Can Large Language Models Replace Statistical Software?

Prof. Dr. Yves Staudt
Institute of Data Analysis, Artificial Intelligence, Visualization and Simulation
University of Applied Sciences of the Grisons
yves.staudt@fhgr.ch

Abstract

Statistical hypothesis testing is a cornerstone of evidence-based medicine and clinical research. Despite its central importance, previous research has consistently shown substantial deficits in statistical literacy among healthcare professionals. At the same time, large language models (LLMs) have demonstrated remarkable capabilities in scientific reasoning and data analysis. This study examines whether LLMs can serve as viable substitutes for conventional statistical software in guiding users through the selection, execution, and interpretation of hypothesis tests. Using a standardized prompt based on real survey data on the association between kick-scooter riding and knee pain in children, we evaluated seven LLMs and compared their outputs with statistical software results. Our findings indicate that none of the evaluated models can currently be considered a viable substitute. Although all models selected the appropriate test, substantial variation was observed in the quality of their explanations and in test execution. Gemini 3.1 Pro Preview, Claude Opus 4.6, and ChatGPT 5.4 Thinking performed strongly in test selection and result interpretation, with Gemini producing the most structured responses. However, none matched statistical software's result in test execution.

1 Introduction

Statistical hypothesis testing is one of the most widely used methodological tools in empirical research. It addresses a fundamental question in scientific inquiry: whether an observed effect or association in a sample reflects a true phenomenon in the underlying population or can instead be attributed to chance (Turner et al., 2020; Zhang, 2025). By quantifying uncertainty and enabling structured decision-making, hypothesis testing has become indispensable to evidence-based medicine, underpinning trial design, therapeutic comparisons, and diagnostic evaluation (Turner et al., 2020; Zhang, 2025).

At the same time, several studies indicate that statistical knowledge among medical researchers and healthcare professionals is often limited (Jenny et al., 2018). Selecting suitable methods, analysing data correctly, and drawing valid conclusions all involve considerable potential for error (Lakhlifi et al., 2023). Insufficient statistical competence may therefore substantially compromise the scientific quality of medical research.

Recent advances in large language models (LLMs) have introduced a useful interface for statistical reasoning and data analysis (Angelis et al., 2023; Thirunavukarasu et al., 2023). Unlike traditional statistical software such as SPSS or R, LLMs can be queried in natural language, which may lower the barrier to entry for users with limited statistical training. This raises the question of whether LLMs can act as intelligent assistants in statistical hypothesis testing.

In this paper, we investigate whether contemporary LLMs-ChatGPT 5.4 Thinking (OpenAI, 2026), Gemini 3.1 Pro Preview (Google DeepMind, 2026), Claude Opus 4.6 (Anthropic, 2026), Llama 4 Scout (Meta AI, 2025), Apertus 8b (Swiss AI Initiative, 2025), Qwen 3.5 9b (Swiss AI Initiative, 2025), and Phi 4 (Abdin et al., 2024)-can serve as substitutes for traditional statistical software in hypothesis testing. We evaluate their ability to select, execute, and interpret hypothesis tests by comparing their outputs with results obtained from conventional statistical software. Our findings indicate that LLMs cannot currently replace standard statistical tools, although they may still serve as useful supportive systems. Among the evaluated models, Gemini, Claude Opus, and ChatGPT showed the strongest overall documentation and performance.

2 Literature Review

2.1 Hypothesis Testing and Statistical Literacy

Hypothesis testing remains a central component of statistical inference in empirical research (O'Dushlaine, 2019). At the same time, methodologists have warned against an overly narrow focus on statistical significance alone, arguing for complementary reporting of effect sizes and confidence intervals (Mark et al., 2016; Patel and Green, 2025). In healthcare, several studies show that professionals often lack the statistical literacy needed to interpret p-values, confidence intervals, and diagnostic statistics correctly (Jenny et al., 2018; Lakhlifi et al., 2023). This is particularly concerning because weaknesses in statistical reasoning may compromise the planning, analysis, and interpretation of scientific studies.

