

Large Language Model Agents for Autonomous Scientific Discovery and Experiment Design

Anais Breton

Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.

anaiswork@uab.edu

Asmianto Halimah

Department of Computer Science, Binghamton University, Binghamton, NY, USA.

halimah1980@binghamton.edu

Tymoteusz Dzierwa

Department of Computer Science, University of North Texas, Denton, TX, USA.

tymoteusz.dzierwa@unt.edu

Abstract

Large language model agents are increasingly proposed as autonomous collaborators in scientific discovery, not merely as interfaces for literature search or code generation but as systems capable of planning experiments, calling domain tools, interpreting results, and revising hypotheses. This paper examines the system-level conditions under which such agents can support credible autonomous scientific discovery and experiment design. It argues that the central challenge is not model capability alone but the integration of reasoning, memory, tool use, provenance, laboratory execution, governance, and long-term maintenance into robust socio-technical infrastructures. The paper reviews the evolution from assistive language models to agentic discovery systems, analyzes layered architectures that connect foundation models to instruments and data repositories, and discusses closed-loop experiment design in chemistry, materials science, biology, and computational domains. It then evaluates infrastructure and sustainability constraints, including compute intensity, energy demand, and technical debt, before addressing robustness, security, reproducibility, fairness, and policy. The analysis suggests that autonomous discovery agents must be evaluated not only by novelty or predictive accuracy but also by reproducibility, auditability, safety, access equity, and contribution to public knowledge. A governance model combining provenance standards, model and dataset documentation, human oversight, and domain-specific validation is essential. The paper concludes that large language model agents can meaningfully expand scientific capacity only when they are embedded in accountable infrastructures that treat discovery as a collective, cumulative, and socially situated process rather than as an isolated optimization problem.

Keywords

large language model agents, autonomous scientific discovery, experiment design, research infrastructure, reproducibility, governance.

1. Introduction

Scientific discovery has long depended on instruments, institutions, and information systems that coordinate observation, hypothesis, experimentation, and validation. The emergence of

large language models has introduced a new possibility: software agents that can read literature, propose hypotheses, write code, invoke simulations, and interpret intermediate results with minimal human prompting. Early work on general-purpose language models demonstrated broad reasoning and task-transfer behavior that suggested new forms of machine assistance in knowledge work [1]. Subsequent surveys of language model agents have described architectures that combine planning, memory, reflection, and tool use, thereby moving beyond single-turn question answering toward persistent goal-directed behavior [2]. These developments have encouraged the view that agents may eventually participate in the full cycle of scientific discovery rather than only in isolated computational subtasks [3].

However, the transition from assistive model to autonomous discovery agent is not a simple matter of increasing model scale or improving benchmark performance. It requires system-level design choices about how agents represent scientific problems, how they access instruments and databases, how they justify claims, how they recover from error, and how their outputs are governed. Historical work on robot scientists already explored closed-loop hypothesis generation and experimentation in biology, showing both the promise and the difficulty of automating scientific reasoning in physical laboratories [12]. Later perspectives on artificial intelligence for science argued that machine learning could amplify discovery by identifying patterns, prioritizing experiments, and integrating heterogeneous data [14]. The contemporary challenge is therefore to understand how large language model agents can inherit these ambitions while avoiding the fragility, opacity, and accountability gaps that accompany generative systems.

This paper focuses on autonomous scientific discovery and experiment design as a system-level problem. It examines architectural patterns, deployment constraints, robustness requirements, fairness concerns, and policy implications. The central argument is that credible autonomous discovery requires more than capable models. It requires an infrastructure of validated tools, structured data, provenance records, human oversight, and domain-specific governance. The sections that follow analyze how such infrastructures can be designed and what trade-offs they introduce.

2. From Assistive Models to Agentic Discovery Systems

The first generation of language model applications in science functioned as assistive interfaces. Researchers used them to summarize papers, generate code, draft explanations, and suggest possible hypotheses. Such uses can be valuable, but they remain bounded by human framing and human validation. Agentic systems differ because they maintain goals over time, select actions, observe outcomes, and modify plans. The ReAct pattern illustrated how reasoning traces can be interleaved with actions in external environments, allowing a model to search, calculate, and interact with tools rather than only produce text [4]. Toolformer and related approaches showed that language models can learn to call application programming interfaces and external utilities, thereby extending their effective competence beyond parametric memory [5]. These capabilities are foundational for scientific agents because scientific work depends on instruments, simulators, databases, and statistical software.

