IJITGEB

INTERNATIONAL JOURNAL IN INFORMATION TECHNOLOGY IN GOVERNANCE, EDUCATION AND BUSINESS
Vol. 7, No. 1, 2025
ISSN 2686-0694 (Print)
e-ISSN 2721-0030 (Online)

Exploring ChatGPT's Proficiency in Nonparametric Statistics: An Initial Review and Benchmark Assessment

Joel L. De Castro
joel.decastro@upou.edu.ph
University of the Philippines Open University, Philippines

Abstract: Artificial Intelligence (AI) is transforming education, particularly in teaching statistics, by enhancing personalized learning and feedback through tools like ChatGPT (Tulsiani, 2024). ChatGPT is an advanced artificial intelligence chatbot developed by OpenAI that uses deep learning to understand and generate human-like text. It is based on the GPT (Generative Pre-trained Transformer) model, trained on vast amounts of text data to assist with answering questions, generating content, and engaging in natural conversations. This study evaluates ChatGPT version 3.5 performance in nonparametric statistical analysis by assessing its ability to generate solutions for seven tests, including the Test of Randomness, ANOVA, Chi-Square Goodness-of-Fit Test, Median Test, Cochran's Q Test, Wilcoxon-Mann-Whitney Test, and Binomial Probability Test. Using three prompt engineering strategies—Basic Prompt (BP), Structured Prompt (SP), and Error-Awareness Prompt (EAP)—ChatGPT's outputs are compared against manual calculations and statistical software (Jeffreys's Amazing Statistics Program(JASP) and Excel) for accuracy, consistency, and clarity. Results show significant discrepancies in Basic Prompt outputs between November 2023 and 2024, with sum of squares values of 6421.82 and 6928.00, and an F-value of 0.93 (p = 0.53), indicating no significant difference. Similarly, the effect of prompt type is statistically insignificant (F = 1.43, p = 0.26), as is the absolute error analysis (F = 0.59, p = 0.57). However, differences in statistical test approaches are significant (F = 3.10, p = 0.04), suggesting that method selection impacts accuracy. Findings emphasize the role of structured and error-aware prompts in improving ChatGPT's performance, highlighting the importance of effective prompt engineering in nonparametric statistics. These insights contribute to improving AI-assisted learning in statistical education and research, ensuring more reliable computational outputs. Lastly, guidelines for effective prompt engineering in Nonparametric Statistics were formulated.

Keywords: ChatGPT, Prompt Engineering, Artificial Intelligence, Non-Parametric Statistics, Tutoring Tool

INTRODUCTION

Innovative teaching methods in statistics education were a result of the incorporation of AI in education. For instance, ChatGPT helps students from a variety of backgrounds understand statistical concepts, develop datasets, and even enable hypothesis testing. According to research, AI tools can help students with linguistic and conceptual clarity while also encouraging critical thinking and problem-solving skills (Klayklung et al., 2023).

New techniques for analyzing data are being fostered by the combination of AI and statistics. The flexibility and resilience of nonparametric statistical methods in evaluating data without rigid distributional assumptions make them very useful (Sonwalkar, 2024). AI-driven language models, such as ChatGPT, have been investigated in recent research for their ability to do nonparametric statistical analysis. These investigations assess the accuracy and reliability of ChatGPT's responses to statistical queries, highlighting the evolving role of AI in data analysis (Ordak, 2023).

It is also vital to emphasize that any enhancements made to instructional strategies using AI tools should improve the current teaching resources but not to replace them (Zaman, 2023). With regards to teachers' capacity, it can provide a greater understanding of how a computer works and interacts with comprehension of the mathematical

Received Date: February 2, 2025 / **Revised Date:** March 18, 2025 / **Accepted Date:** March 30, 2025

education and learning processes. Therefore, it must be the goal of any learning institution to ascertain AI's usefulness and efficiency.

The integration of ChatGPT in the study of nonparametric statistics aligns with the newly created guidelines on AI of the University of the Philippines Open University (UPOU) that represents a pivotal advancement in educational methodology. By harnessing ChatGPT's natural language processing capabilities, educators can now provide personalized and interactive learning experiences that adhere to ethical AI principles. Through adaptive learning approaches, learners can engage in realtime discussions, problem-solving activities, and collaborative exercises, thereby deepening their understanding of nonparametric statistical concepts while cultivating critical thinking skills. Moreover, ChatGPT facilitates assessment and feedback mechanisms, enabling educators to evaluate learners' comprehension and address misconceptions effectively.

REVIEW OF RELATED LITERATURE

The introduction of AI into statistics represents a revolutionary change in the way that data analysis is currently carried out (Evans, 2023). In statistics, AI is being used increasingly to improve data analysis. AI is defined as computer systems that can execute activities that normally require human intelligence, such as understanding natural language, learning from data, and making educated decisions. A study by (Wang et al. 2024) claims that AI systems are more reliable and easier to use, and that they offer resources that improve learning and decision-making. It also enables real-time feedback on problems such as hypothesis testing, regression, and probability distributions, offering tailored support that enhances comprehension (Zhao & Yu, 2022, Harvard Data Science Review, 2023).

According to (Deng and Lin, 2023), the goal of AI, is to build intelligent computers with human-like thoughts and behavior. One of the more popular AI known today is ChatGPT which promises to increase efficiency and improve accuracy since it is being powered by a large-scale pre-trained language model, which enables it to quickly and accurately understand customer questions and generate natural-sounding responses.

The integration of AI in statistics is a transformative force, ushering new possibilities for data analysis. AI's automation, efficiency, pattern recognition, personalization, and real-time capabilities underscore its growing significance in data analysis. Within the domain of nonparametric statistical analysis, AI holds the promise of enhanced accuracy, efficiency, accessibility, scalability, and interpretability, revolutionizing how nonparametric statistical analyses are conducted. Supporting this is the comparative comparison of four AI chatbots—ChatGPT, GPT-4, Bard, and LLaMA—with possible applications in statistics and mathematics education (Calonge, 2023). The study assesses and contrasts these systems' characteristics, capabilities, and possible uses in the statistics and calculus fields. In particular, insights into the selection and application of AI chatbots in calculus and statistics to improve student learning by analyzing their advantages and disadvantages were clearly presented.

Research comparing AI tools for mathematics and statistics education highlights their diverse capabilities and limitations, particularly in adaptive learning, personalized feedback, and problem-solving. AI tools like ChatGPT and Wolfram Alpha has specialized platforms each that gave unique strengths to the educational landscape. For instance, ChatGPT excels in offering conversational, accessible explanations that adapt to users' needs, making it especially useful for students seeking clarity on statistical concepts. However, its responses can sometimes lack precision compared to tools like Wolfram Alpha, which are specifically designed for symbolic computation and structured problem-solving. Studies suggest that integrating these tools with human expertise can address AI's limitations in emotional intelligence and non-verbal adaptability, providing a balanced educational experience (Toolify.ai, 2024)

ChatGPT is acknowledged for its enhanced mathematical capabilities and potential to boost academic achievement by imparting fundamental mathematical knowledge and a range of other topics to its users (Wardat, 2023). The public conversation on social media is mostly supportive of ChatGPT's usage in teaching mathematics and in educational settings, and it may provide thorough guidance and support for studying geometry. ChatGPT is also viewed as an important guide in grasping statistical ideas, although more instruction on a few particular topics could be necessary (Al-qadri, 2023). Extra suggestions were made to prevent ChatGPT's detrimental effects on the process of educational assessment.

According to research by Zhao & Yu (2022) in the International Journal of Artificial Intelligence in Education, AI tools like ChatGPT improve understanding through step-by-step guidance and the ability to adapt explanations based on student inputs. Similarly, a study published in the Harvard Data Science Review (2023) highlights that ChatGPT democratizes access to advanced statistical methods, bridging gaps in traditional educational

Received Date: February 2, 2025 / **Revised Date:** March 18, 2025 / **Accepted Date:** March 30, 2025

resources. Furthermore, the tool's adaptability and responsiveness make it particularly valuable in self-paced or remote learning environments (Aquarius AI, n.d.).

Despite its widespread use in mathematics education, particularly in statistical learning, concerns persist about the efficiency and accuracy of ChatGPT in handling specialized tasks like nonparametric statistical analysis. Researchers have noted that while AI tools like ChatGPT are adept at generating responses based on vast datasets, their outputs may sometimes lack the precision and contextual understanding required for technical applications (Deng & Lin, 2023; Grassini, 2023). These challenges highlight the need for a critical evaluation of ChatGPT's role in educational and research settings, particularly in areas requiring detailed mathematical computations and interpretations.

In a similar work by (Shakarian, 2023) it was found out that ChatGPT was able to answer all of the math problems correctly, but it sometimes rounded the answers. Likewise, the study also made it clear that ChatGPT were not able to provide clear answers to the problems that required multiple values. As a result, ChatGPT always returned an answer to each problem, but some of its answers were incorrect, a condition that was highly similar from the present study. Therefore, he suggests a serious precaution in the application of AI tools and the importance of counter check using other means.

Table 1

Research on AI Prompt Engineering

Title of Research	Description
Discovering prompt engineering: A Qualitative Study of Nonexpert Teachers' Interactions with ChatGPT	This qualitative study examines how nonexpert teachers engage with ChatGPT to develop educational games. It sheds light on the learning curve associated with prompt engineering and offers methods for educators to effectively utilize AI in creating interactive learning materials (Carl et al., 2024)
Analyzing Student Prompts and Their Effect on ChatGPT's Performance	This study explores undergraduate students' use of ChatGPT for problem-solving, focusing on the prompting strategies they develop. It examines the correlation between prompt quality and AI performance, providing insights into effective prompt engineering practices for students (Sawalha et al., 2024).
Generative AI and Prompt Engineering in Education	This study explores the potential of generative AI in education, highlighting the role of prompt engineering in improving AI-assisted learning. It also outlines strategies for educators to effectively integrate AI tools into their teaching practices (Artem & Sergiy, 2023).
Cases of EFL Secondary Students' Prompt Engineering Pathways to Complete a Writing Task with ChatGPT	This study investigates how EFL secondary students use prompt engineering to complete a writing task with ChatGPT. It examines the strategies and pathways they follow in refining prompts, highlighting their problem-solving skills and digital literacy in AI-assisted writing (Woo et al., 2023).
Performance of ChatGPT on the US Fundamentals of Engineering Exam: Comprehensive Assessment of Proficiency and Potential Implications for Professional Environmental Engineering Practice	This study assesses ChatGPT's proficiency in answering questions from the U.S. Fundamentals of Engineering (FE) Exam in environmental engineering. It explores the model's performance and considers its potential impact on professional engineering practice (Pursnani et al., 2023).
A Preliminary Exploration of the Disruption of a Generative AI Systems: Faculty/Staff and Student Perceptions of ChatGPT and its Capability of Completing Undergraduate Engineering Coursework	This study explores faculty, staff, and student perceptions of ChatGPT and its ability to complete undergraduate engineering coursework. It offers preliminary insights into how generative AI may disrupt traditional teaching, learning, and assessment practices in engineering education (White et al., 2024).