2.2 LLMs as Statistical Assistants

LLMs have emerged as potentially useful tools for statistical assistance because of their capabilities in instruction following, reasoning over textual input, and code generation. In principle, an LLM should be able to identify an appropriate statistical test based on data structure, variable types, and assumptions (Fay and Brittain, 2022; Nikolić and Popovic, 2024). However, whether these models can reliably perform statistical calculations and provide valid interpretations remains an open empirical question, particularly in medically relevant contexts (Liu, 2025). This question motivates the present study.

3 Methodology

The aim of this study is to evaluate whether LLMs can serve as viable substitutes for conventional statistical software in the context of hypothesis testing. To this end, we benchmarked the models' outputs for test selection, execution, and interpretation against reference results obtained from statistical software. For this purpose, we defined a standardized prompting procedure, a fixed output structure, and an explicit evaluation framework. This design was intended to ensure comparability, transparency, and reproducibility across all tested models.

3.1 Prompt

All models were evaluated with the same prompt and input data, without model-specific adaptations, to minimize variation unrelated to model capability. Responses were analysed according to the predefined evaluation framework. The benchmark prompt was designed using a chain-of-thought-inspired reasoning framework (Sahoo et al., 2025). In the first part, the research problem and relevant tabular input data were provided. To ensure reproducibility and reuse, the complete prompt was documented in Markdown, with the input data embedded in the same file.

In the second part, the LLM was instructed to analyse the data and perform the hypothesis test. The prompt specified the required output format, the expected data structure, and the handling of missing values. In addition, the models were instructed to solve the task using a zero-shot chain-of-thought strategy. This ensured that all models received the same information under identical conditions.

3.2 Output Structure

To compare model responses systematically, a fixed output structure was imposed on all generated answers. This structure was chosen to improve transparency, traceability, and comparability across models. It was inspired by the classical statistical procedure for hypothesis testing described by (Gonick and Smith, 2005), who conceptualize hypothesis testing as a four-step process: formulation of hypotheses, calculation of the test statistic, calculation of the p-value, and evaluation of the result against a predefined significance level (α).

To assess whether LLMs can perform this process in practice, we adapted the classical procedure into a task-oriented format suitable for prompting and software-based execution. Accordingly, the required output structure consisted of the following four steps:

1. Selection of the appropriate statistical test and formulation of the hypotheses
2. Execution of the statistical test
3. Summary of the statistical results
4. Interpretation of the results

3.3 Model Selection

Seven models were selected to represent different segments of the contemporary LLM landscape:

- **Proprietary high-end models:** ChatGPT 5.4 Thinking (OpenAI, 2026), Claude Opus 4.6

(Anthropic, 2026), and Gemini 3.1 Pro Preview (Google DeepMind, 2026)

- **Open-weight benchmark models:** Llama 4 Scout (Meta AI, 2025) and Qwen 3.5 9b (Qwen Team, 2026)
- **Compact models:** Phi 4 (Abdin et al., 2024) and Apertus 8b (Swiss AI Initiative, 2025)

This selection captures variation in model size, accessibility, and expected performance. By including proprietary, open-weight, and compact models, the benchmark enables a broader comparison of how closely current LLMs approximate conventional statistical software. To reflect realistic usage, all models were accessed through web interfaces. The open-weight and compact models were hosted locally using Ollama, whereas the proprietary models were accessed through the providers' web interfaces.

3.4 Output Evaluation

The benchmark task focused on a defined use case (see Section 4) involving the association between two binary variables. For this type of problem, the appropriate statistical procedure is the χ^2-test of independence. The corresponding hypotheses are:

- Null hypothesis (H_0): There is no relationship between the two categorical variables.
- Alternative hypothesis (H_1): There is a statistically significant relationship between the two categorical variables.

