The Role of Chatbots in Data Analytics: An Evaluation of Functional Abilities

Preeti Patel[1] and Sowgol Shooshtarian[2]

[1] London Metropolitan University, London N7 8DB, UK
p.patel@londonmet.ac.uk

[2] London Metropolitan University, London N7 8DB, UK
sos0651@my.londonmet.ac.uk, shooshtarian.sowgol@gmail.com

Abstract. In the fast-paced evolution of artificial intelligence, in which chatbots are becoming vital tools in many areas, this study is focused on the various functional roles of ChatGPT-4 in data analytics. The primary objective is to assess the technological advancements and capabilities of ChatGPT-4 in analyzing data. The research methodology employs three distinct datasets and the analysis compares various prompt engineering techniques with ChatGPT-4 against human-driven analysis. This comparison is grounded in profile optimization techniques, the CRISP-DM framework, and a structured Data Science Project Plan. The findings reveal the diverse capabilities of ChatGPT-4. On one hand, basic prompts yield quick, initial insights, positioning ChatGPT-4 as a valuable asset for those with limited technical expertise. Conversely, refined and engineered prompts enable ChatGPT-4 to deliver a deeper, more detailed analysis. Yet, when contrasted with traditional human-driven analysis, typically utilizing tried and tested techniques, tools and experience, all combined with a deep knowledge of machine learning and data analysis; the human approach remains unsurpassed in depth, granularity, and accuracy. Conclusively, while ChatGPT-4 stands out for preliminary data explorations and general insights, complicated, nuanced analytical tasks are best tackled using time-tested, traditional data science methodologies and tools. While ChatGPT-4 is versatile, it has not yet matched the expertise of seasoned data analysts.

Keywords: Chatbot, ChatGPT-4, Data Analytics, Generative AI.

1. Introduction

OpenAI's text-generating chatbot, ChatGPT-4 is a conversational AI interface that uses natural language processing and machine learning algorithms to generate high-quality responses to a wide range of queries. In the rapidly evolving landscape of data science, chatbots have emerged as a significant player, revolutionizing the way we interact with data and derive insights. The evolution and utilization of chatbots, particularly ChatGPT, have permeated various sectors, marking a significant shift in how interactions and data analyses are carried out. ChatGPT is an advanced Large Language Model (LLM) that uses deep learning techniques to produce human-like responses to natural language inputs. It is part of the family of generative pre-training transformer (GPT) models developed by OpenAI and is currently one of the largest publicly available LLM. ChatGPT can capture the nuances and intricacies of human language, allowing it to generate appropriate and contextually relevant responses across a broad spectrum of prompts [5].

The emergence and integration of chatbots in data analytics present both opportunities and challenges. They offer potential for enhanced user interaction, democratization of data analysis, and increased efficiency. However, concerns about their accuracy, depth of analysis, and suitability for complex tasks remain. This research aims to address these issues and evaluate the functional abilities of chatbots, with a specific focus on their role in data analytics.

This study is motivated by the need to understand the extent of the lack of comprehensive studies on the accuracy and performance of advanced chatbots such as ChatGPT-4 in processing and analyzing data in comparison to traditional data analysis methods. The study involves the design and implementation of multiple experiments with ChatGPT-4 on three unique datasets, selected for their distinct nature and significance in data science issues. The research problem focuses on evaluating the accuracy and performance of ChatGPT-4 in handling data analysis processes, such as its ability in automating data preparation, analysis, and interpretation tasks, and how these capabilities align with the current and future needs of data analytics.

This paper is structured to first present an overview of the current state of chatbots in data analytics, followed by a detailed examination of ChatGPT-4's abilities. The methodology section describes the experimental approach taken and the multiple experiments conducted with ChatGPT-4 on diverse datasets. The results and discussion will analyze and compare the findings, and the conclusion will summarize the study's implications and potential future directions for chatbot technology in data analytics. In doing so, this paper not only fills a critical gap in existing research but also charts a path forward for the effective utilization of ChatGPT-4 in data analytics, potentially revolutionizing how we approach data-driven decision-making.

2. Related Work

An extensive overview [1] of the technology and applications of chatbots emphasizes the historical development of chatbots, tracing their evolution from rule-based systems to sophisticated AI-driven entities. AI-powered chatbots have adopted technologies such as Natural Language Processing (NLP), Machine Learning (ML), and Artificial Intelligence (AI), facilitating their application in various domains, including data analytics. In the realm of data analytics, chatbots now function as tools for data querying, visualization, and exploration, facilitating conversational interaction with data. They provide an accessible platform for users with limited technical expertise, enabling them to analyze data through intuitive natural language queries. This democratizes data analysis, making it more accessible and user-friendly. Nonetheless, there remain challenges to the widespread adoption of chatbots in data analytics. These include issues of data privacy and security, as well as the need for more sophisticated contextual understanding and more human-like interaction capabilities [1,10, 11].

Past research has addressed the significant challenge of evaluating chatbot performance in a quantitative manner [9] in which the application of data analytics to generate measurable and comparable metrics, thereby fostering an objective evaluation of a chatbot's functional abilities in projects and assessments. In this analysis, key performance indicators (KPIs) are identified such as response time, accuracy of response, and user satisfaction, demonstrating the extensive potential of data analytics to enhance the understanding of chatbot performance. The work further highlights the transformative role of chatbots in data analytics, echoing other works in emphasizing the growing importance of chatbots in facilitating user-friendly data interactions. Jongerius in [9] emphasizes the potential of chatbots to democratize data analysis, in line with findings from [7]. However, Jongerius also underscores potential challenges associated with quantifying chatbot performance, such as data privacy and security concerns, further extending the ongoing discourse on the limitations of chatbot applications in data analytics. Khlaif et al. also address concerns regarding ChatGPT's accuracy and reliability. They note instances where the chatbot may provide responses that lack depth or are not entirely accurate, underscoring the importance of human oversight in data analytics processes [11].

Others have delved into the transformative role of ChatGPT in the data science field [7], emphasizing the significant revolution brought about by AI-assisted conversational interfaces. The authors highlight how ChatGPT can assist data scientists in automating various aspects of their workflow, including data cleaning and preprocessing, model training, and result interpretation. Alshater's paper notes that the ChatGPT's ability to process and simplify complex datasets can be helpful in teaching and learning data analytics, thereby improving overall academic performance [2]. It also can provide new insights and improve decision-making processes by analyzing unstructured data. However, there are also limitations and issues associated with using ChatGPT. For example, while it can save time and resources compared to training a model from scratch, it may not perform well on certain tasks if it has not been specifically trained for them and lead to biased or not concise results. Additionally, the output of ChatGPT may be difficult to interpret, which could pose challenges for decision-making in data science applications. The benefits of using ChatGPT outweigh the costs and can enhance the productivity and accuracy of data science workflows. There is also a suggestion that ChatGPT has become an increasingly crucial tool for intelligence augmentation in data science [7].

