Perspective | **Open Access**

AI Agent

Li. *AI Agent* **2026**, *2*, 34 **DOI:**10.20517/aiagent.2026.37

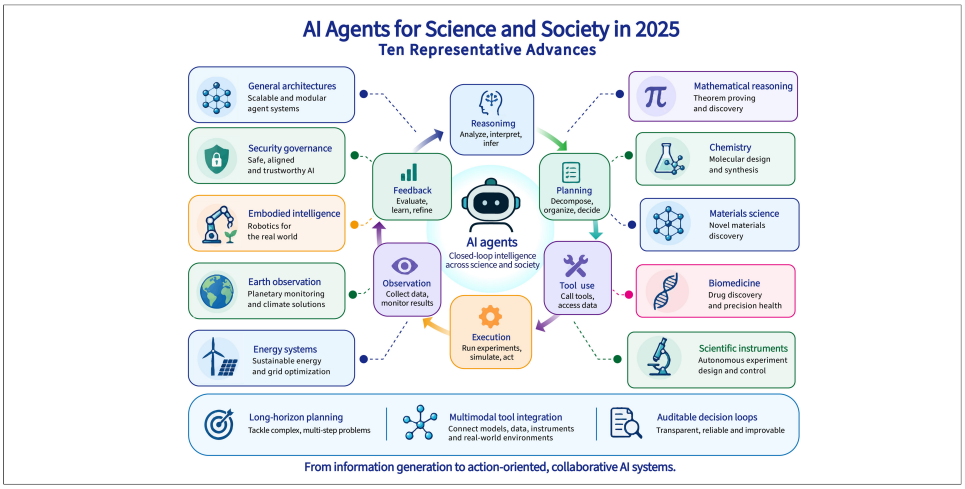
AI agents for science and society in 2025: ten representative advances

Limin Li[1,2]

Citation: Li, L. AI agents for science and society in 2025: ten representative advances. *AI Agent* **2026**, *2*, 34. https://dx.doi.org/10.20517/aiagent.2026.37

Received: 29 Jun 2026 **First Decision:** 22 Jul 2026 **Revised:** 14 Oct 2026 **Accepted:** 22 Oct 2026 **Published:** 18 Nov 2026

Academic Editor: Jianjun Liu **Copy Editor:** Tong Wang **Production Editor:** Tong Wang



Artificial intelligence (AI) agents are transitioning from conversational add-ons to integrated scientific operating layers, reflecting a broader shift toward autonomous systems capable of reasoning, planning, and interacting with external tools[1]. Advances around 2025 instead embed models in structured loops for planning, tool use, feedback, and revision. This Perspective highlights ten representative advances spanning general agent architectures, mathematical reasoning, chemistry, materials science, biomedicine, scientific instrumentation, energy systems, Earth observation, embodied intelligence, and security governance. Collectively, these developments mark a transition from fluent text generation toward controllable, auditable, and increasingly autonomous cycles of reasoning and action.

To identify representative advances in AI agents for science and society, we considered four criteria: scientific significance, agentic functionality, cross-domain representation, and demonstrated progress and influence during 2024-2025. Particular emphasis was placed on systems that move beyond passive information generation by combining planning, tool utilization, feedback-driven adaptation, and iterative decision-making with scientific data, computational resources, physical environments, and human expertise. The selected advances represent major developmental directions rather than a ranked list of individual studies.

[1]Suzhou MatSource Technology Co., Ltd., Suzhou 215000, Jiangsu, China.
[2]Gusu Laboratory of Materials, Suzhou 215000, Jiangsu, China.

Correspondence to: Dr. Limin Li, Suzhou MatSource Technology Co., Ltd., Suzhou 215000, Jiangsu, China. E-mail: hotline@matsourceai.com