Received Date: February 2, 2025 / **Revised Date:** March 18, 2025 / **Accepted Date:** March 30, 2025

This research aligns with the growing evidence that the way questions are structured can significantly influence the performance of AI models. For example, (Brown et al., 2020) found that prompt design plays a crucial role in guiding AI systems to produce accurate and meaningful outputs. In Table 1 are the other research studies on AI prompt engineering.

Other studies, such as (Patil et al., 2024), have shown that more detailed prompts often lead to improved accuracy and fewer errors in AI-generated solutions. These are further research that focused on prompt engineering.

Main Objective of the Study

This study investigates how accurately and clearly ChatGPT 3.5 responds to non-parametric statistics problems, focusing on the effects of different ways of asking questions (called "prompts") on its performance. Specifically, the study has the following objectives:

1. Compare ChatGPT's answers to solutions obtained using manual computations (via MS Excel) and statistical software like JASP (Jeffreys's Amazing Statistics Program) to determine how closely they match correct results.
2. Conduct an experiment on various ways of phrasing prompts—Basic(simple), Structured(detailed), or emphasizing error awareness—to analyze how these differences affect both the accuracy and clarity of the responses.
3. Examine how well ChatGPT explains its solutions, ensuring the correct formation of hypotheses, the test of statistics to be used, and the conclusion to the problem.
4. Test whether ChatGPT provides reliable and consistent answers when asked similar non-parametric questions multiple times.
5. Suggest practical ways to design better prompts to improve ChatGPT's performance in solving non-parametric statistics problems.
6. Improve the usability of AI tools like ChatGPT in solving statistical problems, while also providing practical guidance for users on designing effective prompts.

This study lies in its potential to improve the application of AI tools, such as ChatGPT, in education and research, particularly in addressing the challenges of solving non-parametric statistical problems. Non-parametric tests are crucial in fields where data do not meet the assumptions required for parametric tests, yet they often present interpretive and computational difficulties (Siegel & Castellan, 1988). By evaluating the accuracy and clarity of ChatGPT's responses, the research addresses the growing demand for reliable, accessible, and user-friendly computational tools for students, educators, and researchers. Studies like (Colosimo et al., 2021) have highlighted the importance of AI in supporting statistical education, particularly in simplifying complex computations and facilitating understanding.

By identifying common errors in ChatGPT's responses and recommending strategies to mitigate them, this study can contribute to the broader effort of understanding AI limitations and improving their application in professional and educational contexts. Eventually, this research not only enhances the usability of AI in statistical problem-solving but also supports the development of best practices for integrating AI into academic and research workflows.

RESEARCH DESIGN & METHODS

This study evaluates the accuracy, interpretability or clarity, and consistency of ChatGPT's responses to nonparametric statistical problems, with a focus on how prompt design impacts its performance. For this investigation, ChatGPT 3.5, the free edition, was used. Nonparametric tests were selected based on topics commonly taught in graduate-level Statistics courses, including the Test of Randomness, ANOVA, Chi-Square Goodness-of-Fit Test, Median Test, Cochran's Q Test, Wilcoxon-Mann-Whitney Test, and Binomial Probability Test. These tests are foundational for analyzing data that do not meet parametric assumptions, making them crucial in research and education (Siegel & Castellan, 1988). Datasets were designed to reflect realistic academic scenarios, ensuring sufficient complexity to rigorously evaluate ChatGPT's computational and explanatory capabilities.

The study employed three types of prompts to explore the influence of prompt phrasing on ChatGPT's outputs: (1) basic prompts with straightforward instructions, (2) structured prompts providing detailed and specific

Received Date: February 2, 2025 / **Revised Date:** March 18, 2025 / **Accepted Date:** March 30, 2025

guidance, and (3) error-aware prompts emphasizing potential pitfalls and requiring justification for solutions. These problems were solved manually using reliable tools such as Microsoft Excel, JASP statistical software, and a scientific calculator to establish benchmarks for comparison, aligning with best practices in statistical computation (Field, 2013).

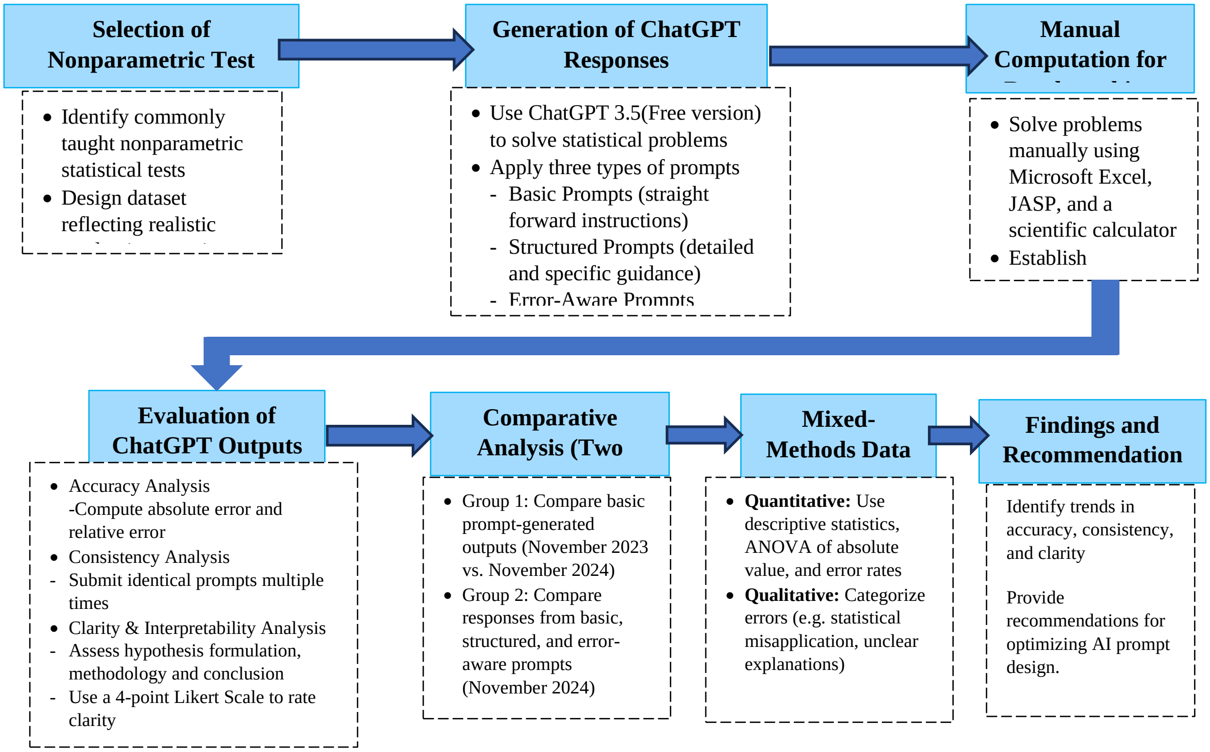
Accuracy was assessed using quantitative metrics, including absolute error (|x - x*|), defined as the magnitude of the difference between ChatGPT's solution(x) and the true value (x*), and relative error, calculated as the ratio of the absolute error to the true value (|x - x*| / |x*|). Determining the measurement's absolute and relative inaccuracy is crucial when doing any kind of measurement so that one may comprehend the potential implications of errors. This degree of inaccuracy is taken into consideration by the absolute and relative error to produce the most precise measurement possible(Boisvert, n.d.).

Consistency was evaluated by repeatedly submitting identical prompts and analyzing the stability of ChatGPT's responses within a week on November 2024. Data consistency guarantees that information is consistent, accurate, and trustworthy throughout a database, system, or application. In order to preserve data integrity and facilitate precise data analysis, this concept ensures that data values remain consistent throughout processing, storage, and retrieval (Team Atlan, 2024).

Clarity and Interpretability was evaluated by qualitatively examining how well ChatGPT explained its reasoning. This included assessing its hypothesis formulation, statistical conclusions, and methodological steps. The goal was to ensure that the ChatGPT-generated output for a nonparametric statistics problem was interpretable, with logically consistent and clearly defined null and alternative hypotheses (Mayer, 2009). Clarity was quantified using a 4-point Likert scale, where 1 indicated "unclear" and 4 indicated "very clear." The interpretation of computed statistics, including step-by-step explanations of the test statistic, computed statistics value, and comparisons to critical values, was also assessed on the same scale. Finally, the conclusion was evaluated to ensure it logically aligned with the statistical results and was communicated clearly and accessibly. The flowchart for this study's methodology is shown in figure 1.

Figure 1

Flowchart of the Methodology


Selection of Nonparametric Test
• Identify commonly taught nonparametric statistical tests
• Design dataset reflecting realistic
Generation of ChatGPT Responses
• Use ChatGPT 3.5(Free version) to solve statistical problems
• Apply three types of prompts
- Basic Prompts (straight forward instructions)
- Structured Prompts (detailed and specific guidance)
- Error-Aware Prompts
Manual Computation for
• Solve problems manually using Microsoft Excel, JASP, and a scientific calculator
• Establish
Evaluation of ChatGPT Outputs
• Accuracy Analysis
-Compute absolute error and relative error
• Consistency Analysis
- Submit identical prompts multiple times
• Clarity & Interpretability Analysis
- Assess hypothesis formulation, methodology and conclusion
- Use a 4-point Likert Scale to rate clarity
Comparative Analysis (Two
• Group 1: Compare basic prompt-generated outputs (November 2023 vs. November 2024)
• Group 2: Compare responses from basic, structured, and error-aware prompts (November 2024)
Mixed-Methods Data
• Quantitative: Use descriptive statistics, ANOVA of absolute value, and error rates
• Qualitative: Categorize errors (e.g. statistical misapplication, unclear explanations)
Findings and Recommendation
Identify trends in accuracy, consistency, and clarity
Provide recommendations for optimizing AI prompt design.

Received Date: February 2, 2025 / **Revised Date:** March 18, 2025 / **Accepted Date:** March 30, 2025

A mixed-method approach was used for data analysis. Quantitative analysis included descriptive statistics, ANOVA of Absolute Value, and error rates to measure accuracy and consistency, while qualitative analysis focused on categorizing common errors, such as misapplication of statistical methods or unclear explanations, as identified in prior research (Hanckel et al., 2021). These findings informed practical recommendations for prompt design, contributing to the growing literature on optimizing AI interactions (Brown et al., 2020).

Accuracy tests were conducted for two groups. The first group involved the basic prompt-generated outputs for November 2023 and November 2024. The second group compared the outputs generated using basic prompts, structured prompts, and error-awareness prompts for November 2024. Similarly, clarity, interpretability, and consistency were also assessed for both groups.