Domain-specific agent systems have moved this pattern into specialized scientific settings. In chemistry, agents augmented with computational tools can plan syntheses, query reaction databases, and reason about molecular properties [6]. Autonomous chemical research systems have demonstrated that language models can coordinate experimental planning and robotic execution in limited but meaningful domains [7]. These examples reveal a crucial architectural insight: the scientific agent is not the model alone. It is a composite system in

which the language model acts as an orchestrator among symbolic tools, domain models, robotic platforms, and human supervisors. The quality of the agent therefore depends on interface design, tool reliability, data schemas, and error handling as much as on the underlying language model.

This shift also changes the nature of evaluation. A model that produces plausible text can be assessed by linguistic fluency or factual accuracy. An agent that designs experiments must be assessed by whether its proposals are physically feasible, statistically sound, ethically permissible, and reproducible. It must also be assessed by whether it knows when to stop, when to ask for help, and when to reject its own hypothesis. These requirements place agentic discovery systems closer to safety-critical engineering than to conventional text generation.

3. System Architecture and Integration in Scientific Workflows

A useful way to understand autonomous discovery agents is as layered systems. The foundation layer includes one or more large language models, possibly fine-tuned for scientific reasoning or equipped with retrieval mechanisms. The orchestration layer manages planning, task decomposition, memory, reflection, and coordination among subagents. The tool layer exposes scientific software, databases, simulators, and laboratory instruments through controlled interfaces. The data layer stores experimental records, provenance metadata, embeddings, and structured results. The execution layer connects digital plans to physical or computational experiments. The governance layer provides access control, audit trails, validation checks, and policy enforcement. Failures at any layer can compromise the entire discovery process.

Integration with existing scientific workflows is particularly difficult. Laboratory information management systems, electronic lab notebooks, instrument controllers, and high-performance computing environments often use incompatible schemas and authentication models. An agent that can generate a protocol in natural language may still be unable to execute it safely or record it in a way that satisfies institutional requirements. The hidden technical debt of machine learning systems offers a cautionary analogy: data dependencies, configuration issues, feedback loops, and undeclared consumers can accumulate invisibly and undermine long-term reliability [18]. Autonomous discovery agents are likely to create similar debt unless provenance, versioning, and interface contracts are treated as first-class design concerns.

Case illustrations from computational biology and materials science show both the value and the limits of integration. AlphaFold demonstrated how a specialized model can transform a scientific field when it is coupled to high-quality data and community validation [9]. Materials discovery efforts have used machine learning to screen vast candidate spaces and prioritize promising compounds [10]. Autonomous laboratories have further linked computational prediction to robotic synthesis and characterization [11]. These systems succeed because they narrow the problem domain, enforce structured representations, and embed human review at critical points. General-purpose language model agents must learn from these constraints rather than assuming that broader language competence automatically yields scientific reliability.

4. Autonomous Experiment Design and Closed-Loop Discovery

Experiment design is where autonomous agents face their most demanding test. A scientific experiment must operationalize a hypothesis, select variables and controls, specify measurements, account for uncertainty, and satisfy safety and ethical constraints. Language

model agents can assist by drafting protocols, generating candidate hypotheses, and writing analysis code. In closed-loop systems, they can also interpret results and propose follow-up experiments. Chemistry agents have shown that tool-augmented models can propose syntheses and evaluate feasibility, while robotic systems have demonstrated automated synthesis and characterization in selected materials domains [6][7][11]. These examples suggest that closed-loop discovery is feasible when the search space is bounded and the evaluation function is reliable.