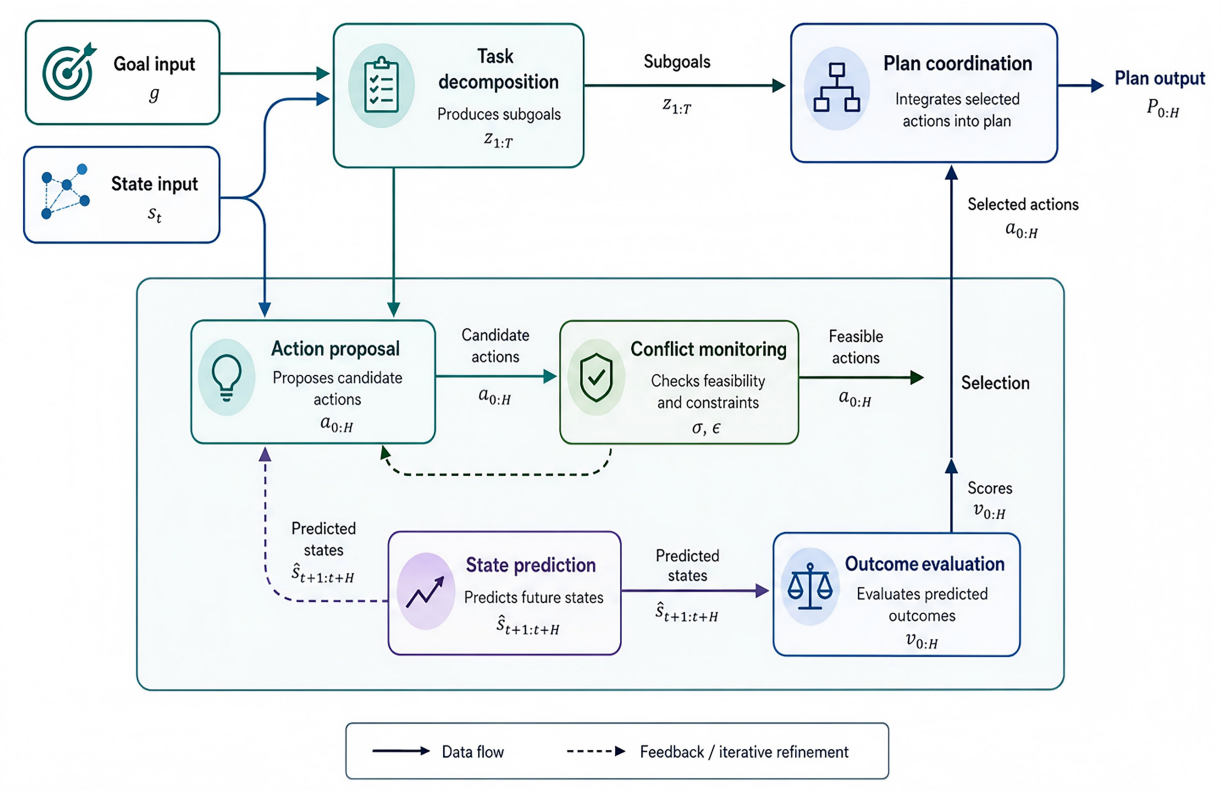


Figure 1. Modular planning architecture linking goal decomposition, action proposal, constraint checking, state prediction, outcome evaluation, and feedback.

FOUNDATIONS AND GENERAL ARCHITECTURES OF AI AGENTS

Modular planning and long-horizon agent architectures

Long-horizon agency exposes a fundamental limitation of monolithic prompting: a single stream of text must simultaneously retain goals, track state, follow constraints, and anticipate long-term consequences. Recent agent architectures have therefore explored explicit reasoning traces, external memory, and iterative interaction mechanisms to improve long-horizon execution[2-4]. Errors may accumulate over extended tasks, motivating inspectable architectures[1,5]. Figure 1 summarizes this emerging design pattern.

Rather than relying on an opaque prompt, modular systems separate functions such as goal decomposition, action proposal, feasibility monitoring, future-state prediction, outcome evaluation, and plan coordination[6,7]. Feedback can be routed back to earlier stages, turning planning into a controlled search process and making intermediate assumptions visible[6,8].

The main significance of this architecture lies in its debuggability and extensibility: individual modules can be tested, replaced, or constrained independently, while the agent can be coupled to simulators, specialized tools, and long-term memory[6-8]. Robust coordination under noisy observations, non-stationary environments, and limited computational resources remains an important challenge[5-8].

