International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V73I4P117

Volume 73 Issue 4, 180-190, April 2025

Original Article

Comparative Evaluation of Statistical Tools in Different ChatGPT Iterations

Pauly Awad[1], Maya Grigolia[2], Soraia Oueida[1], Nour Mostafa[1*]

[1]*College of Engineering and Technology, American University of the Middle East, Kuwait.*
[2]*Independant Researcher*.

[1*]*Corresponding Author : nour.moustafa@aum.edu.kw*

Received: 14 January 2025 Revised: 20 March 2025 Accepted: 27 March 2025 Published: 26 April 2025

Abstract *- ChatGPT, a variant of the Generative Pre-trained Transformer (GPT) models, has emerged as an essential tool for processing, summarizing, and deriving insights from vast textual datasets. This paper delves into the multifaceted role of ChatGPT in data analysis, highlighting its functionalities, challenges, and potential; unlike traditional statistical software that demands specialized expertise and often lacks interpretive capabilities, AI-powered generative tools offer a novel approach, empowering non-experts to execute statistical models and extract results accompanied by interpretations and conclusions. Our investigation evaluates the output accuracy across different iterations of ChatGPT (3.5, 4.0, and the dedicated data analyst tool) for standard statistical inferences, uncovering notable disparities that underscore the necessity of professional validation for reliability. Our findings indicate that while the 3.5 version may exhibit numerous errors and omissions, necessitating thorough scrutiny and debugging by statisticians, version 4.0 demonstrates improved accuracy in most instances. However, the results obtained through the Data Analyst tool exhibit the highest level of nuance and correctness.*

Keywords *- Artificial Intelligence, Natural Language Processing, ChatGPT, Data Analysis, Statistics*.

1. Introduction

Large Language Models (LLMs) such as ChatGPT radically change the data analysis field. These artificial intelligence tools promise to transform how we interact with and develop insights from data, offering previously unheard-of capacities for text generation, understanding, and manipulation. This article explores the potential applications, limitations, and ethical issues surrounding ChatGPT's emerging interaction with data analysis. In addition, a comparison between ChatGPT 3.5, ChatGPT 4.0, and the integrated data analyst tool in ChatGPT 4.0 is provided. Developed by OpenAI, ChatGPT is an LLM trained on a massive dataset of text and code [1]. Recent advancements in AI have yielded powerful LLMs capable of processing and generating human-quality text, translating languages, writing creative content, and answering complex questions informatively [2]. These capabilities have enormous potential to improve workflows for data analysis in terms of efficacy and efficiency. Traditionally, data analysis necessitates manual processes like data cleaning, filtering, and exploration, which can consume significant time and effort. ChatGPT can automate these tasks through its ability to understand natural language. For example, it can promptly detect and rectify data inconsistencies, format and restructure datasets and produce descriptive summaries based on user queries [3]. This automation significantly reduces manual effort, allowing analysts to focus on 2 Not specified, 2024, vol., no. higher-level tasks like model building, interpretation, and analysis. Moreover, ChatGPT's capacity for human readable insights and communication offers a solid chance to popularize data analysis. ChatGPT can enable individuals with low technical expertise to explore and analyze data by providing clear explanations and visuals. This promotes cooperation and well-informed decision-making across various sectors by enabling non-technical stakeholders to obtain insightful knowledge from data. It is necessary to understand the possible obstacles linked to the incorporation of ChatGPT into processes for data analysis. The training data's fundamental bias can produce skewed results, and the LLM's transparent decision-making process can raise questions about interpretability and reliability. Furthermore, it is crucial to pay close attention to ethical considerations related to data privacy and its misuse [4]. More specifically, several challenges need to be addressed, such as:

- Data Privacy and Security: Using ChatGPT for data analysis may additionally contain the sharing of sensitive information. Ensuring that the data is private and safe is critical to prevent unauthorized access or data breaches.

- Bias and Fairness: Language models like ChatGPT can inherit biases found in the trained data. Care should be

taken to cope with and mitigate biases to avoid perpetuating discriminatory effects while using ChatGPT for decision-making processes [5].

- Result Validation: It is imperative not to depend solely on the provided result without validating AI-generated answers. From the outset, it is acknowledged that "ChatGPT can make mistakes. Consider verifying critical information."

- Model Interpretability: The nature of large language models poses challenges for expertise and validating the reasoning behind their responses. Efforts to improve model interpretability are essential in critical data analysis tasks.

- Ethical Considerations: Responsible use of ChatGPT in data analysis requires adherence to ethical guidelines and transparency in decision-making procedures.

ChatGPT has proven promise in enhancing various components of the data analysis procedure, from data querying to reporting and collaboration. However, bias, interpretability, reliability, and ethics must be cautiously addressed in setting up privacy-related situations. Stahl and Eke provided a comprehensive ethics review for ChatGPT and Large Language Models (LLMs), disclosing a wider range of ethical concerns to provide a more balanced perspective [6]. Despite those challenges, ChatGPT can transform data analysis completely. A review study by Kwan Lo explored the challenges associated with using ChatGPT, especially in education. As a result, ChatGPT's performance is ranked using three different classes depending on the subject, ranging from outstanding and satisfactory to unsatisfactory. Although ChatGPT has massive potential, generating incorrect or fake information can bypass plagiarism detector tools and therefore data analysis must be crucial [7]. ChatGPT's capability to comprehend natural language allows it to automate time-consuming procedures, increase accessibility, and improve the understanding of complex data [8]. ChatGPT 4.0 offers incredible advances in understanding language, remembering context, and multimodal capability, making it a more efficient and flexible AI assistant. It excels in memorizing contextual understanding, reducing hallucinations, and delivering more coherent and accurate answers, even in lengthy conversations.

