How effective is an LLM-based Data Analysis Automation Tool? A Case Study with ChatGPT's Data Analyst

Beatriz A. de Miranda[1], **Claudio E. C. Campelo**[1]

[1]Systems and Computing Department
Federal University of Campina Grande (UFCG)
58.109-970 – Campina Grande – PB – Brazil

beatriz.miranda@ccc.ufcg.edu.br, campelo@dsc.ufcg.edu.br

Abstract. *Artificial Intelligence (AI) tools are increasingly becoming integral to analytical processes. This paper evaluates the potential of Large Language Models (LLMs), specifically OpenAI's ChatGPT's Data Analyst, in data analysis. We conducted a structured experiment employing this tool in 36 questions spanning descriptive, diagnostic, predictive, and prescriptive analyses to assess its effectiveness. The study revealed an overall efficiency rate of 86.11%, with robust performance in the descriptive and diagnostic categories but reduced efficacy in the more complex predictive and prescriptive tasks. By discussing the strengths and limitations of a state-of-the-art LLM-based tool in aiding data scientists, this study aims to mark a critical milestone for future developments in the field, particularly as a reference for the open-source community.*

1. Introduction

Among the most significant innovations in Artificial Intelligence (AI) are the Large Language Models (LLMs). Examples include the commercial models from the ChatGPT [1] family [Solaiman et al. 2019, Achiam et al. 2024], developed by OpenAI[2] and Google's [3] Gemini [Team et al. 2023], as well as open-source models like Mistral [Jiang et al. 2023] and LLaMA [Touvron et al. 2023]. These models are revolutionizing communication between humans and machines, demonstrating an exceptional ability to understand and produce human language, and positioning themselves at the forefront of AI innovations [Ouyang et al. 2022]. This article explores the impact of these models on data analysis, with a focus on the ChatGPT's Data Analyst[4], a tool specifically developed to enhance analytical tasks.

Recent studies highlight GPT-4's significant potential and effectiveness in data analysis [Zhang et al. 2023, Ding et al. 2023, Jaimovitch-López et al. 2022], often performing on par with human data analysts [Cheng et al. 2023]. The ChatGPT's Data Analyst, built on the GPT-4 architecture, excels in analytical tasks within a Python environment, producing detailed responses and code outputs. Despite the capabilities, these researches has primarily focused on the GPT API or specific applications of the Data Analyst tool [Daibes and Lima 2024], rather than analyzing the effectiveness of the tool itself. This article addresses this gap by conducting a comprehensive case study to explore the

[1]ChatGPT - https://chat.openai.com/
[2]OpenAI - https://openai.com/
[3]Google - https://www.google.com/
[4]ChatGPT Data Analyst - https://chatgpt.com/g/g-HMNcP6w7d-data-analyst

Data Analyst's potential in data analysis. We evaluate the tool across four data analysis categories – Descriptive, Diagnostic, Predictive, and Prescriptive – using 36 questions of different levels of complexity, showcasing its accuracy, adaptability, and limitations.

Although open-source models have proven to be competitive, often outperforming commercial counterparts in various benchmarks, there remains a scarcity of open tools comparable to ChatGPT's Data Analyst. This gap in the open-source landscape fosters hope for the emergence of a similar community-driven tool. Thus, by assessing the current capabilities of the Data Analyst, our objective is to stimulate the development of similar open technologies by establishing a reference baseline.

To the best of our knowledge, this is the first study that analyzes the performance of the ChatGPT's Data Analyst through a comprehensive case study. Our findings reveal the tool's proficiency in descriptive and diagnostic tasks, while highlighting hurdles in prescriptive and predictive analysis due to the handling of complex data. Significant constraints include data processing capacity, susceptibility to errors, and operational glitches necessitating session restarts. This study establishes a foundation for future research, providing insights into the use of LLMs for data analysis and addressing both its main capabilities and technical limitations in practical scenarios.

2. Related Work

This section presents a review of relevant studies and prior work in the field, providing essential context for the current research.

