Response to Discussion 3: AI Agents for Scientific Inquiry and Human-AI Collaboration

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

We are grateful to Dr. Mert Yuksekgonul and Professor James Zou for their stimulating discussion of our work on LAMBDA [Sun et al., 2025a]. Their expertise in AI agents and biomedical applications provides valuable insights into how intelligent data analysis systems can contribute to scientific discovery. They contextualize LAMBDA within the broader landscape of AI agents for scientific inquiry and raise important considerations regarding human-AI collaboration. We address their key points below.

Our survey on large language model-based agents for statistics and data science [Sun et al., 2025b] provides a comprehensive overview of the current state of the field and the challenges that remain to be addressed.

1 AI Agents for Scientific Inquiry

Drs. Yuksekgonul and Zou provide an excellent framework for understanding the role of AI agents in scientific discovery, organizing their discussion around the main stages of the data science lifecycle: problem definition and exploration, hypothesis generation, design and execution of analyses, interpretation of results, and communication. This framework helps situate LAMBDA's current capabilities and identify opportunities for enhancement.

Regarding hypothesis generation, the authors note that LAMBDA is not designed or evaluated for hypothesis generation, and this would be a good direction for further work. We agree that hypothesis generation represents an important frontier for AI agents in science. Recent work has demonstrated that AI systems can generate ideas judged as more novel than those from human experts, though achieving genuine novelty requires open-ended search strategies and intrinsic motivation to explore under-examined but promising directions, balanced by safeguards against hallucinations.

For experimental design, given a hypothesis, the challenge shifts to operationalizing it into an experimental or analytical design. This may involve dataset construction or curation, metric selection, and baseline definition, while reasoning about causal structure, interventions, and potential confounders. LAMBDA fits naturally in this space, taking user inquiries, translating them into concrete setups, executing analyses, and iterating based on intermediate results to tighten the loop between hypothesis and evidence.

On data collection and analysis, rigorous data handling is essential to test hypotheses effectively. This includes robust data processing pipelines, correct execution of analyses, and appropriate modeling or statistical techniques. LAMBDA's open-source model usage and its Programmer-Inspector loop address some concerns by providing interpretable outputs and reducing data exposure risks. However, challenges remain in ensuring reliability, reproducibility, and privacy, especially with proprietary models.

For drawing conclusions and communication, synthesizing results into defensible claims requires linking conclusions to evidence, quantifying uncertainty, and situating findings within existing work. While LLMs can produce fluent narratives, maintaining fidelity to data and analyses is critical. LAMBDA partially addresses this by tying its conclusions directly to the analyses it executes and structuring them into verifiable formats, reducing the risk of unsupported or overstated claims.

We appreciate the authors' recognition that AI agents excel at combining and applying existing tools but are less capable of developing entirely new methodologies from scratch. Human researchers are uniquely positioned to create cutting-edge innovations, while AI agents can deploy these methods across datasets and applications. This complementary relationship helps address the challenge that many end-users either do not know which new tools to use or struggle with how to use them.

Similarly, LLMs are generally stronger in breadth than in depth, having ingested vast amounts of text across disciplines. Human experts excel at deep specialization that drives creative insights within a domain. The most fruitful collaborations will therefore combine the breadth of LLMs with the depth of human expertise.

At the same time, LLMs and agents remain prone to error, often because they lack full understanding of the context behind a dataset. If an agent is asked to analyze data without information on how it was collected and pre-processed, it may make invalid assumptions about biases or confounders. To mitigate this, it is crucial to capture and provide comprehensive metadata in formats digestible by agents and to routinely audit their assumptions and outputs to ensure validity.

2 Human-AI Interaction

Drs. Yuksekgonul and Zou raise an important research agenda: how to develop AI data agents that are good collaborators for human researchers. Current LLMs are optimized to be passive responders rather than active collaborators. When faced with ambiguities in the task, LLMs often make assumptions rather than asking the user to clarify. This can lead to wasted efforts and wrong conclusions.

