Discussion of "LAMBDA: A Large Model Based Data Agent"

David Donoho
Department of Statistics, Stanford University, Stanford, CA

This paper [4] hit the streets at a pivotal moment. We're currently amid a surge in enthusiasm around AI for data analysis, accompanied by a broader societal conversation about AI's impact on white-collar work and potential job displacement.

The paper poses an important question: how should we evaluate AI systems for data analysis? There are many competitors in this space right now, including some of the dominant tech companies with essentially infinite resources. But the biggest question that emerges, for me, is: What ought academics be doing in this space? Should they be pursuing projects like LAMBDA? And how should academics evaluate and validate the significance of such efforts?

The paper we're discussing released in May 2025—is well-written, extremely clear, and personally inspiring. A noteworthy detail: if you check Google Trends, you'll see a marked uptick in searches for "AI data analysis" around the time this paper was released. Perhaps LAMBDA contributed to that spike, or perhaps it is simply part of a broader wave.

The Commercial Landscape and its Rhetoric

At the same time, venture capital is pouring tens of billions of dollars into agentic AI—particularly in the narrower domain of agentic AI for data analysis, which it has identified as a likely sweetspot for attracting well-off customers. This investment is being accompanied by strong marketing rhetoric: claims that these systems will fully replace white-collar jobs in finance, healthcare, government, and beyond.

For instance, the startup causaLens[1] claims their tool will make data science "frictionless" or even "effortless" though of course, this claim is aspirational, in service of monetizing the workflow. Notably, the sales rhetoric isn't positioning the system as a copilot or assistant; they're marketed as autonomous agents, "full teammates" capable of independent action. While some messaging softens this claim by suggesting that data scientists won't disappear, but rather be "elevated," the underlying narrative is that a major transformation is underway.

This vision is echoed by other startups. AccelData[2], for example, promotes its agentic AI to identify patterns too overwhelming for humans. They claim this technology will shift professionals from tactical report building to strategic advising, with new roles in governance and ethics.

Hex[3], a startup that has received Series C funding, offers another illustrative case. Their product reimagines notebooks like Jupyter or Colab, integrating AI assistance what they call "Hex Magic" to help automate the analytics process, combining code, database access, and visuals.

Similarly, Matillion[4] claims to automate data engineering, modeling, and preparation without requiring users to know SQL or coding at all. Their promise: "virtual data analysts" that transform the process from manual to autonomous, dramatically increasing productivity.

[1] `https://causalens.com`
[2] `https://www.acceldata.io`
[3] `https://hex.tech`
[4] `https://www.matillion.com`

Naturally, VC marketing benefits from the perception that this is an epochal transformation. You'll often see alarmist headlines like claims that "500,000 jobs" have already been lost due to AI—though such figures are usually breezily speculative or cherry-picked.

On the other hand, there are observers and experts who are not part of the AI marketing ecosystem—unlike companies such as Amazon, OpenAI, Google, or Meta. When you read mainstream media coverage, including articles like those in the Harvard Gazette[5], there's a clear pattern: the headlines and opening paragraphs emphasize doom-and-gloom scenarios and sweeping claims about AI's disruptive potential. But when you read further or hear from actual subject-matter experts, the tone becomes far more measured. Often, the message is simply: we don't really know yet how disruptive AI will be.

Evaluating LAMBDA and its Competition

For those of us in academia, a central concern is whether AI tools like LAMBDA are genuinely poised to replace data scientists or at least aspects of the work we currently assign to graduate students as homework, particularly in regression and classification modeling. The LAMBDA paper frames itself as aiming to do precisely that. Of course, there are voices suggesting this won't be a matter of replacement, but rather elevation: professionals will shift to higher-level, more strategic roles. As MIT economist and Nobel laureate Daron Acemoglu [1] has pointed out, the productivity gains from AI might be far more modest than some forecasts perhaps a 1% boost to GDP.

The LAMBDA paper presents a methodological framework for documenting what the authors have achieved. They evaluate performance across a range of datasets many drawn from the UCI Machine Learning Repository and report results showing that the system's automated analysis is essentially on par with traditional, human-conducted analysis.