There appears to be considerable value in using ChatGPT in healthcare, particularly in relation to improving accessibility to medical information and services [5]. The use of chatbots enhances overall patient satisfaction through instant responses and service availability, thus offering a promising outlook for chatbot integration in healthcare. It is also argued that ChatGPT has made significant strides in facilitating efficient interaction with complex medical datasets, thereby improving the ease and accessibility of data analysis. By transforming data interpretation into a more conversational format, ChatGPT offers the potential to democratize data analytics across various professional spheres, not just medicine. The study by Kalla et al. underscores the versatility of ChatGPT in various fields, including but not limited to healthcare, cyber security, and business. They note that the chatbot's ability to provide quick, informed responses can help to automate some aspects of different fields [10]. However, the authors also identify notable limitations, specifically, the potential for inaccuracies and biases in complex data interpretation. They point out that while ChatGPT has made substantial advancements in the handling of natural language queries, there is room for improvement in the context of intricate data analysis [5, 10]. Moreover, Nah et al. notes that ChatGPT's output could potentially exhibit reflections of any inaccuracies, disproportionate representation of information sources, or inherent biases present within the training dataset. [14]

Chatbots have been leveraged in educational contexts, enabling students and educators to engage with complex data in a more accessible, conversational manner [15]. They illustrate the functional abilities of chatbots in managing and interpreting educational data in interactive manner, making them vital tools in assessments and project management. This underlines the wider applicability of chatbots in diverse fields requiring data analysis and interpretation, beyond the educational context. [6] In the context of data analytics, the authors draw attention to the use of chatbots in personalized learning, data-driven student assessments, and automated feedback, all of which rely heavily on the analysis and interpretation of large data sets. This illustrates how chatbots can enhance the efficiency and precision of data analysis, reinforcing their crucial role in this field. However, potential limitations and highlights include issues around data privacy and the potential for inaccuracies or misunderstandings in chatbot responses, particularly in complex or ambiguous queries [6, 14, 15]. These concerns align with discussions in the wider literature around the ethical and practical challenges associated with chatbot use in data analytics.

Chatbots have become essential tools in project management, providing intuitive, interactive platforms for complex data analysis. This aligns with the role of chatbots in data analytics by enhancing the understanding of intricate project data, facilitating decision-making processes, and promoting project efficiency [3]. Significantly, the chatbots' ability to provide real-time data insights, ensuring more efficient tracking and monitoring of project progress. This functionality could lead to more accurate and timely decision-making, reinforcing the broader role of chatbots in data analytics. Moreover, potential challenges are associated with the implementation of chatbots, including the need for ongoing training and development to ensure chatbots can adequately understand and respond to user queries. These considerations offer further insights into the evaluation of chatbots' functional abilities in data analytics and the need for continuous improvement [3].

The role of chatbots, particularly ChatGPT, in the data science community is widely praised. It is suggested that the most significant advantage lies in the ability of these AI models to automate and streamline data processing, leading to improved efficiency and effectiveness in data analysis. Liu et al in [13] provide a comprehensive analysis of the prospects of large language models such as ChatGPT highlighting that with the increasing complexity and sophistication of these models, their ability to generate insightful data and analytics is continually growing.

As chatbots become more accurate, reliable, and able to understand complex queries, they will become more widely adopted in data analytics projects. Automating tasks, providing personalized assistance, answering direct questions and providing assessment of skills are valuable to data analysts.

3. Methodology

Data science, an interdisciplinary domain, has transformed the way data is analyzed and interpreted. This research employs the Cross-Industry Standard Process for Data Mining (CRISP-DM) as its foundational methodology. The research also leverages modern tools, with a particular emphasis on ChatGPT-4, to explore its capabilities in data analytics. We adopt an inductive and empirical approach, rooted in experimentation. The aim is to conduct experiments using ChatGPT-4 on three distinct datasets to assess its adaptability and efficiency in diverse data scenarios. For this research, the focus is exclusively on ChatGPT-4, a state-of-the-art chatbot based on the GPT-4 architecture. Renowned for its ability to generate human-like text and its prowess in understanding context, ChatGPT-4 will be the focus of our experiments. Beta features of ChatGPT-4 include plug-ins and Advanced Data Analysis which allows users to retrieve real-time information. This is a critical issue since ChatGPR-3.5 has access to the data before 2021 [8]. Furthermore, the recently introduced feature allowing for instruction customization enhances prompt accuracy. ChatGPT-4 plug-ins are additional capabilities or tools that can be integrated with the

ChatGPT-4 model to enhance its functionality. Plug-ins have been selected for their utility in data analysis and include Data Interpreter, Notable, WebPilot and Wolfram.

3.1. Datasets

Three datasets are selected from the different domains of finance, health and energy; enabling scope for comparison. Each dataset provides a unique perspective on the capabilities of ChatGPT-4. Fig. 1, shows the framework of this methodology.

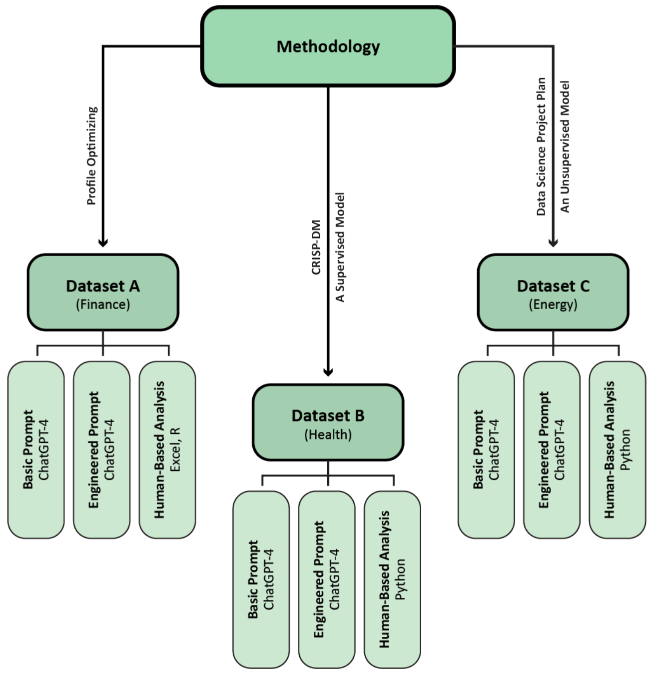


Fig. 1. The framework of methodology.