Its reasoning and problem-solving capabilities have also improved, enabling the tackling of complex math, coding, and analysis issues with greater logical precision. Multimodal capability makes it possible to process and understand both text and images, making it open to greater usage in various applications. In coding, ChatGPT 4.0 writes more efficient, optimized code while delivering real-time debugging and step-by-step explanations. Personalization is another strong aspect, with the ability to learn to adapt to the user's preferences regarding tone, style, and role-playing responses to ensure higher engagement and intuitiveness. It also has the added benefit of improved knowledge updates, real-time web browsing, and improved handling of mysterious issues. Efficiency has also improved through higher speed and scalable handling to tackle complex issues smoothly. Safety and ethics have also improved through higher moderation of the content, protection of privacy, and minimization of bias. Multilingual fluency has also seen a widespread improvement, delivering higher translation precision and natural-sounding feedback across languages. ChatGPT 4.0 integrates smoothly with third-party tools with higher plugin and API support to deliver higher automation and workflow efficiency. With these improvements collectively, ChatGPT 4.0 is not an AI chatbot but an evolved and flexible assistant that can perform a wide range of tasks more accurately, efficiently, and intuitively.

This paper explores these features in further detail, discussing the difficulties that arise when using them in particular data analysis contexts. ChatGPT has the potential to revolutionize various fields by gathering significant information and promoting informed decision-making through ethical development and implementation. The data analysis capabilities of ChatGPT 4.0 were evaluated by Huang et al. against those of the traditional bio-statistical programs, such as SAS, SPSS, and R. As a case study, the dataset used was related to epidemiological research. Findings state that ChatGPT 4.0 is an effective tool for statistical analysis. However, it revealed some limitations regarding result consistency, which need to be advocated in future research [9].

This paper is divided into six sections. Section 2 is dedicated to a literature review on ChatGPT and data analysis. Section 3 presents a research methodology. Section 4 lists the analysis and results of the discussed cases. In section 5, limitations and future research are discussed. Finally, Section 6 is designated for the conclusion and recommendations to ensure that ChatGPT is effectively used.

2. Literature Review

ChatGPT, an effective language model developed via OpenAI, has recently attracted massive attention for its capacity to generate human-like text responses in natural language conversations. ChatGPT can also be a valuable tool in statistical analysis. ChatGPT is based on the GPT-3 structure, a generative pre-skilled transformer model that performs numerous language know-how and technology obligations. It has a hundred seventy-five billion parameters and may generate regular and contextually applicable textual content using activities. The version's versatility has led to its adoption in various programs, including chatbots, content material technology, and records analytics. The authors of this paper explore various uses of ChatGPT, focusing on its capacity to manage tasks like text classification and answering complex questions in several domains such as education, medicine and communication [10].

One of the primary applications of ChatGPT in data analysis is assisting analysts in querying and exploring datasets. Researchers have developed interactive structures that allow customers to ask questions about their facts in natural language. ChatGPT interprets and retrieves applicable information from databases or statistics resources (e.g., SQL databases and spreadsheets) [11]. For instance, you can ask ChatGPT questions like, "What is the average sale for a certain product in Q2 2023?". Moreover, data preprocessing is an essential step in data analysis for cleaning, transforming, and integrating your data. ChatGPT can help automate this procedure by generating code snippets for common data preprocessing tasks, including handling missing values, scaling functions, or encoding qualitative variables. This can minimize time and reduce human mistakes in the process of data preparation [12]. ChatGPT can be used to generate human-readable reports from data analysis outcomes. It can remodel statistical findings, visualizations, and insights into concise and coherent narratives, making it more straightforward for non-technical stakeholders to understand and interpret the output. This can improve data-driven decision-making within companies [13].

Furthermore, ChatGPT can facilitate collaboration among data analysts and domain experts by acting as a conversational partner during the analysis process. Analysts can discuss hypotheses, validate findings, and brainstorm insights with ChatGPT, which can offer suggestions based on its vast knowledge of language models [14]. A comprehensive survey of ChatGPT with its applications, challenges and advancements is presented by Nazir and Wang [15]. The authors discussed the revolutionary process of ChatGPT creation for several real-world applications and presented this key technology tool's limitations and ethical aspects. They also highlighted the different characteristics associated with ChatGPT models, especially the massive capabilities related to language understanding. A study conducted by Sem et al. showed that ChatGPT can be used as auxiliary software for researchers in qualitative research data analysis [16]. Xing et al. explored in their study the use of ChatGPT in learning and proved its efficiency in teaching statistics and data analytics [17]. Bengesi et al. significantly contributed to the study of General Artificial Intelligence (GAI). They focused on understanding the theoretical and mathematical principles that power GAI models and the various tasks these models can perform. They also investigated the current limitations of GAI and discussed its potential for future development [18].

Moreover, GAI demonstrates extensive capabilities in data analysis, as extensively discussed in various literature covering areas such as data exploration, interpretation, and collaboration. These literature findings have been grouped in Table 1 for clarity and easy reference. As natural language processing technology continues to improve over time, ChatGPT and similar models will probably play an increasingly significant role in data analysis, improving efficiency and accessibility on this subject. Alawida et al. explored the impact of ChatGPT on the Natural Language Processing (NLP) field in their work. The authors analysed the different aspects of this technology, including the training data used to develop ChatGPT, its ability to generate human-like text and its capabilities in NLP. They also examined its various applications, such as language translation, dialogue generation and text summarization. Additionally, they discussed limitations and addressed ethical and privacy aspects related to its use. Finally, this study compares ChatGPT to other existing language generation models to provide a thorough analysis of its impact and potential [14]. The application of ChatGPT, specifically to education, is explored in the study by Zayoud et al. The authors presented the massive advantages and impacts of its application to learning strategies, including opportunities and recommendations for safely using this generative language model. The following section will describe, discuss, and analyze the case studies addressed using ChatGPT to perform data analysis.