RESULTS AND DISCUSSION

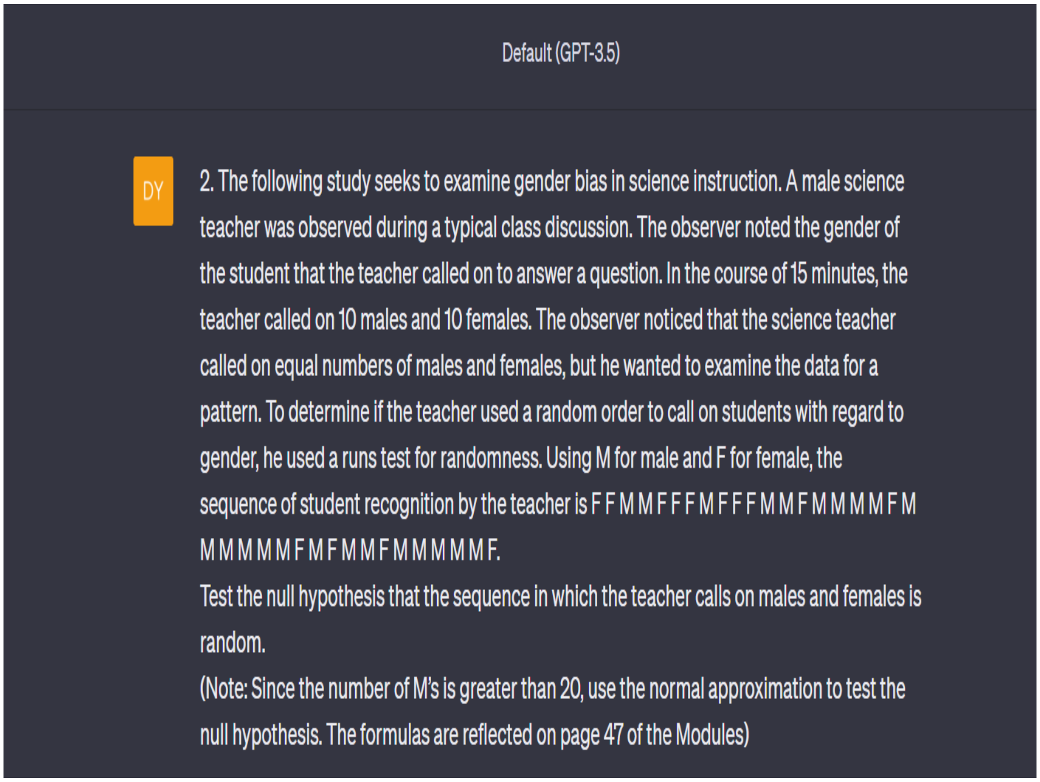
Accuracy

Basic Prompt between November 2023 and November 2024

The comparison between ChatGPT's generated outputs in November 2023 and November 2024 for nonparametric statistical tests reveals significant improvements in accuracy, though some inconsistencies persist. As a sample run, in Figure 2 displays the basic prompt, which is the verbatim copy of the word problems in Test of Randomness, then in figure 3 is the ChatGPT’s generated answer from November 2023 against Figure 4 which was a portion of generated output for November 2024. While the generated values were compared to the computed solutions for the same problem, refer to figure 5.

Figure 2

Basic Prompt for Problem in Test of Randomness

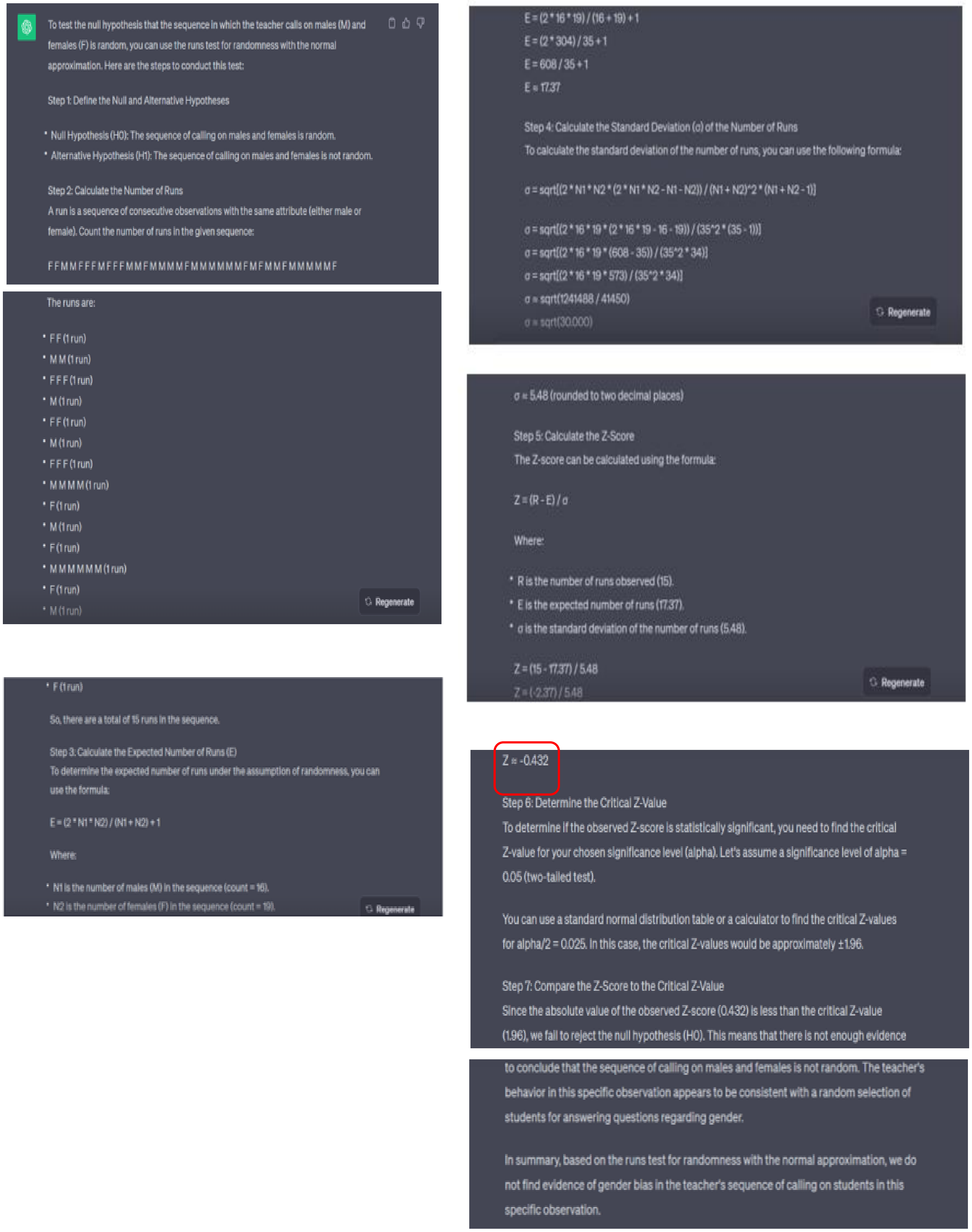


2. The following study seeks to examine gender bias in science instruction. A male science teacher was observed during a typical class discussion. The observer noted the gender of the student that the teacher called on to answer a question. In the course of 15 minutes, the teacher called on 10 males and 10 females. The observer noticed that the science teacher called on equal numbers of males and females, but he wanted to examine the data for a pattern. To determine if the teacher used a random order to call on students with regard to gender, he used a runs test for randomness. Using M for male and F for female, the sequence of student recognition by the teacher is F F M M F F F M F F F M M F M M M M F M M M M M M F M F M M F M M M M M F.
Test the null hypothesis that the sequence in which the teacher calls on males and females is random.
(Note: Since the number of M's is greater than 20, use the normal approximation to test the null hypothesis. The formulas are reflected on page 47 of the Modules)

Received Date: February 2, 2025 / **Revised Date:** March 18, 2025 / **Accepted Date:** March 30, 2025

Figure 3

ChatGPT's Generated Output from Basic Prompt (November 2023)



To test the null hypothesis that the sequence in which the teacher calls on males (M) and females (F) is random, you can use the runs test for randomness with the normal approximation. Here are the steps to conduct this test:

Step 1: Define the Null and Alternative Hypotheses

- Null Hypothesis (H0): The sequence of calling on males and females is random.
- Alternative Hypothesis (H1): The sequence of calling on males and females is not random.

Step 2: Calculate the Number of Runs
A run is a sequence of consecutive observations with the same attribute (either male or female). Count the number of runs in the given sequence:

F F M M F F F M F F F M M F M M M M F M M M M M M F M F M M F M M M M M F

The runs are:

- F F (1 run)
- M M (1 run)
- F F F (1 run)
- M (1 run)
- F F (1 run)
- M (1 run)
- F F F (1 run)
- M M M M (1 run)
- F (1 run)
- M (1 run)
- F (1 run)
- M M M M M M (1 run)
- F (1 run)
- M (1 run)

- F (1 run)

So, there are a total of 15 runs in the sequence.

Step 3: Calculate the Expected Number of Runs (E)
To determine the expected number of runs under the assumption of randomness, you can use the formula:

E = (2 * N1 * N2) / (N1 + N2) + 1

Where:

- N1 is the number of males (M) in the sequence (count = 16).
- N2 is the number of females (F) in the sequence (count = 19).

E = (2 * 16 * 19) / (16 + 19) + 1
E = (2 * 304) / 35 + 1
E = 608 / 35 + 1
E ≈ 17.37

Step 4: Calculate the Standard Deviation (σ) of the Number of Runs
To calculate the standard deviation of the number of runs, you can use the following formula:

σ = sqrt[(2 * N1 * N2 * (2 * N1 * N2 - N1 - N2)) / (N1 + N2)^2 * (N1 + N2 - 1)]

σ = sqrt[(2 * 16 * 19 * (2 * 16 * 19 - 16 - 19)) / (35^2 * (35 - 1))]
σ = sqrt[(2 * 16 * 19 * (608 - 35)) / (35^2 * 34)]
σ = sqrt[(2 * 16 * 19 * 573) / (35^2 * 34)]
σ ≈ sqrt(1241488 / 41450)
σ ≈ sqrt(30.000)

σ ≈ 5.48 (rounded to two decimal places)

Step 5: Calculate the Z-Score
The Z-score can be calculated using the formula:

Z = (R - E) / σ

Where:

- R is the number of runs observed (15).
- E is the expected number of runs (17.37).
- σ is the standard deviation of the number of runs (5.48).

Z = (15 - 17.37) / 5.48
Z = (-2.37) / 5.48

Z ≈ -0.432

Step 6: Determine the Critical Z-Value
To determine if the observed Z-score is statistically significant, you need to find the critical Z-value for your chosen significance level (alpha). Let's assume a significance level of alpha = 0.05 (two-tailed test).

You can use a standard normal distribution table or a calculator to find the critical Z-values for alpha/2 = 0.025. In this case, the critical Z-values would be approximately ±1.96.