To test LAMBDA beyond the datasets used in the paper, I turned to a newly released dataset: the Electricity Cost Prediction dataset[6], posted during the week of July 2025. It contains ∼10,000 rows and ∼10 variables—well within the scope of a typical supervised learning task. Its main advantage is that it serves as a truly out-of-sample dataset unseen during the development of LAMBDA—while still fitting squarely within the type of task addressed in a typical master's-level regression and classification curriculum, such as that found in James, Witten, Hastie, and Tibshirani [3].

I engaged Dr. Elena Belogolovsky to test LAMBDA on this dataset. She has a PhD in Management and Behavioral Sciences and possibly represents the archetype of skilled professional who has to do data analysis but doesn't have an advanced statistics degree. I also engaged Andrew Donoho, an experienced information technology person to advise her, in case of any difficulties getting the system going.

We tested LAMBDA on this new dataset and compared its performance to other platforms. While LAMBDA had some successes, we also encountered issues. Notably, when Dr. Belogolovsky requested a report, it ignored our novel dataset and returned results based on unrelated pre-existing datasets from the UCI repository. This suggests the system may have defaulted to cached or training data rather than dynamically analyzing our input perhaps due to semantic similarity or model behavior at inference time.

Despite these hiccups, the core interaction was informative. We asked a structured set of questions to evaluate the system:

- Which features are most important for predicting electricity cost?
- Can a linear regression model accurately predict cost?

[5] `https://news.harvard.edu/gazette/story/2025/07/will-your-job-survive-ai/`

[6] `https://www.kaggle.com/datasets/shalmamuji/electricity-cost-prediction-dataset`

- Does model performance improve when grouped by structure type? Are nonlinear models significantly better?

The system produced usable results in many cases, but in others, we received irrelevant outputs or execution errors, suggesting fragility in dynamic context handling.

To establish a comparative baseline, we also tested Google Colab with Gemini 2.5[7], which recently added support for data analysis through natural language. We uploaded the same dataset and were able to ask structured questions; get code suggestions and automatically execute analysis. Initially, we assumed this was a form of out-of-sample testing, but we later discovered that the dataset may have been included in Gemini's training data, since it was hosted on Kaggle, which is owned by Google. So, while performance was strong, it wasn't a fully arms-length benchmark test.

While LAMBDA was developed as an academic research project, in the time since its release, comparable tools have rapidly come to market. For example, Google's Gemini 2.5 now integrates directly into Colab, providing a powerful AI assistant for data analysis as a no-cost add-on. Notably, this feature appeared just as the LAMBDA paper was circulating, and it offers a persuasive, text-driven interface that produces end-to-end analyses.

We also explored Grok, which can generate high-level summaries and, when asked to perform statistical analysis, will produce relevant code snippets. However, unlike Colab or LAMBDA, users must still run the code manually, meaning the experience is not entirely seamless or hands-free.

We also tested Shortcut AI[8], which embeds AI assistance directly into Excel spreadsheets, suggesting that agentic AI is now permeating even basic productivity software. LLM systems are claimed to often hallucinate and make unspoken unwarranted assumptions about user intent. For example, when we asked Shortcut AI to analyze data, it returned an executive report with strategic recommendations like improving recycling programs to reduce costs even though we had not mentioned cost reduction as a goal. This reflects a kind of premature autonomy, where the AI projects intent rather than responding directly to user queries.

A Proper Role for Academic Research: Building Evaluations, Not Products

LAMBDA is designed by an academic team that has accomplished something both impressive and timely. Their work with LAMBDA speaks directly to the broader narrative that AI is poised to transform the nature of employment across many sectors including what it means to be a data scientist. LAMBDA demonstrates that some tasks traditionally assigned to a graduate student in regression or classification modeling can now be automated, at least in part.

