*Athens Journal of Sciences - Volume 13, Issue 3, September 2026 – Pages 171-188*

# Assessing Accuracy and Reliability of AI-Generated Statistical Interpretations

*By Dulanjalee Devage Dona**

*Artificial intelligence (AI) tools are rapidly used in academic research and industry to generate content and assist with analytical tasks. Because of the recent advances in generative AI, these tools can help with statistical data analysis, including data preprocessing, exploratory analysis, statistical modeling, and interpretation of results. However, concerns remain regarding the reliability and accuracy of AI-generated statistical outcomes and interpretations. This study evaluates the effectiveness and reliability of AI-assisted statistical analysis by comparing the performance of ChatGPT, Claude AI, and Google Gemini using a structured evaluation framework. The analysis was conducted using two datasets from the education and healthcare domains, which are among the most prominent data-driven fields. The study examines how effectively these AI tools select appropriate statistical methods, perform the analysis, and interpret the statistical outputs across multiple dimensions, including statistical accuracy, interpretive quality, domain appropriateness, limitation awareness, and consistency. The study concludes that AI can effectively support statistical analysis but should not replace careful human evaluation. Reliable AI-assisted interpretations still depend on well-structured prompts, critical review, and human oversight. These findings provide practical guidance for researchers, educators, and data practitioners and emphasize the importance of using AI responsibly in both statistics' education and real-world data analysis.*

**Keywords:** *artificial intelligence, generative AI, large language models (LLMs), statistical analysis, data interpretations*

## Introduction

Artificial intelligence (AI) is a technology that enables computers and machines to replicate human abilities of learning, understanding, problem-solving, decision-making, creativity, and autonomy (Kavlakoglu et al. 2024). The rapid development of technology, along with major global crises such as the COVID-19 pandemic, has transformed both education and research. As a result, there is now a stronger need to incorporate artificial intelligence into these systems (BenMessaoud et al. 2025). Therefore, these AI tools are rapidly used in academic research, business, healthcare, and education. With the recent improvement of generative AI and large language models (LLMs), these tools can generate text, images, and videos and assist with analytical tasks such as data processing and interpretation. In statistical data analysis, AI tools such as ChatGPT, Claude AI, and Google Gemini are becoming more frequently used to assist researchers and data practitioners in handling data-related tasks that traditionally require advanced statistical expertise.

*Assistant Professor, University of Wisconsin, USA.

https://doi.org/10.30958/ajs.13-3-1 doi=10.30958/ajs.13-3-1

The developments and growing availability of these AI systems offer many advantages, including efficiency, faster interpretation of results, and statistical support for non-experts. Moreover, current AI systems can support data preprocessing, descriptive analysis, correlation analysis, hypothesis testing, statistical modeling, and interpretation of statistical results in natural language (Demir & Yavuz 2026). A key feature of these tools is how users interact with them. With models such as ChatGPT, this is accomplished through prompts in which humans provide natural language instructions to direct AI toward a particular activity, such as learning and analysis (Slobodsky et al. 2026). As a result, these technologies are becoming integrated into research, development, and data-driven decision-making processes.

Despite these advantages, it is uncertain that AI-generated statistical explanations will be precise. While AI models are effective at processing and organizing large amounts of data, they tend to be stronger at identifying patterns and handling data than at truly understanding what those patterns mean or what is happening within the data (Fehlmann & Kranich 2025). Although AI models can produce fluent and convincing explanations, their outputs include inaccurate numerical summaries, inappropriate statistical methods, unsupported conclusions, and misleading interpretations (Nagarkar et al. 2026, Peters & Chin-Yee 2025).

As students, academics, and analysts use AI more often, it is important to assess how well these tools interpret statistical results in comparison to human analysis. Most existing studies on AI-assisted statistical analysis have focused mainly on the technical abilities of language models, while giving less attention to the consistency and reliability and overall quality of their interpretations when compared with accepted analytical standards (Keller et al. 2026). Therefore, this study examines the effectiveness and reliability of AI-assisted statistical analysis through a comparative analysis of ChatGPT, Claude AI, and Google Gemini. The study evaluates how well AI tools can select appropriate statistical methods and conduct analyses, as well as how well they interpret the outcomes. The evaluation framework is designed based on multiple dimensions, including statistical accuracy, interpretive ability, domain appropriateness, limitation awareness, and consistency. Two datasets were used for the study to represent education and healthcare, where statistical methods are frequently applied. By applying a structured evaluation framework across datasets from more prominent domains, this study aims to provide a systemic assessment of the strengths and limitations of AI-assisted statistical analysis.

The remainder of this paper is organized as follows. Section 2 reviews existing literature on the use of large language models in statistical analysis. Section 3 describes the research methodology and the evaluation framework applied in the study. Section 4 presents and discusses the findings of the comparative analysis, while Section 5 concludes the study by summarizing the key findings, outlining the study's limitations, and suggesting directions for future work.

## Literature Review

Research interest in assessing the accuracy, reliability, and quality of AI-generated outputs has become increasingly important due to the rapid growth of artificial intelligence in data-driven fields. The literature review focuses on the

foundations and capabilities of large language models, the reliability of AI-generated outputs, AI in statistical analysis, the applications of AI in healthcare and educational data analysis and the methodological framework used to assess AI performance. These topics provide the theoretical and research foundation for the present study.