To solve this task correctly, a model had to identify the appropriate test, determine the observed joint frequencies in a contingency table, derive the expected frequencies under the assumption of independence, calculate the test statistic and the corresponding p-value, and evaluate the result against the specified significance level. Because the calculation of frequencies depends on the sample size, correct determination of the sample size was included as an additional evaluation criterion. Finally, the model was expected to provide an interpretation of the result in the context of the use case.

Based on this setup, model outputs were evaluated according to the following criteria:

1. Selection of the statistical test
2. Formulation of the null and alternative hypotheses
3. Determination of the sample size
4. Presentation of the observed joint frequencies
5. Presentation of the expected joint frequencies
6. Provision of calculation details
7. Verification of test assumptions
8. Reporting of the p-value and other relevant statistical values
9. Identification or discussion of the significance level
10. Evaluation of the test result
11. Interpretation of the result

Each criterion was assessed for every evaluated LLM and later summarized in tabular form in the Results section (Section 5). All responses were additionally required to follow a structured Markdown format, reproduced in the Appendix, in order to support consistent documentation and systematic comparison.

To assess output quality across these criteria, we defined an ordinal grading scheme. This scheme captures graded differences in performance rather than relying on a binary correct/incorrect judgment. This was necessary because LLM responses frequently contained partial reasoning, conceptually correct intermediate steps, or errors that only became apparent during the application of the statistical procedure. The grading framework was therefore designed to distinguish between theoretical understanding, practical execution, and complete omission of a criterion.

Because the central research question of this study is whether LLMs can function as an alternative to statistical software, the grading scheme places particular emphasis on methodological correctness and successful application to the concrete use case. In this sense, the benchmark does not only assess whether a model mentions a relevant concept, but whether it applies the statistical procedure correctly and produces results consistent with the software-based reference solution.

All outputs were evaluated manually by the author with formal training in statistics, using the predefined criteria and grading rubric. To strengthen the consistency of the assessment, each response was systematically compared with the corresponding reference results produced by conventional statistical software.

The following grades were used:

- **A:** Criterion addressed correctly and applied correctly
- **B:** Criterion addressed correctly with minor application errors

- **C:** Criterion addressed correctly in principle, but affected by follow-up errors in application
- **D:** Criterion addressed correctly in principle, but with substantial application errors
- **E:** Theoretical understanding present, but no meaningful application to the use case
- **F:** Criterion treated incorrectly at the theoretical level
- **G:** Criterion not addressed

Grade A represents the highest level of performance, whereas Grade G represents the lowest. Grade G indicates that a criterion was omitted entirely. Grade F indicates that the model failed at the conceptual or methodological level. Grade E indicates that the model identified the relevant theoretical concept but did not successfully apply it to the given task. Grades A to D reflect increasing levels of correctness in execution. The distinction between Grades C and D is that Grade C was assigned when an error primarily resulted from an earlier mistake, whereas Grade D was assigned when the criterion itself was applied incorrectly.

4 Use Case

This study uses data from (Crottogini et al., 2026), who surveyed primary school students in grades 3 to 6 (aged 8 to 12 years) in Chur, Switzerland, about their kick-scooter use and knee health. The dataset comprises 134 complete responses, with a gender distribution of 51.5% female and 48.5% male. Most participants were 11 years old. Overall, 85% of respondents reported using a kick scooter either regularly or occasionally, and 41% reported regular use.

For the present benchmark, the analysis was restricted to the two variables *knee pain* and *kick-scooter riding*. Both variables were encoded as binary variables (yes/no). The observed joint frequencies are shown in Table 1.