Yet, the context surrounding this project is shifting rapidly. Since LAMBDA's release, hundreds of millions of dollars in venture capital have been poured into agentic AI startups and products. These industry players have vast engineering teams, massive data infrastructure, and a pace of iteration that far exceeds what academic labs can typically sustain. It seems likely that commercial systems will soon—if they haven't already—surpass LAMBDA in raw capability.

So what, then, is the role of academia? Are academic researchers meant to compete head-to-head with corporate labs in building full-stack AI systems? Or is there another, perhaps more intellectually meaningful, contribution they can make?

One area where academic work can make a lasting and principled impact is in evaluation. In the current AI development environment, success is often defined by performance on

[7] `https://colab.research.google.com`

[8] `https://www.shortcut.com/agents`

benchmarks—well-defined tasks with accompanying datasets, standardized metrics, and automated evaluation pipelines. These benchmarks are frequently accompanied by public leaderboards, allowing developers to compare models across competitors on a shared footing.

Designing such benchmarks is not trivial, and it's an area where academics are well positioned to lead. Benchmarks can define what success should look like and thereby shape the direction of the field. If AI systems for data science are going to be widely deployed—whether for modeling, analysis, or decision-making—then the academic community should take responsibility for articulating what we value in those systems: rigor, interpretability, fairness, and reproducibility.

If commercial tools continue to evolve without academic input, they will be guided solely by market incentives. We already see signs of this: models that hallucinate strategic recommendations and tools that optimize for surface-level insights.

Rather than attempting to outcompete VC-backed platforms with limited resources, academic teams could instead focus on building meaningful, principled challenge problems—benchmarks, datasets, and evaluation frameworks that are aligned with the professional standards of data science and statistics. This would not only steer commercial development toward more useful and responsible directions but also ensure that our expertise as a community continues to shape the tools of the future.

LAMBDA itself offers a glimpse of what this might look like. It created a structured evaluation using datasets from the UCI Machine Learning Repository and reported results across key metrics such as R^2, classification error etc. While it's possible perhaps even likely that LAMBDA's performance will be surpassed by new entrants, the evaluation suite it proposed may remain useful. If future systems adopt these benchmarks, then LAMBDA's legacy might lie not just in its modeling capabilities, but in the criteria it helped define for what counts as success in agentic data science.

There's a recent precedent here as well: Chatbot Arena, developed at UC Berkeley, has become a go-to benchmarking framework for large language models [2]. By providing a crowdsourced, competitive arena for chatbot evaluation, the academic team created a neutral, principled standard one that was rapidly adopted by industry and became a kind of lingua franca for performance comparison. This shows how academia can shape not just research, but practice and market behavior—if it focuses on what it does best.

Conclusion

LAMBDA is an impressive academic project that shows how far agentic AI can go in replicating routine data analysis tasks. VC-funded competitors are moving fast, with significantly greater resources and access to private infrastructure than any academic lab. Rather than competing directly, academics can add the most value by defining rigorous evaluation standards and shaping the benchmarks that govern the field. The academic community should take the lead in setting the agenda for what good data science looks like in the era of AI, before others—guided by different incentives—define it for us.

References

[1] Acemoglu, D. (2025). The simple macroeconomics of AI. *Economic Policy*, 40(121), 13-58.

[2] Chiang, W. L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., ... & Stoica, I. (2024, March). Chatbot arena: An open platform for evaluating LLMs by human preference. *In Forty-first International Conference on Machine Learning* `https://arxiv.org/pdf/2403.04132`

[3] James, G. G. M., Witten, D., Hastie, T. J., Tibshirani, R. J., & Taylor, J. E. (2023). *An introduction to statistical learning: with applications in Python*.

[4] Sun, M., Han, R., Jiang, B., Qi, H., Sun, D., Yuan, Y., & Huang, J. (2025). Lambda: A large model based data agent. *Journal of the American Statistical Association*, 1–13.

[5] Sun, M., Han, R., Jiang, B., Qi, H., Sun, D., Yuan, Y., & Huang, J. (2025). A survey on large language model-based agents for statistics and data science. *The American Statistician*, 1–14.