3. Methodology

This analysis aims to assess the consistency and reliability of ChatGPT's results compared to the gold standard provided by traditional statistical methodologies. To achieve this, a comprehensive comparison was conducted across all available ChatGPT versions 3.5, 4.0, and 4.0 data analyst features. Additionally, case studies were developed focusing on the most commonly used statistical inference methods, including Simple Linear Regression (SLR) and hypothesis testing for two sample means. In the first case, a Simple Linear Regression (SLR) was performed between the score and the hours studied. This allowed us to assess ChatGPT’s capability to handle key aspects like model fitting, parameter estimation, and residual analysis. In this context the focus was on the regression model to check the capabilities of ChatGPT. The second case involved hypothesis testing for two sample means using electricity accessibility data sourced from the World Bank. The data was segmented into two groups (developed and developing countries), and a two-sample t-test was conducted to compare the means. ChatGPT’s outputs, including regression coefficients, test statistics, and p-values, were compared to those produced by standard statistical software such as Excel and Minitab. The procedure involved detailed data preparation, prompt engineering to structure requests to ChatGPT, and parallel analyses with traditional tools to ensure a rigorous comparison. Discrepancies between ChatGPT’s results and those from traditional methods were examined, particularly in cases where assumptions (e.g., normality or variance equality) were violated or when complex issues such as multicollinearity appeared in the data. Error analysis was conducted to identify the root causes of any inconsistencies, providing a thorough evaluation of ChatGPT’s potential as a reliable statistical analysis tool. The methodological approach taken in this analysis allows for a comprehensive assessment of ChatGPT’s reliability and

consistency when compared to traditional statistical software. By leveraging both simulated and real-world data and utilizing various statistical techniques, the strengths and limitations of ChatGPT across different versions were analyzed. The results provide insight into the feasibility of using ChatGPT as a tool for statistical analysis, particularly in exploratory and educational contexts.

4. Analysis and Results

This section is dedicated to analysing the findings and discussing the results. Two case studies are considered. The first depends on a simple linear regression analysis, and the second concerns hypothesis testing for two samples of the means and variances using actual data from the World Bank.

Table 1. ChatGPT capability in data analysis

ChatGPT in Data Exploration	
Natural Language Querying	**Data Visualization**
ChatGPT has been leveraged to facilitate statistical exploration via natural language querying. Users can interact with their datasets in simple language, asking complex questions and receiving significant insights. ChatGPT bridges the space between information and users, allowing a more accessible and intuitive approach to exploration [19].	In addition to answering questions, ChatGPT can generate data visualizations. Changing data into comprehensible narratives and visual representations help convey complex insights to non-technical stakeholders, enhancing records-driven choice-making [20].
ChatGPT in Data Interpretation	
Automated Insights	**Explaining Model Outputs**
ChatGPT can help data analysts by automatically generating insights from the given data. It identifies patterns, anomalies, and potential correlations, offering a valuable beginning for further analysis. This can keep analysts a huge amount of time inside the initial stages of exploration [21].	Understanding model decisions is essential in machine learning. ChatGPT may generate human-readable explanations for complex version outputs, improving model transparency and interpretability. This is especially vital in critical applications where model decisions impact real-world outcomes [22].
ChatGPT in Data Collaboration	
Collaborative Analysis	**Automated Documentation**
ChatGPT serves as a collaborative partner for facts analysts and domain experts. It enables discussions, hypothesis testing, and brainstorming sessions, fostering interdisciplinary collaboration and knowledge sharing [23].	ChatGPT can automate the documentation of data analysis techniques. It generates clean and concise summaries of analysis steps, ensuring that insights are appropriately documented and accessible to group members and stakeholders [24].

4.1. Case 1: Simple Linear Regression (SLR) Analysis using the ChatGPT Different Versions (3.5, 4.0, and Data Analyst)

The first example is about conducting simple linear regression. The same example using the same prompt will be given to ChatGPT 4.0 and the data analyst tool. The analysis cannot be done using ChatGPT 3.5 since this version does not support uploads, and the data file will be uploaded as an Excel file. So, ChatGPT 3.5 gave the example with the data set as a text command.

The AI-generated results are compared with the output of statistical software. Simple linear regression is used to study the relationship between two quantitative variables. It mathematically models the endogenous or dependent variable and the exogenous or independent variable as a linear equation [25]. The data set in Table 2 represents simulated scores of students in a statistics course for a given number of study hours in the semester. The scatterplot of the data used is illustrated in Figure 1.