2.1. Applications of LLMs in Data Analysis

The specialization of LLMs for specific domain tasks underscores their adaptability and potential to be tailored to unique needs, highlighting the versatility of these models. The studies [Ding et al. 2023, Cheng et al. 2023] play a crucial role in understanding the effectiveness of LLMs in data analysis tasks. These studies evaluate the accuracy and efficiency of GPT-3 and GPT-4 in data annotation and analysis, respectively, and both demonstrate efficacy in their respective areas, providing valuable insights for our case study on the expected performance of LLMs like the ChatGPT Data Analyst in similar analytical operations.

Furthermore, as [Sharma et al. 2023] proposes, LLMs can be customized to automate data transformations in specific industries, such as the energy sector. This adaptation enabled the model to perform complex data transformations, significantly reducing time and effort compared to traditional methods. This study's relevance to our work lies in its practical demonstration of LLMs' capability to efficiently manage and transform data in specific, real-world scenarios, an essential factor in the data analysis process.

2.2. LLMs in Analytical Task Automation

Research into the automation of analytical tasks through LLMs is highlighted in the studies by [Nasseri et al. 2023, Jaimovitch-López et al. 2022]. These studies demonstrate how LLMs can streamline and automate data preparation and manipulation, achieving effective results across essential tasks in the analytical workflow. This aligns with our investigation into how the ChatGPT Data Analyst can facilitate similar processes, thereby enhancing the efficiency and accuracy of analyses.

Moreover, the study conducted by [Zhang et al. 2024] benchmarks various data science agents, including LLMs, to evaluate their performance across diverse analytical tasks. This research is pivotal for future studies as it provides a comparative framework to assess the ChatGPT's Data Analyst's performance relative to other data science agents. Additionally, [Kasetty et al. 2024, Liu et al. 2024] highlights the challenges LLMs face in complex reasoning, which are essential for the advanced tasks of the ChatGPT's Data Analyst. Updates in the GPT-4 architecture [Achiam et al. 2024, Bubeck et al. 2023] suggest potential improvements, setting the stage for future advancements in LLMs for data analysis.

3. Methodology

Our evaluation methodology is based on four data analysis categories: Descriptive, Diagnostic, Predictive, and Prescriptive. We defined 36 questions of three different levels of complexity: Basic, Moderate, and Challenging.

This section provides an overview of the methods employed to conduct this study, encompassing the dataset used and its preprocessing approach; the rationale behind the definition of the questions explored in the case study; and the evaluation criteria.

3.1. Interaction with Data Analyst

The code used for our experiments is publicly available on GitHub[5]. The process for interacting with the Data Analyst is shown in Figure 1. It starts with uploading the input data and submitting the data preprocessing prompt into the chat. Then, a data analysis question is selected and prompted into the chat.

For each question, the tool provides a written response along with the corresponding Python code. We then copy the generated code and execute it on Google Colab to assess its response and determine its correctness. If an error occurs during the code execution, the response is deemed incorrect. If the code runs without errors, we evaluate whether the result is satisfactory or unsatisfactory. An example of an unsatisfactory result would be the tool failing to adequately address the question, generating code that does not meet the specified requirements. While this specific case did not occur in our experiments, there was an instance where the model generated by the tool exhibited low accuracy in the Predictive Analysis question of moderate level 3, which was consequently classified as an incorrect response.

3.2. Dataset

The dataset for this study comprises anonymized academic records from students at the Federal University of Campina Grande (UFCG), produced by the system that monitors students' academic progress. It initially contained 150,703 records across 34 columns, totaling 37.9MB. To optimize the performance of the tool, which is recommended for datasets under 10MB, a stratified sample was taken. This sample, constituting 20% of the total data based on the course sector column, reduced the dataset to 8.2MB with 30,130 records, maintaining all original columns and ensuring representation across all sectors.

