Response to Discussion 5: Academic Evaluation and the Commercial Landscape

Team LAMBDA
The Hong Kong Polytechnic University
Hong Kong SAR, China

We are deeply grateful to Professor David Donoho for his thoughtful and thought-provoking discussion of our work on LAMBDA [Sun et al., 2025a]. As a towering figure in statistics and data science, his perspective carries particular weight, and his questions about the role of academia in the era of AI-powered data analysis are both timely and profound. We address his key concerns below, particularly regarding the commercial landscape, performance comparisons, and the proper role of academic research.

We have recently developed DSAEval, a comprehensive framework for evaluating data science agents on a wide range of real-world data science problems [Sun et al., 2026]. This benchmark represents our commitment to advancing rigorous evaluation standards in the field, addressing many of the concerns raised by Professor Donoho about the proper role of academic research in AI development.

1 The Commercial Landscape and Evaluation Context

Professor Donoho provides an insightful analysis of the current commercial landscape, where venture capital is pouring billions into agentic AI, accompanied by strong marketing rhetoric promising "frictionless" or "effortless" automation. He correctly notes that the biggest question for academics is what they should be doing in this space and how they should evaluate and validate their significance.

We agree that the commercial rhetoric often promises "frictionless" autonomy beyond present capabilities, intensifying competition. However, we believe there remains a crucial and distinct role for academic research in this ecosystem. While commercial systems may have greater resources and engineering teams, academic research has distinct strengths: it can originate novel ideas, serve the public interest, and set rigorous, transparent evaluation standards that shape the field.

We appreciate Professor Donoho's assistant Dr. Elena Belogolovsky's comparison of LAMBDA with several commercial products on the analysis of the Electricity Cost Prediction dataset. Her finding that Google Colab Data Science Agent performed best on this particular dataset, with LAMBDA ranking second, provides valuable comparative information.

Regarding the performance gap between LAMBDA and Google Colab on the Electricity Cost Prediction dataset, the underlying models are a key factor. The evaluation used

GPT-4o-mini as LAMBDA's base model, while the Colab Data Science Agent relied on the substantially more powerful Gemini 2.5 Pro. For example, GPT-4o-mini scores 40.2 on the GPQA Benchmark, whereas Gemini 2.5 Pro achieves 84 on the diamond set. This foundational capability gap likely explains much of the difference in analytical performance.

Moreover, we have verified that LAMBDA completes the task successfully when equipped with comparably powerful models. This suggests that much of the performance differential can be attributed to base model capability rather than fundamental architectural limitations.

2 The Problem of Irrelevant Reports

Professor Donoho raises an important observation that when Dr. Belogolovsky requested a report on the Electricity Cost Prediction dataset, LAMBDA returned results based on unrelated pre-existing datasets from the UCI repository rather than analyzing the novel dataset. This suggests the system may have defaulted to cached or training data rather than dynamically analyzing the input.

We take this observation seriously. The observation that LAMBDA produced an irrelevant report is informative. Professor Donoho suggested this may reflect an out-of-distribution (OOD) issue, since the dataset was released after the knowledge cutoff dates of the GPT-4/GPT-5 series. However, in our implementation, model outputs are driven primarily by the chat history, so we cannot rule out limitations in model capability, including possible OOD effects.

We hypothesize that the irrelevance may also stem from deficiencies in context construction, including retrieval and prompt assembly. To investigate this, we compared two approaches for structuring the reporting module's context: placing the generation instructions above the chat history (the original setting) and placing the instructions below the chat history (the new setting).

The results indicate that the new context configuration improves the robustness of report generation. Placing the generation instruction at the end of the dialogue context appears to be a more effective strategy. However, factors such as the model's underlying knowledge, its instruction-following ability, and the overall length of the dialogue may also influence performance to a lesser extent.

Since LAMBDA is designed to be compatible with most LLMs, using the latest and most capable models remains an effective way to mitigate such issues. We are also exploring enhanced context construction mechanisms that more reliably incorporate user-provided data.

3 The Role of Academic Research: Building Evaluations

Professor Donoho poses a fundamental question: "What ought academics be doing in this space? Should they be pursuing projects like LAMBDA?" He suggests that one area where academic work can make a lasting and principled impact is in evaluation rather than product development.

We find this perspective compelling and aligned with our own views on the proper division of labor between academia and industry. Designing benchmarks is not trivial, and it is 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, the academic community should take responsibility for articulating what we value in those systems: rigor, interpretability, fairness, and reproducibility.

