AI Agents for Data Science: a Discussion of "LAMBDA: A Large Model Based Data Agent"

Mert Yuksekgonul, James Zou

Stanford University

1 Introduction

An AI agent is an autonomous system that can interact with its environment, reason about tasks, take actions often by invoking external tools or APIs and adapt its behavior based on feedback. In scientific contexts, such agents could evolve from simply answering questions to contributing meaningfully to discovery: proposing hypotheses, designing and executing analyses, assessing uncertainty, and presenting defensible claims.

Early indications of these agentic capabilities are promising. With appropriate scaffolding and instructions, agents can write and execute code, perform search and retrieval over literature and datasets, use domain-specific tools, and engage in basic self-correction and verification. While still limited, these early developments signal an important shift from passive chat interfaces towards systems able to handle broader aspects of the scientific workflow under human guidance. The paper "LAMBDA: A Large Model Based Data Agent" [Sun et al., 2025a] illustrates further progress in this direction.

2 AI agents for scientific inquiry

To contextualize the role of AI agents in science, it is helpful to structure the 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. In the following, we consider each stage in turn, outlining how an AI data scientist agent could contribute and how systems like LAMBDA align with these contributions.

Hypothesis generation. Researchers have begun exploring AI systems that support hypothesis generation from observed data. For example, Si et al. [2024] report that ideas generated by AI were judged as more novel than those from human experts, and Yamada et al. [2025] describe AI-generated hypotheses accepted in workshop papers at machine learning conferences. Achieving genuine novelty will require open-ended search strategies and intrinsic motivation to explore under-examined but promising directions, balanced by safeguards against hallucinations. This involves combining exploration with retrieval and cross-checking against trusted sources, as well as efficient synthesis of relevant literature to identify substantive gaps [Singh et al., 2025]. LAMBDA is not designed or evaluated for hypothesis generation, and this would be a good direction for further work.

Design of experiments. 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. Recent work has explored agentic systems for adaptive experiment design in domains such as biology [Swanson et al., 2024] and chemistry [Bran et al., 2023]. LAMBDA fits naturally here taking user inquiries, translating them into concrete setups, executing analyses, and iterating based on intermediate results to tighten the loop between hypothesis and evidence.

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. Examples of AI agents for such tasks include reanalyzing public biological datasets to uncover new findings [Alber et al., 2025], using case-based reasoning from past Kaggle solutions to guide iterative experiments [Guo et al., 2024], and falsifying hypotheses via data-driven analysis in varied domains [Huang et al., 2025]. There remain challenges in ensuring reliability, reproducibility, and privacy, especially with proprietary models. LAMBDA's open-source model usage and its Programmer-Inspector loop address some of these concerns by providing interpretable outputs and reducing data exposure risks.

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. Moreover, it's easy for LLMs to be overly confident in framing its findings. 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.

3 Human-AI Interaction

It is important to recognize both the limitations of AI agents and their complementarities with human researchers. Current agents excel at combining and applying existing tools, but they are less capable of developing entirely new methodologies from scratch.

Human researchers, by contrast, are uniquely positioned to create cutting-edge innovations such as a novel method for multi-omics integration or the next generation of AlphaFold while AI agents can deploy these methods across datasets and applications. This helps address a persistent challenge in method development: many end-users either do not know which new tools to use or struggle with how to use them. Here, the agent can serve as the interface between novel methods and a heterogeneous user community.

Similarly, LLMs are generally stronger in breadth than in depth, having ingested vast amounts of text across disciplines. Human experts, meanwhile, 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. 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.

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.

Because of the complementarities discussed above, an important research agenda is how to develop AI data agents that are good collaborators for human researchers. Current LLMs are optimized to directly answer queries rather to engage in multi-term collaboration with human users [Wu et al., 2025]. For example, 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. Recent efforts like CollabLLM proposed multi-turn training paradigm with new objectives to explicitly encourage LLM to engage with the user and ask questions where helpful.

LLMs can also provide high-quality feedback to other LLMs [Yuksekgonul et al., 2025], 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.

Beyond execution, AI agents can serve educational purposes [Wang et al., 2024], 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.

4 Conclusion

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 that are useful for data analysis and interpretation. Fully realizing the vision of AI data scientist 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.

References

Samuel Alber, Bowen Chen, Eric Sun, Alina Isakova, Aaron J Wilk, and James Zou. Cellvoyager: Ai compbio agent generates new insights by autonomously analyzing biological data. *bioRxiv*, pages 2025–06, 2025.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

Siyuan Guo, Cheng Deng, Ying Wen, Hechang Chen, Yi Chang, and Jun Wang. Ds-agent: Automated data science by empowering large language models with case-based reasoning. *arXiv preprint arXiv:2402.17453*, 2024.

Kexin Huang, Ying Jin, Ryan Li, Michael Y Li, Emmanuel Candès, and Jure Leskovec. Automated hypothesis validation with agentic sequential falsifications. *arXiv preprint arXiv:2502.09858*, 2025.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*, 2024.

Amanpreet Singh, Joseph Chee Chang, Chloe Anastasiades, Dany Haddad, Aakanksha Naik, Amber Tanaka, Angele Zamarron, Cecile Nguyen, Jena D Hwang, Jason Dunkleberger, et al. Ai2 scholar qa: Organized literature synthesis with attribution. *arXiv preprint arXiv:2504.10861*, 2025.

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.

Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. The virtual lab: Ai agents design new sars-cov-2 nanobodies with experimental validation. *bioRxiv*, pages 2024–11, 2024.

Rose E Wang, Ana T Ribeiro, Carly D Robinson, Susanna Loeb, and Dora Demszky. Tutor copilot: A human-ai approach for scaling real-time expertise. *arXiv preprint arXiv:2410.03017*, 2024.

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. Collabllm: From passive responders to active collaborators. *arXiv preprint arXiv:2502.00640*, 2025.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Pan Lu, Zhi Huang, Carlos Guestrin, and James Zou. Optimizing generative ai by backpropagating language model feedback. *Nature*, 639(8055):609–616, 2025.