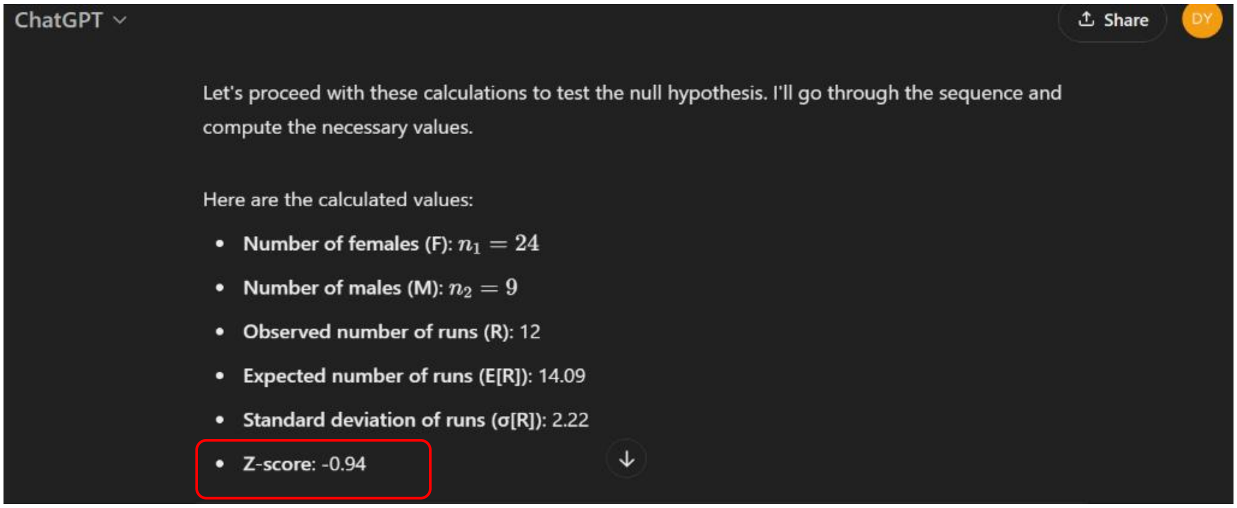
Step 7: Compare the Z-Score to the Critical Z-Value
Since the absolute value of the observed Z-score (0.432) is less than the critical Z-value (1.96), we fail to reject the null hypothesis (H0). This means that there is not enough evidence to conclude that the sequence of calling on males and females is not random. The teacher's behavior in this specific observation appears to be consistent with a random selection of students for answering questions regarding gender.

In summary, based on the runs test for randomness with the normal approximation, we do not find evidence of gender bias in the teacher's sequence of calling on students in this specific observation.

Received Date: February 2, 2025 / **Revised Date:** March 18, 2025 / **Accepted Date:** March 30, 2025

Figure 4

Part of ChatGPT's Generated Output from Basic Prompt (November 2024)



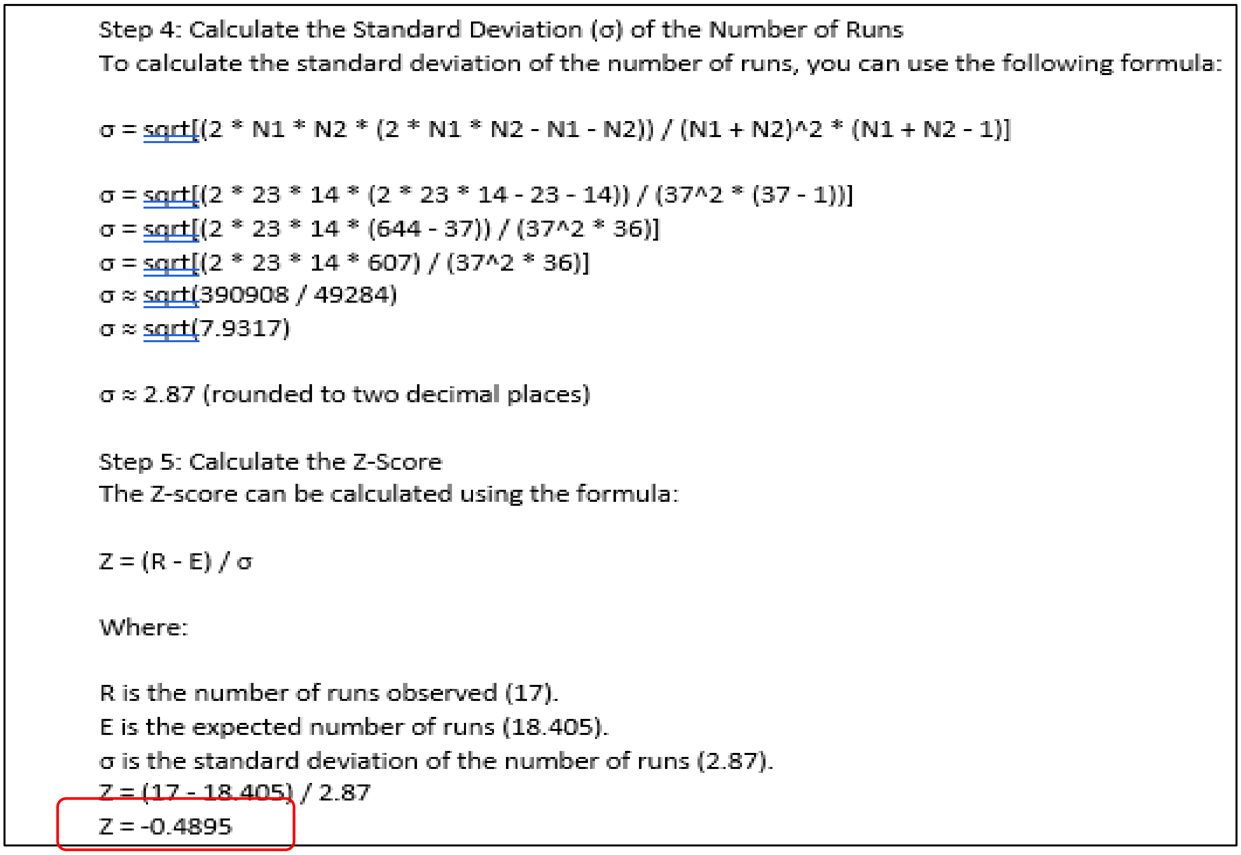
Let's proceed with these calculations to test the null hypothesis. I'll go through the sequence and compute the necessary values.

Here are the calculated values:

- **Number of females (F):** $n_1 = 24$
- **Number of males (M):** $n_2 = 9$
- **Observed number of runs (R):** 12
- **Expected number of runs (E[R]):** 14.09
- **Standard deviation of runs (σ[R]):** 2.22
- **Z-score:** -0.94

Figure 5

Portion of Manual Computation using Calculator



Step 4: Calculate the Standard Deviation (σ) of the Number of Runs
To calculate the standard deviation of the number of runs, you can use the following formula:

σ = sqrt[(2 * N1 * N2 * (2 * N1 * N2 - N1 - N2)) / (N1 + N2)^2 * (N1 + N2 - 1)]

σ = sqrt[(2 * 23 * 14 * (2 * 23 * 14 - 23 - 14)) / (37^2 * (37 - 1))]
σ = sqrt[(2 * 23 * 14 * (644 - 37)) / (37^2 * 36)]
σ = sqrt[(2 * 23 * 14 * 607) / (37^2 * 36)]
σ ≈ sqrt(390908 / 49284)
σ ≈ sqrt(7.9317)

σ ≈ 2.87 (rounded to two decimal places)

Step 5: Calculate the Z-Score
The Z-score can be calculated using the formula:

Z = (R - E) / σ

Where:

R is the number of runs observed (17).
E is the expected number of runs (18.405).
σ is the standard deviation of the number of runs (2.87).
Z = (17 - 18.405) / 2.87
Z = -0.4895

Table 2

Generated Output using Basic Prompt (November 2023 VS November 2024)

		Basic Prompt			Basic Prompt			MS Excel/JASP
		December 2023			November 2024			
Test		**ChatGPT Generated Value**	**Absolute Error |GV-CV|**	**Relative Error (%) (AE/CV)**	**ChatGPT Generated Value**	**Absolute Error |GV-CV|**	**Relative Error (AE/CV)**	**Computed Value**
1. Test of Randomness	Expected Runs (E)	17.37	1.04	0.06	14.09	4.32	0.23	18.41
	Standanrd Variation(σ)	5.48	2.61	0.91	2.22	0.65	0.23	2.87
	Z- score	-0.43	0.06	-0.12	-0.94	0.45	-0.92	-0.49
2. ANOVA	MSB	0.10	0.12	0.56	1.388	1.17	5.31	0.22
	MSW	0.080	0.07	9.00	0.029	0.02	2.65	0.01
	F Value	12.00	15.95	0.57	47.55	19.60	0.70	27.95
	Critical Value	3.35	0.00	0.00	3.88	0.53	0.16	3.35
3. Chi-square (χ^2) Goodness-of-fit Test	Chi-Square Value	47.81	25.45	1.14	9.08	13.28	0.59	22.36
	Critical Value	9.49	0.00	0.00	9.49	0.00	0.00	9.49
4. Median Test	Chi-square value	0.00	0.00	#DIV/0!	0.67	0.67	0.00	0.00
	degree of freedom	1.00	0.00	0.00	1.00	0.00	0.00	1.00
	Critical value	3.81	0.00	0.00	3.841	0.03	0.01	3.81
5. Cochran Q test	Q value	78.80	78.47	237.79	-33.80	34.13	103.42	0.33
	Degree of freedom	2.00	0.00	0.00	2.00	0.00	0.00	2.00
	critical Value	5.99	0.00	0.00	5.99	0.00	0.00	5.99
6. Wilcoxon-Mann-Whitney Test	Sum of Male Ranks	387.00	214.00	1.24	173.00	0.00	0.00	173.00
	Sum of Female Ranks	365.00	75.00	0.26	292.00	2.00	0.01	290.00
	U value	365.00	310.00	5.64	53.00	2.00	0.04	55.00
	critical value	64.00	0.00	0.00	64.00	0.00	0.00	64.00
7. Binomial Probability Test	P(X ≥ 6)	0.0392	0.020	0.99	0.0214	0.0017	0.09	0.02
	P(X ≥ 7)	0.0080	0.005	1.29	0.0043	0.0008	0.23	0.00

The data shows a comparison of ChatGPT’s performance in generating statistical values across two time periods, December 2023 and November 2024 as shown in Table 2. For each statistical test, key metrics such as the generated values, absolute errors (the difference between ChatGPT's output and the computed value), and relative errors (the percentage of the error compared to the computed value) are analyzed. By looking at the errors in December 2023 and November 2024, it becomes evident whether ChatGPT's accuracy has improved or not over time.

Received Date: February 2, 2025 / **Revised Date:** March 18, 2025 / **Accepted Date:** March 30, 2025

In some tests, such as the Test of Randomness and ANOVA, the reduction in both absolute and relative errors suggests that ChatGPT became more reliable in performing these calculations by November 2024. On the other hand, if errors remained high or unchanged in specific tests, such as the Chi-Square Goodness-of-Fit Test or Cochran Q-Test, this indicates areas where ChatGPT continues to struggle in generating accurate statistical outputs. Moreover, the presence of consistent discrepancies in critical values across time periods points to potential limitations in how ChatGPT interprets or applies statistical formulas.

The comparison reveals that while there are improvements in certain statistical computations, there remain challenges in ensuring consistent accuracy across all tests. This analysis highlights the importance of cross-verifying ChatGPT's results with traditional tools like MS Excel or JASP, especially when making decisions based on statistical data. These trends reflect OpenAI's continued optimization of numerical accuracy and reasoning in newer iterations of their models (OpenAI, 2024; Brown et al., 2023).