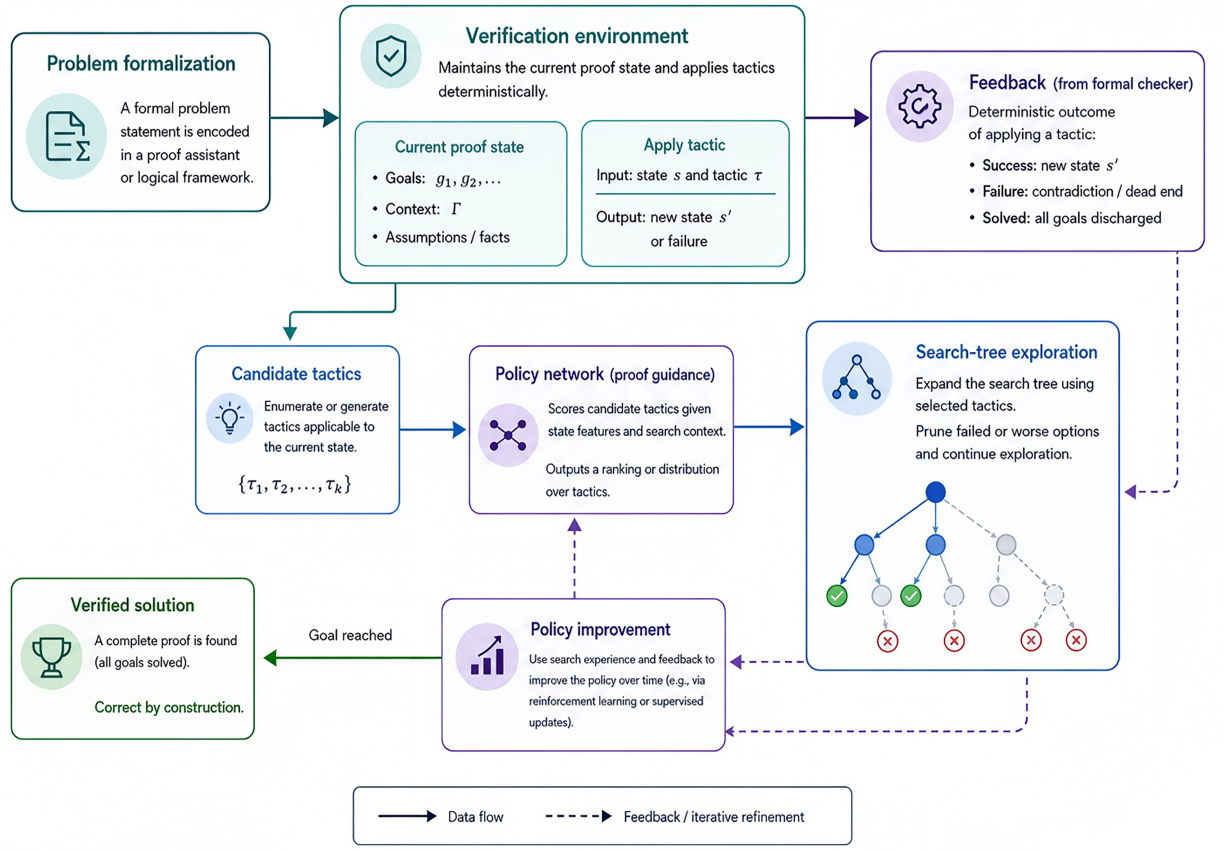


Figure 2. Formal reasoning loop combining problem formalization, tactic selection, verification feedback, search, and policy improvement.

MATHEMATICAL REASONING AND FORMAL METHODS

Verifiable reasoning through formal methods and search

Mathematical agents require verification protocols that go beyond semantic plausibility. Because a single-step error can invalidate a long logical chain, recent research has coupled large language models (LLMs) with proof assistants, formula checkers, and explicit search algorithms to enforce correctness through deterministic environments rather than statistical plausibility alone[9-11]. The verification loop is summarized in Figure 2.

A typical workflow formalizes a natural-language problem into a symbolic representation, proposes proof tactics, evaluates them within a verification engine, and expands the search over valid proof states. Checker feedback identifies progress, failure, or proof completion. These trajectories can subsequently improve search policies and iterative reasoning.

This framework shifts agentic reasoning toward auditable computation and is relevant to theorem proving, program verification, symbolic derivation, and other scientific workflows governed by hard constraints. Remaining bottlenecks include the cost of formalization, expanding search spaces, and the interpretability gap between machine-checkable proofs and human mathematical intuition.