Table 1: Observed contingency table for knee pain and kick-scooter use

		Kick-scooter riding		
		No	**Yes**	**Total**
Knee pain	**No**	14	55	69
	Yes	6	59	65
	Total	20	114	134

This use case was selected because it represents a simple but realistic hypothesis-testing scenario in which correct test selection, computation, and interpretation can be assessed against a clear software-based reference. This use case assesses how well LLMs can identify and analyse the relationship between these two variables by means of hypothesis testing. As a reference standard, the corresponding χ^2-test of independence was performed in both Python and R. The tested hypotheses were:

- H_0: There is no association between knee pain and kick-scooter riding.
- H_a: There is an association between knee pain and kick-scooter riding.

As described in Section 3.2, a key step is the correct derivation of the expected joint frequencies and test statistics. The expected frequencies for the contingency table are reported in Table 2, and the resulting statistical values are summarized in Table 3. Both Python and R applied Yates' continuity correction, commonly used for χ^2-tests with one degree of freedom.

Table 2: Expected contingency table for knee pain and kick-scooter use under the null hypothesis of independence

		Kick-scooter riding	
		No	**Yes**
Knee pain	**No**	10.3	58.7
	Yes	9.7	55.3

Table 3: Summary of the statistical values of the χ^2-test with Yates' continuity correction

Statistic	Value
χ^2 value	2.41
p-value	0.12
Degrees of freedom	1

Using a significance level of $\alpha = 0.05$, the null hypothesis cannot be rejected. Accordingly, the reference analysis does not provide sufficient evidence for an association between knee pain and kick-scooter riding in this sample.

For comparison, the χ^2-test can also be calculated without Yates' continuity correction. In that case, the test statistic increases to $\chi^2 = 3.22$, and the p-value decreases to 0.07 (see Table 4). Although the null hypothesis still cannot be rejected at the 5% significance level, the result lies closer to the significance threshold. This illustrates the impact of Yates' correction in small 2×2 contingency

tables and highlights why correct handling of this issue is relevant when evaluating LLM-generated analyses.

Table 4: Summary of the statistical values of the χ^2-test without Yates' continuity correction

Statistic	Value
χ^2 value	3.22
p-value	0.07
Degrees of freedom	1

5 Results

We assessed the seven models' structured outputs according to the criteria defined in Section 3.2. Table 5 summarizes the resulting grades. Across criteria, Gemini, Claude Opus, and ChatGPT documented and performed best, whereas Llama 4 and Qwen 3.5 performed worst.

Overall, all models were able to identify the task as a hypothesis-testing problem involving two binary variables and selected the χ^2-test as the appropriate procedure. All models also formulated the null and alternative hypotheses correctly and recognized the significance level of $\alpha = 5\%$. Substantial differences emerged, however, in the execution of the analysis, especially in determining the correct population size, constructing the observed and expected frequency tables, checking assumptions, computing the relevant statistical values, and drawing the conclusion.

Only one model showed limitations already at the stage of data recognition (Figure 1). Llama 4 identified the variables as binary, but its description of the data structure was less complete than that of the other systems.

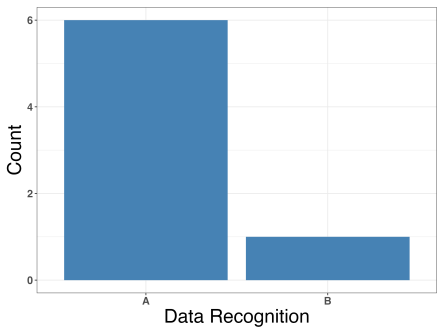

Figure 1: Recognition of the data format by the evaluated LLMs.

A clearer separation between stronger and weaker models emerged in the determination of the population size. As shown in Figure 2, only three models-Gemini, ChatGPT, and Claude Opus-correctly identified the sample size. The remaining four models produced incorrect values, indicating hallucinated or inconsistent reconstructions of the input data.

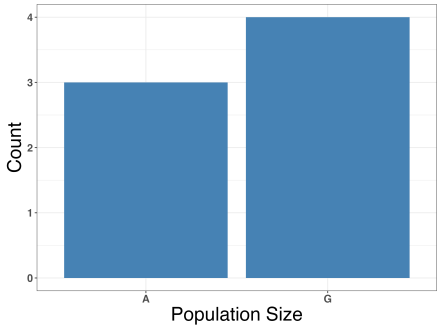

Figure 2: Performance of the LLMs in determining the population size.