- Dataset A: This dataset is focused on structured data of five stocks in the US market downloaded from the website www.finance.yahoo.com. The dataset provides the stock prices for five tickers: “AVGO”, “OXY”, “FSLR”, “CAH”, and “LW" from 01/06/2020 to 01/05/2023.
- Dataset B: This dataset contains medical measurements related to heart health downloaded from https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset?resource=download.
- Dataset C: A synthetic dataset is generated by “Noteable” plug-ins of ChatGPT-4 based on a prompt and requirements and afterwards, the CSV file of this dataset is downloaded for further investigation in Python.

3.2. Prompt Engineering and Evaluation

Prompt engineering is crucial to maximize ChatGPT-4's potential. It involves crafting input prompts to obtain precise and relevant responses. An adeptly engineered prompt can steer ChatGPT-4 to generate outputs that resonate with the specific prerequisites of a data science task. Conversely, ambiguous, or poorly structured prompts may yield generic or misaligned responses, potentially compromising the quality of data analysis.

ChatGPT-4 offers two primary avenues for prompt engineering: its user interface (UI) and its API. The UI provides an interactive platform for manual prompt input and refinement, facilitating real-time response observation. This approach is particularly beneficial during the nascent stages of experimentation. Conversely, the API provides a more automated, scalable methodology, allowing for programmatic prompt input and response retrieval. For this research, prompt engineering was conducted via the UI.

For the two types of prompts, baseline prompt and engineered prompts, a systematic approach was taken for each dataset in order to aid comparison. A baseline prompt is a generic prompt that asks ChatGPT-4 to analyze the dataset without any specific guidance. This type of prompt is more open-ended and general. It gives the model freedom to provide a broad answer without specific guidelines. The response might be more generic and might not delve deep into specific details or insights. Engineered prompts allow the design of a series of prompts that guide ChatGPT-4's analysis, focusing on specific aspects of the dataset or asking for types of insights. This prompt is crafted with precision, providing clear instructions to the model about what is expected in the response. It guides the model towards generating a more detailed and specific output, tailored to the user's requirements. The most effective prompts are relevant, clear, complex, scalable and include role-play. After designing a set of prompts for each dataset, the chatbot is fed with the dataset for setup. The chatbot is then queried using these prompts, and the results are evaluated based on metrics like precision, recall, execution speed, and relevance.

4. Results

This section evaluates the analytical capabilities of ChatGPT-4 across three distinct datasets: A, B, and C. Each dataset offers unique challenges, from portfolio optimization to heart health metrics and energy performance certifications. The approach is to assess ChatGPT-4's responses using basic and engineered prompts, and then benchmark its findings against human-based analyses. The goal is to understand ChatGPT-4's capabilities in the data analysis field and its potential for future applications.

4.1. Dataset A: Stock Market Analysis

Using the structured data of five stocks in the US market, ChatGPT-4 was tasked with guiding potential investors in choosing the most efficient portfolio. The analysis incorporates various metrics, performance indicators, and visual representations. The basic prompt, shown in Fig. 2, gives back a general exploration of the dataset, identifying potential errors, and then proceeding to extract actionable insights.

Prompt: "Check this dataset for errors in data and provide insights."

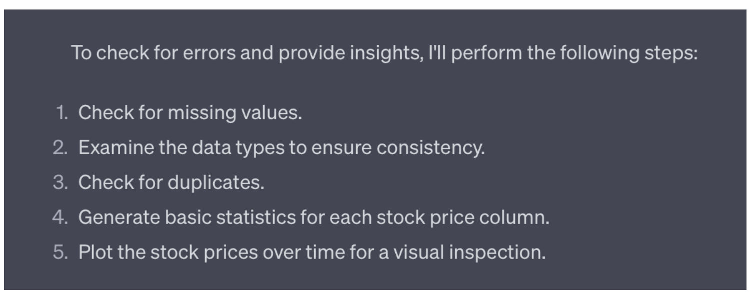


Fig. 2. ChatGPT-4 answer for the prompt.

One of the indications of this analysis was the visualization of stock price trends over time for each company. This was accompanied by the presentation of the Efficient Frontier curve and its insights—a key tool in financial analysis that portrays the best possible return for a given level of risk (See Figs. 3 and 4).

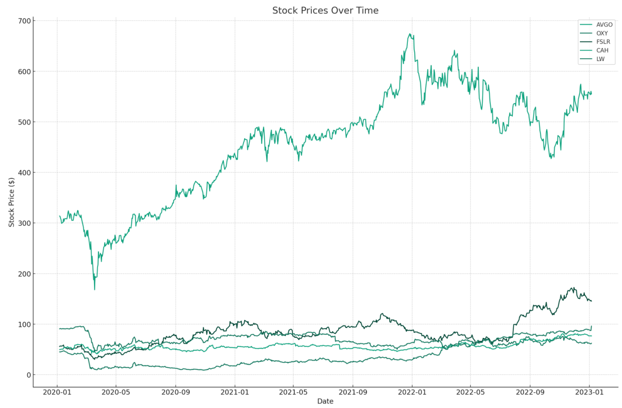


Fig. 3. Stock prices over time for each company.

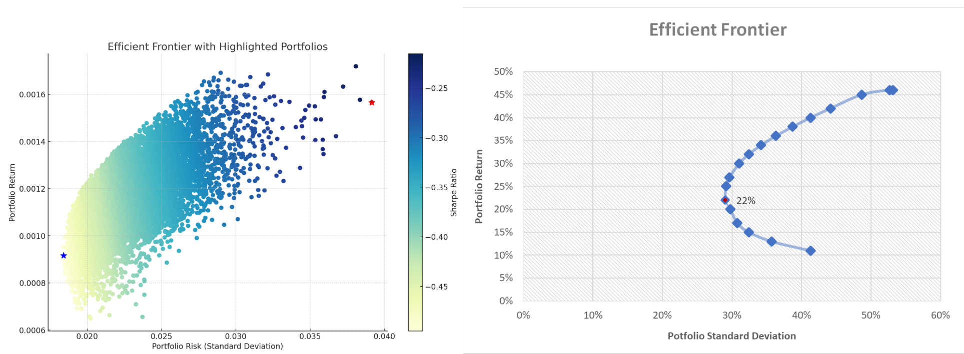
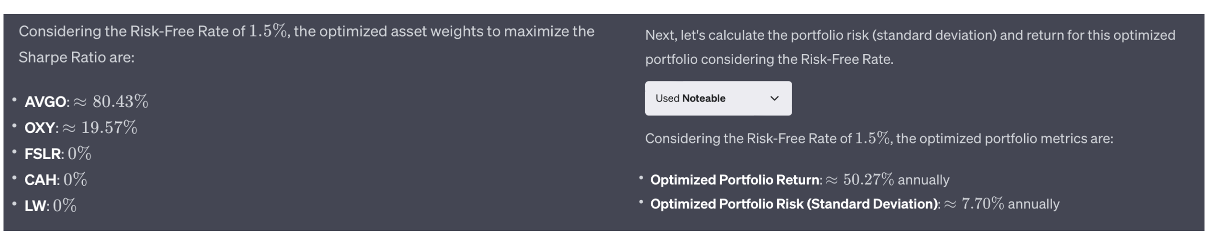


Fig. 4. Plotted Efficient Frontier Curve by ChatGPT-4 on the left and plotted in Excel by a data analyst on the right.