[5]GitHub Repository - https://github.com/beatrizadm/chatgpts-data-analyst

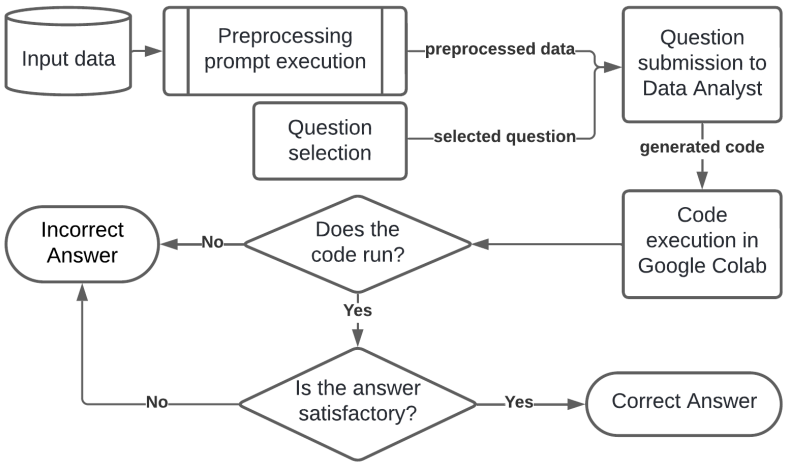

Figure 1. Block diagram

3.3. Preprocessing

After creating the dataset sample, preprocessing was necessary to ensure clearer and more coherent results. The selected period for student entries spans 14 years, from the academic semesters 2006.1 to 2019.2, chosen for its completeness, which facilitates a detailed analysis. The semester of 2020.1 was excluded as it was canceled due to the SARS-CoV-2 pandemic. "Exemption" enrollments were removed because they involve students being exempted from courses due to prior knowledge or equivalent credits, which means they pass automatically without grade evaluation. Enrollments marked as "In progress" were also excluded because these students do not yet have final grades, and their status could be pass, fail, or absent, affecting the integrity of the data.

This structured approach involved loading the dataset, applying the standard preprocessing steps, and initiating new chat sessions for each question to ensure independent responses. This methodology ensured consistent and identical datasets across all questions. Thus, the standard preprocessing prompt was: *Analyze the data and clean the columns as follows: enrollment_period: from 2006.1 to 2019.2; enrollment_type: remove "Exemption"; status: remove "In Progress".*

3.4. Descriptive Analysis

Descriptive data analysis offers a understanding of data by detailing it and is typically the initial step in data analysis. Basic questions are tackled using straightforward data comprehension techniques, including summarization, averaging, and finding minimums and maximums. Moderate questions involve more complex tasks, including the creation of new columns, percentage calculations, and the analysis of distributions, as well as averages of maximums and minimums. For Challenging questions, advanced metrics such as correlation, entropy, and skewness are employed.

Basic Questions

1. What is the number of students by type of admission?
2. What is the average credit hours taken in the first semester?
3. What are the maximum and minimum ages of dropout?

Moderate Questions

1. What is the proportion of dropouts by year of admission?
2. What is the distribution of grades by enrollment period?
3. Consider that the overall average of a student is defined as the average of all non-null grades in the courses the student has taken. Who is the student with the fewest semesters attended and the highest overall average who dropped out as a graduate?

Challenging Questions

1. Consider that the overall average of a student is defined as the average of all non-null grades in the courses the student has taken. Also, consider that there are two semesters each year, the first semester of the year is characterized by ending with ".1" and the second semester of the year ends with ".2", for example, in the year 2015 there are the periods 2015.1 and 2015.2. Based on these definitions, what is the correlation between the overall average of the students and enrolling in each semester of the year (first and second)?
2. How has the entropy of the distribution of students by academic sector changed over the last 5 recorded semesters?
3. Consider that the overall average of a student is defined as the average of all non-null grades in the courses the student has taken. What is the degree of skewness in the distribution of students' overall averages, and how does it affect the overall academic performance?

3.5. Diagnostic Analysis

Diagnostic data analysis usually follows descriptive analysis and aims to ascertain the underlying causes of observed phenomena. For Basic questions, basic metrics like percentage, frequency, and percentile are employed. Moderate difficulty questions involve hypothesis testing, the definition of indices that require further processing, and the coefficient of variation. For Challenging questions, advanced statistical methods are used, including analysis of variance, normality tests, tests for homogeneity of variances, and non-parametric tests.

Basic Questions

1. What is the pass rate by course?
2. How has the frequency of dropout changed over the periods?
3. For students who entered through affirmative action, what is the 70th percentile of the age of entry?