These errors in population size were reflected directly in the observed contingency tables (Figure 3). All models that failed to determine the correct sample size also failed to reconstruct the observed joint frequencies correctly. In contrast, the observed frequency tables generated by ChatGPT, Gemini, and Claude Opus matched the reference results shown in Table 1.

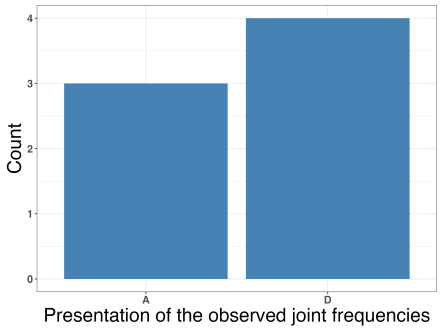

Figure 3: Performance of the LLMs in determining the observed joint frequencies.

Performance deteriorated further when the models were required to derive the expected joint frequencies (Figure 4). Llama 4 and Apertus did not provide expected frequencies at all. Qwen 3.5 produced expected frequencies that were not grounded in its own preceding contingency table, indicating a substantial inconsistency in the analytical chain. Phi 4 derived expected frequencies from an already incorrect observed table and therefore produced a follow-up error. Only ChatGPT, Gemini, and

Table 5: Summary of the evaluation of LLM performance across the predefined criteria.

Model	ChatGPT 5.4 Thinking	Gemini 3.1 Pro-Preview	Claude Opus 4.6	Llama 4 Scout 17b	Apertus 8b	Phi 4	Qwen 3.5 9b
Data Recognition	A	A	A	B	A	A	A
Selection of statistical test	A	A	A	A	A	A	A
Formulation of the null and alternative hypotheses	A	A	A	A	A	A	A
Determination of the sample size	A	A	A	G	G	G	G
Presentation of the observed joint frequencies	A	A	A	D	D	D	D
Presentation of the expected frequencies	A	A	A	G	G	C	E
Provision of calculation details	G	A	A	F	G	C	E
Verification of test assumption	A	A	A	G	G	G	G
Report of the p-value and other relevant statistical values	B	B	B	G	D	C	E
Identification or discussion of the significance level	A	A	A	A	A	A	A
Evaluation of the results	A	A	A	F	C	A	C
Interpretation of the results	B	A	A	G	A	A	B

Claude Opus generated expected frequencies consistent with the software-based reference reported in Table 2.

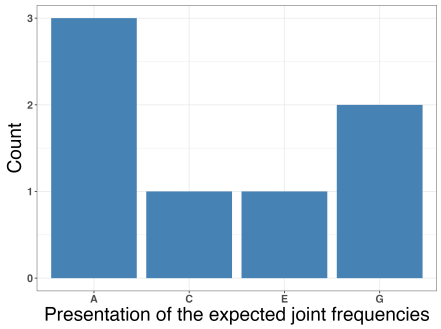

Figure 4: Performance of the LLMs in determining the expected joint frequencies.

The provision of calculation details also varied markedly across models (Figure 5). This criterion is particularly important because it allows expert users to verify whether the reported statistical values follow coherently from the data and formulas. Gemini and Claude Opus provided transparent and traceable intermediate steps. ChatGPT reported the key results but did not disclose intermediate calculations, which reduced transparency but did not prevent interpretation. Llama 4 failed to connect the underlying statistical procedure to the concrete use case, while Apertus provided insufficient detail to reconstruct the analysis. Qwen 3.5 presented correct formulas in principle, but the numerical values did not consistently follow from the earlier steps, which limited the trustworthiness of the output.