Average Daily Returns and Daily Standard Deviations, the Correlation Matrix and the Covariance Matrix of Each Asset are also calculated. Furthermore, the iterative nature of the analysis, especially when factoring in varying risk-free rates, underscores the versatility and adaptability of ChatGPT-4 in handling complicated financial datasets. To recalculate the "Sharpe Ratio" of the portfolio, another detailed prompt is given to calculate this value for equally weighted assets portfolio and optimized portfolio.

Prompt: *"We have a portfolio with all the above assets. Use an appropriate Solver function to determine the portfolio risk and the percentage of investment in each asset in your portfolio for a target return of your choice. first try it with equally weighted assets in the portfolio and then optimize the weights to have the highest "Sharpe Ratio"."* (See Fig. 5).


Considering the Risk-Free Rate of 1.5%, the optimized asset weights to maximize the Sharpe Ratio are:

- **AVGO**: $\approx 80.43\%$
- **OXY**: $\approx 19.57\%$
- **FSLR**: 0%
- **CAH**: 0%
- **LW**: 0%

Next, let's calculate the portfolio risk (standard deviation) and return for this optimized portfolio considering the Risk-Free Rate.

Used **Noteable**

Considering the Risk-Free Rate of 1.5%, the optimized portfolio metrics are:

- **Optimized Portfolio Return**: $\approx 50.27\%$ annually
- **Optimized Portfolio Risk (Standard Deviation)**: $\approx 7.70\%$ annually

Fig. 5. Optimized portfolio's weights of assets, return and standard deviation

Capital Market Line (CML) is plotted through a prompt by ChatGPT-4. (See Fig. 6) To verify the result, prompts are optimized and revised iteratively. Each prompt yielded varied outcomes, and after several iterations, the free versions of the "Data Interpreter" and "Noteable" analysis plugins were finished.

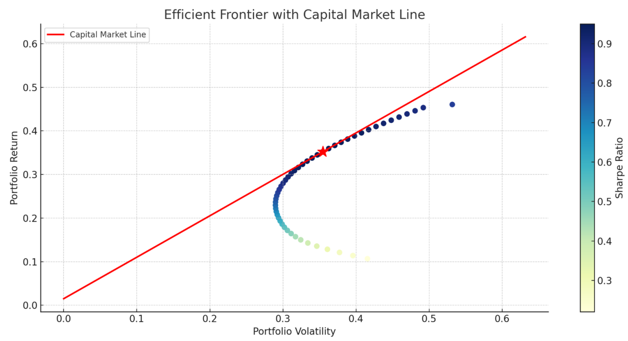


Fig. 6. Capital Market Line (CML) plotted by ChatGPT-4

To calculate the Beta Coefficient for each asset, the "SPY" ticker, representing the S&P 500, is used as a benchmark. UI prompts for ticker details and prices are ineffective for ChatGPT-4's calculation. Hence, an updated "SPY" price dataset was uploaded again, prompting the calculation of the Beta Coefficient for each asset (See Fig. 7).

Here are the computed β values for the first five assets:

- **AVGO**: $\beta = 1.29$
- **OXY**: $\beta = 1.52$
- **FSLR**: $\beta = 0.92$
- **CAH**: $\beta = 0.78$
- **LW**: $\beta = 0.95$

Fig. 7. Beta Coefficient for asset

5% value at risk (VaR) is also crucial as it measures the potential loss in value of a risky asset that is calculated by ChatGPT-4. However, ChatGPT-4 attempts for estimating the volatility of assets using ARCH/GARCH fail.

4.2. Dataset B: Heart Health Analysis

ChatGPT-4 was engaged to automate the six steps of the CRISP-DM process using a structured heart health dataset. A supervised model is assessed in this experiment. The analysis aims to understand the data, identify patterns, and model the results. A prompt is used for data extraction and analysis, with "Business Understanding" results in Fig. 8.

Prompt: *"This dataset needs to be analyzed. Consider CRISP-DM as your main methodology to do that. Conduct each step separately on this dataset and give back your result."*

1. Business Understanding

In this phase, we need to understand the business problem and objectives. Since I don't have explicit information about the business context for this dataset, I'll make assumptions based on the name of the file and its contents. The file is named "heart", which suggests that the dataset might be related to cardiovascular health or heart disease.

Fig. 8. ChatGPT-4 answer for "Business Understanding"

Beginning with an in-depth exploration, the data is examined for attributes, anomalies, and trends. Box plots detect outliers in continuous variables, like cholesterol (See Fig. 9). Yet, ChatGPT-4 avoided manipulating data at this stage due to the following reason: *“we'll keep the outliers intact for now. If they prove problematic during modeling, we'll address them accordingly.”*

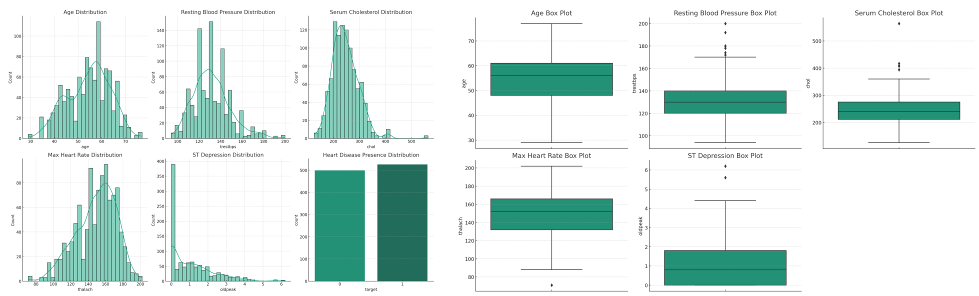


Fig. 9. Exploring the data (left) exploring the data with box plots to detect outliers (right) by ChatGPT-4

The dataset's dependent variable is "Target", indicating if a patient has developed heart disease or not. ChatGPT-4 considers feature scaling for Logistic Regression analysis, essential for magnitude-sensitive algorithms. Model assessment includes a confusion matrix, Receiver Operating Characteristic (ROC) curve, Area Under the Curve (AUC), and classification report (See Figs. 10 and 11).