We have recently tried to build a reliable, comprehensive benchmark for data agents. Our preliminary efforts involve constructing benchmark data derived from a corpus of statistical learning textbooks and supplemented by highly-voted, complex datasets from platforms such as Kaggle. This work has culminated in the development of DSAEval [Sun et al., 2026], which represents our most recent contribution to the field of data agent evaluation. Currently, we have collected more than 2,000 datasets encompassing over 10,000 tasks across diverse domains, covering various data modalities, problem domains, and task types.

Evaluating data agents requires broad coverage across domains, data types, and analytical methodologies. Because different disciplines rely on specialized analytic strategies, expanding benchmarks with domain-specific datasets is essential for assessing the generality and robustness of systems like LAMBDA.

As Professor Donoho noted, competitive platforms such as Chatbot Arena can capture real users' preferences and feedback. We believe such benchmarks are important for advancing rigor and reproducibility in the field.

Moving forward, we will release open benchmarks and process-based evaluation suites for data agents; conduct user studies on collaboration and learning outcomes in data science education; and propose standards for data-quality audits and reproducible reporting. Our aim is to complement industry's scale with academic rigor, transparency, and public goods that benefit the entire ecosystem.

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. The Chatbot Arena, developed at UC Berkeley, has become a go-to benchmarking framework for large language models. This shows how academia can shape not just research but practice and market behavior if it focuses on what it does best.

4 Engineering Considerations and Production Readiness

Professor Donoho's discussion also highlights engineering considerations that are critical for bridging the gap between academic prototypes and production-ready systems. We acknowledge that the transition from an academic prototype to a production-ready system requires rigorous engineering optimization.

At present, the core of LAMBDA is implemented using the Jupyter Python kernel. Restricting LAMBDA to a Python-only environment limits its adoption within the broader statistical community, where R remains a dominant language. We acknowledge this limitation and share this concern. We are actively working on expanding the supported programming languages and software within the environment. In future versions, we will integrate a formal

sandbox that will enable LAMBDA to execute shell commands as well as Python, R, Julia, SQL, and perform file-retrieval actions.

Moreover, the hybrid programming workflow suggested by other discussants can also be naturally supported by this sandbox framework. For example, an agent could perform data preprocessing in Python and conduct statistical modeling in R (using specialized packages such as `lme4`) within the same session. Such capabilities would significantly enhance the flexibility and applicability of LAMBDA.

We also recognize that agents with "Computer Use" capabilities pose significant risks, such as accidentally deleting critical user files. However, data agents like LAMBDA operate under a different risk profile. They generally do not require, nor should they be granted, direct control over the host operating system. A virtualized sandbox environment restricted to code execution is sufficient to meet analytical needs while isolating the system from destructive actions.

Privacy remains a critical bottleneck for commercial or API-based agents. These systems often require data uploads or allow the LLM to inspect raw data contents, creating inherent leakage risks. A trade-off solution is to deploy open-source LLMs locally, but this requires powerful hardware and incurs higher electricity costs. Thus, developing API-based solutions with privacy preservation remains an important research direction.

5 Conclusion

We thank Professor David Donoho for his insightful discussion that challenges us to think more deeply about the role of academia in the AI era. His perspective on the importance of building rigorous evaluation frameworks rather than competing directly with well-resourced commercial systems resonates with our vision for LAMBDA's ongoing development.

We acknowledge that academic resources are far more limited than those of venture-capital-backed startups building data-analysis agents, and that academia faces structural hurdles in matching industry-grade engineering and infrastructure. Even so, we believe that academic research can make lasting contributions by setting standards for rigor, transparency, and evaluation that shape the entire field. Our work on LAMBDA [Sun et al., 2025a] and our broader survey of the field [Sun et al., 2025b] exemplify this commitment to academic rigor and public-good research.

References

Maojun Sun, Ruijian Han, Binyan Jiang, Houduo Qi, Defeng Sun, Yancheng Yuan, and Jian Huang. Lambda: A large model based data agent. *Journal of the American Statistical Association*, pages 1–13, 2025a.

Maojun Sun, Ruijian Han, Binyan Jiang, Houduo Qi, Defeng Sun, Yancheng Yuan, and Jian Huang. A survey on large language model-based agents for statistics and data science. *The American Statistician*, pages 1–14, 2025b.

Maojun Sun, Yifei Xie, Yue Wu, Ruijian Han, Binyan Jiang, Defeng Sun, Yancheng Yuan,

and Jian Huang. Dsaeval: Evaluating data science agents on a wide range of real-world data science problems. *arXiv preprint arXiv:2601.13591*, 2026.