CHEMICAL SCIENCES AND AUTONOMOUS DISCOVERY

Multi-agent orchestration and closed-loop chemical research

Autonomous chemistry requires coordination across diverse representations and infrastructures, including literature, reaction logic, experimental protocols, instruments, measurements, and predictive models. Recent

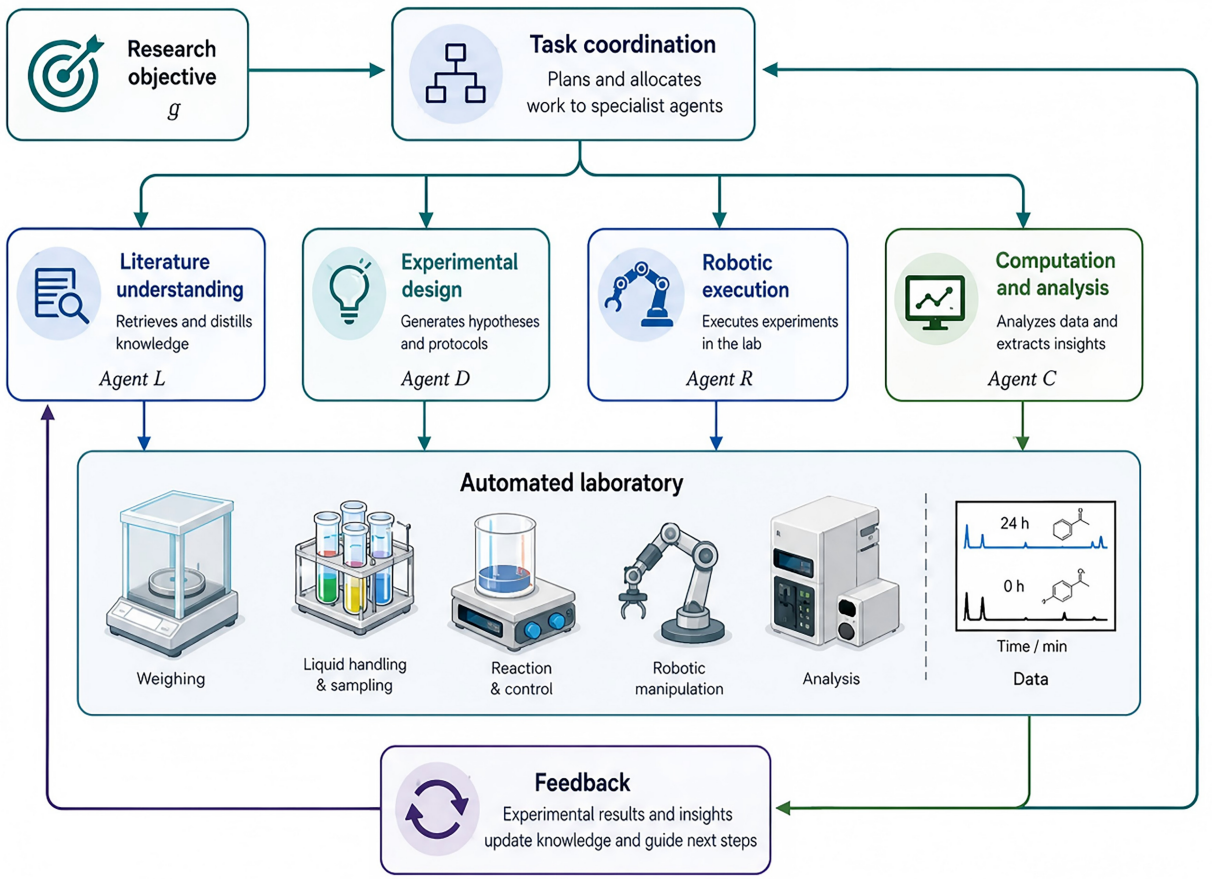


Figure 3. Multi-agent chemistry workflow connecting research planning, computational analysis, laboratory operations, and experimental feedback.

methodologies have therefore moved beyond single-agent chatbots toward multi-agent and tool-integrated workflows. Figure 3 illustrates how literature understanding, computation, robotic execution, and data analysis can be connected within a closed-loop workflow[12-14].

Specialist agents can perform localized tasks: one may extract prior knowledge from literature, another generate hypotheses or synthetic protocols, while computational or robotic modules carry out prediction, synthesis, characterization, or analysis. DigCat provides catalyst data and AI-enabled design resources[15], Cloud Synthesis describes catalyst design with synthesis and feedback[16], and DIVE extracts literature-figure data for hydrogen-storage materials discovery[17]. Robotic laboratory studies demonstrate experimental execution and feedback[13,14]. Practical challenges remain in safe instrument control, exception handling, protocol transferability, data quality assurance, and reproducibility.