Moderate Questions

1. Determine if the mode of admission has a significant impact on graduation and dropout rates.
2. Consider that the Course Difficulty Index is calculated as the average of the difference between the average grades of the course and the overall average of all courses. What is the difficulty index of the course Differential and Integral Calculus I?
3. How consistent are student grades over time?

Challenging Questions

1. Does the type of admission statistically significant influence the number of semesters until graduation?
2. Is there a statistically significant difference in grades between students enrolled in Normal and Extracurricular modes?
3. Consider that the overall average of a student is defined as the average of all non-null grades in the courses the student has taken. Does affirmative action, the mode of admission, and the student's gender have any significant influence on their overall average?

3.6. Predictive Analysis

Predictive data analysis aims to anticipate future events. This study incorporates the Chain of Thoughts technique [Wei et al. 2022] to involves the tool proposing three potential solutions to a problem and selecting the most suitable one. Basic questions are expected to use simple models like Linear Regression, Moderate questions use more robust models such as Random Forest, while Challenging questions are addressed with sophisticated techniques, including advanced Regression, Classification, and Neural Networks.

Basic Questions

1. Consider that the overall average of a student is defined as the average of all non-null grades in the courses the student has taken. Do the age of entry and age of dropout have any significant influence on their overall average? Define 3 options for solving this issue and follow the best one.
2. What is the probability of a student's exit mode being graduated versus dropped out, based on the mode of admission, period of admission, and academic status? Define 3 options for solving this issue and follow the best one.
3. Is it possible to classify students into categories of academic performance (e.g., high, medium, low) based on their final averages and course workload? Define 3 options for solving this issue and follow the best one.

Moderate Questions

1. Consider that the overall average of a student is defined as the average of all non-null grades in the courses the student has taken. What is the probability of a student with an overall average below 7.0 and more than 3 failures being approved in the next course? Define 3 options for solving this issue and follow the best one.
2. Is it possible to identify patterns of similarity between students with exit modes of dropped out and graduated, considering the mode of admission and course workload? Define 3 options for solving this issue and follow the best one.
3. Is it possible to determine a student's exit mode based on characteristics such as the number of credits taken, type of enrollment, and course status? Define 3 options for solving this issue and follow the best one.

Challenging Questions

1. Through advanced analysis, can we obtain a student's performance in PROGRAMMING II, based on their performance in PROGRAMMING I and PROGRAMMING LABORATORY I? Define 3 options for solving this issue and follow the best one.

2. Through advanced analysis, is it possible to predict the mode of admission based on the period of admission, gender, affirmative action, and age of entry? Define 3 options for solving this issue and follow the best one.
3. Through advanced analysis, can we track a student's academic trajectory over time, using their sequence of grades and their course status (such as Passed and Failed), to anticipate the possibility of a lockout situation in the future? Define 3 options for solving this issue and follow the best one.

3.7. Prescriptive Analysis

Prescriptive analysis forecasts potential future scenarios, to suggest specific actions to enhance outcomes. The tool is directed to recommend three potential solutions to a problem, and selecting the most suitable one through the Chain of Thoughts technique [Wei et al. 2022]. Simple AI models and statistical methods are expected for Basic questions, more complex temporal analyzes for Moderate questions, and Challenging questions are expected to be addressed using more advanced models (including deep learning models), as the tool is required to perform through an advanced analysis.

Basic Questions

1. Based on the history of failures due to absences, how many cases will occur in the next period of the course LABORATORY OF PROGRAMMING I? Define 3 options for solving this issue and follow the best one.
2. What are the variables that most impact the differentiation between passed and failed students? Define 3 options for solving this issue and follow the best one.
3. What is the performance of the bottom 10% of students in mathematics sector courses? Define 3 options for solving this issue and follow the best one.

Moderate Questions

1. How does the profile of a student who took extra-curricular courses, never failed, and never dropped out significantly influence their academic success? Define 3 options for solving this issue and follow the best one.
2. What is the graduation rate forecast for the next year based on past trends? Define 3 options for solving this issue and follow the best one.
3. Is it possible to determine student dropout trends over time, based on historical patterns of academic performance and types of enrollment? Define 3 options for solving this issue and follow the best one.