Table 2. Data used for the simple linear regression model

Study Hours	35	51	28	35	33	44	44	39
Score	90	93	77	75	76	86	82	82

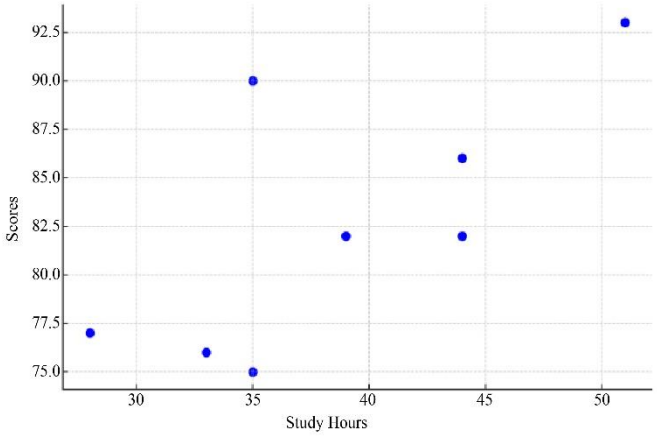

Fig. 1 Scatterplot: Study Hours vs. Score

The aim is to find the SLR regression model between the score, as a dependent variable, and the study hours, as an independent variable. First, the analysis was performed using statistical software, followed by the data analyst tools of ChatGPT 4.0, ChatGPT 4.0, and ChatGPT 3.5. Based on statistical software, the regression equation is Score = 58.3 + 0.631 Study Hours. The following command was used to run the same analysis in different versions of ChatGPT: "The following data set represents the scores students achieved in a statistics course for a given number of study hours in the semester. Can you find the regression equation?" The second command was, "What does it mean?" The analysis was done quickly, supplemented with the associated Python code. The data was uploaded as an Excel file in the Data Analyst tool existing in ChatGPT 4.0. The answer started with a brief explanation of the simple linear regression, and ChatGPT 4.0 returned the same equation as the Minitab software. After the regression equation, the following prompt was to describe the result. The output of ChatGPT 4.0 was very professional and provided a detailed explanation of the slope, intercept, and overall model.

After that, the same was performed in ChatGPT 4.0 without using the Data Analyst tool. The answer provided by ChatGPT 4.0 was similar to the data analyst tool and returned the same regression equation. Regarding the interpretation of the result, ChatGPT 4.0 only interpreted the meaning of the slope and the intercept without mentioning any information about the overall model. The last analysis was done using the accessible version of ChatGPT, which is ChatGPT 3.5. First, the data upload file is not supported in this version. ChatGPT 3.5 was given the value in the prompt, and the same question was asked to determine the regression equation, and the result was surprising. ChatGPT 3.5 first showed how to find the regression equation, but unfortunately, the answer was wrong, and the regression equation found was the following: Score = 0.6471*Study Hours + 72.0286. The calculation needed to be more accurate for the slope and, therefore, for the intercept. The result interpretation could have been more professional than ChatGPT4 and the data analyst.

In summary, referring to Table 3, the ChatGPT 3.5 version showcased limitations in computational accuracy and could not interpret or analyze data correctly. The main takeaway is that while ChatGPT 3.5 could handle basic inquiries, its performance in complex analytical tasks, such as regression analysis, was found not appropriate. The inaccuracies in calculations and subsequent analysis underscore the need for improvements in handling numerical data and statistical interpretation. A significant advancement over its predecessor, ChatGPT 4.0, demonstrated the ability to accurately calculate regression equations and interpret coefficients precisely. However, it fell short of offering model interpretations, indicating room for further enhancement in comprehensively understanding and explaining the outcomes of statistical models. The improvements in accuracy and partial interpretative abilities mark a step forward in integrating AI with data analytics.

The data analyst tool emerged as the most adept tool, providing accurate calculations and coefficient interpretations and delivering comprehensive insights into the model's behavior and implications. Its ability to encompass all aspects of regression analysis, from equation accuracy to insightful interpretation, underscores its superiority in specialized analytical tasks. The ultimate solution is integrating advanced AI models, like ChatGPT 4.0, with specialized data analyst tools. This hybrid approach combines the strengths of both the intuitive, conversational interface of ChatGPT and the analytical precision of data analyst tools. By leveraging the advancements in AI within the framework of specialized analytics tools, users can achieve a balance between user-friendly interactions and rigorous data analysis. Such integration promises enhanced accuracy, deeper insights, and a more holistic understanding of complex data analysis tasks, especially within a paid version.

Table 3. Comparison of SLR analysis between statistical software and different iterations of ChatGPT (NA Stands for not available – A for available – C for correct – I for incorrect)

	Software	Chat GPT 3.5	Chat GPT 4	ChatGPT 4 Data Analyst
Regression Equation	C	I	C	C
Coefficients interpretation	NA	A and C	A and C	A and C
Model Interpretation	NA	NA	NA	A and C

4.2. Case 2: Hypothesis Testing on the Difference in means using the ChatGPT Different Versions (3.5, 4.0, and Data Analyst Tool)

This example concerns the hypothesis testing for two samples for the means and variances using actual data from the World Bank about electricity accessibility, which was analyzed using ChatGPT 3.5, ChatGPT 4.0, ChatGPT 4.0 data analyst tool, and statistical software. Sustainable Development Goal (SDG) number 7 aims to ensure access to affordable, reliable, and sustainable energy for all. The share of the global population with access to electricity increased from 78 percent in 2000 to 91 percent in 2021 [26]. Aggregated measures, like means, hide the discrepancies within the data. Behind the impressive improvement in this indicator, it is interesting if different groups of countries experience equal accessibility. The World Bank data has been used to compare the percentage of the population with access to electricity in high-income vs. low-income countries. When testing hypotheses about two sample means, the researcher needs to choose the proper testing methodology based on available information about the variances of the two populations. There are three possible scenarios for comparing two sample averages, requiring a solid understanding of assumptions and manual decision-making for accurate results.

Also, the analyst must interpret the results effectively to derive meaningful and correct insights. Additionally, considering factors like the sample size, significance level, and type of data distribution can play a critical role in the accuracy of the test results. After checking the data's independence and normality, the analysis of two sample means can take shape in three different scenarios, depending on the information available about variances. If population variances are unspecified, a test for the ratio of two variances based on the two samples is necessary to correctly determine the model for the two-sample means hypothesis. Refer to Figure 2 for a clear guide on choosing the correct test method.