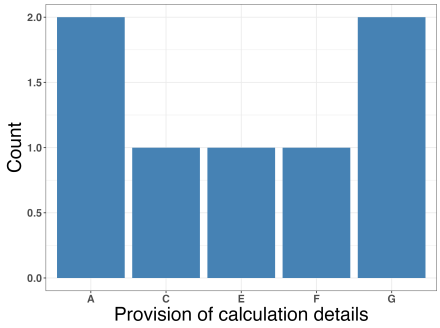

Figure 5: Performance of the LLMs in providing calculation details for the selected test.

Before interpreting the test outcome, it is necessary to verify whether the assumptions of the χ^2-test are satisfied. As shown in Figure 6, only Gemini, ChatGPT, and Claude Opus explicitly checked the minimum assumptions of the test. The remaining models did not assess whether the minimum expected cell frequency requirement was met.

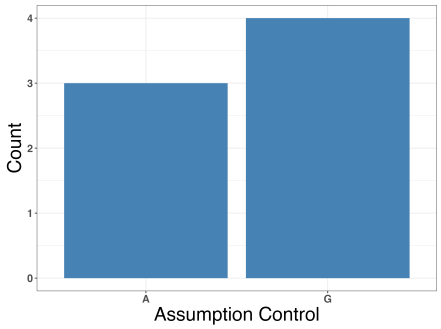

Figure 6: Performance of the LLMs in checking the assumptions of the χ^2-test.

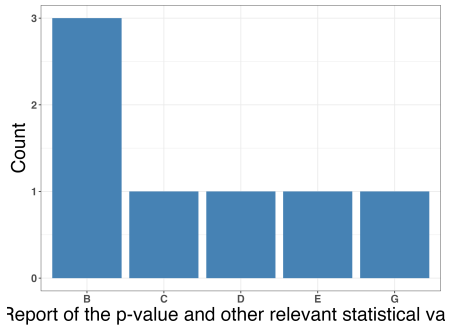

Figure 7: Performance of the LLMs in reporting the χ^2 statistic, p-value, and degrees of freedom.

The most critical criterion concerned the reporting of the final statistical values, namely the χ^2 statistic, the p-value, and the degrees of freedom (Table 3). These values are essential because they allow an independent reader to verify the correctness of the analysis. Among all evaluated models, Gemini came closest to reproducing the reference calculations, as it was the only model to explicitly mention Yates' correction. However, neither Gemini nor any other model reproduced the software output exactly. Instead, Gemini, Claude Opus, and ChatGPT reported values corresponding to the χ^2-test without Yates' continuity correction, that is, the values shown in Table 4 rather than those from Table 3.

This distinction is important because the omission of Yates' correction shifted the result closer to the conventional significance threshold. In the present use case, the corrected analysis yields $\chi^2 = 2.41$ and $p = 0.12$, whereas the uncorrected analysis yields $\chi^2 = 3.22$ and $p = 0.07$. Thus, the stronger models were able to reconstruct an analytically plausible result, but not the exact reference result produced by conventional software. Phi 4 correctly propagated its own internal calculations, but these were based on an incorrect population size and therefore did not correspond to the actual dataset. Llama 4 did not provide usable final statistics, while the outputs of Apertus and Qwen 3.5 lacked sufficient consistency for reliable verification.

These differences also affected the evaluation and interpretation of the test results (Figures 8 and 9). Gemini and Claude Opus provided the strongest overall interpretations, including discussion of the small number of observations in the "no kick-scooter riding" category. ChatGPT also produced a largely coherent evaluation, although with less detail. In contrast, weaker models were unable to derive a reliable conclusion because earlier errors in data reconstruction or statistical execution propagated into the final interpretation.

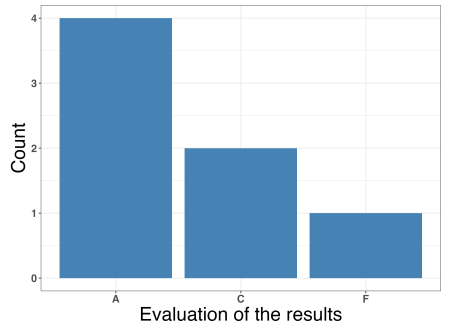

Figure 8: Performance of the LLMs in evaluating the result of the χ^2-test.