Challenging Questions

1. Through advanced analysis, based on dimensionality reduction of students' academic characteristics, how can we predict which students are at risk of dropping out? Define 3 options for solving this issue and follow the best one.
2. Through advanced analysis, is it possible to generate new student profiles that maximize the probability of graduation, based on the characteristics of previously graduated students? Define 3 options for solving this issue and follow the best one.
3. Through advanced analysis, is it possible to identify critical moments in a student's academic trajectory, such as periods where the risk of dropout is higher, based on sequences of grades and academic situations? Define 3 options for solving this issue and follow the best one.

3.8. Metrics

In our analysis of ChatGPT's Data Analyst's performance, each response – comprising both textual content and code – is evaluated and categorized as either "Correct" or "Incorrect" based on its ability to comprehensively address the posed questions. We also monitor for Warnings, which are compiler or execution alerts indicating sub-optimal practices. These do not halt the process but could be problematic. It is important to note that the tool currently operates with Python 3.11.8 and Pandas 1.5.3, while the latest versions are 3.12.3 and 2.2.2, respectively; this version discrepancy could affect the accuracy of the generated code. Additionally, we observed Interference issues, such as technical disruptions from connection problems or unexpected errors, which may necessitate initiating a new chat session or briefly pausing the analysis for recalculations before continuation.

4. Results

4.1. Descriptive Analysis

During the Descriptive Analysis, all questions were answered correctly, using appropriate data and applying the requested metrics. There were no issues with warnings or interferences. In the Moderate Question 2, a box plot was generated to support the analysis. Table 1 presents the results obtained.

Level	Question	Answer		Problem	
		Correct	Incorrect	Warning	Interruption
Basic	1	X			
	2	X			
	3	X			
Moderate	1	X			
	2	X			
	3	X			
Challenging	1	X			
	2	X			
	3	X			

Table 1. Results of the Descriptive Analysis

4.2. Diagnostic Analysis

During the Diagnostic Analysis, all questions were answered correctly, with appropriate data and metrics. However, an interruption due to a connection failure occurred in the Challenging-level Question 3, requiring part of the analysis to be rerun, without affecting the final results. Additionally, four warnings were recorded. The first two, relating to the Moderate Questions 1 and 3, generating "SettingWithCopyWarning" from the *Pandas* library. While not directly causing errors, these warnings are significant as changes not reflected in the original data frame can lead to unexpected data.

The third warning emerged during the Challenging Question 2 while attempting to apply the Student's T-test. During the Shapiro-Wilk normality test, the *SciPy* library issued a warning that the p-value may be inaccurate for samples larger than 5000, yet

the analysis proceeded. However, the absence of homogeneity of variances prevented the application of the Student's T-test. Although this warning did not directly affect the results, it still warrants attention. The final warning occurred during the visualization of box plots for the Challenging Question 3, where a "UserWarning" from *Matplotlib* indicated that tick positions should be set before formatting the labels; this did not impact the visualization result. Table 2 summarizes the results of the Diagnostic Analysis.

Level	Question	Answer		Problem	
		Correct	Incorrect	Warning	Interruption
Basic	1	X			
	2	X			
	3	X			
Moderate	1	X		X	
	2	X			
	3	X		X	
Challenging	1	X			
	2	X		X	
	3	X		X	X

Table 2. Results of the Diagnostic Analysis

4.3. Predictive Analysis

In the Predictive Analysis, the tool correctly answered seven questions, missed two, presented warnings in four, and experienced interferences in two. In the Basic and Moderate questions, it often opted for straightforward approaches and simplified data modeling, which, although elementary, satisfied the requirements. With proper specification in Challenging questions, there was a transition from using simple regressions to employing Random Forests, allowing for more sophisticated analyses.

The first incorrect response occurred while addressing the Moderate Question 3, where a Random Tree model yielded only 54.5% accuracy. This triggered an "UndefinedMetricWarning" from scikit-learn, indicating multiple errors and unsatisfactory performance. Despite attempts to improve performance by expanding the feature set and addressing class balancing issues, compatibility problems with the *imblearn* library persisted. Further attempts to retrain the model did not change the accuracy. Additionally, an interference resulted in the session being closed, ultimately leading to the failure of the task.