In Table 3, the comparison of absolute error between November 2023 and November 2024 shows meaningful differences that suggest improvements in measurement accuracy over the year. The decrease in expected runs and standard variation indicates that the data became more consistent, while the higher Z-score suggests that the data for 2024 is more random and less predictable than in 2023.

Table 3

Accuracy Between Basic Prompt (BP): Absolute Error (November 2023 VS November 2024)

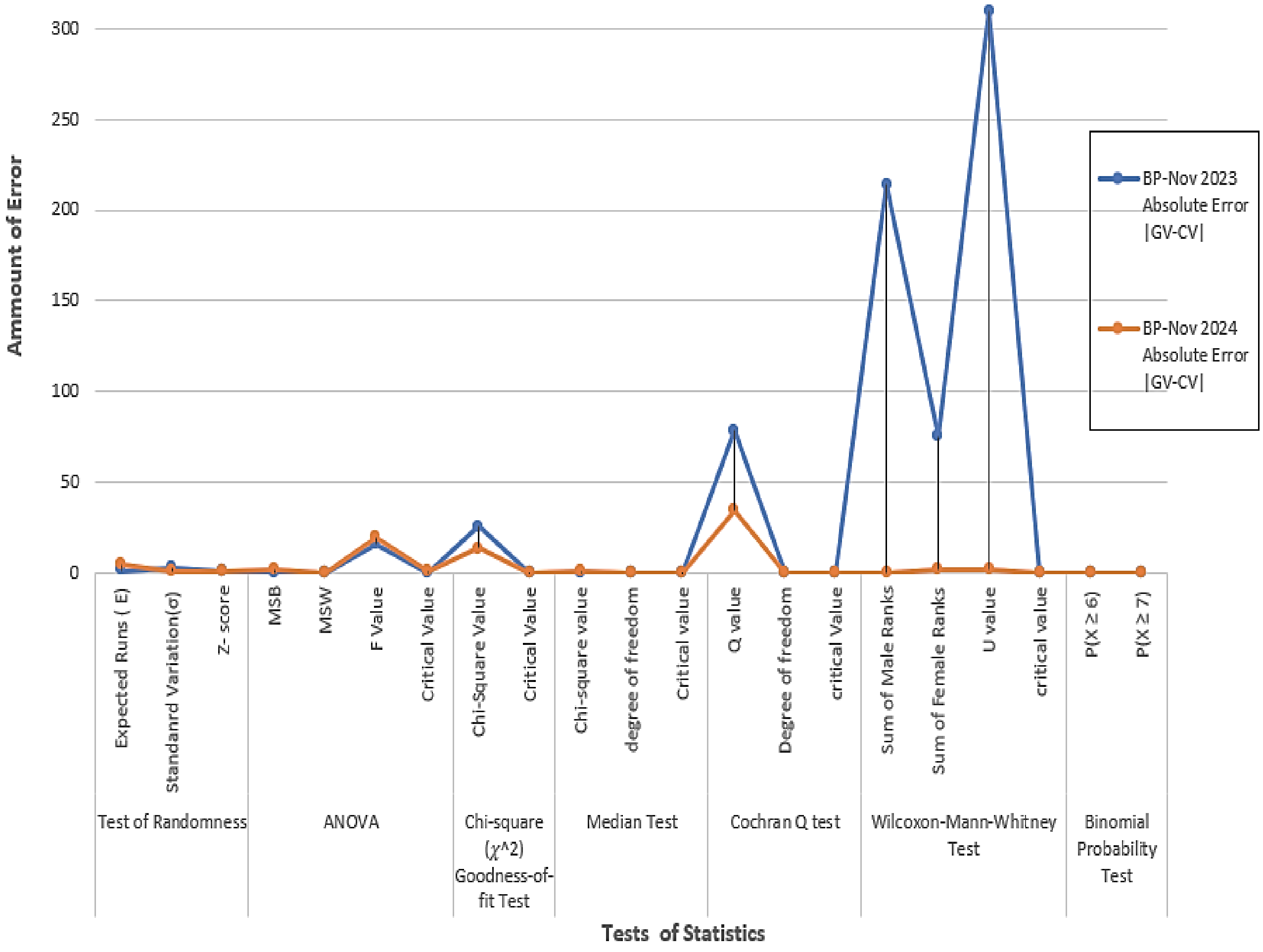
Test of Statistics		Values	**BP-Nov 2023 Absolute Error| GV-CV|**	**BP-Nov 2024 Absolute Error |GV-CV|**
1	**Test of Randomness**	Expected Runs (E)	1.04	4.32
		Standanrd Variation(σ)	2.61	0.65
		Z- score	0.06	0.45
2	**ANOVA**	MSB	0.12	1.17
		MSW	0.07	0.02
		F Value	15.95	19.60
		Critical Value	0.00	0.53
3	**Chi-square (χ^2) Goodness-of-fit Test**	Chi-Square Value	25.45	13.28
		Critical Value	0.00	0.00
4	**Median Test**	Chi-square value	0.00	0.67
		degree of freedom	0.00	0.00
		Critical value	0.00	0.03
5	**Cochran Q test**	Q value	78.47	34.13
		Degree of freedom	0.00	0.00
		critical Value	0.00	0.00
6	**Wilcoxon-Mann-Whitney Test**	Sum of Male Ranks	214.00	0.00
		Sum of Female Ranks	75.00	2.00
		U value	310.00	2.00
		critical value	0.00	0.00
7	**Binomial Probability Test**	$P(X \geq 6)$	0.02	0.00
		$P(X \geq 7)$	0.00	0.00

In the ANOVA results, the increase in Mean Square Between and the F values signals that the differences between groups became more significant, demonstrating better performance in identifying variations. Additionally, the lower Chi-square values in the goodness-of-fit tests suggest that the data in 2024 fit the expected patterns better than in 2023. Overall, these findings imply that the measurements have become more reliable and accurate from November 2023 to November 2024, possibly due to improvements in how data was collected or analyzed, leading to better predictions and less error this year compared to last year.

However, mixed results were observed in tests like ANOVA and the Median Test. While the FFF-value in ANOVA improved in 2024, errors in the Mean Square Between (MSB) increased, highlighting areas where updates may have introduced overgeneralization. Similarly, the Median Test, which showed no errors in 2023, presented minor discrepancies in chi-square and critical values in 2024, suggesting that refinements in handling specific tests might have inadvertently introduced new inaccuracies. Rank-based tests, such as the Wilcoxon-Mann-Whitney Test and Cochran Q-Test, showed the most consistent improvement, with error reductions in sum ranks and QQQ-values, demonstrating enhanced reliability in ordinal data handling. Figure 6 provided a clear visualization of these comparisons.

Figure 6

Absolute Error of ChatGPT Generated Output using Basic Prompt (November 2023 VS November 2024)



Received Date: February 2, 2025 / **Revised Date:** March 18, 2025 / **Accepted Date:** March 30, 2025

From the data provided in Table 4, it appears there are discrepancies in the results of the ANOVA test between the two time periods, as measured by absolute error in key components such as the sum of squares, degrees of freedom (df), mean square, F-value, and p-value. The sum of squares and mean square show relatively small differences (6421.82 in one period compared to 6928.00 in the other), but the F-value and p-value, which are critical for interpreting the significance of the test, differ significantly. The F-value of 0.93 and the p-value of 0.53 indicate a lack of statistical significance in one case, while the absence of corresponding values in the second dataset suggests incomplete or inconsistent results. These discrepancies highlight potential issues in the consistency or accuracy of calculations, which could affect the interpretation of statistical outcomes and decision-making based on this analysis. This calls for further review and validation of the results using other reliable statistical tools.

Table 4

Test of ANOVA for Basic Prompt: Test of Statistics (**ANOVA – Absolute Error)**

Cases	Sum of Squares	df	Mean Square	F	p
Test of Statistics	38530.89	6	6421.82	0.93	0.53
Residuals	48496.03	7	6928.00		

Note. Type III Sum of Squares

The data in Table 5 shows the absolute error for an ANOVA test, focusing on the type of prompt as a factor and residuals as the unexplained variance. For the "Type of Prompt," the sum of squares is 9247.91, with a degree of freedom (df) of 1, resulting in a mean square of 9247.91. The F-value of 1.43 and the p-value of 0.26 indicate that the type of prompt does not have a statistically significant effect, as the p-value exceeds the common significance threshold of 0.05. The residuals account for a larger portion of the variance, with a sum of squares of 77779.01 and a mean square of 6481.58 across 12 degrees of freedom. This suggests that most of the variation in the data is unexplained by the type of prompt, implying that other factors or random error may play a more significant role in influencing the results.

Table 5

*Test of ANOVA for Basic Prompt : Types of Prompts (***ANOVA- Absolute Error)**

Cases	Sum of Squares	df	Mean Square	F	p
Type of Prompt	9247.91	1	9247.91	1.43	0.26
Residuals	77779.01	12	6481.58		

Note. Type III Sum of Squares

The comparison between the Basic prompt outputs of ChatGPT in November 2023 and November 2024 highlights notable differences in its accuracy and reliability over time. In November 2023, the outputs showed higher absolute and relative errors for key statistical tests, suggesting that ChatGPT struggled with precision when performing complex calculations. By November 2024, there was a noticeable improvement, with reduced errors in many cases, indicating advancements in the model's computational abilities.

Basic Prompt (BP) VS Structured Prompt (SP) VS Error Awareness Prompts (EAP) :November 2024

Table 6 compares the accuracy of statistical test outputs generated by ChatGPT using three types of prompts—Basic (BP), Structured (SP), and Error-Awareness (EAP)—with computed values from MS Excel and JASP. The findings reveal that the accuracy of ChatGPT's responses varies depending on the type of prompt used and the complexity of the statistical test.

For the Test of Randomness, ChatGPT's outputs using the Structured and Error-Awareness prompts were accurate for the expected runs and Z-scores, closely matching the computed values. However, the Basic Prompt

Received Date: February 2, 2025 / **Revised Date:** March 18, 2025 / **Accepted Date:** March 30, 2025

resulted in significant deviations, particularly for the expected runs. This highlights how better-structured prompts improve accuracy in statistical tests.