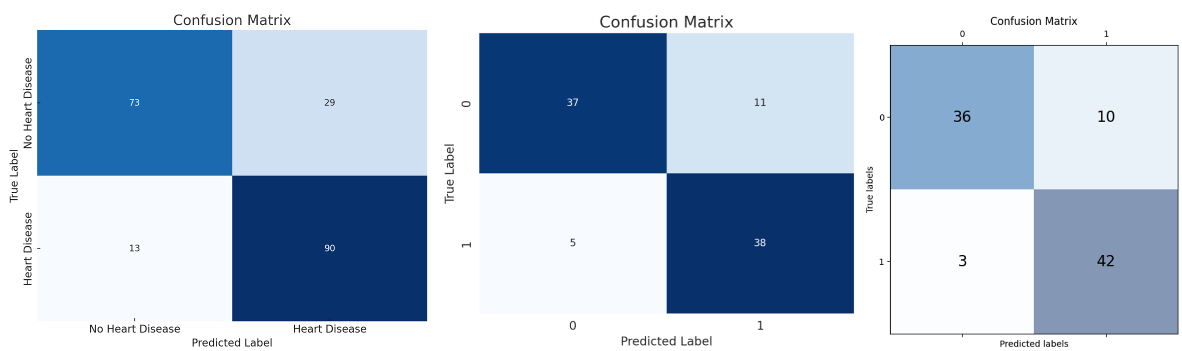


Fig. 10. Confusion matrix for logistic regression model: left: pre-outlier basic prompt, middle: engineered prompt and right: human-based analysis

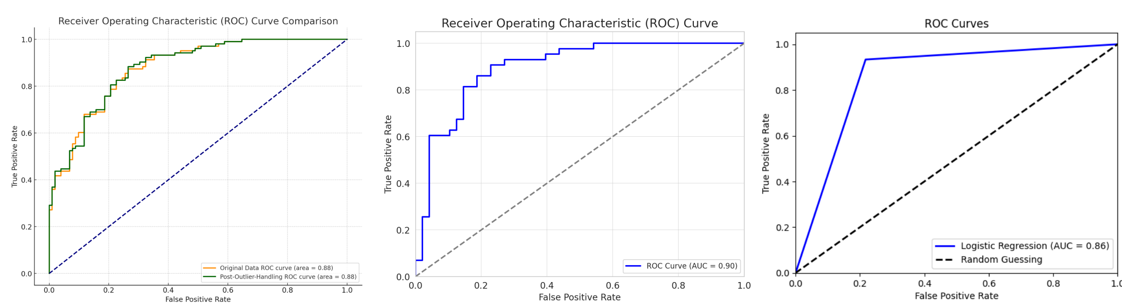


Fig. 11. ROC curve for logistic regression model: left: pre-outlier and post-outlier basic prompts, middle: engineered prompt and right: human-based analysis. The AUC for the engineered prompt is 4% lower in ChatGPT-4's report compared to the image.

4.3. Dataset C: Synthetic Energy Performance Analysis

Highlighting the innovative capabilities of ChatGPT-4, this section focused on the creation and analysis of a synthetic dataset. Therefore, for a broader understanding of its features, this dataset is created by ChatGPT-4 plug-ins called “Noteable”. Moreover, this experiment will use a deeper methodology than CRISP-DM called Data Science Project Plan and an unsupervised model is assessed in this experiment. ChatGPT-4 checks the attributes of the dataset and then, it checks the data for any missing values and summarizes the statistics. (See Fig. 12)

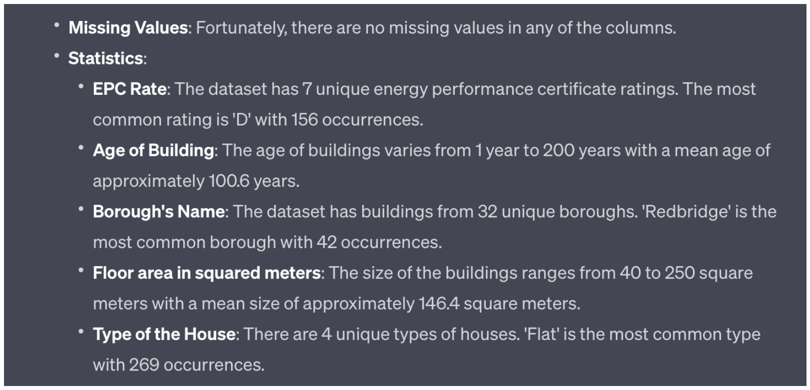

- **Missing Values**: Fortunately, there are no missing values in any of the columns.
- **Statistics**:
  - **EPC Rate**: The dataset has 7 unique energy performance certificate ratings. The most common rating is 'D' with 156 occurrences.
  - **Age of Building**: The age of buildings varies from 1 year to 200 years with a mean age of approximately 100.6 years.
  - **Borough's Name**: The dataset has buildings from 32 unique boroughs. 'Redbridge' is the most common borough with 42 occurrences.
  - **Floor area in squared meters**: The size of the buildings ranges from 40 to 250 square meters with a mean size of approximately 146.4 square meters.
  - **Type of the House**: There are 4 unique types of houses. 'Flat' is the most common type with 269 occurrences.

Fig. 12. Checking null values and providing basic statistics by ChatGPT-4

ChatGPT-4 determines that feature selection or engineering isn't needed, but categorical variables are required to be converted to numerical. It visualizes distributions and box plots (See Fig. 13).