*Foundations and Capabilities of Large Language Models*

The emergence of large language models is one of the most important developments in the field of artificial intelligence. Fundamentally, LLMs are computer systems that have been trained on enormous text data to understand, generate, and work with human language across a wide range of tasks. Stryker (2021) describes LLMs as deep learning models that can recognize, summarize, translate, predict, and generate text using large datasets.

The foundations of modern LLMs come from several key contributions. Bommasani et al. (2021) introduced the idea of "foundation models," which are large models trained on broad datasets and then adapted for different tasks. Xiao and Zhu (2025) also provide a comprehensive foundation on how modern LLMs are constructed using pre-training, generative modeling, prompting, alignment, and inference. In the pre-training stage, models learn from large-scale text sources by predicting patterns in language without being given specific tasks. After this stage, they generate responses step by step based on learned probability patterns. This makes them fundamentally different from traditional software like SPSS, R or Python which produce exact numerical results through deterministic calculations (Bender et al. 2021, Frieder et al. 2023).

The models such as ChatGPT GPT-5, Claude Sonnet 4.6 and Gemini 3 Flash used in this study are all foundation models trained on broad datasets rather than built for specific statistical tasks (Bommasani et al. 2021). While this gives them strong general capabilities, it also introduces certain limitations. These include hallucinated information, numerical approximation errors, and weak domain-specific reasoning, all of which are evaluated in a structured way in this study.

*Reliability and Evaluation of AI-Generated Outputs*

Reliability is the center of this study and has been widely discussed in most recent research on AI evaluation. Chang et al. (2024) highlight that large language models should not be evaluated using a single metric. Instead, they must be assessed across multiple dimensions such as accuracy, reasoning ability, robustness, and domain-specific performance. This supports the use of a multidimensional evaluation framework in this study rather than relying on overall accuracy alone.

A major issue affecting reliability is hallucination, where models generate information that is incorrect or not supported by the data. Ji et al. (2023) categorize hallucinations as intrinsic errors (contradicting source information) and extrinsic errors (introducing unsupported content). In a statistical context, this can appear as numerical statistics or results that do not exist in the dataset. Alongside this, studies by Frieder et al. (2023) show that LLMs often struggle with multi-step statistical reasoning and may produce plausible but incorrect numerical outputs. Similarly,

Nagarkar et al. (2026) report variability in model performance across different types of statistical tasks.

The significance of multidimensional assessment is further emphasized by broader evaluation frameworks. Liang et al. (2022) introduce the HELM framework that evaluates models across accuracy, robustness, fairness, and efficiency and support the idea that no single metric adequately reflects the reliability. Guo et al. (2023) and Hong et al. (2023) also emphasize the importance of consistency, reliability, and output stability when assessing AI systems.

*AI in Statistical Analysis*

Research on the use of large language models in statistical analysis shows both capabilities and limitations. Demir and Yavuz (2026) found that different AI models vary significantly in how well they identify statistical relationships, choose appropriate measures, and interpret results. These differences highlight the importance of comparing models, which is also the focus of this study.

In addition, Peters and Chin-Yee (2025) found that LLMs often overgeneralize findings and make claims that go beyond what data support. This is especially relevant to overclaiming errors in statistical reporting. Pearl and Mackenzie (2018) emphasize the importance of distinguishing between correlation and causation. Their work is directly relevant to determining whether AI models appropriately characterize the directionality and nature of statistical relationships.

*Applications for AI in Healthcare and Educational Data Analysis*

AI has been widely studied in healthcare, especially in relation to how reliably it can interpret clinical and statistical data. Thirunavukarasu et al. (2023) show that while large language models can identify patterns and correlations in medical data, they often struggle to place these findings within proper clinical context or significance frameworks. This gap between statistical output and clinical meaning is an important concern and is directly addressed in this study through the domain appropriateness in the evaluation framework.

In education, research shows similar opportunities and challenges in using AI for data interpretation. Kasneci et al. (2023) note that while tools like ChatGPT can support educational analysis, they often struggle to translate statistical results into insights that are meaningful in a learning or teaching context. Moreover, Holmes et al. (2022) further emphasize that AI systems used in education must meet high standards of accuracy, fairness, and responsible interpretation. They highlight that errors in AI-generated insights can directly affect educational decisions and outcomes. These concerns align closely with the domain appropriateness and limitation of awareness used in the evaluation framework of this study.

*Evaluation Framework for AI Outputs*

Research on AI evaluation methods provides the foundation for the framework used in this study. Liu et al. (2023) introduced G-EVAL, which is a rubric-based approach that assesses AI-generated text using multiple criteria instead of a single

score. Demir and Yavuz (2026) also developed a rubric-based multidimensional measurement framework to evaluate the performance of generative AI models. Their findings show that structured and multidimensional evaluation methods align more closely with human judgment, which supports the approach used in this study.

The theoretical basis for this framework also comes from psychometric measurement and rating scale design. Messick (1989) emphasizes that validity is not just about whether a tool measures something, but also whether the results are meaningful and appropriate to interpret. This idea directly supports the evaluation dimensions used in this study, such as interpretive quality and domain appropriateness. In addition, Likert (1932) established the use of rating scales for measuring qualitative judgments. This is later applied by Demir and Yavuz (2026) in their evaluation framework.