The incorrect response to Challenging Question 3 was primarily due to hallucinations, which caused the model to generate incorrect results. It performed calculations on non-existent columns, which were neither part of the original data nor created by the model, highlighting a significant limitation of the tool. These errors, exacerbated by interferences during the process, ultimately led to the model's failure.

Warnings also emerged during the analysis. The first appeared while addressing the Basic Question 2, with a FutureWarning from scikit-learn about parameter renaming. The second occurred in another Basic Question 3, manifesting as a "SettingWithCopyWarning" from Matplotlib related to dataset modification. The third warning, associated

with the Moderate Question 2, was a "WARNING: matplotlib.legend" message. None of these warnings impacted their respective analyses. For details on the Predictive Analysis results, refer to Table 3.

Level	Question	Answer		Problem	
		Correct	Incorrect	Warning	Interruption
Basic	1	X			
	2	X		X	
	3	X		X	
Moderate	1	X			
	2	X		X	
	3		X	X	X
Challenging	1	X			
	2	X			
	3		X		X

Table 3. Results of the Predictive Analysis

4.4. Prescriptive Analysis

In the Prescriptive Analysis, six questions were answered correctly, while three were incorrect, accompanied by two warnings and four interferences. The tool consistently applied straightforward options and simplified data modeling for Basic and Moderate level questions, meeting the requirements adequately. However, for more challenging questions, such as those requiring advanced analyses, the tool employed complex models like Autoencoder, Hidden Markov Model (HMM), and Random Forest, showcasing a notable increase in sophistication compared to the Predictive Analysis.

The first incorrect response occurred in addressing the Moderate Question 1, where attempts with Linear Regression and Random Forest yielded unsatisfactory results. Interruptions because of technical issues during hyper-parameter adjustment via cross-validation further compounded the problem, resulting in an inadequate model and a marked inaccuracy.

On the other hand, the Moderate Question 2 was successfully resolved despite interruptions during data processing. During the implementation of the ARIMA model using the *statsmodels* library, warnings were encountered but did not compromise the analyses. These included two "ValueWarning" alerts, which are expected given the format of our dataset, and a "FutureWarning", indicating that this condition will be treated as an error in future versions of the library.

A significant limitation of the tool was highlighted during the Challenging Questions 1 and 3, where interferences were reported due to missing libraries in the tool's environment, such as *TensorFlow*, *Keras*, and *hmmlearn*. As a result, the analysis was halted, and the tool generated code that, when executed in a different environment, resulted in execution errors. For a summary of the Prescriptive Analysis results, please refer to Table 4.

Level	Question	Answer		Problem	
		Correct	Incorrect	Warning	Interruption
Basic	1	X			
	2	X			
	3	X			
Moderate	1		X		X
	2	X		X	X
	3	X			
Challenging	1		X	X	X
	2	X			
	3		X		X

Table 4. Results of the Prescriptive Analysis

4.5. Discussion

The tool's performance across 36 questions was evaluated, where 9 questions were posed for each category. The summary of the obtained results is presented in Figure 2, by showing the number of responses and issues identified per category.