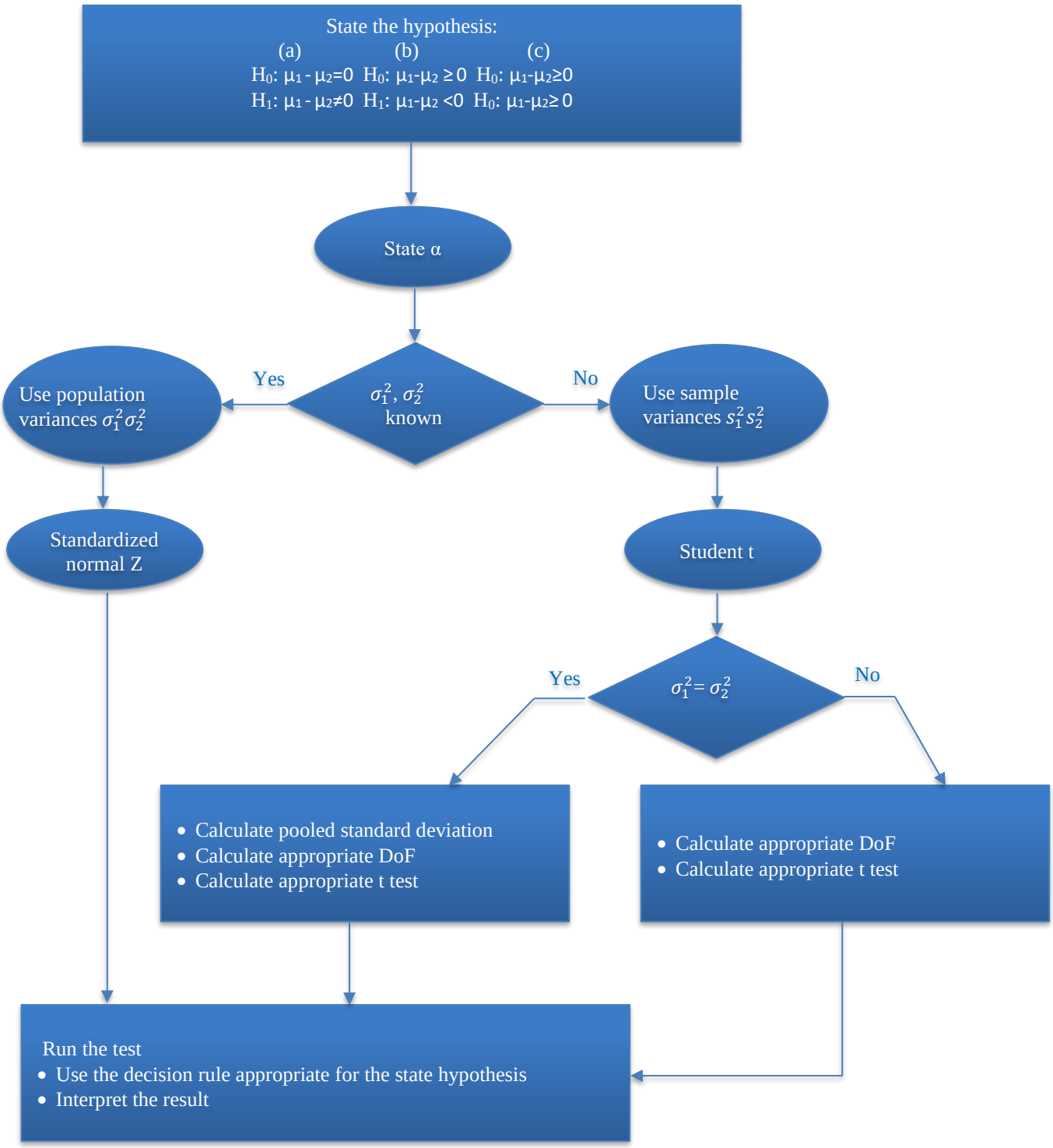

Fig. 2 Simplified flowchart for selecting the appropriate hypothesis test

The simplified flowchart for selecting the appropriate hypothesis test is presented to compare two sample means where the populations are independent and normally distributed. The World Bank assigns the world's economies to four income groups: low, lower-middle, upper-middle, and high-income. Classifications are based on the previous year's GNI per capita. GNI measures are expressed in United States dollars (USD). The latest classification is given in Table 4. The latest complete data about the % access to electricity (2021) was used. The dataset was downloaded from the Word Bank Atlas for Sustainable Development Goals 2023, and the data were matched with the country's income classification. Of 214 countries, 81 are high-income, 53 are upper-middle income, 54 are lower-middle income, and 26 are low-income countries. Table 4 and Figure 3 summarize the descriptive information about the variable. Lower-middle income averages are compared to upper-middle income averages. Two alternative commands were created for ChatGPT. The first one requests running a two-sample mean test using the provided summarized data to see if Gen-AI can identify the need for testing variances first. Meanwhile, the second command provides instructions for testing variance and utilizing the result to correctly choose the two-sample test procedure.