### *Identified Gaps in Literature*

The review of previous studies reveals several important gaps in the existing literature on AI-assisted statistical analysis. Although current research has proposed general frameworks for evaluating large language models, limited attention has been given to assess the reliability of AI-generated statistical interpretations across specific domains such as healthcare and education within the same study. Most existing studies also depend on automated scoring methods or public ratings, with limited use of direct supervision of a human expert.

Another important gap is the lack of detailed classification of errors made by AI systems during statistical interpretation. Although earlier research acknowledges that AI models can make mistakes in reasoning and interpretation, these errors are rarely organized into clear and practical categories for deeper analysis. Furthermore, very few studies compare multiple AI models under the same conditions using identical datasets and prompts. Because of this, it can be difficult to determine whether differences in performance are caused by the models themselves or by differences in the evaluation setting. This study addresses these gaps by applying a structured and multidimensional evaluation framework to compare AI-generated statistical outputs and interpretations across models and domains.

## Methodology

The study employs a comparative analysis design, where the same data sets are independently analyzed by three different AI tools and then outputs are compared using predefined evaluation criteria. Comparative analysis is carried out in several stages, including data selection, AI analysis, and evaluation.

### *Datasets Selection*

For comparative analysis, two data sets were selected from the Kaggle website to represent the fields of education and medicine, where statistical methods are frequently applied. Each data set contained more than two quantitative variables, at

least one categorical variable, and more than 50 observations, making them appropriate for comprehensive statistical analysis.

The first dataset is the Kaggle *Student Academic Performance Dataset*, which includes variables related to students' academic performance factors such as age, gender, study hours per day, attendance percentage, parental education, internet access, and participation in extracurricular activities collected from 5,000 students. The academic performance is measured by subject-wise scores in mathematics, science, and English along with a final percentage, performance level, and pass/fail status (Sonawane n.d.).

The second dataset is a health-related dataset obtained from the Kaggle website titled *UCI Heart Disease Data.* The dataset contains 14 attributes describing patients' characteristics and various health indicators associated with heart disease (Redwan n.d.).

*AI Analysis*

We used three commonly used AI tools for comparative analysis. These three AI tools are built upon large language models (LLMs). Unlike traditional statistical software such as SPSS, R, or Python, which rely on mathematically exact computations, LLMs produce responses by predicting patterns in language based on probabilities learned during pre-training (Bender et al. 2021, Frieder et al. 2023). For the purpose of this study, we selected three LLM-based AI tools that represent different design philosophies and capabilities (Table 1).

**Table 1.** *Description of LLM-based AI tools*

| AI tool | Model Version | Developer | Design Orientation |
|---|---|---|---|
| Claude | Sonnet 4.6 | Anthropic | General-purpose reasoning language LLM |
| ChatGPT | GPT-5 Data Analyst | Open AI | LLM with data analysis specialization |
| Gemini | 3 Flash | Google DeepMind | Light weight fast-inference LLM |

*Source:* Anthropic (2025), Google DeepMind (2025), OpenAI (2025).

We selected these models considering their design orientations. Together, they make it possible to compare systems that are built with very different goals. Then we submitted the datasets in their raw, unprocessed CSV format with no pre-cleaning, variable explanations, or contextual hints provided along with them. This was intentional, since we need to assess each model's independent analytical capability without any advantage of guided context. We applied the same prompt across all three AI models and both data sets with no variation:

> "*As a statistician, analyze this data set and provide a complete analysis report, including descriptive analysis, correlation, significance tests, and statistical interpretation for each of the outputs.*"

Using the same prompt for every model removes prompt engineering as a confounding factor. This avoids differences in output caused by the wording of the

instructions. Moreover, the expert framing "*as a statistician*" provides enough professional context for the model to respond using its natural analytical style, depth, and behavior. It also reflects a realistic scenario, since a non-expert user would usually upload a dataset and request a standard analysis. As a result, findings have become more practical and relevant.

*Evaluation Framework*

The evaluation framework is the central analytical component of this study. We used a clear and structured way to compare the statistical interpretations produced by three AI models, Claude Sonnet 4.6. ChatGPT GPT-5 Data Analyst and Gemini 3 Flash, created by human expert analysts. Rather than relying on subjective impressions, we used a framework that uses measurable, criteria-based assessment to evaluate each interpretation across several dimensions of quality and reliability.

The evaluation framework is built on three theoretical foundations that support a comprehensive and academic assessment process.

I. Psychometric Evaluation Theory
The framework draws on psychometric theories of validity introduced by Samuel Messick (1989). From this perspective, the statistical interpretations are evaluated not only for true correctness but also for how meaningful, complete, and appropriate the conclusions are in relation to the data.

II. AI Output Evaluation Literature
This framework is also informed by recent research on large language model evaluation conducted by Chang et al. (2024) and Guo et al. (2023). They argued that AI-generated outputs should be assessed across multiple dimensions rather than accuracy alone. As a result, this framework includes criteria such as faithfulness to the data, relevance, coherence, and the presence or absence of hallucinated information, which is now widely recognized as an essential component of LLM evaluation.