MATERIALS SCIENCE AND INTELLIGENT DESIGN

Physics-grounded agents for materials design and simulation

Materials discovery is inherently a high-dimensional, multimodal reasoning problem. Viable candidates must satisfy criteria including thermodynamic stability, kinetic accessibility, structural symmetry, synthetic feasibility, and target performance. General-purpose LLMs still show limitations in domain-specific physical reasoning, quantitative reliability, and robust materials-property prediction, motivating their integration with curated materials data and physics-based computational tools[18]. Figure 4 summarizes a representative physics-aware materials-agent architecture.

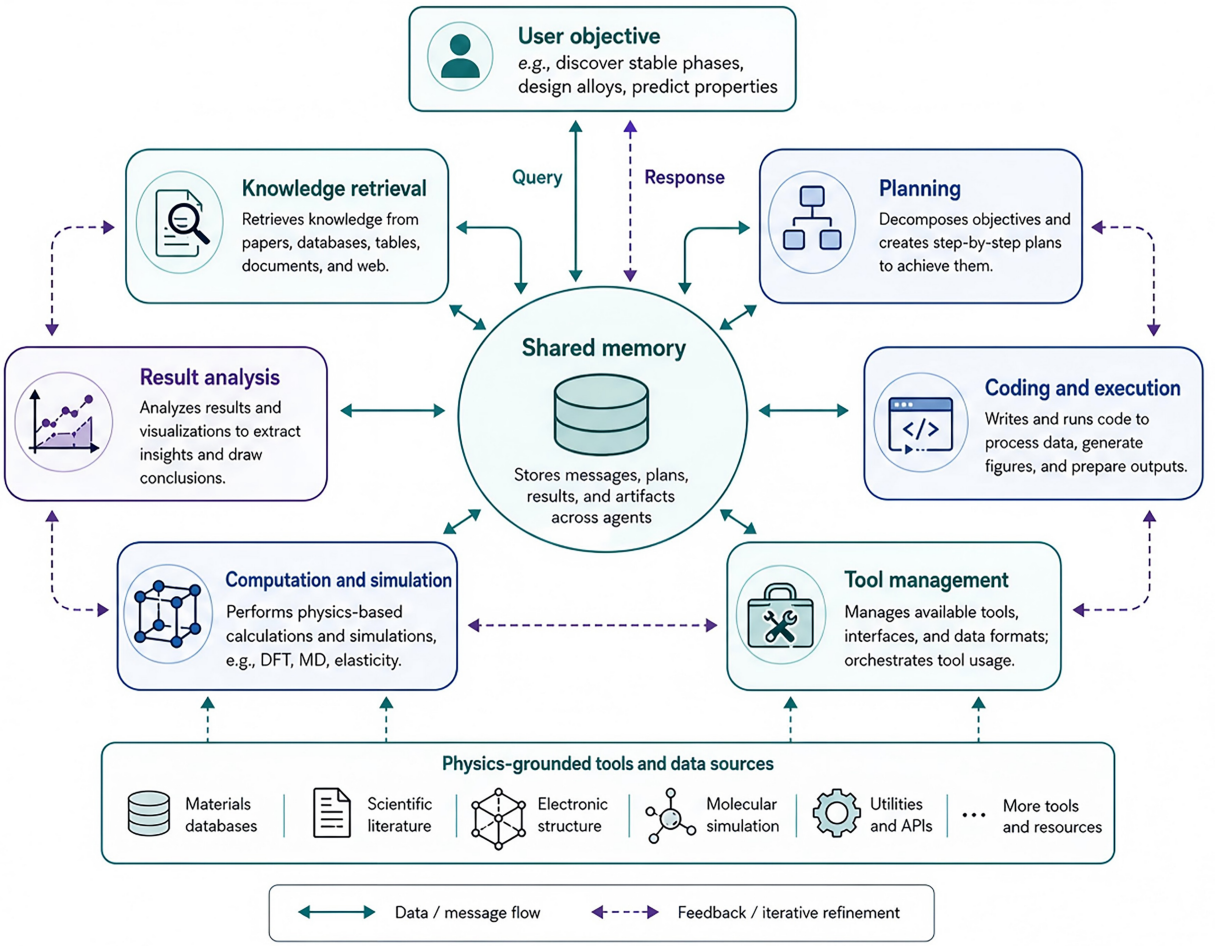


Figure 4. Physics-aware materials-agent architecture connecting retrieval, shared memory, planning, computation, simulation, and analysis.