Table 4. World Bank country classifications by income level; % of population with access to electricity, 2021

Group	Country classifications by income level July 1, 2022, for FY23	Number of countries	Average % of population with access to electricity (2021)
Low income	<1085	26	41.37%
Lower-middle income	1086-4255	53	81.41%
Upper-middle income	4256-13205	52	96.46%
High income	>13205	81	99.85%

Source: World Bank, authors calculations

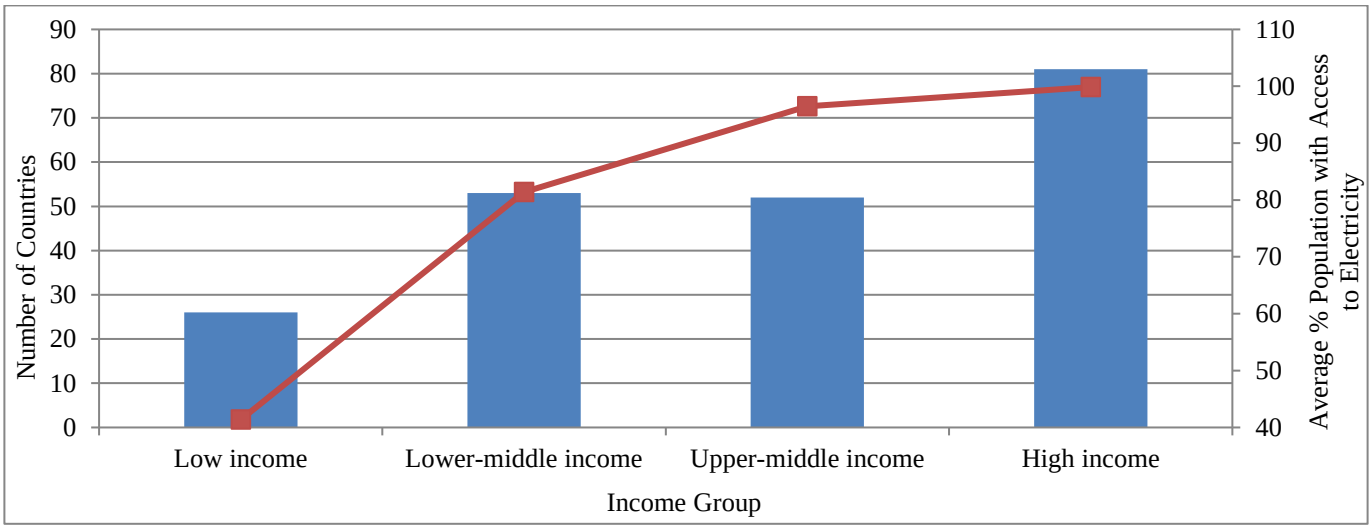

Fig. 3 World bank country classifications by income level and access to electricity (2021)

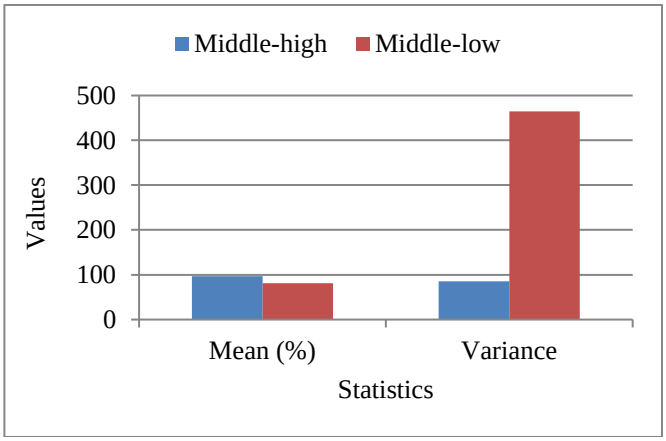

Fig. 4 Comparison of mean and variance: middle-high vs. middle-low-income countries

Table 5. Access to electricity summary statistics for middle-high income and middle-low income countries

	Middle-high	Middle-low
Mean (%)	96.465	81.411
Variance	85.548	464.214
Sample size	52	53

The information given to the chat was as follows: Sustainable Development Goal (SDG) number 7 aims to ensure access to affordable, reliable, and sustainable energy for all. The share of the global population with access to electricity increased from 78 percent in 2000 to 91 percent in 2021 (Atlas of Sustainable Development Goals 2023, World Bank). Aggregated measures, such as means, may obscure discrepancies within the data. Behind the impressive improvement in this indicator, it is intriguing to explore

whether different groups of countries experience equal accessibility to electricity. Table 5 and Figure 4 provide the means of electricity access (in %) for middle-high and middle-low country groups, along with their variances.

By comparing statistical software capabilities after running Command N1, it is evident that while ChatGPT 3.5 had its strengths in basic calculations and providing discussions, it struggled with detailed calculations such as degrees of freedom in t-tests and critical values. ChatGPT 4.0 represents a significant improvement, offering detailed calculations and explanations for results. On the other hand, the data analyst tool showcased remarkable performance, delivering the most thorough and precise solutions akin to traditional statistical software. However, neither version of ChatGPT could identify the need to test variances and provide a discussion of choosing the correct method for the two-sample mean test. The commands used for the previous case are presented in Table 6. Table 7 summarizes the output provided by the statistical software ChatGPT3.5, ChatGPT4.0, and the Gen AI data analyst tool using Command N1.

Table 6. Gen-AI commands for the two samples mean hypothesis testing case

Command N1 to ChatGPT to test using the provided summarized data, aiming to see if Gen AI can identify the need for testing variances first.	Command N2 to ChatGPT instructions for testing variance and utilizing the result to correctly choose the two-sample test procedure.
Is there enough evidence to claim that the electricity access in middle-high and middle-low-income countries is different? use a 5% significance level. Show calculations, display the final output, and write detailed interpretations.	Is there enough evidence to claim that the electricity access in middle-high and middle-low-income countries is different? Use a 5% significance level. Make sure to use the correct t-test by conducting the two variances equality test before comparing two means. Show calculations, display the final output, and write detailed interpretations.