III. Statistical Reporting Standards
To ensure the interpretations are assessed against accepted professional standards, the framework incorporates established statistical guidelines. These include the Publication Manual of the American Psychological Association, as well as the STROBE standards used in healthcare research. It also references the standards developed by the American Educational Research Association for educational data interpretation.

*Evaluation Dimensions*

We established the evaluation framework to assess interpretations generated by AI models using five core evaluation dimensions. Each dimension is broken down into clear, measurable sub-criteria to apply the assessment consistently across all cases. These dimensions are based on the above theoretical foundations.

Dimension 1: Statistical Accuracy

Statistical accuracy is the most important dimension of the evaluation framework. It focuses on whether the reported numerical summaries are correct and whether the appropriate analytical methods were used to reach those results. This dimension is treated as the baseline for any interpretation since accurate statistics are essential for creditability (Messick 1989, Ji et al. 2026).

This dimension is especially important when assessing outputs from large language models. These models do not actually compute values in the way that statistical software does. Instead, they generate responses based on patterns in data, which can sometimes lead to numerical inaccuracies. Therefore, every output that AI tools generate needs to be carefully checked and verified (Frieder et al. 2023, Bender et al. 2021).

**Table 2.** *Sub-criteria for Dimension 1*

| Sub-Criterion | Description |
|---|---|
| Numerical Correctness | Are reported numerical values accurate and consistent with the dataset? |
| Appropriate Method Selected | Was the correct statistical method applied given the data type and research question? |
| Distributional Accuracy | Are distributional properties (skewness, kurtosis, normality) correctly identified? |
| Hallucination Detection | Are any statistics cited that do not exist in or cannot be derived from the dataset? |

*Source:* Frieder et al. (2023), Field (2018), Ji et al. (2023).

Dimension 2: Interpretive Quality

Interpretive quality assesses the clarity, depth, and logic of the narrative that accompanies the statistics. A result can be technically accurate but still weak if the explanation is confusing, superficial, or goes beyond what the data actually supports. This is also an important part of strong scientific communication (American Psychological Association 2020).

Interpretive quality is particularly important when evaluating LLM outputs. Studies have shown that LLMs may produce numerically accurate results with exaggerated conclusions and interpretations of the underlying data (Bang et al. 2023, Liu et al. 2023).

**Table 3.** *Sub-criteria for Dimension 2*

| Sub-Criterion | Description |
|---|---|
| Completeness | Are all statistically significant findings identified and analyzed? |
| Analytical Depth | Does the interpretation go beyond restating numbers to explain meaning? |
| Logical Coherence | Are interpretive conclusions logically consistent with the statistical findings? |
| Contextual Relevance | Are findings interpreted within the appropriate domain context? |
| Casual Language Discipline | Does the interpretation avoid implying causation from correlational data? |

*Source:* Chang et al. (2024), Guo et al. (2023), Liu et al. (2023), Pearl & Mackenzie (2018), Thirunavukarasu et al. (2023).

Dimension 3: Domain Appropriateness

Domain appropriateness focuses on whether the interpretation aligns with the expectations and standards of the specific field, such as education or healthcare. It goes beyond statistical context by considering whether the conclusions make sense in context. For example, results may be statistically significant but have little practical or clinical importance or vice versa (Thirunavukarasu et al. 2023).

This dimension is especially relevant for LLM evaluation in applied domains, as research has shown that LLMs trained on broad, general data can sometimes apply statistical reasoning in a generic way, without fully accounting for the specific standards or evidentiary thresholds required in clinical or educational research (Holmes et al. 2022, Kasneci et al. 2023).

**Table 4.** *Sub-criteria for Dimension 3*

| Sub-Criterion | Description |
|---|---|
| Domain Terminology | Is domain-appropriate language used correctly? |
| Threshold Awareness | Are domain-specific significance thresholds addressed? |
| Population Relevance | Are findings discussed in relation to the relevant population? |
| Practical Implication | Does the interpretation identify practice-relevant implications? |

*Source:* Holmes et al. (2022), Kasneci et al. (2023), Kung et al. (2023), Thirunavukarasu et al. (2023).

Dimension 4: Limitation Awareness

Limitation awareness assesses whether AI tools clearly recognize the limits of the analysis. This includes data constraints, methodological weaknesses, assumptions underlying statistical techniques, and possible sources of error. Acknowledging these limits is a basic part of good scientific practice and is expected in all credible research reporting according to the STROBE Statement (2007). For AI-generated interpretations, this dimension also reflects something deeper: how well the model understands and communicates uncertainty.

**Table 5.** *Sub-criteria for Dimension 4*

| Sub-Criterion | Description |
|---|---|
| Sample Size Acknowledgement | Is the adequacy of sample size discussed in relation to statistical power? |
| Statistical Assumptions Awareness | Does interpretation acknowledge and appropriately discuss assumptions underlying the statistical methods? |
| Confounding Identification | Are potential confounding variables identified? |
| Data Quality Concerns | Are issues such as missing values, outliers, or class imbalance acknowledged? |
| Practical Implication | Does the interpretation identify practice-relevant implications? |

*Source:* Elm et al. (2007), Kadavath et al. (2022), Messick (1989), Pearl & Mackenzie (2018).

