https://doi.org/10.1038/s42256-026-01183-2

Multi-agent AI systems need transparency

Agentic artificial intelligence (AI) frameworks are in vogue. However, implementing such systems in scientific research workflows requires clear motivations and explanations, given the risk of wasting computational as well as human resources.

Now that large language models (LLMs) are adopted in most scientific domains, the focus is shifting to multi-agent AI systems, which can be loosely defined as ensembles of LLMs that interact with each other, that have access to tools and databases and that can solve tasks in a multi-step way. Arguably, such virtual teams made up of LLMs could perform the type of reasoning and planning that is required for research, with the added benefit that they can communicate with human collaborators in natural language[1]. A growing number of research groups are excited by the idea, and the past year has seen reports in different domains on implementations of multi-agent LLM frameworks that can semi-autonomously generate and refine research ideas.

For example, researchers reported last summer that a virtual team of LLM agents, with humans in the loop, could discover new antibodies to bind to the recent variants of SARS-CoV-2[2]. In the framework, an LLM-powered AI agent, the 'Principal Investigator', receives a research question from a human researcher, and in response assembles a team of LLM-based researchers with different expertise. The Principal Investigator guides the LLM experts through virtual research meetings to decide on promising research directions and design solutions, and the findings are sent to the human researcher for further feedback. As a result, the framework came up with 92 potential antibodies, and after experimental testing by the authors, two new promising candidates were identified. The authors highlight the importance of the diversity of expertise in the LLM team.

In another example, researchers developed a multi-agent LLM approach for materials screening. This framework involves a supervisor LLM that receives a research objective as input and assigns tasks to a team of specialist LLMs, in this case ones that can perform atomistic calculations, allocate computational resources and fix errors[3].

Several companies are taking such concepts from ad hoc designs to off-the-shelf tools intended for fully automated discovery. Examples are automated science frameworks such as AI Scientist[4] and Kosmos[5] from start-up companies Sakana AI and Edison Scientific, respectively. These platforms can take an open-ended research question and run a full research cycle from literature search, data analysis and hypothesis generation to writing up scientific papers. AI Scientist so far focuses on machine learning research, while the developers of Kosmos report discoveries in several disciplines. In addition, Google Research developed 'AI co-scientist', powered by the Gemini 2.0 LLM, which is intended to be a collaborator, continually generating, reviewing and refining research hypotheses.

There is substantial excitement about the potential for multi-agent LLM platforms to accelerate scientific discovery, as such systems could race through research cycles, developing, evaluating and testing research hypotheses much faster than humans. But such approaches are not without challenges. One major concern is the complexity and increasingly high computational cost of large scale multi-agent systems. A related problem is that it is unclear how to systematically test and evaluate them. Are comprehensive ablation and baseline comparisons even possible? Moreover, LLMs suffer from prompt fragility, in that subtly different inputs lead to different results, an effect that will be magnified by using ensembles of LLMs. Moreover, LLMs make odd mistakes and can make up facts (hallucinations), which means that human oversight — scientists in the loop — is needed at different stages.

To address such concerns, when using and reporting on multi-agent LLM approaches, researchers should dedicate special efforts towards transparency. The motivation behind the framework and the specific design choices should be clearly discussed. Documentation on workflows, model versions, prompts and reasoning chains should be provided, and the level of human oversight and manual steps that are required should be transparently reported. Although comprehensive evaluations, ablation analyses and baseline comparisons will be challenging, attempts to test how specific components contribute to the framework could be expected, and at a minimum, authors should include a comparison to single-agent LLM approaches. Finally, authors should ideally address the overall question of whether the computational cost and complexity are justified and, for example, contrast the automated system with manual, expert-lead workflows in terms of speed and accuracy.

The possibility of accelerating science with multi-agent LLM systems is certainly tantalizing. At the same time, there is a fine line between the expectation that such systems could be valuable collaborators that save human researchers time and effort and the risk that researchers are wasting resources on an endless cycle of checking and curating output from complex LLM systems. At this early stage of their development, we recommend gaining clarity about the role and value of multi-agent LLM tools in the research workflow.

Published online: 27 January 2026

References

1. Xin, H., Kitchin, J. R. & Kulik, H. J. *Nat. Mach. Intell.* **7**, 1373–1375 (2025).
2. Swanson, K. et al. *Nature* **646**, 716–723 (2025).
3. Wang, Z. et al. Preprint at https://doi.org/10.48550/arXiv.2507.14267 (2025).
4. Lu, Z. et al. Preprint at https://doi.org/10.48550/arXiv.2408.06292 (2024).
5. Mitchener, L. et al. Preprint at https://doi.org/10.48550/arXiv.2511.02824 (2025).