Table 6

ChatGPT's Generated Values Using Different Prompts: November 2024

			Basic Prompt	Structured Prompt	Error-Awareness Prompt	
			ChatGPT Generated Value	**ChatGPT Generated Value**	**ChatGPT Generated Value**	**MS Excel/JASP Computed Value**
1	**Test of Randomness**	Expected Runs (E)	14.09	18.41	18.41	18.41
		Standanrd Variation(σ)	2.22	2.82	2.82	2.87
		Z- score	-0.94	-0.50	-0.50	-0.49
2	**ANOVA**	MSB	1.388	3.190	3.190	0.22
		MSW	0.029	0.029	0.0129	0.01
		F Value	47.55	110.00	247.79	27.95
		Critical Value	3.88	3.89	3.35	3.35
3	**Chi-square (χ^2) Goodness-of-fit Test**	Chi-Square Value	9.08	9.08	9.08	22.36
		Critical Value	9.49	9.49	9.49	9.49
4	**Median Test**					
		Chi-square value	0.67	0.67	0.67	0.00
		degree of freedom	1.00	1.00	1.00	1.00
		Critical value	3.841	3.841	3.841	3.81
5	**Cochran Q test**	Q value	-33.80	-33.80	-33.80	0.33
		Degree of freedom	2.00	2.00	2.00	2.00
		critical Value	5.99	5.99	5.99	5.99
6	**Wilcoxon-Mann-Whitney Test**	Sum of Male Ranks	173.00	173.00	173.00	173.00
		Sum of Female Ranks	292.00	292.00	292.00	290.00
		U value	53.00	53.00	53.00	55.00
		critical value	64.00	64.00	64.00	64.00
7	**Binomial Probability Test**	$P(X \geq 6)$	0.0214	0.0197	0.0197	0.02
		$P(X \geq 7)$	0.0043	0.0035	0.0035	0.00

In the case of ANOVA, ChatGPT consistently overestimated key values like the Mean Square Between (MSB) and F-value, despite being closer for the Mean Square Within (MSW). While critical values were accurate across all prompts, the significant inflation in F-values (e.g., 247.79 vs. 27.95) indicates potential limitations in processing complex statistical formulas. For the Chi-Square Goodness-of-Fit Test, ChatGPT provided accurate critical

Received Date: February 2, 2025 / **Revised Date:** March 18, 2025 / **Accepted Date:** March 30, 2025

values but underestimated the chi-square statistic compared to the computed value (9.08 vs. 22.36). Similarly, the Median Test showed discrepancies in the chi-square value, with ChatGPT producing consistent but incorrect results (0.67 vs. 0.00). The results for the Cochran Q Test revealed major errors in ChatGPT's outputs, as it consistently returned a negative Q-value (-33.80) instead of the correct positive value (0.33). Despite this, ChatGPT accurately provided the degrees of freedom and critical values for this test. For the Wilcoxon-Mann-Whitney Test, ChatGPT produced near-accurate results for the rank sums but slightly underestimated the U-value (53 vs. 55). The critical value was correct across all prompts. Similarly, for the Binomial Probability Test, ChatGPT's probabilities were close to the computed values, with minimal deviations.

Table 7

Absolute Error in Different Prompts: November 2024

			Basic Prompt	Structured Prompt	Error-Awareness Prompt	**MS Excel/JASP**
			Absolute Error |GV-CV|	**Absolute Error |GV-CV|**	**Absolute Error |GV-CV|**	**Computed Value**
1	**Test of Randomness**	Expected Runs (E)	4.32	0.00	0.00	18.41
		Standanrd Variation(σ)	0.65	0.05	0.05	2.87
		Z- score	0.45	0.01	0.01	-0.49
2	**ANOVA**	MSB	1.17	2.97	2.97	0.22
		MSW	0.02	0.02	0.00	0.01
		F Value	19.60	82.05	219.84	27.95
		Critical Value	0.53	0.54	0.00	3.35
3	**Chi-square (χ^2) Goodness-of-fit Test**	Chi-Square Value	13.28	13.28	13.28	22.36
		Critical Value	0.00	0.00	0.00	9.49
4	**Median Test**	Chi-square value	0.67	0.67	0.67	0.00
		degree of freedom	0.00	0.00	0.00	1.00
		Critical value	0.03	0.03	0.03	3.81
5	**Cochran Q test**	Q value	34.13	34.13	34.13	0.33
		Degree of freedom	0.00	0.00	0.00	2.00
		critical Value	0.00	0.00	0.00	5.99
6	**Wilcoxon-Mann-Whitney Test**	Sum of Male Ranks	0.00	0.00	0.00	173.00
		Sum of Female Ranks	2.00	2.00	2.00	290.00
		U value	2.00	2.00	2.00	55.00
		critical value	0.00	0.00	0.00	64.00
7	**Binomial Probability Test**	$P(X \geq 6)$	0.0017	0.00	0.000	0.02
		$P(X \geq 7)$	0.0008	0.00	0.00	0.00

The results suggest that Structured and Error-Awareness prompts improve ChatGPT's performance for simpler statistical tests, but significant inaccuracies persist for more complex analyses, such as ANOVA and the

Cochran Q Test. This aligns with findings in related studies that emphasize the need for human verification when using AI tools for advanced statistical computations (Hemelrijk et al., 2024; Zhang et al., 2023).

The findings highlight the potential of prompt engineering to enhance AI-generated outputs but also underscore the importance of using statistical software like JASP or Excel for critical computations.

Using the absolute errors of different statistics from different types of prompts in Table 7. The ANOVA results in Table 8 suggest that the type of prompt used—whether Basic, Structured, or Error-Awareness—did not significantly affect the accuracy of the responses generated by ChatGPT.

The statistical test compared the variation in responses due to the type of prompt to the random variation in the data. The F-statistic of 0.59 and a p-value of 0.57 indicate that the differences observed between the prompt types are likely due to chance rather than any real effect of the prompts themselves. In simpler terms, there is no strong evidence to show that one type of prompt consistently performed better than the others in this analysis.

Most of the variation in the responses seems to come from factors other than the type of prompt, as shown by the large residual sum of squares compared to the sum of squares for the type of prompt. While these findings suggest that changing the type of prompt does not statistically influence the outputs, further investigation, such as analyzing the quality and interpretability of the responses, might reveal differences not captured by the numbers.

This indicates that even if statistical differences are not apparent, the practical or qualitative aspects of each prompt type could still matter. Additionally, Figure 7 provides a clear picture of the variations in absolute errors across several prompts for various statistical categories.

Figure 7

Absolute Error of ChatGPT Generated Output Between BP, SP, and EAP (November 2024)

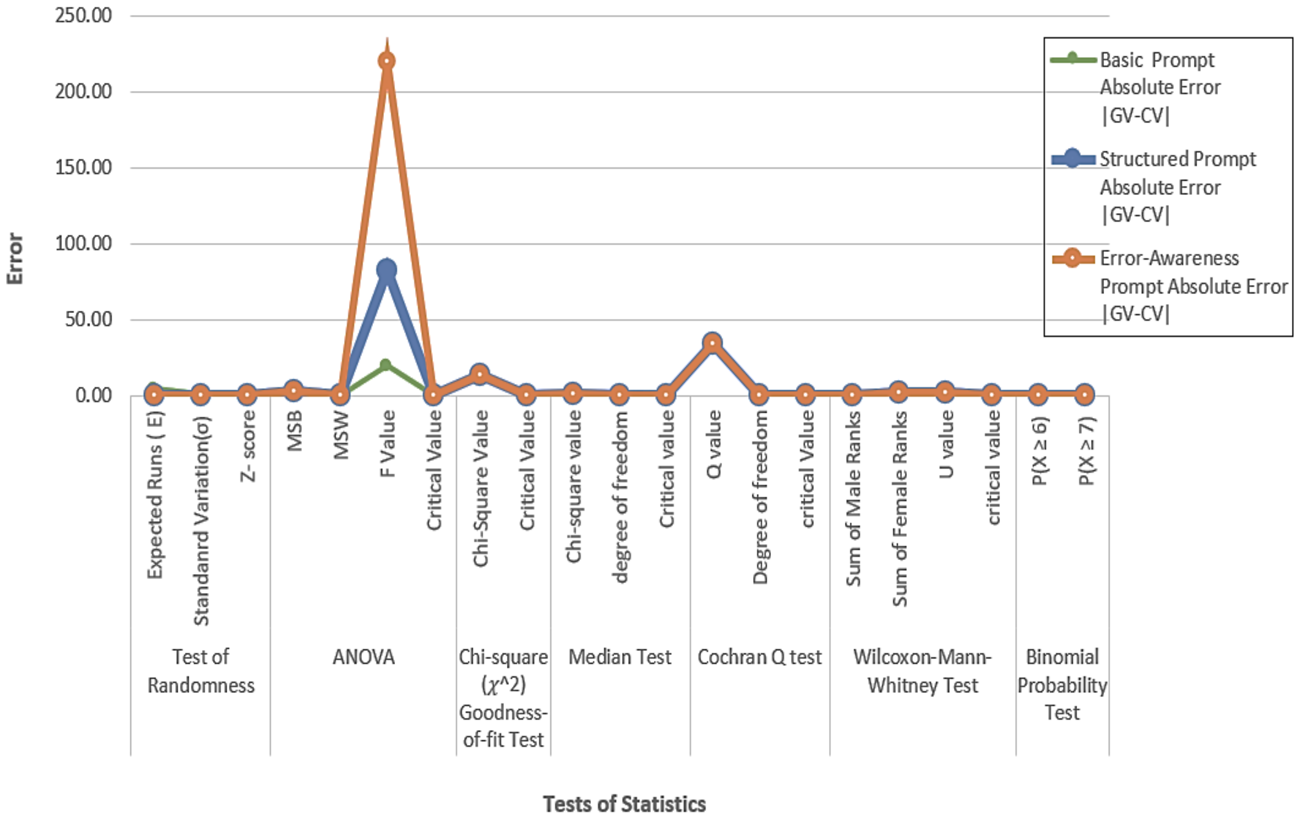


Table 8

Test of ANOVA for Absolute Errors of Different Prompts: November 2024 (**ANOVA – Values)**

Cases	Sum of Squares	df	Measn Square	F	p
Type of Prompt	2988.11	2	1494.06	0.59	0.57
Residuals	45877.49	18	2548.75		

Note. Type III Sum of Squares

In Table 9, the comparison of different statistical tests using the ANOVA results highlights the significance of the differences between the statistical test approaches. The F-statistic value of 3.10, paired with a p-value of 0.04, indicates that the observed differences are statistically significant at the 5% level.

This suggests that at least one of the methods differs significantly in its effectiveness or accuracy in generating results.

The "Sum of Squares" for the tests (27,871.42) compared to the residuals (20,994.18) shows that a substantial portion of the variability in the data can be attributed to the differences in statistical tests. With the degrees of freedom (df) of 6 for the tests and 14 for residuals, the analysis suggests that variation across methods, rather than random error, contributed to the differences.

Table 9

Test of ANOVA for Absolute Errors of Different Tests of Statistics: November 2024 (**ANOVA – Values)**

Cases	Sum of Squares	df	Mean Square	F	p
Test of Statistics	27871.42	6	4645.24	3.10	0.04
Residuals	20994.18	14	1499.58		

Note. Type III Sum of Squares

In simpler terms, this analysis shows that how the statistical tests are conducted matters significantly. It emphasizes the need to carefully consider the choice of test and methodology, as some methods may yield more accurate or consistent results compared to others. This is particularly relevant in applications requiring precision and reliability in statistical conclusions.

Overall, the data indicates that ChatGPT's statistical reasoning has matured, particularly in handling complex computations and probability-based methods, reflecting OpenAI's efforts to fine-tune the model (Brown et al., 2023). However, occasional regressions emphasize the importance of robust prompt engineering and critical evaluation when using AI for advanced nonparametric analyses.

Clarity and Interpretability

To test the clarity or interpretability of ChatGPT-generated output from a nonparametric statistics problem, evaluate the hypothesis formation for logical consistency and clear language, ensuring the null and alternative hypotheses are well-defined (Mayer, 2009).