Dimension 5: Consistency and Reliability

Consistency and reliability check how stable AI-generated interpretations are when the same prompts and dataset are used more than once. This dimension focuses on a key concern in using language models for evidence-based work that their outputs

are not fixed and can be changed from one run to another (Ouyang et al. 2022, Liang et al. 2022).

Unlike traditional statistical software, which produces the same result every time for a given input, LLMs generate responses using probabilistic sampling. Because of this, the same prompt can produce lightly different or sometimes significantly different responses across runs. In high-risk areas like healthcare and education, this inconsistency can raise serious reliability issues and needs to be measured carefully (Bommasani et al. 2021, Shen et al. 2023).

**Table 6.** *Sub-criteria for Dimension 5*

| Sub-Criterion | Description |
|---|---|
| Model Consistency | Does the model produce substantively equivalent outputs across three separate runs? |
| Numerical stability | Are reported statistical values consistent across runs? |
| Interpretive stability | Does the interpretation remain consistent across runs? |
| Prompt Sensitivity | How significantly does output shift with minor prompt rewording? |

*Source:* Bang et al. (2023), Bommasani et al. (2021), Liang et al. (2022), Liu et al. (2023), Ouyang et al. (2023), Shen et al. (2023).

*Scoring Rubric*

Each sub-criterion is scored on a 1-5 Likert scale using the following rubric applied uniformly by all evaluators (Table 7).

**Table 7.** *Scoring Rubric*

| Score | Label | Description |
|---|---|---|
| 5 | Excellent | Criterion fully and precisely met without meaningful gaps |
| 4 | Good | Criterion mostly met with minor imprecision |
| 3 | Adequate | Criterion partially met with noticeable gaps |
| 2 | Weak | Criterion minimally addressed with significant errors |
| 1 | Poor | Criterion not met with incorrect outputs |

*Source:* Likert (1932).

Based on this scoring rubric, we calculated the AI Reliability Score (ARS) for each AI model per dataset, combining performance across all five dimensions.

$$ARS = (Statistical\ Accuracy \times 0.35) + (Interpretive\ Quality \times 0.30) + (Domain\ Appropriateness \times 0.20) + (Limitation\ Awareness \times 0.10) + (Consistency\ \&\ Reliability \times 0.05)$$

For each dimension, the final score is calculated using the average of its sub-criterion scores before applying it to the overall formula. The resulting ARS can range from 1.00, representing the lowest score, to 5.00, representing the highest (Table 8).

**Table 8.** *ARS Interpretation Bands*

| ARS Range | Reliability Classification |
|---|---|
| 4.50 – 5.00 | Highly reliable and appropriate for independent analytical use |
| 3.50 – 4.49 | Moderately reliable and appropriate with more human oversight |
| 2.50 – 3.49 | Partially reliable and requires substantial human verification |
| 1.50 – 2.49 | Unreliable and not suitable for evidence-based use |
| 1.00 – 1.49 | Critically unreliable and actively misleading |

*Source:* Liu et al. (2023).

The ARS interpretation bands provide a clear way to classify the overall reliability of AI-generated interpretations. These categories help to determine whether an interpretation is appropriate for independent use, requires human oversight, or is too unreliable for evidence-based applications.

In addition to quantitative scoring, all AI-generated interpretations are also examined through a qualitative error classification process. This process classifies identified errors into a structured taxonomy based on the type of issue present in the interpretation (Table 9).

**Table 9.** *Error Taxonomy*

| Error Type | Definition |
|---|---|
| Hallucination | Statistics or findings cited that do not exist in the dataset |
| Misinterpretation | Correct statistics reported but incorrectly explained |
| Omission | Statistically significant finding present in the data but not identified |
| Under-contextualization | Correct statistics reported but no domain-relevant meaning provided |

*Source:* Chang et al. (2024), Guo et al. (2023), Liu et al. (2023), Kasneci et al. (2023), Frieder et al. (2023), Thirunavukarasu et al. (2023), Ji et al. (2023).

We recorded, categorized, and counted each detected error for each model and dataset. This produces a clear error profile for every AI system and shows how often different types of mistakes occur.

Overall, this evaluation framework assesses AI reliability using five dimensions and a mix of quantitative and qualitative methods. This approach ensures transparent, evidence-based, and comparable results across models, addressing gaps in prior research that often relied on limited evaluation methods.

## Results and Discussion

This section presents the findings of the comparative evaluation of statistical outputs and interpretations generated by three AI models across the two datasets. Results are reported across five evaluation dimensions and using the *AI Reliability Score (ARS)* as the composite performance metric. In addition, qualitative error classification findings are reported separately through error taxonomy analysis.

*Results for Student Academic Performance Dataset*

First, we calculated each model's scores on five dimensions for the student academic performance dataset and the descriptive statistics of scores are presented in Table 10.