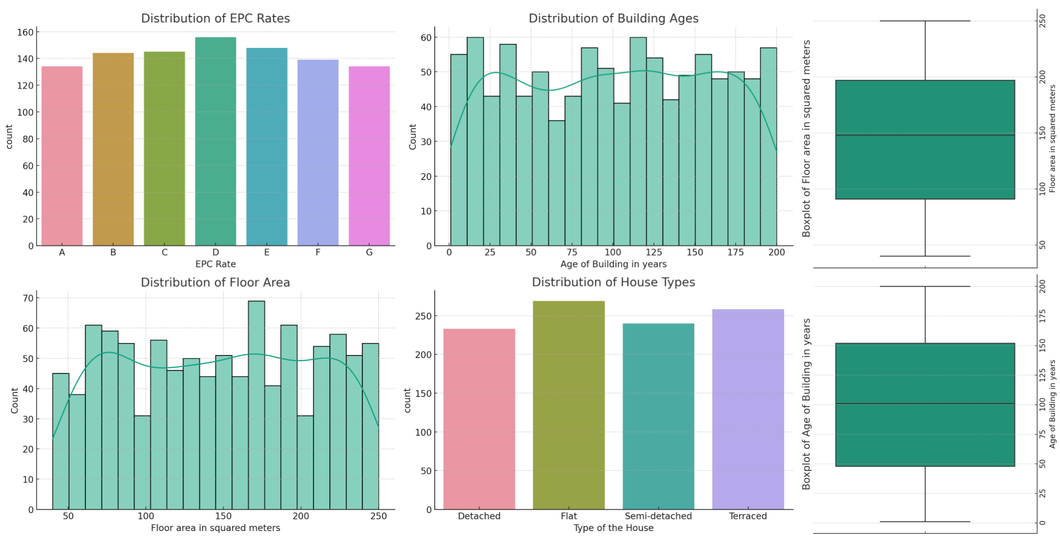


Fig. 13. ChatGPT-4 visualization such as distribution and identifying outliers

For the predictive model and data mining phase, ChatGPT-4 selects K-Means and scales the data. Both basic and engineered prompts identified 4 as the optimal cluster count. While basic prompts did not perform any performance metrics evaluations, Silhouette Score or Elbow Method, engineered prompts did. Based on these methods, ChatGPT-4 chooses four clusters. In human-based analysis, the optimal number of clusters is evaluated by plotting the Elbow Method for both SSE (Sum of Squared Errors) and WCSS (Within-Cluster Sum of Squares) (See Fig. 14).

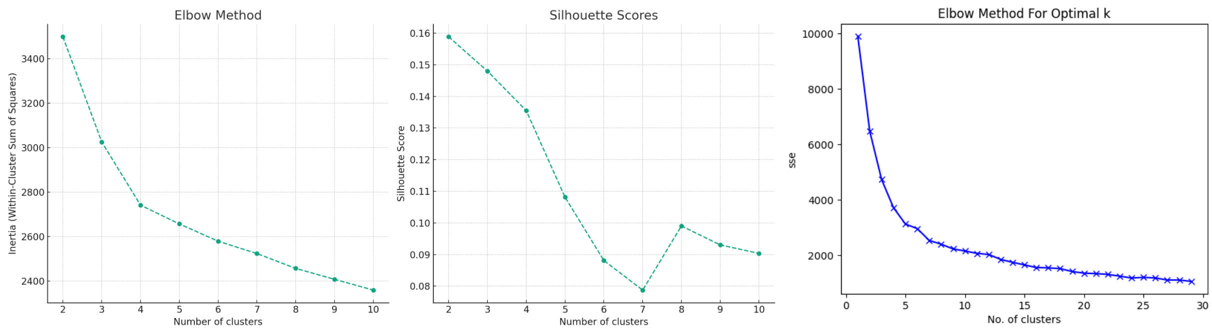


Fig. 14. Elbow Method (Left) and Silhouette Score (Middle) by ChatGPT-4 (Top), SSE Elbow method (Right) using Python

Since ChatGPT-4 did not provide all the required evaluation of the model's performance, a separate prompt was crafted to achieve these values. To find the best metrics for evaluation, different numbers of clusters are tested and assessed (k= 2, 4, 6), by key metrics in the human-based analysis. Table 1 summarizes all models' performance metrics. Although the six-cluster model performs the best, the four-cluster model is selected for consistency with other models.

Table 1. Comparison of metrics evaluation for each approach

Approach	**k**	**Calinski-Harabasz Index**	**Davies-Bouldin Index**	**Inertia**	**Silhouette score**
Basic prompt	4	--	--	--	--
Detailed prompt	4	183.30	1.92	2740.78	0.135
Human-based	2	701.49	1.16	4296.92	0.34
Human-based	4	498.59	1.42	3293.97	0.27
Human-based	*6*	558.21	1.09	2652.67	0.38

5. Discussion

This section serves as a bridge between the raw results obtained and their interpretation in the context of broader research. This chapter delves deep into the themes of 'comparative approach', 'efficiency and depth', 'integration', 'accuracy and trust', and more. Each theme is explored in relation to the results obtained from different datasets, providing a comprehensive understanding of the capabilities and limitations of the ChatGPT-4 in data analysis.

5.1. Dataset A

The results present an exploration of ChatGPT-4's capabilities in automating the stages of the portfolio optimization.

Comparative Approach: The three analysis methods offer a gradient in terms of complexity and depth. The basic prompt approach provides a surface-level analysis, ensuring data integrity and preliminary insights. The engineered prompt method dives deeper, leveraging specific plug-ins and detailed questions to extract more granular data insights. Moreover, several principles, such as clarity, precision, and role-play, were considered to enhance the effectiveness of these prompts. The human-based approach, being the most detailed, serves as the accuracy benchmark for this article.

Efficiency and Depth: The chatbot-driven approaches, especially the basic prompt, are efficient for quick insights and data validation. They enable users without a deep technical background to access and understand data. In contrast, the human-based method in Excel and R offers depth and is more suitable for in-depth analysis and generating metrics.

Integration: The engineered prompt approach highlights ChatGPT-4's ability to integrate with platforms like Google Docs. Such integrations can streamline data uploading and preprocessing stages, making the analysis more seamless. However, there are some limitations for using ChatGPT-4. ChatGPT-4 has a limited service and cap for the number of queries and messages during a specific period of time. For example, every 3 hours, it is possible to send 50 messages. This restriction might slow analysis. Additionally, timeout errors can disrupt analysis if there is a delay between prompts, requiring dataset reloading and metric recomputation.

Accuracy and Trust: The human-based method, as the benchmark, is compared against ChatGPT-4's accuracy and reliability. In the basic prompt approach, while less detailed, the data had fewer errors and values such as correlation and variance-covariance matrix aligned with the benchmark. However, the engineered prompts had more errors, occasionally failed to answer, and certain technical analyses, like the Beta Coefficient, did not match the benchmark.

Dataset A Recommendations: Chatbots like ChatGPT-4, with their expanding capabilities and integrations, signal a bright future. They can swiftly handle initial data analysis stages, especially preliminary stages, allowing analysts to tackle more complex analyses. Recent advancements such as Excel-Python integration, can handle more advanced data analysis which expedite and simplify the process [12]. Moreover, using the OpenAI's API playground for engineered prompting is suggested for further and deeper research. This feature allows users to have more options to customize their prompts such as adjusting the model's temperature to influence its randomness.

5.2. Dataset B

The results explore ChatGPT-4's capabilities to automate the six CRISP-DM stages using a heart health dataset.

Comparative Approach: The three analysis methods vary in complexity and depth. The basic prompt provides preliminary observations. The engineered prompt goes further with specialized tools and precise queries for enhanced data interpretations. This is attributed to tailored prompts that are clear, precise, and include a role-play (Data Analyst) for each CRISP-DM step. Meanwhile, the detailed human-based method establishes the benchmark in this study.