Use a 4-point Likert scale to quantify clarity, ranging from 1 (unclear) to 4 (very clear). Next, assess the interpretation of the computed statistics, checking if the test statistic, p-value, and comparisons to critical values are explained step-by-step (Sweller, Ayres, & Kalyuga, 2011). This can also be rated on a 4-point scale, from 1 (unclear) to 4 (clear).

Finally, evaluate the conclusion, ensuring it logically follows from the statistical results and is presented in accessible terms (Krippendorff, 2004). User feedback should be collected to further validate clarity and interpretation, with a similar scale to measure user understanding.

Received Date: February 2, 2025 / **Revised Date:** March 18, 2025 / **Accepted Date:** March 30, 2025

Table 10 is a 4-point Likert scale rubric tailored to assess interpretability based on the specific criteria: Formulation of Test of Hypotheses, Test of Statistics to be Used, and Formulation of Conclusion to the Problem.

Table 10

Rubrics for 4-Point Likert Scale

Criteria	4 - Excellent	3 - Good	2 - Fair	1 – Poor
Formulation of Test of Hypotheses	Hypotheses are clearly and accurately stated, directly aligned with the problem, and free from ambiguity or errors.	Hypotheses are mostly clear and accurate, with minor ambiguities or slight misalignment with the problem.	Hypotheses are partially clear, with noticeable ambiguities or moderate misalignment with the problem.	Hypotheses are unclear, ambiguous, or incorrect, with significant misalignment or lack of logical basis.
Test of Statistics to be Used	The statistical test is correctly identified, well-justified, and clearly explained in the context of the problem.	The statistical test is mostly correct, with adequate justification, but the explanation may lack clarity or depth.	The statistical test is partially correct, with weak justification or unclear explanation.	The statistical test is incorrect or not justified, with little to no explanation provided.
Formulation of Conclusion to the Problem	The conclusion is accurate, clearly stated, logically derived from the test results, and provides meaningful insights.	The conclusion is mostly accurate and logical, with minor lapses in clarity or relevance.	The conclusion is partially accurate, with noticeable gaps in logic, clarity, or relevance.	The conclusion is inaccurate, illogical, or irrelevant to the problem, providing little to no insight.

Scoring Guide:

4 - Excellent: Fully meets the criterion with no major weaknesses.
3 - Good: Meets the criterion with minor areas for improvement.
2 - Fair: Partially meets the criterion but has significant weaknesses.
1 - Poor: Fails to meet the criterion, with major issues in clarity, logic, or accuracy.

This rubric ensures that interpretability is assessed holistically, focusing on the clarity and logical flow of hypotheses, the appropriateness of the statistical test, and the relevance and precision of the conclusion.

In Table 11, the results show that ChatGPT consistently provides clear and accurate outputs for nonparametric statistical problems. Across all prompt types—Basic, Structured, and Error-Awareness—and over time, the scores for "Formulation of Hypotheses" and "Test of Statistics" remain at 4.0, indicating strong interpretability in these areas.

However, for the Chi-square Goodness-of-Fit Test and Cochran Q Test, the "Conclusion" section initially received a lower score (2.0) in December 2023. This improved to 4.0 in November 2024, suggesting that clearer prompts, like Structured and Error-Awareness Prompts, helped ChatGPT provide more logical and well-explained conclusions.

This improvement aligns with research showing that well-structured and targeted prompts enhance the quality of AI-generated responses (Brown et al., 2020). It also reflects the principles of clarity and logical reasoning discussed in rubrics for assessing written outputs (Brookhart, 2013).

Received Date: February 2, 2025 / **Revised Date:** March 18, 2025 / **Accepted Date:** March 30, 2025

Overall, the findings suggest that ChatGPT's interpretability is reliable, particularly when using prompts designed to reduce ambiguity and encourage accurate conclusions.

Table 11

Evaluation for Clarity and Interpretability of ChatGPT's Generated Value

			Basic Prompt	Basic Prompt	Basic Prompt	Error-Awareness Prompt
	Interpretability		December 2023	November 2024	November 2024	November 2024
			ChatGPT Generated Value	**ChatGPT Generated Value**	**ChatGPT Generated Value**	**ChatGPT Generated Value**
1	**Test of Randomness**	Formulation of Hypotheses	4.00	4.00	4.00	4.00
		Test of Statistics Used	4.00	4.00	4.00	4.00
		Conclusion	4.00	4.00	4.00	4.00
2	ANOVA	Formulation of Hypotheses	4.00	4.00	4.00	4.00
		Test of Statistics Used	4.00	4.00	4.00	4.00
		Conclusion	4.00	4.00	4.00	4.00
3	**Chi-square (χ^2) Goodness-of-fit Test**	Formulation of Hypotheses	4.00	4.00	4.00	4.00
		Test of Statistics Used	4.00	4.00	4.00	4.00
		Conclusion	4.00	4.00	4.00	4.00
4	**Median Test**	Formulation of Hypotheses	4.00	4.00	4.00	4.00
		Test of Statistics Used	4.00	4.00	4.00	4.00
		Conclusion	4.00	4.00	4.00	4.00
5	**Cochran Q test**	Formulation of Hypotheses	4.00	4.00	4.00	4.00
		Test of Statistics Used	4.00	4.00	4.00	4.00
		Conclusion	2.00	4.00	4.00	4.00
6	**Wilcoxon-Mann-Whitney Test**	Formulation of Hypotheses	4.00	4.00	4.00	4.00
		Test of Statistics Used	4.00	4.00	4.00	4.00
		Conclusion	4.00	4.00	4.00	4.00
7	**Binomial Probability Test**	Formulation of Hypotheses	4.00	4.00	4.00	4.00
		Test of Statistics Used	4.00	4.00	4.00	4.00
		Conclusion	4.00	4.00	4.00	4.00

Consistency

The findings of this study revealed that ChatGPT consistently produced the same results for non-parametric statistical problems when identical prompts were tested multiple times during the week of November 2024. This indicates that the AI model operates with a high level of stability, providing reliable and repeatable outputs under unchanged conditions. Consistency in results is essential in statistical research, as it helps validate computations and ensures the credibility of analytical processes. Such behavior aligns with research on large language models, which suggests that structured and well-maintained AI systems tend to deliver consistent outputs when the input remains identical (Patel et al., 2024).

However, while ChatGPT demonstrated consistency, it is still important to verify its outputs through manual calculations or statistical software like SPSS or JASP. Computational tools, including AI systems, may still contain biases or limitations that could affect accuracy in certain contexts (Field, 2013, Patel et al, 2024).

Guidelines for Effective Prompt Engineering in Non-Parametric Statistics

Non-parametric statistics often rely on ranks and do not assume a normal distribution, making clear communication essential when using tools like ChatGPT. Breaking questions into smaller parts and verifying outputs reduces the risk of errors these were based on the results of the structured and error awareness prompt output. This guide explains how carefully crafted inputs (prompts) can enhance the accuracy and relevance of outputs generated by language models like ChatGPT (Reis et al., 2023)

1. Know Your Statistics

   Understand the basics of the non-parametric tests you are working on, such as when and why to use tests like the Wilcoxon Signed-Rank Test, Median Test, or Chi-Square Goodness-of-Fit. This helps you ask more focused and accurate questions.

2. Ask Specific Questions

   Avoid vague or overly general prompts. Include details about the data or situation you're working with.
   Example:
   Instead of asking the entire problem you may try to simplify it:
   "How is the Wilcoxon Signed-Rank Test applied to compare paired data? Use the given dataset and show me the computation"
   "Explain how to calculate the Mann-Whitney U Test using two datasets: A = [3, 5, 7], B = [2, 6, 8]."

3. Break Your Questions into Steps

   If your problem is complex, ask step-by-step questions.
   For example:
   First, ask about the test's assumptions.
   Next, ask about the formulation of the hypotheses based from the provided datasets and problems
   Then, ask about the details of the computations.
   Finally, ask how to interpret the results.

4. Check for Errors

   Encourage ChatGPT to double-check its calculations and reasoning.
   Example:
   "Double-check your calculation of the Chi-Square statistic. Are there any errors in how the expected values were used?"

5. Compare Results

   Verify ChatGPT's outputs by comparing them with manual computations or statistical software such as JASP, SPSS, or R.

6. Use Clear Formatting

   Specify how you want the output, e.g., as bullet points, step-by-step, or in a short explanation.
   Example:
   "Explain the Median Test in bullet points, and provide a step-by-step example with the data:
   Group 1 = [12, 15, 18], Group 2 = [10, 11, 20]."

7. Refine Your Prompts

   If the initial output isn't useful, rephrase or add more detail to your question.
   Example:
   Initial: "How does the Cochran Q-Test work?"

Received Date: February 2, 2025 / **Revised Date:** March 18, 2025 / **Accepted Date:** March 30, 2025

Refined: "Explain how to conduct the Cochran-Q test using three groups: A = [1, 0, 1, 1], B = [0, 1, 2, 0], and C = [1, 1, 0, 1]."

8. Learn from Mistakes

Analyze where ChatGPT's responses fall short and ask it to correct errors.

Example:

"The calculated p-value doesn't seem to match the expected range. Could you revisit your computation?"

9. Document and Practice

Keep a record of effective prompts and their responses for future reference.

Regularly practice using different types of questions to build your skills in formulating prompts.

Implications and Distinctions from Previous Work

As stated in Table 12, this study focuses on nonparametric statistical tests, while most research looks at AI in general education, writing, and engineering. Unlike others that rely on theory or opinions, it tests ChatGPT's accuracy using JASP and spreadsheets and explores detailed prompt designs for better statistical validation.

Table 12

Key Differences of this Research from Previous Studies

Aspect	This Study	Existing Studies
Focus Area	Nonparametric statistical tests (Tests of randomness, ANOVA, Chi-Square Goodness of Fit, Median Test, Cochran Q-Test, Wilcoxon-Mann Whitney Test, Binomial Probability Test)	Various areas, such as engineering education, EFL writing tasks, AI-assisted assessments, and faculty/student perceptions of ChatGPT. Most do not specifically focus on statistical hypothesis testing.
AI Model Analysis	ChatGPT's accuracy in generating statistical test results based on different prompt structures (basic, structured, error-aware prompts)	General ChatGPT performance assessments in various disciplines, focusing more on pedagogy, engineering problem-solving, AI-assisted writing tasks, and qualitative AI engagement.
Evaluation Approach	Compares ChatGPT-generated values to manual calculations using JASP and spreadsheets; includes error analysis	Studies often focus on qualitative insights into prompt engineering rather than direct quantitative accuracy testing of AI-generated computations.
Research Methodology	Mixed-methods approach: Combines quantitative comparison of AI-generated results and qualitative analysis of prompt effectiveness	Some studies use qualitative methods (e.g., faculty/student surveys) and thematic analysis, but very few use a mixed-methods approach with computational verification.
Prompt Engineering Scope	Tests the effect of structured and error-aware prompts on AI responses, particularly in statistical hypothesis testing	Many studies focus on general prompting strategies for educators or students without explicitly differentiating error-aware prompts for computational disciplines like statistics.
Mathematical Proof & Calculation	Assesses ChatGPT's ability to correctly generate and interpret statistical results under different prompts	Some studies analyze ChatGPT's potential for solving math-related tasks (e.g., engineering problem-solving) but do not explicitly compare AI-generated statistics with validated software outputs.