**Table 10.** *Descriptive Statistics of Scores for Student Academic Performance Dataset*

| **Error Type** | **Claude Sonnet 4.6** | **ChatGPT GPT-5** | **Gemini 3 Flash** |
|---|---|---|---|
| Statistical Accuracy | 5.0 | 4.0 | 4.0 |
| Interpretive Quality | 4.6 | 3.6 | 3.8 |
| Domain Appropriateness | 5.0 | 5.0 | 5.0 |
| Limitation Awareness | 4.8 | 3.8 | 3.6 |
| Consistency and Reliability | 4.0 | 3.8 | 4.0 |
| ARS Score | 4.81 | 4.05 | 4.10 |

According to Table 10, Claude Sonnet 4.6 achieved the highest ARS score (4.81), followed by Gemini 3 Flash (4.10) and ChatGPT GPT-5 (4.05). Across the statistical accuracy dimension, all three models produced mostly accurate numerical results. However, ChatGPT and Gemini showed limited discussion of the distributional properties of the variables compared to Claude. Claude also demonstrated stronger methodological completeness by checking distributions and correctly considering assumptions for statistical tests such as t-tests and linear regression. In addition, Claude correctly identified *the class* variable as categorical, while the other two models treated it as numerical in parts of their analysis even though it is not a key variable.

In terms of interpretive quality, Claude provided clearer and more structured scientific explanations with stronger depth. However, it was less effective in linking statistically significant results to their practical meaning within the domain context. On the other hand, ChatGPT and Gemini were better at connecting results to real-world implications, although their methodological handling was weaker. Gemini also tended to use more advanced and technical language, which sometimes made the interpretation less accessible to a general audience. About domain appropriateness, all three AI tools performed well and demonstrated a good understanding of the educational context of the dataset. For the consistency and reliability dimension, Claude and Gemini outperformed ChatGPT by producing more stable and methodologically consistent outputs across repeated runs.

*Results for Heart Disease Dataset*

Next, we calculated each model scores across five evaluation dimensions for the heart disease dataset, and the descriptive statistics of these scores are shown in Table 11.

**Table 11.** *Descriptive Statistics of Scores for Heart Disease Dataset*

| Error Type | Claude Sonnet 4.6 | ChatGPT GPT-5 | Gemini 3 Flash |
|---|---|---|---|
| Statistical Accuracy | 5.0 | 4.0 | 4.0 |
| Interpretive Quality | 4.8 | 3.2 | 3.6 |
| Domain Appropriateness | 5.0 | 4.0 | 4.2 |
| Limitation Awareness | 4.8 | 3.2 | 3.6 |
| Consistency and Reliability | 4.0 | 3.8 | 4.0 |
| ARS Score | 4.87 | 3.67 | 3.88 |

As shown in Table 11, Claude Sonnet 4.6 achieved the highest ARS score (4.85), outperforming Gemini Flash 3 (3.88) and ChatGPT GPT-5 (3.67) across most evaluation dimensions for the heart disease dataset. Compared to the education dataset, both ChatGPT and Gemini showed weaker methodological performance in the healthcare domain. Although all three AI tools produced largely accurate numerical values, ChatGPT and Gemini performed only limited statistical analyses and did not fit a proper regression model for deeper analysis. In contrast, Claude provided a more complete statistical analysis with detailed interpretations and explanations relevant to the heart disease context.

Claude also demonstrated stronger methodological awareness by checking distributional properties and normality assumptions before applying statistical techniques. When assumptions were not fully satisfied, Claude suggested appropriate non-parametric alternatives. ChatGPT and Gemini did not discuss distributions of variables or statistical assumptions in sufficient detail. In addition, both ChatGPT and Gemini introduced a new variable named “heart disease," which was not originally present in the dataset. They attempted to conduct analyses based on this variable. This represents an example of hallucination, where the models generated unsupported information not derived from the dataset.

In terms of interpretive quality, Claude performed better by providing clearer explanations and interpretations related to clinical relevance and the healthcare domain. Compared to the performance in education data analysis, both ChatGPT and Gemini showed weaker performance in healthcare analysis and provided more general interpretations with less meaningful discussion related to the healthcare domain. However, all three models showed a limited discussion on confounding variables and did not sufficiently address their possible impact on the findings. For consistency and reliability, Claude and Gemini produced relatively stable outputs across repeated runs.

Table 12 presents a summary of the comparative performance of three AI models across the education and healthcare datasets. Claude Sonnet 4.6 consistently achieved the highest performance with a mean ARS of 4.84. The very small variability between the two datasets (0.04) indicates that Claude’s performance was highly stable and reliable across different domain contexts. This suggests that Claude not only performs well in individual tasks but also maintains consistent quality regardless of the dataset type.

**Table 12.** *Summary of the Comparison Results*

| Model | ARS of Education Dataset | ARS of Healthcare Dataset | Mean ARS | Standard Deviation |
|---|---|---|---|---|
| Claude Sonnet 4.6 | 4.81 | 4.87 | 4.84 | 0.04 |
| ChatGPT GPT-5 | 4.05 | 3.67 | 3.86 | 0.27 |
| Gemini 3 Flash | 4.10 | 3.88 | 3.99 | 0.16 |

In contrast, ChatGPT GPT-5 showed the lowest overall performance, with a mean of 3.86. Its higher variability (0.27) indicates less reliability and is more sensitive to changes in domain context. Gemini 3 Flash performed slightly better than ChatGPT overall, with a mean ARS of 3.99, but still showed moderate variation between data sets.