In the comparison of statistical software capabilities, while running command N2, it is evident that ChatGPT 3.5 struggled with conducting the two-sample variance test, as it only displayed the formulas without performing the calculations. This limitation highlights a significant gap in functionality compared to ChatGPT 4.0 and the data analyst tool generated by Gen-AI. ChatGPT 4.0 and the data analyst tool provided more robust solutions, offering accurate and nuanced results for the two-sample variance test. However, qualitative differences in the solutions become apparent when comparing ChatGPT 4.0 with the data analyst tool. While both tools offer correct test statistics for the two-sample variance test, the data analyst tool outperforms ChatGPT 4.0 in readability and providing detailed discussion. Additionally, the data analyst tool excels in delivering the correct decision for choosing the appropriate two-sample mean test, whereas ChatGPT 4.0 fails to do so. Looking further into the comparison, ChatGPT 3.5 lacks most of the outputs. At the same time, ChatGPT 4.0 and the data analyst tool share the capability to accurately determine whether to reject the two-sample mean test and provide interpretations and conclusions accordingly. Table 8 summarizes Not Specified, 2024, vol., no. 11, the output provided by the statistical software, ChatGPT3.5, ChatGPT4.0, and the Gen-AI data analyst tool using Command N2.

Table 7. Output based on command N1 (NA stands for not available – A for available – C for correct)

	Software	**ChatGPT 3.5**	**ChatGPT 4**	**ChatGPT 4 Data Analyst**
Identifies the need for variances test	NA	NA	NA	NA
Test Statistics for two sample means test	A and C	A and C	A and C	A and C
The P-value for two sample means test	A and C	NA	NA	A and C
Decision to reject two sample means test	A and C	A and C	A and C	A and C
Interpretation and conclusion for two sample means test	NA	A and C	A and C	A and C

The results of the two case studies represented in this work support the hypothesis stating that the latest versions of AI tools, mainly ChatGPT4.0 and the Data Analyst tool, are capable of handling statistical analysis compared to the previous versions, like ChatGPT 3.5. That was evident in the accuracy of the regression equation and the interpretation of the hypothesis testing results. ChatGPT 3.5, while able to calculate basic equations, demonstrated inaccuracies and a lack of deep interpretation. For instance, ChatGPT 4.0, especially when enhanced with the Data Analyst tool, produced correct calculations and adequate interpretations. The study's findings support the hypothesis that AI tools are evolving toward greater accuracy in statistical computations. Moreover, the performance of ChatGPT in statistical analysis is in line with previous research, which suggests that AI's accuracy in dealing with numerical data improves as the models evolve [27]. Similar conclusions have been drawn from prior research evaluating AI-based tools with statistical

software, indicating that while AI is powerful, it still lags in several contexts, such as hypothesis testing and model interpretation [28]. However, including these tools can enhance AI's functionality, bringing its performance closer to that of specialized software such as Minitab, R, and Python. The results of this study proved practical effects, especially in the field of data analysis and education. The accuracy of ChatGPT 4.0 and other AI tools can perform regression analysis and hypothesis testing that can be used as an additional tool in both academic and professional environments. Educators can benefit from these tools to teach complex statistical concepts, while businesses may use them for quick data analysis.

Table 8. Output based on command N2 (NA stands for not available – A for available – C for correct – W for wrong)

	Software	ChatGPT 3.5	ChatGPT 4	ChatGPT 4 Data Analyst
Test Statistics for two variances	A and C	W	A and C	A and C
Decision to choose the correct two-sample means test	NA	W	A and C	A and C
Test Statistics for two sample means test	A and C	NA	A and C	A and C
The P-value for two sample means test	A and C	NA	A and C	A and C
Decision to reject two sample means test	A and C	NA	A and C	A and C
Interpretation & conclusion two sample means test.	A and C	NA	A and C	A and C

5. Limitation and Future Work

Several limitations were identified in this study. First, our analysis was restricted to only two statistical methods: simple linear regression and two-sample hypothesis testing. So, the generalization of the findings may be limited as other statistical models were not tested. Second, the inaccuracies found in ChatGPT 3.5 demonstrate the risks of depending on earlier iterations of AI for accurate data analysis. The study opens up various future research.

Lastly, further research is needed to explore the capacity of AI to handle complex statistical models, such as but not limited to multiple regression and non-parametric tests. Furthermore, future studies may examine the performance of other AI tools, such as Claude AI and Gemini, and compare them with both traditional statistical software and the latest AI models. Also, the applicability of AI in corporate and educational sectors may be improved by strengthening its capacity to fully understand and explain statistical results in a simple manner. In summary, the findings of this study showed that AI tools represent a significant advancement. However, the observed limitations emphasize the need for continued improvements. As AI continues to evolve, its applications in various fields, including education and business, will continue to expand, provided that its limitations are acknowledged and addressed.

6. Conclusion

In conclusion, ChatGPT stands as a fundamental development in the realm of artificial intelligence, specifically in its application to data analysis. Its proficiency in handling and interpreting large volumes of data efficiently showcases its potential to revolutionize how we approach data-driven tasks. This capacity for practical text analysis, combined with the promise of future enhancements, positions ChatGPT as a transformative force in the ever-evolving data analytics landscape. Based on the conducted analysis, ChatGPT 4.0 represents an improvement over its predecessor in certain aspects of fundamental data analysis. However, it still falls short compared to the data analyst tool.

The data analyst tool provides correct and detailed output, enhances readability, and offers superior decision-making capabilities. Even though the paper proved that it relies not only on ChatGPT responses, technical knowledge and analysis skills are still needed to validate the responses generated by such tools, especially for non-statistician users. The future iterations of ChatGPT should offer more accurate data analysis capabilities, advancing more informed decision-making across various sectors. Realizing the full potential of this technology necessitates a continuous commitment to several key areas. First, technical refinement is essential to enhance accuracy, speed, and adaptability to various data contexts. Secondly, ethical considerations must be rigorously addressed to ensure that the deployment of such AI tools aligns with social values and norms, particularly in areas concerning privacy and data security. ChatGPT's capabilities can be effectively and responsibly leveraged to adapt to an increasingly data-centric world. More research and studies are generally needed to explore the potential of ChatGPT in data analysis.