The hardest problems arise when evaluation is ambiguous or expensive. In many scientific fields, ground truth is not immediately available, experiments are noisy, and negative results are underreported. An agent optimized for novelty may generate many plausible but untestable hypotheses. An agent optimized for experimental success may exploit known reactions and avoid genuine discovery. A recent full-scenario agentic scaling effort points toward broader task coverage and multi-domain adaptability, but it also highlights the difficulty of maintaining reliability across heterogeneous scientific environments [15]. The design challenge is therefore to balance exploration and exploitation, novelty and reproducibility, and autonomy and human oversight.

Closed-loop discovery also raises questions about credit, error, and recovery. If an agent proposes an experiment that fails, the failure may be scientifically informative, but only if the system records the reasoning, parameters, and environmental conditions. If an agent accidentally produces a hazardous compound or misuses an instrument, accountability cannot rest solely with the model. Robust experiment design therefore requires safety envelopes, simulation-first validation, human approval gates, and rollback mechanisms. The more autonomous the loop becomes, the more important it is to encode constraints before action rather than to rely on post hoc review.

5. Infrastructure, Deployment, and Sustainability

Deploying large language model agents for science requires substantial infrastructure. Inference, retrieval, fine-tuning, simulation, and data storage all consume computational resources. In laboratory settings, agents must interact with instruments that may have long operation times, limited throughput, and strict safety requirements. In computational settings, agents may need access to high-performance computing clusters, specialized accelerators, and large scientific datasets. The infrastructure must support reproducibility, which means capturing model versions, prompts, tool versions, data snapshots, and execution environments. Without such records, an apparently successful discovery may be impossible to reproduce or audit.

Sustainability is both an environmental and an institutional concern. Large-scale model training and inference have significant energy and carbon implications, and scientific users must weigh those costs against the value of the resulting discoveries. At the same time, the less visible sustainability problem is maintenance. Scientific software often depends on fragile grant-funded components, undocumented scripts, and discontinued services. An autonomous discovery agent that relies on dozens of external tools may become unusable within a few years if those tools are not maintained. Sustainable design therefore favors modular interfaces, open standards, and community governance rather than tightly coupled proprietary pipelines [16]. It also favors shared infrastructure that can serve multiple laboratories rather than duplicating expensive resources.

Deployment models will likely vary by domain. Some fields may use cloud-hosted agents with centralized governance and strong auditability. Others may require on-premises systems to protect sensitive data or to control instrument latency. Federated approaches could allow institutions to share models and tools while retaining local control over data and physical experiments. Each model involves trade-offs among cost, privacy, performance, and control. The key is not to select one universal architecture but to make the trade-offs explicit and to design governance mechanisms that match the risk profile of the scientific domain.

6. Robustness, Security, and Reproducibility

Robustness is a precondition for trustworthy autonomy. Language model agents can fail through hallucination, tool misuse, prompt injection, data contamination, and unstable planning. In scientific contexts, such failures are not merely inconvenient. They can waste scarce experimental resources, damage instruments, produce misleading conclusions, or create safety hazards. Ethical and social risk analyses of language models have identified a wide range of harms related to misinformation, bias, privacy, and malicious use [17]. These risks are amplified when agents are given access to laboratory equipment or sensitive data. Security must therefore be designed into the architecture, including authentication, authorization, sandboxing, rate limiting, and anomaly detection.

Reproducibility requires more than storing code. It requires capturing the full context of an experiment, including the agent's goals, prompts, tool calls, intermediate results, and human interventions. Datasheets for datasets and model cards for model reporting provide useful templates for documenting the assumptions, limitations, and intended uses of data and models [19][20]. In agentic systems, similar documentation should extend to tool registries, action policies, and evaluation protocols. Provenance standards such as persistent identifiers, semantic metadata, and immutable logs can help link claims to evidence. Reproducibility also depends on negative results and failed trials being recorded, because selective reporting distorts the scientific record.

A robust agent should be able to recognize uncertainty and request human assistance. It should distinguish between a failed experiment, a flawed hypothesis, and a tool malfunction. It should maintain multiple hypotheses rather than collapsing prematurely onto a single explanation. These capabilities are not guaranteed by scale. They require explicit training, evaluation, and system design. Robustness testing should include adversarial scenarios, distribution shifts, instrument failures, and ambiguous instructions. In high-stakes domains, redundancy and independent verification may be necessary before an agent's output is accepted as scientific evidence.