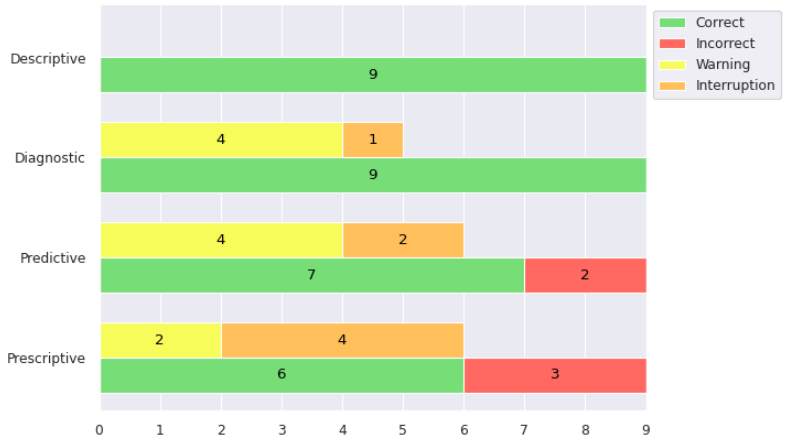

Figure 2. Summarization of the analysis results

With an accuracy rate of 86.11%, the tool exhibited excellent performance in descriptive and diagnostic analyses, accurately answering all questions in these categories. However, challenges arose in the predictive and prescriptive analyses, where some responses were incorrect. Moreover, a significant number of warnings and interferences were recorded across the diagnostic, predictive, and prescriptive analyses. Despite the good overall performance, the error rate of 13.89%, warnings rate of 27.78%, and interruptions rate of 19.44% highlight concerns regarding reliability and operational stability, particularly in the more complex predictive and prescriptive analyses that demand greater computational resources. Key limitations identified include a 10MB data processing cap and the absence of support for robust libraries such as hmmlearn, imblearn, Keras, and TensorFlow. Additionally, the tool was prone to hallucinations and operational failures,

which can necessitate session restarts, thereby affecting the efficiency and effectiveness of the analytical process.

5. Conclusion

This study evaluated the ChatGPT Data Analyst, an LLM-based tool that assists in data analysis, addressing the four categories of analysis: Descriptive, Diagnostic, Predictive, and Prescriptive. It was found that, although LLMs offer transformative potential in data analysis automation, they face significant technical barriers, such as integration with advanced libraries and management of large volumes of data. The study contributes valuable perspectives to the literature, enriching the theoretical and practical knowledge about the application of LLMs.

For future work, we plan to fine-tune the GPT-4 model with the aim of minimizing errors and warnings. We also intend to apply the methodology of this study to other datasets to verify the consistency of the results. Additionally, we will explore the model's propensity to produce hallucinations in different scenarios. Finally, we will use the evaluation paradigm proposed by [Zhang et al. 2024] to assess where the ChatGPT Data Analyst stands in relation to other data science agents. These investigations are crucial to expand our understanding of the applicability and limitations of LLMs in the automation of data analysis.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2024). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Cheng, L., Li, X., and Bing, L. (2023). Is gpt-4 a good data analyst? *Journal of Artificial Intelligence Research*, Findings of the Association for Computational Linguistics: EMNLP 2023:9496—-9514.

Daibes, M. and Lima, B. B. (2024). Cracking the heart code: using chatgpt's data analyst feature for cardiovascular imaging research. *The International Journal of Cardiovascular Imaging*, pages 1–2.

Ding, B., Qin, C., Liu, L., Chia, Y. K., Li, B., Joty, S., and Bing, L. (2023). Is gpt-3 a good data annotator? pages 11173–11195.

Jaimovitch-López, G., Ferri, C., Hernández-Orallo, J., Martínez-Plumed, F., and Ramírez-Quintana, M. J. (2022). Can language models automate data wrangling? *Machine Learning*, 112:2053—-2082.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7b.

Kasetty, T., Mahajan, D., Dziugaite, G. K., Drouin, A., and Sridhar, D. (2024). Evaluating interventional reasoning capabilities of large language models. *arXiv preprint arXiv:2404.05545*.

Liu, X., Wu, Z., Wu, X., Lu, P., Chang, K.-W., and Feng, Y. (2024). Are llms capable of data-based statistical and causal reasoning? benchmarking advanced quantitative reasoning with data.

Nasseri, M., Brandtner, P., Zimmermann, R., Falatouri, T., Darbanian, F., and Obinwanne, T. (2023). Applications of large language models (llms) in business analytics – exemplary use cases in data preparation tasks. 14059:182–198.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Sharma, A., Li, X., Guan, H., Sun, G., Zhang, L., Wang, L., Wu, K., Cao, L., Zhu, E., Sim, A., Wu, T., and Zou, J. (2023). Automatic data transformation using large language model - an experimental study on building energy data. pages 1824–1834.

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., and Wang, J. (2019). Release strategies and the social impacts of language models. *CoRR*, abs/1908.09203.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, b., Xia, F., Chi, E., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Zhang, H., Dong, Y., Xiao, C., and Oyamada, M. (2023). Large language models as data preprocessors.

Zhang, Y., Jiang, Q., Han, X., Chen, N., Yang, Y., and Ren, K. (2024). Benchmarking Data Science Agents. *arXiv e-prints*, page arXiv:2402.17168.