Efficiency and Depth: ChatGPT-4 offers a time-saving solution for initial data analyses. The chatbot, especially using the basic prompt, provides rapid insights and data validation, making it user-friendly for those without technical background. While it may not match the depth of human analysts, its speed in data processing is invaluable for early project stages. Compared to the "human-based", ChatGPT-4 excels in many data analysis areas but can struggle with many such as data visualization, possibly due to text-platform limitations on visualizing complex data.

Integration: The integration of chatbots in data science workflows, as demonstrated by this experiment, could significantly streamline, and automate various stages of the data analysis process. By handling tasks such as data preprocessing, modeling, and evaluation, chatbots can free up experts to focus on higher-level analytical tasks, interpretation, and strategy formulation.

Accuracy and Trust: Using the human-based method as the benchmark, the basic prompt approach was not deep, particularly in “Business Understanding” and “Evaluation” phases. Moreover, it was the least accurate among all three approaches. The visualizations show incremental accuracy from basic to detailed prompts, with human analysis as the benchmark. While ROC curves for both chat-based analysis are notably close, their documented values were lower, indicating caution when relying on ChatGPT-4 visuals that do not match the report. This discrepancy further suggests higher value of human analysis (See Fig. 15).

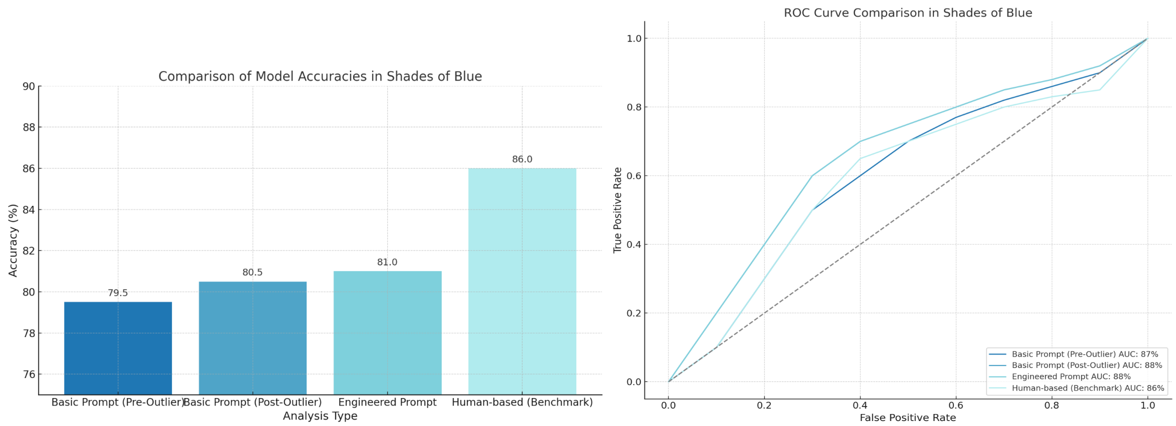


Fig. 15. Comparing accuracy and ROC curves of three different approaches

Conversely, the engineered prompts encountered errors or failed to respond altogether. Therefore, it needed several iterations to show results that can be used. The accuracy of the model increased slightly, but there was still a huge gap between this value and human-based model. Additionally, there was a lack of technical analysis, such as how results might differ with an 80-20% training-testing data split, or a report detailing feature importance after its application.

Limitations: While the findings are promising, it is essential to consider potential limitations. Some easy tasks such as visualizing confusion matrices take a long time for ChatGPT-4 and may lead to errors. Furthermore, some visualizations such as a confusion matrix are plotted empty of any results. This suggests that while ChatGPT-4 might be proficient at data analysis, it might face challenges with more complex visualization tasks. Additionally, any errors might necessitate restarting the entire process or even initiating a new chat, potentially inconveniencing the user.

Logistic Regression does not require data standardization due to its insensitivity to variable magnitudes. Yet, during the Chat-GPT4 analysis, data standardization was sometimes applied, suggesting the chosen model required it. This is important as preprocessing steps should be chosen based on the specific machine learning model in use. Overprocessing or unnecessary preprocessing can sometimes skew data or waste computational resources.

Dataset B Recommendations: In this experiment the dataset was small, ChatGPT-4 has some limitations regarding large scale datasets. There are new services like Google's new service which allows users to edit various programming languages, including SQL, Python, and Spark, to conduct analytics and machine learning tasks on a massive scale. Such services simplify user experience, facilitate data discovery, exploration, analysis, and prediction [17]. Moreover, as machine learning models and platforms evolve, there is a potential for future versions of platforms like ChatGPT to overcome the highlighted limitations. Integration with visualization libraries or enhanced capabilities can make them more versatile for end-to-end data science tasks.

5.3. Dataset C

The results thoroughly examine ChatGPT-4's ability to generate datasets and automate the seven stages of the Data Science Project Plan using a residential houses' energy efficiency dataset.

Comparative Approach: Analytically, the three approaches offer distinct depths and nuances. The basic prompt serves as an introductory exploration, providing a broad overview of the dataset. The engineered prompt technique, based on its precise principles and data analyst role-play, provides more complex insights and a better understanding of the data. The human-based analysis, deeply established traditional methods, stands as the most comprehensive, ensuring a thorough and detailed examination of the dataset and serves as the benchmark.

Efficiency and Depth: The basic prompt, while simple, gives quick insights, making it great for those new to data analysis. ChatGPT-4 is fast and covers many areas of data study, but it does not match the detailed understanding a human analyst brings. While the human-driven approach is seen as the most thorough, ChatGPT-4 struggles in some parts, especially showing data visually. This might be because text-based and complex visuals are harder to capture.

Integration: Datasets sometimes lack adequate sample sizes. Chatbots like ChatGPT-4 can generate synthetic data. Following guidelines, synthetic data, as evidenced in contemporary research, provides a robust mechanism for data scientists to engage with centrally archived data, all while preserving confidentiality of the data [16]. Moreover, incorporating chatbots into data science routines can effectively optimize and automate several aspects of the data evaluation process such as data preprocessing, exploring data, and general data profiling. This automation enables data experts to focus on deeper analysis and interpretations, and strategy creation.

Accuracy and Trust: Table 1 shows that the basic prompt approach yielded no results. Therefore, the result is unreliable. Comparing the detailed prompt to human-based methods, while the first set has tighter clusters as indicated by the lower inertia, the latter outperforms in terms of the Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index. This suggests that the human-based approach has better-defined and more distinct clusters than the first. Moreover, by comparing different metrics of optimal clusters, the model that has six clusters has the best evaluation among all models. Therefore, none of ChatGPT-4's analyses met the desired accuracy.