7. Fairness, Governance, and Policy

Autonomous discovery agents raise fairness questions at multiple levels. Access to advanced models, compute, and robotic laboratories is unevenly distributed across countries, institutions, and disciplines. If agentic discovery becomes concentrated in well-resourced settings, it may widen existing inequalities in scientific capacity. Fairness also concerns data. Models trained predominantly on English-language literature and well-funded research areas may underrepresent local knowledge, indigenous perspectives, and low-resource scientific communities. Governance must therefore consider not only safety but also distributive justice and the right to participate in the production of knowledge.

Policy frameworks will need to address authorship, intellectual property, liability, and publication norms. If an agent contributes to a discovery, how should credit be assigned? If an

agent makes an error, who is responsible? If an agent generates a patentable invention, who owns it? Current scientific norms assume human accountability, but agentic systems blur the boundary between tool and collaborator. Governance models should require human principal investigators to retain responsibility for experimental design, safety, and interpretation. They should also require disclosure of agent involvement, tool versions, and validation procedures. Regulatory approaches may vary by domain, but common principles of transparency, auditability, and contestability can provide a baseline.

Governance should be adaptive rather than static. As agents become more capable, the appropriate level of oversight may change. A tiered approach could require stricter review for physical experiments, hazardous materials, human subjects, and dual-use research, while allowing greater autonomy for low-risk computational exploration. International coordination will be important because scientific data and models cross borders. Funding agencies, journals, professional societies, and laboratories all have roles to play in setting standards. The goal is not to prevent autonomy but to make it accountable. Trustworthy autonomous discovery depends on institutions that can verify claims, assign responsibility, and correct errors.

8. Evaluation and Benchmarking

Evaluating autonomous discovery agents is difficult because scientific success is multidimensional. Accuracy on a benchmark may not predict the ability to design a useful experiment. Novelty may be difficult to verify, especially when prior knowledge is dispersed across languages and disciplines. Reproducibility may require years of follow-up work. Cost, safety, and environmental impact also matter. Evaluation frameworks should therefore combine computational benchmarks with expert review, laboratory trials, and longitudinal studies. They should assess not only whether an agent can solve a predefined task but also whether it can identify worthwhile problems and recognize when a problem is beyond its competence.

Benchmarks should include adversarial and open-ended settings. In closed-world tasks, agents can be scored against known answers. In open-ended discovery, the ground truth may not exist yet. Evaluation may instead focus on process quality: whether hypotheses are falsifiable, protocols are complete, controls are appropriate, and provenance is sufficient for reproduction. The AI Scientist and related efforts have provoked debate about whether fully automated discovery is desirable or even meaningful, underscoring the need for evaluation criteria that reflect scientific values rather than only performance metrics [8]. Materials discovery platforms provide another model, where computational predictions are validated through synthesis and characterization [10]. Cross-domain comparisons can reveal which architectural choices generalize and which remain domain-specific.

Evaluation should also consider human-machine team performance. In many cases, the most effective configuration may not be full autonomy but dynamic collaboration, where the agent handles routine tasks and the human provides judgment, creativity, and ethical oversight. Metrics for such collaboration include trust calibration, workload, error detection, and knowledge transfer. The goal is not to replace scientists but to expand the space of questions they can ask and the speed at which they can test them.

9. Cross-Domain Comparisons and Forward-Looking Perspectives

Different scientific domains present different constraints for agentic discovery. Chemistry and materials science offer structured representations, simulation tools, and robotic platforms, making them relatively amenable to closed-loop automation [6][7][10][11]. Biology combines

high-throughput data with complex causal mechanisms, ethical review, and variability across organisms, making full autonomy more difficult. Computational fields such as machine learning may allow faster iteration, but they also suffer from benchmark overfitting and reproducibility problems. Social science and medicine introduce human subjects, privacy, and interpretive complexity that require strong governance and community engagement. A general-purpose discovery agent must therefore be adaptable without pretending that all domains share the same epistemic standards.