DIVE extracts materials data from scientific figures[17], while Merchant *et al.* demonstrate large-scale learning for materials stability prediction[19]. Materials databases, including the Materials Project[20], and catalytic data resources[15] can support screening and property prediction.

A shared memory or coordination layer can connect knowledge retrieval, inverse design, code execution, simulation, and result analysis, allowing the agent to function as an orchestration layer rather than an isolated predictor. Cloud Synthesis links catalyst design to experimental feedback[16], while DIVE connects materials data extraction to inverse design[17]. Important challenges remain in uncertainty quantification, reliable tool interfaces, efficient simulations, and integration with experimental validation.

BIOMEDICINE AND THERAPEUTIC INNOVATION

Tool-constrained biomedical experimentation

Biomedical design operates under strict biological, ethical, and safety constraints, where parameter changes can affect pharmacokinetics, target selectivity, and experimental feasibility. Biomedical foundation models have advanced protein structure prediction[21-23]. Figure 5 summarizes a tool-constrained framework in which biological objectives are translated into structured plans and computational tasks.

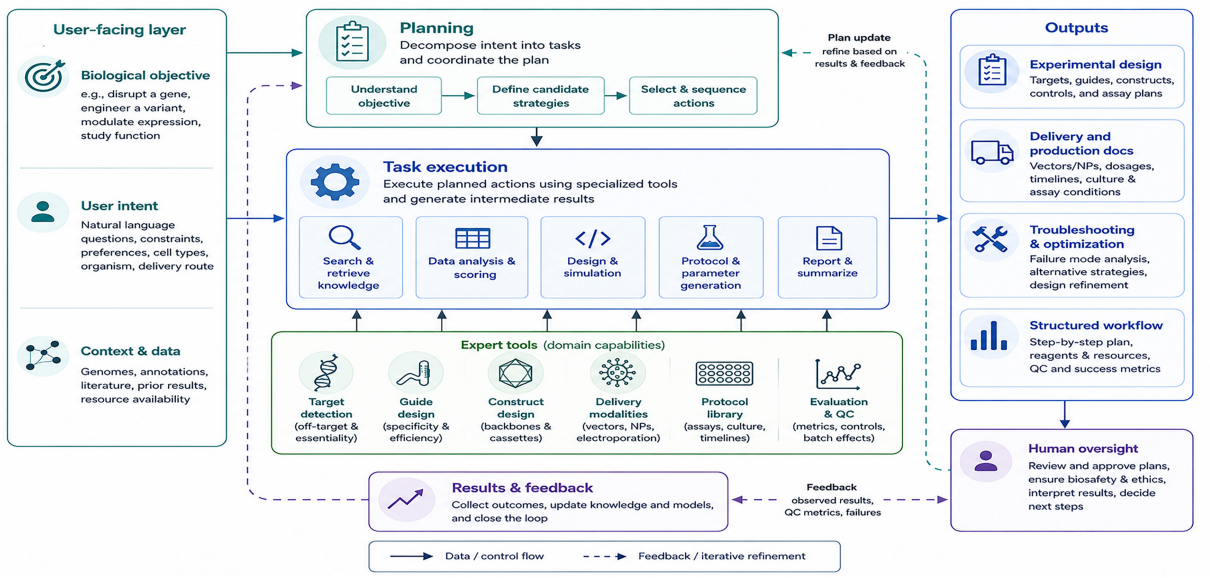


Figure 5. Biomedical-agent framework linking biological objectives, planning, specialized computational tools, feedback, and human oversight. QC: Quality control; NPs: nanoparticles.

The agent can decompose an objective into target selection, sequence or structure design, protocol matching, evaluation, and failure-mode prediction. Biomni integrates biomedical reasoning with tools and code execution[24], while BioDiscoveryAgent proposes genetic perturbation experiments[25].

This paradigm reflects a broader shift from standalone language models toward biomedical agents that coordinate domain-specific tools under explicit experimental constraints. Its practical value depends on biological validation, human supervision, laboratory safety, and ethical review.

