Mathematica
Progress Together

Trusting Your LLM: Building a Framework for Evaluating LLMs on Complex Statistical Data Sets

AUTHORS
Andrés Nigenda, Camille Shao, Dallas Dotter

Background

Most public LLM benchmarks do not test a model's ability to filter microdata, apply survey weights, or compute valid population estimates—core tasks in statistical practice. As a result, existing benchmarks offer limited insight into whether LLMs can perform applied statistical reasoning on complex, policy-relevant data sets.

Objective

Develop a cloud-native evaluation framework to assess how accurately and transparently LLMs perform applied statistical analysis on complex, survey-based data.

Methods

- Built 35 validated question/answer pairs leveraging the American Community Survey (ACS) Public Use Microdata Sample (PUMS)
- Evaluated models under two tooling constraints: reasoning-only and code execution
- Compared model outputs to verified benchmark estimates using quantitative and qualitative metrics
- Executed all runs in a secure, serverless cloud environment (AWS Lambda, Bedrock, OpenAI API, LangChain)

Findings

- Code-execution workflows substantially improved accuracy, with performance varying by model family.
- Reasoning-only workflows produced large errors and served primarily as a data-leakage check.
- Statistical competence differed meaningfully across models, particularly for causal inference tasks.
- Small differences in model reasoning behavior led to large differences in analytic outcomes.

Table 1. Accuracy of LLMs when allowed to analyze data through code execution[1]

LLM	Descriptive tasks	Causal inference
Claude Sonnet 4	88%	91%
Claude Sonnet 4.5	92%	73%
ChatGPT 4o	75%	64%
ChatGPT 5.1	88%	100%

[1] Accuracy is measured as the proportion of answers that exactly match verified benchmark values.

Table 2. Model error rates by question type, with and without code access[2]

	Descriptive		Causal inference	
LLM	Reasoning only	Code execution	Reasoning only	Code execution
Claude Sonnet 4	104%	3%	188%	18%
Claude Sonnet 4.5	92%	4%	179%	54%
ChatGPT 4o	111%	14%	199%	25%
ChatGPT 5.1	100%	4%	196%	0%

[2] Error is measured using symmetric mean absolute percentage error, which captures the average percentage difference between model estimates and verified values. Lower values indicate more accurate results.

Conclusions

- Organizations considering LLM-powered analysis should treat data set access, tool invocation, and executable workflows as prerequisites for trustworthy output.
- Differences in statistical competence across models remain substantial, especially for causal inference tasks.
- Data governance and documentation quality materially affect model performance; poorly labeled variables or ambiguous metadata degrade analytic accuracy.
- Structured frameworks that combine real data, controlled tooling, and reproducible execution enable evidence-based assessment of LLM analytic readiness.

Next steps

- Develop additional prompting strategies with templates, particularly for multi-step causal inference tasks
- Extend automation of the evaluation framework to support scalable deployment across data sets, models, and analytic tasks

Whether strengthening data systems, advancing analytics, or applying artificial intelligence responsibly, Mathematica is a trusted modernization partner, helping clients deliver efficient, effective programs that improve public well-being. Learn more at Mathematica.org.

For more information, contact anigendazarate@mathematica-mpr.com

Mathematica®
Progress Together