Forward-looking systems may combine multiple agents with specialized roles. A planner agent might decompose a research question, a retrieval agent might gather literature, a simulation agent might test hypotheses, a safety agent might screen for hazards, and a critic agent might evaluate reproducibility. Multi-agent architectures can improve robustness by distributing reasoning and enabling checks and balances. However, they also increase coordination costs, failure modes, and the difficulty of attribution. The design of agent societies will be as important as the design of individual agents. Governance must ensure that responsibility does not disappear into the interactions among autonomous components.

The long-term promise of large language model agents in science lies in their potential to democratize access to sophisticated methods, accelerate hypothesis testing, and integrate fragmented knowledge. The long-term risk is that they amplify existing inequalities, produce unreliable claims at scale, and erode public trust in science. The difference between these outcomes will depend on infrastructure and institutions. Open standards, shared benchmarks, transparent documentation, and inclusive governance can help steer development toward public benefit. If agentic discovery remains a proprietary race among a few well-resourced actors, its benefits may be narrow and its risks widely distributed.

10. Conclusion

Large language model agents are becoming plausible participants in scientific discovery and experiment design, but their value depends on system-level design rather than model capability alone. This paper has argued that autonomous discovery requires layered architectures that connect reasoning models to validated tools, structured data, laboratory instruments, and human oversight. It requires robust provenance, reproducibility, security, and safety mechanisms. It requires sustainable infrastructure that can be maintained over time and governed in the public interest. It also requires fairness considerations that address access, representation, and accountability.

The most credible path forward is not full autonomy without oversight but graduated autonomy within accountable infrastructures. Low-risk computational tasks may be delegated to agents with minimal review, while physical experiments, sensitive data, and dual-use research require stronger controls. Evaluation should assess process quality, reproducibility, and societal impact alongside novelty and accuracy. Governance should be adaptive, domain-specific, and internationally coordinated. If these conditions are met, large language model agents can expand scientific capacity in ways that are robust, equitable, and trustworthy. If they are ignored, agentic discovery may produce a surplus of plausible claims and a deficit of reliable knowledge.

References

1. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023).

Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712.

2. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J.-R. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), 186345.
3. Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., ... & Gui, T. (2023). The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864.
4. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629.
5. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 68539-68551.
6. Bran, A. M., Cox, S., Schilter, O., Baldassari, C., White, A. D., & Schwaller, P. (2024). Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5), 525-535.
7. Boiko, D. A., MacKnight, R., Kline, B., & Gomes, G. (2023). Autonomous chemical research with large language models. *Nature*, 624(7992), 570-578.
8. Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., & Ha, D. (2024). The AI Scientist: Towards fully automated open-ended scientific discovery. arXiv preprint arXiv:2408.06292.
9. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583-589.
10. Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., & Cubuk, E. D. (2023). Scaling deep learning for materials discovery. *Nature*, 624(7990), 80-85.
11. Szymanski, N. J., Rendy, B., Fei, Y., Kumar, R. E., He, T., Milsted, D., McDermott, M. J., Gallant, M., Cubuk, E. D., Merchant, A., Kim, H., Jain, A., Bartel, C. J., Persson, K., Zeng, Y., & Ceder, G. (2023). An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624(7990), 86-91.
12. King, R. D., Whelan, K. E., Jones, F. M., Reiser, P. G., Bryant, C. H., Muggleton, S. H., Kell, D. B., & Oliver, S. G. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427(6971), 247-252.
13. Sparkes, A., Aubrey, W., Byrne, E., Clare, A., Khan, M. N., Liakata, M., Markham, M., Rowland, J., Soldatova, L. N., Whelan, K. E., Young, M., & King, R. D. (2010). Towards robot scientists for autonomous scientific discovery. *Automated Experimentation*, 2(1), 1.

14. Gil, Y., Greaves, M., Hendler, J., & Hirsh, H. (2014). Amplify scientific discovery with artificial intelligence. *Science*, 346(6206), 171-172.
15. Jiang, H., Xin, G., Huang, Y., Zhu, G., Zhang, J., Lin, X., ... & Cao, Z. (2026). AgentOmnia: Scaling Agentic Models for Full-Scenario Applications. *arXiv preprint arXiv:2607.23124*.
16. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
17. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.
18. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*, 28, 2503-2511.
19. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86-92.
20. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220-229.