References

[1] Introducing ChatGPT, OpenAI, 2022. [Online]. Available: https://openai.com/blog/chatgpt/

[2] Tom B. Brown et al., "Language Models Are Few-Shot Learners," *arXiv*, pp. 1-75, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[3] Fiona Fui-Hoon Nah et al., "Generative AI and ChatGPT: Applications, Challenges, and AI-Human Collaboration," *Journal of*

Information Technology Case and Application Research, vol. 25, no. 3, pp. 277-304, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] Miles Brundage et al., "The Malicious use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," *arXiv*, pp. 1-101, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[5] Emily M. Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, Canada, pp. 610-623, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] Bernd Carsten Stahl, and Damian Eke, "The Ethics of ChatGPT-Exploring the Ethical Issues of an Emerging Technology," *International Journal of Information Management*, vol. 74, pp. 1-14, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Chung Kwan Lo, "What is the Impact of ChatGPT on Education? A Rapid Review of the Literature," *Education Sciences*, vol. 13, no. 4, pp. 1-15, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[8] Partha Pratim Ray, "ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope," *Internet of Things and Cyber-Physical Systems*, vol. 3, pp. 121-154, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] Yeen Huang et al., "Evaluating ChatGPT-4.0's Data Analytic Proficiency in Epidemiological Studies: A Comparative Analysis with SAS, SPSS, and R," *Journal of Global Health*, vol. 14, pp. 1-10, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Yiheng Liu et al., "Summary of ChatGPT-Related Research and Perspective Towards the Future of Large Language Models," *Meta-Radiology*, vol. 1, no. 2, pp. 1-14, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Julio Christian Young, and Makoto Shishido, "Investigating OpenAI's ChatGPT Potentials in Generating Chatbot's Dialogue for English as a Foreign Language Learning," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, pp. 65-72, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Muhammad Usman Hadi et al., "A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage," *Authorea Preprints*, pp. 1-56, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[13] Anna R. McAlister, Saleem Alhabash, and Jing Yang, "Artificial Intelligence and ChatGPT: Exploring Current and Potential Future Roles in Marketing Education," *Journal of Marketing Communications*, vol. 30, no. 2, pp. 166-187, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[14] Moatsum Alawida et al., "A Comprehensive Study of ChatGPT: Advancements, Limitations, and Ethical Considerations in Natural Language Processing and Cybersecurity," *Information*, vol. 14, no. 8, pp. 1-23, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] Anam Nazir, and Ze Wang, "A Comprehensive Survey of ChatGPT: Advancements, Applications, Prospects, and Challenges," *Meta-Radiology*, vol. 1, no. 2, pp. 1-12, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] Mert Şen, Şevval Nur Şen, and Tuğrul Gökmen Şahin, "A New Era for Data Analysis in Qualitative Research: ChatGPT!," *Shanlax International Journal of Education*, vol. 11, pp. 1-15, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[17] Yixun Xing, "Exploring the use of ChatGPT in Learning and Instructing Statistics and Data Analytics," *Teaching Statistics*, vol. 46, no. 2, pp. 95-104, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Staphord Bengesi et al., "Advancements in Generative Ai: A Comprehensive Review of Gans, GPT, Autoencoders, Diffusion Model, and Transformers," *IEEE Access*, vol. 12, pp. 69812-69837, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[19] Xuanming Zhang et al., "GrounDialog: A Dataset for Repair and Grounding in Task-Oriented Spoken Dialogues for Language Learning," *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 300-314, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Carlo Perrotta, *Natural Language Generation and the Automation of Pedagogical Communication*, World Yearbook of Education 2024, Taylor & Francis Group, pp. 54-69, 2023. [Google Scholar] [Publisher Link]

[21] Isha Kondurkar, Akanksha Raj, and D. Lakshmi, *Modern Applications with a Focus on Training ChatGPT and GPT Models: Exploring Generative AI and NLP*, Advanced Applications of Generative AI and Natural Language Processing Models, IGI Global, pp. 186-227, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[22] Rich Caruana et al., "Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission," *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Sydney, NSW, Australia, pp. 1721-1730, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[23] Tiziano Labruna et al., "Unraveling ChatGPT: A Critical Analysis of AI-Generated Goal-Oriented Dialogues and Annotations," *International Conference of the Italian Association for Artificial Intelligence*, Rome, Italy, pp. 151-171, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[24] Richard Brath, and Craig Hagerman, "Automated Insights on Visualizations with Natural Language Generation," *25th International Conference Information Visualisation*, Sydney, Australia, pp. 278-284, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[25] Ivonne Nuñez et al., "Designing A Comprehensive and Flexible Architecture to Improve Energy Efficiency and Decision-Making in Managing Energy Consumption and Production in Panama," *Applied Sciences*, vol. 13, no. 9, pp. 1-20, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[26] Atlas of Sustainable Development Goals, World Bank Group, 2023. [Online]. Available: https://datatopics.worldbank.org/sdgatlas?lang=en

[27] Stefano A. Bini, "Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care?" *The Journal of Arthroplasty*, vol. 33, no. 8, pp. 2358-2361, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[28] John D. Kelleher, and Brendan Tierney, *Data science*, MIT press, 2018. [Google Scholar] [Publisher Link]