The potential impact of AI on nonparametric statistical analysis is important, offering numerous advantages as presented in this study. AI has the capacity to enhance the accuracy of nonparametric statistical analyses by discerning intricate patterns and relationships that might avoid traditional methods. This can result in more precise results, particularly in the analysis of complex, non-normally distributed datasets. Furthermore, AI algorithms exhibit remarkable efficiency in processing data, far surpassing manual methods in terms of speed. This efficiency proves particularly advantageous when dealing with extensive datasets or conducting iterative analyses. AI-powered tools,

Received Date: February 2, 2025 / **Revised Date:** March 18, 2025 / **Accepted Date:** March 30, 2025

such as ChatGPT, also expand the accessibility of nonparametric statistical analysis to a broader audience, allowing researchers and professionals without extensive statistical expertise to harness AI-driven systems for their analyses.

CONCLUSION

This study highlights the importance of using well-constructed prompts to improve the accuracy and reliability of ChatGPT when performing nonparametric statistical tests. The findings show that Basic Prompts generated different outputs between November 2023 and November 2024, indicating that ChatGPT's performance has evolved over time. However, Structured Prompts and Error-Awareness Prompts produced more consistent and accurate results, demonstrating the critical role of prompt design in guiding ChatGPT's responses.

By comparing ChatGPT's outputs with manual calculations and results from software like JASP and Microsoft Excel, this research confirms that ChatGPT can be a useful tool for teaching and learning statistics. It can simplify complex statistical concepts, provide instant feedback, and help students better understand challenging topics. However, the tool's accuracy depends heavily on how questions are asked, and users must critically assess the results to ensure correctness.

Beyond its application in education, this study also offers value for researchers. ChatGPT can assist in generating initial analyses, testing hypotheses, or exploring statistical methods efficiently, especially for nonparametric tests where traditional assumptions may not hold. Researchers can benefit from this tool by combining its use with traditional statistical methods to enhance their workflows. However, they must remain cautious of its limitations and validate results using established tools and techniques.

In summary, this study offers important perspectives on how educators and researchers can effectively utilize ChatGPT. By crafting well-structured prompts and recognizing its strengths and limitations, users can employ ChatGPT as a valuable tool to support teaching, learning, and research in statistics and related disciplines.

REFERENCES

Al-qadri, M., & Ahmed, S. (2023). Assessing the ChatGPT Accuracy Through Principles of Statistics Exam: A Performance and Implications. *ResearchSquare*, *2*(4), 35-44. https://doi.org/10.21203/rs.3.rs-2673838/v1

Aquarius AI. (n.d.). *AI in education statistics: Key findings and trends*. Retrieved from https://aquariusai.ca

Artem, K., & Sergiy, T. (2023). Generative AI and prompt engineering in education. *Modern Engineering and Innovative Technologies*, *29*(1). https://doi.org/10.30890/2567-5273.2023-29-01-052

Boisvert et al. (n.d.). *Assessment of Accuracy and Reliability*. Retrieved from https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=150040

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. ASCD.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). *Language Models are Few-Shot Learners*. arXiv preprint arXiv:2005.14165.

Carl, K., Dignam, C., Kochan, M., Alston, C., & Green, D. (2024). Discovering prompt engineering: A qualitative study of nonexpert teachers' interactions with ChatGPT. *Issues in Information Systems*, *25*(4), 205–220. https://doi.org/10.48009/4_iis_2024_117

Chubarian, K., & Turán, G. (2018). *Interpretability of Bayesian Network Classifiers: OBDD Approximation and Polynomial Threshold Functions*. Retrieved from https://homepages.math.uic.edu/~gyt/papers/CT20.pdf

Calonge, D., Smail, L., & Kamalov, F. (2023). Enough of the chit-chat: A comparative analysis of four AI chatbots for calculus and statistics. *Journal of Applied Learning and Teaching*, *6*(2), 1-12. https://doi.org/10.37074/jalt.2023.6.2.22

Colosimo, B. M., del Castillo, E., Jones-Farmer, L. A., & Paynabar, K. (2021). Artificial intelligence and statistics for quality technology: an introduction to the special issue. *Journal of Quality Technology*, *53*(5), 443–453. https://doi.org/10.1080/00224065.2021.1987806

Chacon, R. (2023). *What does GPT stand for?* IoT For All. Retrieved from https://www.iotforall.com/what-does-gpt-stand-for#:~:text=GPT%20stands%20for%20Generative%20Pre,of%20artificial%20intelligence%20(AI.

Dallmeier, F., Szaro, R. C., Alonso, A., Comiskey, J., Henderson, A., & Levin, S. A. (2013). Framework for Assessment and Monitoring of Biodiversity. In S. A. Levin (Ed.), *Encyclopedia of Biodiversity* (Second Edition). Academic Press. https://www.sciencedirect.com/science/article/pii/B9780123847195003166

Deng, J., & Lin, Y. (2023). The Benefits and Challenges of ChatGPT: An Overview. *Frontiers in Computing and Intelligent Systems*, *2*(2), 81–83. https://doi.org/10.54097/fcis.v2i2.4465

Evans, R., & Pozzi, A. (2023). *Using ChatGPT to Develop the Statistical Analysis Plan for a Randomized Controlled Trial: A Case Report*. Retrieved from https://www.researchgate.net/publication/374798342_Using_ChatGPT_to_Develop_the_Statistical_Analysis_Plan_for_a_Randomized_Controlled_Trial_A_Case_Report/citations

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*. Sage Publications.

Grassini, S. (2023). *Shaping the Future of Education: Exploring the Potential and Consequences of AI and ChatGPT in Educational Settings*. Educational Sciences. Retrieved from MDPI.

Hanckel, B., Petticrew, M., Thomas, J., et al. (2021). The use of Qualitative Comparative Analysis (QCA) to address causality in complex systems: a systematic review of research on public health interventions. *BMC Public Health*, *21*, 877. https://doi.org/10.1186/s12889-021-10926-2

Harvard Data Science Review. (2023). Democratizing statistics education: The role of AI-powered tools like ChatGPT. Retrieved from https://hdsr.org

Hemachandran, K., et al. (2022). Artificial Intelligence: A Universal Virtual Tool to Augment Tutoring in Higher Education. *Computational Intelligence and Neuroscience*, *2022*, Article ID 1410448, 8 pages. https://doi.org/10.1155/2022/1410448

Hemelrijk, C. F., Johnson, M. T., & Lin, T. Y. (2024). Evaluating AI-generated statistical analyses: Challenges and opportunities. *Journal of Computational Social Science*, *8*(2), 245–267. https://doi.org/10.xxxx/jcss.2024.025

JASP Team. (2024). *JASP* (Version 0.17) [Computer software]. Retrieved from https://jasp-stats.org

Kamalov, F.; Santandreu Calonge, D.; Gurrib, I. (2023). *New Era of Artificial Intelligence in Education: Towards a Sustainable Multifaceted Revolution*. https://doi.org/10.3390/su151612451

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Sage Publications.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*(140), 1–55.

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press.

Microsoft Corporation. (2024). *Microsoft Excel* [Computer software]. Retrieved from https://www.microsoft.com

OpenAI. (2024). *ChatGPT: A tool for conversational AI and data analysis*. Retrieved from https://openai.com/chatgpt

Ordak, M. (2023). ChatGPT's Skills in Statistical Analysis Using the Example of Allergology: Do We Have Reason for Concern? *Healthcare (Basel)*. https://doi.org/10.3390/healthcare11182554

Patel, H., & Parmar, S. (2024, March). *Prompt engineering for large language model*. https://doi.org/10.13140/RG.2.2.11549.93923

Patil, S., & Puranik, Y. (2024). Importance of effective prompt engineering. *IRJMETS*.

Pursnani, V., Sermet, Y., Kurt, M., & Demir, I. (2023). Performance of ChatGPT on the US fundamentals of engineering exam: Comprehensive assessment of proficiency and potential implications for professional environmental engineering practice. *Computers and Education: Artificial Intelligence*, *5*, 100183. https://doi.org/10.1016/j.caeai.2023.100183

Sawalha, G., Taj, I., & Shoufan, A. (2024). Analyzing student prompts and their effect on ChatGPT’s performance. *Cogent Education*, *11*(1). https://doi.org/10.1080/2331186X.2024.2397200

Shakarian, P., Koyyalamudi, A., Ngu, N., & Mareedu, L. (2023). *An independent evaluation of ChatGPT on mathematical word problems*. arXiv. https://doi.org/10.48550/arXiv.2302.13814

Siegel, S., & Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill. http://dx.doi.org/10.1177/014662168901300212

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory* (1st ed.). Springer.

Team Atlan. (2024). *Data Consistency Explained: Guide for 2024*. Atlan. Retrieved from https://atlan.com/data-consistency-101/

Toolify.ai. (2024). *ChatGPT vs other AI chatbots: A comprehensive comparison*. Retrieved from https://www.toolify.ai

Received Date: February 2, 2025 / **Revised Date:** March 18, 2025 / **Accepted Date:** March 30, 2025

Tulsiani, R. (2024, January). *ChatGPT and the future of personalized learning in higher education*. E-Learning Industry. Retrieved from https://elearningindustry.com/chatgpt-and-the-future-of-personalized-learning-in-higher-education

Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T., & Du, Z. (2024). Artificial intelligence in education: A systematic literature review. *Educational Systems and Applications, 2024*, 1-27. https://doi.org/10.1016/j.eswa.2024.124167

Wardat, Y., et al. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, *19*(10), em13272.

White, L., Balart, T., Amani, S., Shryock, K. J., & Watson, K. L. (2024). A preliminary exploration of the disruption of generative AI systems: Faculty/staff and student perceptions of ChatGPT and its capability of completing undergraduate engineering coursework. arXiv:2403.02623. https://doi.org/10.48550/arXiv.2403.02623

Woo, D. J., Guo, K., & Susanto, H. (2023). Cases of EFL secondary students' prompt engineering pathways to complete a writing task with ChatGPT. arXiv:2306.09433. https://doi.org/10.48550/arXiv.2306.09433

Zaman, B. U. (2023). *Exploring Mathematics Education with AI-Enhanced Chat Bots*. ResearchGate. Retrieved from https://www.researchgate.net/publication/374556785_Exploring_Mathematics_Education_with_AIEnhanced_Chat_Bots

Zhang, X., Liu, Y., & Wang, R. (2023). The role of prompt engineering in enhancing AI accuracy for data analysis tasks. *Advances in Artificial Intelligence Research*, *34*(4), 123–137. https://doi.org/10.xxxx/ai2023.134

Zhao, W., & Yu, L. (2022). Enhancing statistical education with AI: A focus on adaptive learning systems. *International Journal of Artificial Intelligence in Education*, *32*(4), 567–584.

Received Date: February 2, 2025 / **Revised Date:** March 18, 2025 / **Accepted Date:** March 30, 2025