Based on the mean ARS values in Table 12, Claude Sonnet 4.6 falls within the 4.5-5.00 range, meaning it is highly reliable and suitable for independent analytical use. In contrast, ChatGPT GPT-5, and Gemini 3 Flash fall in the 3.5-4.49 range, indicating that they are moderately reliable and require human oversight.

*Error Taxonomy Results*

Table 13 provides the frequency of each error type recorded for Claude Sonnet 4.6, ChatGPT GPT-5, and Gemini 3 Flash during the evaluation process.

**Table 13.** *Error Taxonomy*

| Error Type | Claude Sonnet 4.6 | ChatGPT GPT-5 | Gemini 3 Flash |
|---|---|---|---|
| Hallucination | 0 | 1 | 1 |
| Misinterpretation | 1 | 1 | 1 |
| Omission | 0 | 4 | 4 |
| Under-contextualization | 2 | 4 | 4 |
| **Total Errors** | 3 | 10 | 10 |

According to Table 13, Claude Sonnet 4.6 produced the fewest errors overall, with only three errors compared to ten errors each for ChatGPT GPT-5 and Gemini 3 Flash. The most notable hallucination error occurred in the healthcare dataset, where both ChatGPT and Gemini introduced a new variable called *heart disease* that was not present in the original dataset. All three models also made misinterpretation errors in the education dataset by describing a correlation coefficient of approximately 0.5 as a strong linear relationship, when it would generally be considered a moderate relationship.

The largest differences between the models were seen in omission and contextualization errors. Unlike Claude, both ChatGPT and Gemini failed to check important assumptions required for linear regression and significance testing in the education dataset. Furthermore, they failed to fit a linear regression model and to overlook key significance testing assumptions in the healthcare dataset. They further provided less domain-specific discussion, particularly in healthcare analysis since their interpretations were often general and lacked clinical context.

## Conclusions

This study evaluated the effectiveness and reliability of AI-assisted statistical analysis by comparing ChatGPT GPT-5, Claude Sonnet 4.6, and Gemini 3 Flash using a structured evaluation framework across education and healthcare datasets. The results showed that all three AI tools were able to provide generally accurate statistical results and useful interpretations. However, there were clear differences in their overall performance. Claude Sonnet 4.6 performed well in both education and healthcare datasets, achieving high scores across statistical accuracy, interpretability, domain appropriateness, limitation awareness, and consistency. Claude earned the highest mean ARS score, which indicates high reliability and appropriateness for analytical applications independent of a user. GPT-5 and Gemini 3 Flash were rated as having moderate reliability and appropriate for use with human oversight. Although both models provided useful analyses, they showed a weakly methodological completeness with less detailed interpretations and sometimes presented unsupported findings.

### *Limitations*

Several limitations should be considered when interpreting the findings of the study. First, the evaluation was limited only to three AI models, and these models were not specifically designed for statistical tasks. Second, the analysis was conducted using only two datasets from the education and healthcare domains. Therefore, these findings may not fully represent the wide variety of data types and analytical scenarios encountered in real-world applications. In addition, the study relied mainly on a single prompt, with only minor modifications used to assess consistency. Different prompt designs may have produced different results. The study also evaluated the models based on initial responses without using follow-up prompts to correct errors, address omissions, or improve interpretations. Therefore, the findings reflect how the models perform in a single-round analysis rather than in an interactive setting where users can refine and guide the analysis.

### *Future Research*

This work can be expanded in the future by considering more AI tools and datasets from additional domains. Different experimental designs could be used to study the impact of prompt engineering, iterative interactions, and feedback-based improvement on the quality of AI-generated statistical analyses. Future studies could also examine model performance on more complex data structures such as time series, hierarchical, and high-dimensional datasets, which require different analytical approaches. Another important direction would be to compare general-purpose AI models such as ChatGPT, Claude, and Gemini with AI tools specifically designed for statistical data analysis tasks such as DataRobot, IBM SPSS Modeler, and Julius AI. Future evaluations may also consider factors such as efficiency of computation, response time, and resource requirements in addition to analytical quality to provide a more complete understanding of the use of AI-assisted statistical analysis in real-world applications.

Overall, this study shows that although AI tools can significantly assist with statistical analysis, they are not yet reliable enough to replace human decision-making. Their effectiveness is mostly dependent on how well-structured prompts are used to guide AI tools and how closely their outputs are examined and analyzed. Human supervision is still important to identify errors, ensure methodological accuracy, and provide appropriate context. These results provide useful information for researchers, educators, and data practitioners on responsible and controlled use of AI in statistical analyses. Further, this study emphasizes the need to treat artificial intelligence in both academic, research, and real-world statistical applications as a supporting tool rather than a fully independent analyst.

## Acknowledgments

The author would like to acknowledge the use of artificial intelligence tools in the completion of this study. Since the focus of this research is the reliability of AI-generated interpretations, ChatGPT, Claude AI, and Google Gemini were used as part of the data analysis and comparative evaluation process. These tools were applied to generate statistical outputs and interpretations across models. All AI-generated outputs were carefully reviewed, verified, and critically assessed by the author.

## References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014) *Standards for Educational and Psychological Testing*. American Educational Research Association.

American Psychological Association (2020) *Publication Manual of the American Psychological Association* 7th Edition. American Psychological Association.