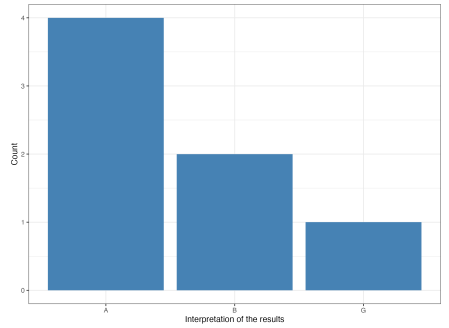

Figure 9: Performance of the LLMs in interpreting the χ^2-test in relation to the use case.

6 Conclusion, Limitations, and Outlook

Statistical hypothesis testing remains one of the central methodological tools in medical research and evidence-based medicine (Turner et al., 2020). At the same time, previous work has shown that statistical literacy among medical researchers and healthcare professionals is often limited (Jenny

et al., 2018), which may negatively affect the selection of appropriate methods, the correct analysis of data, and the derivation of valid conclusions (Lakhlifi et al., 2023). In this context, LLMs have emerged as a potentially useful new interface for statistical reasoning and hypothesis testing (Fay and Brittain, 2022; Nikolić and Popovic, 2024). If such models were able to serve as reliable substitutes for conventional statistical software, statistical analyses could become more accessible through natural-language interaction and thereby support broader and more efficient use of quantitative methods in medical research.

The present study evaluated whether contemporary LLMs can serve as a viable substitute for conventional statistical tools in a standardized hypothesis-testing task. Our results show that, despite remarkable capabilities, none of the evaluated models-ChatGPT 5.4 Thinking, Gemini 3.1 Pro Preview, Claude Opus 4.6, Llama 4 Scout, Apertus 8b, Qwen 3.5, and Phi 4-can currently be considered a viable substitute for statistical software. The main reason is that none of the models reproduced the full reference analysis generated by conventional software, most notably because none applied Yates' continuity correction, which is relevant for small 2×2 contingency tables with one degree of freedom.

At the same time, the evaluated models differed substantially in quality. Gemini achieved the strongest overall documentation and performance, followed closely by Claude Opus and ChatGPT. These models were generally able to identify the correct test, formulate the hypotheses appropriately, and provide plausible analyses and interpretations. However, they still deviated from the software-based reference in critical aspects of statistical execution. ChatGPT was less transparent than Gemini and Claude Opus because it provided fewer calculation details and a less nuanced discussion of the data distribution. Gemini and Claude Opus therefore appear more useful in practice when expert users require both methodological guidance and interpretability.

The open-weight and compact models performed less reliably. Llama 4 Scout and Qwen 3.5 showed substantial weaknesses in reconstructing the dataset and reporting the necessary statistical values, which makes them unsuitable for reliable use in the present task. Apertus 8b and Phi 4 showed partial methodological understanding and were able to provide some correct theoretical elements, but their analyses were undermined by incorrect assumptions about the underlying data. These findings suggest that smaller or less capable models may still offer limited educational or explanatory value, but they cannot currently support statistical analysis with sufficient reliability.

Taken together, the findings indicate that current LLMs are better understood as supportive assistants than as replacements for conventional statistical software. They may help users identify suitable tests, structure an analysis, and interpret outputs, but the calculations should still be performed by statistical tools. Statistical literacy therefore remains necessary to verify analyses and ensure high-quality medical research.