Limitations: Several errors occurred while using ChatGPT-4, instances with no responses, the need for reruns, and occasional loss of computed results. These resulted in iterations to achieve the desirable result. Another issue was when ChatGPT-4 wanted to choose the optimal number of clusters, the basic-prompt approach did not provide performance evaluations or reasons why it chose the certain number of clusters. Moreover, the more detailed-prompt approach did not compare different methods such as SSE and WCSS and did not show why it chose a certain number of clusters.

Dataset C Recommendations: Using the synthetic dataset for this experiment was on purpose to challenge the ability of ChatGPT-4. However, a real-world dataset, especially when the dataset is extremely large, is more challenging. It is suggested to use real-world data that is freely accessible for further challenges. For example, some institutes such as Allen Institute for AI (AI2) has provided a vast repository of unstructured (text-based) data [4].

6. Conclusion

The exploration of chatbots' role, especially ChatGPT-4, in data science has revealed a diverse landscape. ChatGPT-4, with its varying analytical approaches, offers a spectrum of insights ranging from preliminary to in-depth. While the basic prompt approach provides swift insights suitable for users without extensive technical expertise, the engineered prompts delve deeper, but have limitations for sophisticated metrics, correlations and complex calculations. Given the occasional errors and discrepancies observed in ChatGPT-4's analysis, it is crucial for users to continuously monitor and validate the results, especially when relying on the platform for detailed prompts and complex calculations. The ideal approach for data analysis might be a blend of chatbot-driven methods for preliminary tasks and human expertise for in-depth analysis. This collaborative approach ensures accuracy, depth, and nuance in data interpretation. Organizations and researchers should leverage the integration capabilities of ChatGPT-4 with repository platforms to streamline the data uploading and preprocessing stages. The potential of chatbots in creating synthetic data, a feature that can enhance sample sizes without compromising data confidentiality can also be leveraged. This capability, combined with ChatGPT-4's automation capability, can free human experts to focus on more complex analytical tasks.

While ChatGPT-4 demonstrated efficiency in providing insights, there were instances where its results and accuracy did not align with the benchmark set by human-based methods, especially in basic prompts. Users might also encounter operational challenges such as time out errors, limitations on the number of queries, word limitations, and the need to reload datasets, which can disrupt the analytical process. ChatGPT-4 faces challenges in data visualization tasks, indicating that it might not yet be a complete substitute for platforms that specialize in advanced visual representations. The platform's decision-making process, such as in tasks like optimal cluster selection, lacked comprehensive evaluation and clarity, necessitating a cautious approach when relying on it for complex analytical tasks.

Overall, the integration of chatbots in data science workflows promises a streamlined and efficient approach. However, recognizing its limitations is crucial. As chatbot technology matures, the next stage lies in a wholly collaborative approach, where chatbots handle preliminary tasks, setting the stage for human experts to ensure accuracy and nuance in data interpretation.

7. References

1. Adamopoulou, E., Moussiades, L.: Chatbots: History, technology, and applications. Machine Learning with Applications (2020). https://doi.org/10.1016/j.mlwa.2020.100006, last accessed 14/12/2023.

2. Alshater, M.: Exploring the Role of Artificial Intelligence in Enhancing Academic Performance: A Case Study of ChatGPT. SSRN (2022).

3. Cīrule, D., Bērziša, S.: Use of Chatbots in Project Management. In: Damaševičius, R., Vasiljevienė, G. (eds) Information and Software Technologies. ICIST 2019. Communications in Computer and Information Science, vol. 1078, Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30275-7_4.

4. Coldewey, D.: AI2 drops biggest open dataset yet for training language models. TechCrunch (2023). https://techcrunch.com/2023/08/18/ai2-drops-biggest-open-dataset-yet-for-training-language-models/, last accessed 14/12/2023.

5. Dave, T., Athaluri, S.A., Singh, S.: ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Frontiers in Artificial Intelligence (2023). https://doi.org/10.3389/frai.2023.1169595.

6. Gill, S.S. et al.: Transformative effects of CHATGPT on modern education: Emerging era of AI Chatbots. Internet of Things and Cyber-Physical Systems, Vol. 4 (2024), pp. 19–23. https://doi.org/10.1016/j.iotcps.2023.06.002

7. Hassani, H., Silva, E. S.: The Role of ChatGPT in Data Science: How AI-Assisted Conversational Interfaces Are Revolutionizing the Field. Big Data Cognitive Computing, 7(2), 62–71 (2023). doi: 10.3390/bdcc7020062.

8. Hamilton, K.: ChatGPT Will Now Have Access To Real-Time Info From Bing Search. Forbes (2023). https://www.forbes.com/sites/katherinehamilton/2023/05/23/chatgpt-will-now-have-access-to-real-time-info-from-bing-search-report-says/, last accessed 14/12/2023.

9. Jongerius, C.: Quantifying Chatbot Performance by using Data Analytics. Master's thesis (2023).

10. Kalla, D., Smith, N., Samaah, F., Kuraku, S.: Study and Analysis of Chat GPT and its Impact on Different Fields of Study. International Journal of Innovative Science and Research Technology, Vol. 8, Issue 3 (March 2023).

11. Khlaif, Z., Mousa, A., Hattab, M., Itmazi, J., Hassan, A., Sanmugam, M., Ayyoub, A.: The Potential and Concerns of Using AI in Scientific Research: ChatGPT Performance Evaluation. JMIR Med Educ (2023). https://mededu.jmir.org/2023/1/e47049, DOI: 10.2196/47049, last accessed 14/12/2023.

12. Lardinois, F.: Microsoft is bringing Python to Excel. TechCrunch (2023). https://techcrunch.com/2023/08/22/microsoft-is-bringing-python-to-excel/, last accessed 14/12/2023.

13. Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., Ge, B.: Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models (2023).

14. Nah, F.F.-H., Zheng, R., Cai, J., Siau, K., & Chen, L.: Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. Journal of Information Technology Case and Application Research, Vol. 25, No. 3 (2023), pp. 277-304. https://doi.org/10.1080/15228053.2023.2233814

15. Okonkwo, C.W., Ade-Ibijola, A.: Chatbots applications in education: A systematic review. Computers & Education: Artificial Intelligence (2021). https://doi.org/10.1016/j.caeai.2021.100033.

16. Singh, K.: Synthetic Data — key benefits, types, generation methods, and challenges. Towards Data Science. https://towardsdatascience.com/synthetic-data-key-benefits-types-generation-methods-and-challenges-11b0ad304b55, last accessed 14/12/2023.

17. Wiggers, K.: Google launches BigQuery Studio, a new way to work with data. TechCrunch (2023). https://techcrunch.com/2023/08/29/google-launches-bigquery-studio-a-new-way-to-work-with-data/, last accessed 14/12/2023.