generating text from rdf data. In *Proceedings of INLG*.

Jiaqi Guo, Zecheng Zhan, Yan Gao, Yan Xiao, Jian-Guang Lou, Ting Liu, and D. Zhang. 2019. Towards complex text-to-sql in cross-domain database with intermediate representation. *ArXiv*, abs/1905.08205.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2023. Medalpaca – an open-source collection of medical conversational ai models and training data.

Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022. Chart-to-text: A large-scale benchmark for chart summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4005–4023, Dublin, Ireland. Association for Computational Linguistics.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. Text generation from knowledge graphs with graph transformers. In *Proceedings of ACL*.

Xingxuan Li, Yutong Li, Shafiq Joty, Linlin Liu, Fei Huang, Lin Qiu, and Lidong Bing. 2023a. Does gpt-3 demonstrate psychopathy? evaluating large language models from a psychological perspective.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, and You Zhang. 2023b. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge.

Yuyu Luo, Nan Tang, Guoliang Li, Chengliang Chai, Wenbo Li, and Xuedi Qin. 2021. Synthesizing natural language to visualization (nl2vis) benchmarks from nl2sql benchmarks. In *Proceedings of the 2021 International Conference on Management of Data, SIGMOD Conference 2021, June 20–25, 2021, Virtual Event, China*. ACM.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint arXiv:2303.15621*.

Vibhu O. Mittal, Johanna D. Moore, Giuseppe Carenini, and Steven Roth. 1998. Describing complex charts in natural language: A caption generation system. *Computational Linguistics*, 24(3):431–467.

David Noever and Matt Ciolino. 2023. Professional certification benchmark dataset: The first 500 jobs for large language models.

OpenAI. 2023. Gpt-4 technical report.

Jiexing Qi, Jingyao Tang, Ziwei He, Xiangpeng Wan, Yu Cheng, Chenghu Zhou, Xinbing Wang, Quanshi Zhang, and Zhouhan Lin. 2022. RASAT: Integrating relational structures into pretrained Seq2Seq model for text-to-SQL. In *Proceedings of EMNLP*, pages 3215–3229, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

André Ribeiro, Afonso Silva, and Alberto Rodrigues da Silva. 2015. Data modeling and data analytics: A survey from a big data perspective. *Journal of Software Engineering and Applications*, 08:617–634.

Chenhui Shen, Liying Cheng, Yang You, and Lidong Bing. 2023. Are large language models good evaluators for abstractive summarization?

Zhongxiang Sun. 2023. A short survey of viewing large language models in legal aspect.

Chun-Wei Tsai, Chin-Feng Lai, H. C. Chao, and Athanasios V. Vasilakos. 2015. Big data analytics: a survey. *Journal of Big Data*, 2:1–32.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Hong Lin. 2023a. Ai-generated content (aigc): A survey.

Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. 2023b. Large language models are diverse role-players for summarization evaluation. *arXiv preprint arXiv:2303.15078*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023c. Bloomberggpt: A large language model for finance.

Tao Yu, Rui Zhang, Heyang Er, Suyi Li, Eric Xue, Bo Pang, Xi Victoria Lin, Yi Chern Tan, Tianze Shi, Zihan Li, Youxuan Jiang, Michihiro Yasunaga, Sungrok Shim, Tao Chen, Alexander Fabbri, Zifan Li, Luyao Chen, Yuwen Zhang, Shreya Dixit, Vincent Zhang, Caiming Xiong, Richard Socher, Walter Lasecki, and Dragomir Radev. 2019a. CoSQL: A conversational text-to-SQL challenge towards cross-domain natural language interfaces to databases. 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 1962–1979, Hong Kong, China. Association for Computational Linguistics.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.

Tao Yu, Rui Zhang, Michihiro Yasunaga, Yi Chern Tan, Xi Victoria Lin, Suyi Li, Heyang Er, Irene Li, Bo Pang, Tao Chen, Emily Ji, Shreya Dixit, David Proctor, Sungrok Shim, Jonathan Kraft, Vincent Zhang, Caiming Xiong, Richard Socher, and Dragomir Radev. 2019b. SParC: Cross-domain semantic parsing in context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4511–4523, Florence, Italy. Association for Computational Linguistics.

Zaixiang Zheng, Yifan Deng, Dongyu Xue, Yi Zhou, Fei YE, and Quanquan Gu. 2023. Structure-informed language models are protein designers.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.

A Example Code

```python
import sqlite3
import matplotlib.pyplot as plt

# Connect to the database
conn = sqlite3.connect("aircraft.sqlite")
cursor = conn.cursor()

# Query to get the count of each winning aircraft
query = '''
SELECT a.Aircraft, COUNT(m.Winning_Aircraft) as wins
FROM aircraft a
JOIN match m ON a.Aircraft_ID = m.Winning_Aircraft
GROUP BY a.Aircraft
ORDER BY wins DESC
'''

# Execute the query and fetch the results
cursor.execute(query)
results = cursor.fetchall()

# Extract the labels and values for the chart
labels = [row[0] for row in results]
values = [row[1] for row in results]

# Save the labels and values to data.txt
with open("data.txt", "w") as f:
    for label, value in zip(labels, values):
        f.write(f"{label}: {value}\n")

# Create the pie chart
plt.pie(values, labels=labels, autopct="%1.1f%%")
plt.title("Proportion of Wins by Aircraft")

# Save the chart to figure.pdf
plt.savefig("figure.pdf")

# Close the database connection
conn.close()
```

Figure 2: An example of a complete code generated by GPT-4. This code is to answer the question shown in Table 6.

Figure 2 shows an example code generated by GPT-4. First, we can notice that GPT-4 is capable of writing SQL queries with several commands, such as JOIN, GROUP BY, ORDER BY to extract the required data. Second, GPT-4 knows how to use multiple Python packages including sqlite and matplotlib, which help to connect the databases and draw charts respectively. Third, GPT-4 can understand the requirement in the question to save the data and figure into the correct files accordingly. Last but not least, it can also generate comments understandable by readers, which is aligned with the goal of helping human data analysts to boost their daily performance. In the case when the wrong code is generated, a human analyst can easily understand which part goes wrong with the aid of the comments.