This study has several limitations. First, the benchmark was restricted to a single use case involving a χ^2-test of independence. The generalizability of the findings to other forms of hypothesis testing, more complex datasets, or other analytical settings therefore remains uncertain. Second, the study used a single standardized prompt based on a zero-shot chain-of-thought strategy. Different prompting strategies may lead to different levels of performance and should therefore be investigated systematically. Third, the evaluation was conducted by the author, although model outputs were cross-checked against software-based reference results to increase consistency.

Future work should extend this benchmark in three directions. First, a broader range of statistical tasks should be evaluated, including t-tests, nonparametric tests, regression models, and multivariable analyses. Second, future studies should compare different prompting strategies, including iterative prompting, tool-augmented prompting, and workflows in which LLMs interact directly with statistical software. Third, it would be valuable to examine not only whether LLMs can reproduce statistical results, but also whether they improve the analytical performance of human users in realistic research settings. Such work would help clarify whether the most promising role of LLMs lies not in replacing statistical software, but in augmenting statistical reasoning and decision-making in practice.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero

Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. 2023. Chatgpt and the rise of large language models: the new ai-driven infodemic threat in public health. *Frontiers in Public Health*, 11.

Anthropic. 2026. Introducing Claude Opus 4.6. https://www.anthropic.com/news/claude-opus-4-6. Accessed: 2026-05-04.

Armando Crottogini, Bianca Schenk, and Yves Staudt. 2026. Rolling towards resilience: The impact of scooter riding on pediatric knee health.

M Fay and E Brittain. 2022. Statistical hypothesis testing in context. pages –.

Larry Gonick and Woollcott Smith. 2005. *The Cartoon Guide to Statistics*. HarperCollins.

Google DeepMind. 2026. Gemini 3.1 Pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/. Accessed: 2026-05-04.

Mirjam Annina Jenny, Niklas Keller, and Gerd Gigerenzer. 2018. Assessing minimal medical statistical literacy using the quick risk test: a prospective observational study in germany. *BMJ Open*, 8:e020847.

Camille Lakhlifi, François-Xavier Lejeune, Marion Rouault, Mehdi Khamassi, and Benjamin Rohaut. 2023. Illusion of knowledge in statistics among clinicians: evaluating the alignment between objective accuracy and subjective confidence, an online survey. *Cognitive Research: Principles and Implications*, 8:23.

Siyu Liu. 2025. Modern developments in hypothesis testing with emphasis on computational techniques. *Theoretical and Natural Science*, pages –.

D Mark, Kerry Lee, and F Harrell. 2016. Understanding the role of p values and hypothesis tests in clinical research. *JAMA cardiology*, 1 9:1048–1054.

Meta AI. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Model card: https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct.

Božana Nikolić and Tamara Popovic. 2024. Hypothesis testing and statistical test selection: Fundamentals of statistics in clinical studies - part ii. *Medical review*, pages –.

OpenAI. 2026. Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/. Accessed: 2026-05-04.

C. O'Dushlaine. 2019. *Hypothesis testing and confidence intervals*, pages 523–626.

Sharad Patel and Adam Green. 2025. Death by p-value: the overreliance on p-values in critical care research. *Critical Care*, 29:73.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5. Accessed: 2026-05-04.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2025. A systematic survey of prompt engineering in large language models: Techniques and applications.

Swiss AI Initiative. 2025. Apertus: Democratizing open and compliant LLMs for global language environments. *arXiv preprint arXiv:2509.14233*. Developed by EPFL, ETH Zurich, and CSCS.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature Medicine*, 29:1930–1940.

D Turner, H Deng, and T Houle. 2020. Statistical hypothesis testing: Overview and application. *Headache: The Journal of Head and Face Pain*, 60:– .

Sirui Zhang. 2025. Analysis and optimization of the applicability of hypothesis testing methods. *Advances in Operation Research and Production Management*, pages –.

7 Appendix

The complete benchmark prompt, the input data, and all model outputs are available in a closed GitHub https://github.com/staudtyves/llm_my_statistical_software repository for the purpose of reproducibility and further inspection. Access to the repository can be granted by the authors upon request.