Anthropic (2025) *Claude models overview: Sonnet 4.6.* Anthropic PBC.

Bang Y, Cahyawijaya S, Lee N, Dai W, Su D, Wilie B, et al. (2023) A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* [Nusa Dua, Bali], 675–718.

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM FAccT Conference*, 610–623.

BenMessaoud F, Agrebi MH, Ash E (2025) FazBoard: An AI-Educational Hybrid Intelligent Teaching & Learning System. *Athens Journal of Technology & Engineering* 12(1): 29–40.

Bommasani R, Hudson DA, Aditi E, Altman R, Arora S, Kattell S, et al. (2021) *On the opportunities and risks of foundation models*. Stanford Center for Research on Foundation Models.

Chang Y, Wang X, Wang J, et al. (2024) A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology* 15(3): 1–45.

Demir F, Yavuz U (2026) Evaluating the Performance of Generative Artificial Intelligence Models in Multidimensional Data Analysis Tasks: A Comparative Study with Large Language Models. *Discover Computing* 29(1): 20.

Elm E, Altman DG, Egger M, Pocock SJ, Gøtzscheet PC, Vandenbroucke JP (2007) The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. *PLOS Medicine* 4(10): e296.

Fehlmann T, Kranich (2025) How to Teach Literacy to Artificial Neural Networks Making AI Intelligent by Learning from Other Disciplines. *Athens Journal of Sciences* 12(1): 59–70.

Field A (2018) *Discovering statistics using IBM SPSS statistics*. 5th Edition. SAGE Publications.

Frieder S, Pinchetti L, Griffiths RR, Salvatori T, Lukasiewicz T, Petersen PC, et al. (2023) Mathematical capabilities of ChatGPT. *Advances in Neural Information Processing Systems* 36.

Google DeepMind (2025) *Gemini model family: Flash.* Google DeepMind.

Guo Z, Jin R, Liu C, Huang Y, Shi D, Supryadi, et al. (2023) *Evaluating Large Language Models: A Comprehensive Survey*. arXiv.

Holmes W, Porayska-Pomsta K, Holstein K, Sutherland E, Baker T, Shum SB, et al. (2022) Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education* 32(Apr): 504–526.

Hong Y, Lian J, Xu L, Min J, Wang Y, Freeman LJ, et al. (2023) Statistical perspectives on reliability of artificial intelligence systems. *Quality Engineering* 35(1): 56–78.

Ji W, Yuan W, Getzen E, Cho K, Jordan MI, Mei S, et al. (2026) *An Overview of Large Language Models for Statisticians*. The American Statistician.

Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. (2023) Survey of hallucination in natural language generation. *ACM Computing Surveys* 55(12): 1–38.

Kadavath S, Conerly T, Askell A, Henighan T, Drain D, Perez E, et al. (2022) *Language models (mostly) know what they know*. arXiv.

Kasneci E, Seßler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. (2023) ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103(Apr): 102274.

Kavlakoglu C, Stryker E (2024) *What Is Artificial Intelligence (AI)?* IBM.

Keller A (2026) *Expanding the AI Evaluation Toolbox with Statistical Models*. NIST AI NIST AI 800-3, National Institute of Standards and Technology.

Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health 2(2): e0000198.

Liang P, Bommasani R, Lee T, Tsipras D, Soylu D. Yasunaga M, et al. (2022) *Holistic evaluation of language models*. arXiv.

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

Liu Y, Iter D, Xu Y, Wang S, Xu R, Zhu C (2023) *G-Eval: NLG evaluation using GPT-4 with better human alignment.* arXiv.

Messick S (1989) *Validity*. In RL Linn (eds.), *Educational measurement*. 3rd Edition. New York, NY: American Council on Education/Macmillan.

Nagarkar C, et al. (2026) *Can LLM Reasoning Be Trusted? A Comparative Study: Using Human Benchmarking on Statistical Tasks*. arXiv.

OpenAI (2025) *GPT-5 and ChatGPT capabilities.* OpenAI.

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. (2022) Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35: 27730–27744.

Pearl J, Mackenzie D (2018) *The book of why: The new science of cause and effect.* Basic Books.

Peters U, Chin-Yee B (2025) Generalization Bias in Large Language Model Summarization of Scientific Research. *Royal Society Open Science* 12(4): 241776.

Redwan K (n.d.) *UCI heart disease data*. Kaggle. https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data (Accessed March 21, 2026).

Shen T, Jin R, Huang Y, Liu C, Dong W, Guo Z, et al. (2023). *Large language model alignment: A survey*. arXiv.

Slobodsky P, et al. If You Can't Beat It, Join It! Teaching and Learning Mathematics with ChatGPT and Key Prompts to Stimulate Self-Learning. *Athens Journal of Education* 13(1): 185.

Sonawane S (n.d.) *Student performance dataset*. Kaggle. https://www.kaggle.com/datasets/suvidyasonawane/student-performance-dataset (Accessed March 18, 2026).

Stryker C (2021) *What Are Large Language Models (LLMs)?* IBM.

Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW (2023) Large language models in medicine. *Nature Medicine* 29(Jul): 1930–1940.

Xiao T, Zhu J (2025) *Foundations of Large Language Models*. ArXiv.