We agree that this represents a critical area for improvement. Recent efforts like CollabLLM propose multi-turn training paradigms with new objectives to explicitly encourage LLMs to engage with the user and ask questions where helpful. Rather than unilaterally imputing missing data, for example, the agent would ask, "Do you prefer simple mean imputation or a more robust KNN approach?" This promotes the nuanced, conversational interactions essential for rigorous analysis.

The authors also note that LLMs can provide high-quality feedback to other LLMs, generating counterexamples, testing assumptions, and exploring parameter variations to strengthen conclusions. LAMBDA demonstrates an element of this through its iterative Programmer-Inspector loop, which could be expanded to reason more deeply about experimental design.

Regarding educational applications, beyond execution, AI agents can serve educational purposes, explaining analytical choices, surfacing statistical assumptions, and offering alternative interpretations to human users. LAMBDA's structured reports make these processes transparent. Incorporating interactive visualizations, derivations, and literature links could enhance this role. Agents could also simulate research workflows for training, guiding learners through problem definition, hypothesis generation, and analysis in realistic scenarios.

Professor Xiao-Li Meng's concept of "Mindware Agents" aligns closely with this educational vision. Mindware agents are tools designed not merely to generate outputs but to enhance users' data intelligence. This aligns with Drs. Yuksekgonul and Zou's view of agents as educational tools that simulate research workflows. We embrace this perspective and aim to design LAMBDA to nudge users toward better statistical practice.

We plan to implement a "Data Minder" as a dedicated Quality Control Agent that uses "Reverse Prompting." Rather than waiting for user input, it proactively presents a checklist of data-quality questions, for example, "What is the provenance of this dataset?" and "Are there potential selection biases in the collection process?" By embedding these nudges into the workflow, data quality assessment becomes a routine, integral part of the analysis.

The authors' observation that LLMs can hallucinate and make unspoken unwarranted assumptions about user intent is well-taken. For example, when asked to analyze data, agents might return reports with strategic recommendations that were not requested. This reflects a kind of premature autonomy where the AI projects intent rather than responding directly to user queries. Our confidence-based retrieval system and enhanced human-in-the-loop mechanisms are designed to mitigate these risks.

3 Complementarities and Future Directions

The discussion by Drs. Yuksekgonul and Zou highlights the complementarities between AI agents and human researchers. Current agents excel at combining and applying existing tools but are less capable of developing entirely new methodologies. Human researchers excel at deep specialization that drives creative insights within a domain. The most fruitful collaborations combine the breadth of LLMs with the depth of human expertise.

For example, a researcher might uncover a novel domain-specific insight, and the LLM could then systematically identify opportunities to apply that insight across other fields. This synergy represents the vision we are working toward with LAMBDA.

AI data scientist agents represent a shift toward computational partners in research, contributing to hypothesis generation, experiment design, data analysis, and results communication. Data science is an especially promising application domain for AI agents because current agents are particularly advanced in coding, math, and literature retrieval skills useful for data analysis and interpretation. Fully realizing the vision of AI data scientists will require further progress in reliability, adaptability, and integration with human workflows.

An important open challenge is to develop agents that are optimized to collaborate with human researchers. This requires careful interface design so that interacting with the

AI is intuitive and does not itself become a burden, and perhaps new training paradigms for AI agents so that they learn to defer to human input and explain their reasoning in understandable terms.

If successful, such human-AI collaboration could accelerate discovery in data-rich fields: the human provides contextual understanding and ethical oversight, while the AI provides speed, breadth of knowledge, and tireless exploration. LAMBDA's current human-in-loop feature is a foundational step in this journey. By expanding the ways in which humans can guide and partner with the AI, future systems will further amplify human value in the analytical process, ensuring that AI serves as a force multiplier for human intelligence rather than a mere replacement.

4 Conclusion

We thank Dr. Mert Yuksekgonul and Professor James Zou for their thoughtful discussion that situates LAMBDA within the broader context of AI agents for scientific inquiry. Their insights on hypothesis generation, experimental design, and human-AI collaboration are valuable guides for our ongoing research. We are committed to developing LAMBDA into a system that not only executes data analysis tasks but also serves as an educational partner and collaborative tool that enhances users' data intelligence.

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.