PHYSICS, ASTRONOMY, AND SPACE EXPLORATION

Natural-language scientific instrument control

Modern scientific instruments are increasingly powerful but often rely on specialized, manually intensive control interfaces. Natural language can lower this barrier only if user intent is translated faithfully into structured, safe, and machine-readable commands. Recent work has explored LLM-based agents as semantic interfaces between scientists and instrument-control software, with explicit validation and feedback[26,27]. The control loop shown in Figure 6 illustrates this interaction.

A representative workflow parses user intent into goals and constraints, generates executable commands, and uses returned measurements or device status to guide subsequent adjustments. The LLM therefore serves as a reasoning and translation layer rather than an unconstrained controller.

The same principle can extend to accelerators, microscopes, beamlines, telescopes, and robotic measurement platforms. Reliable deployment requires safety interlocks, operator authority, validated command schemas, and deterministic fallback mechanisms[26,27].

ENERGY SYSTEMS AND DECARBONIZATION

Autonomous data-to-algorithm loops for energy systems

Modern energy systems must adapt continuously to variable generation and demand while converting operational data into reliable control and forecasting decisions. Data-driven methods have shown value for prediction, optimization, and adaptive energy management[28]. Figure 7 outlines an iterative workflow linking task formulation, system data, optimization, evaluation, and feedback.

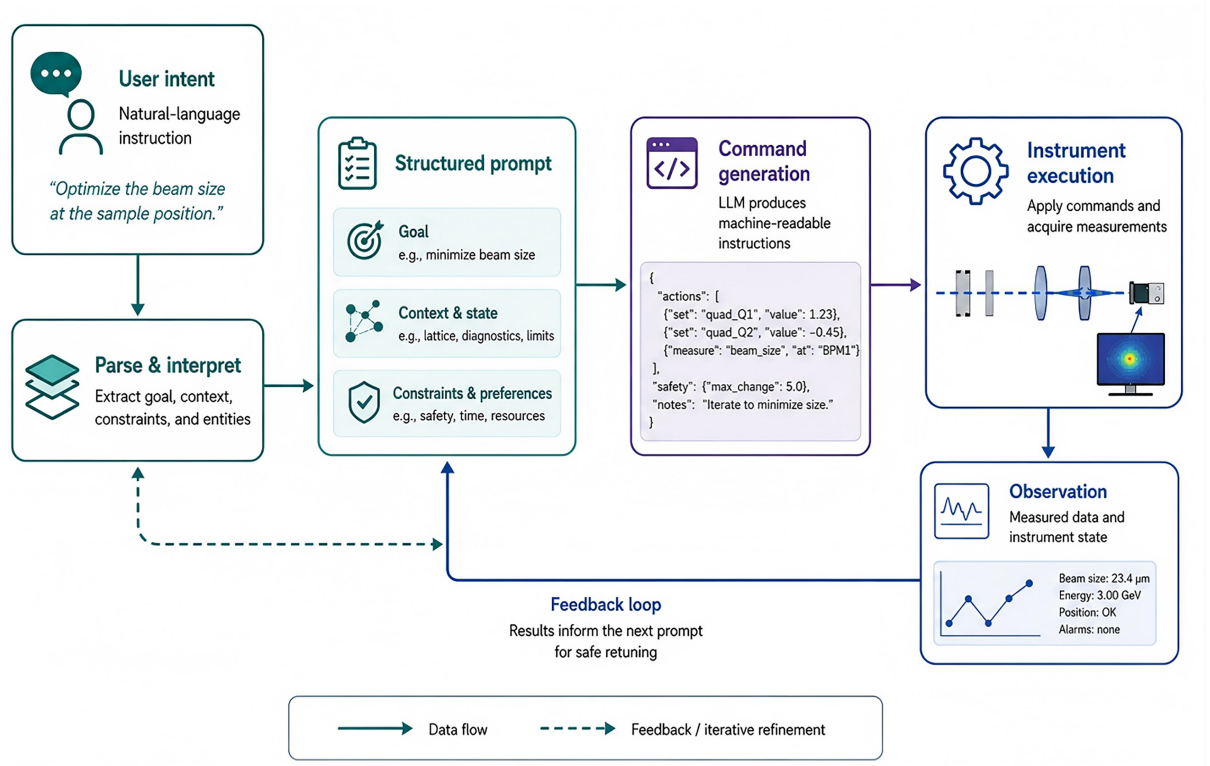


Figure 6. Workflow of scientific instrument control, from user-intent interpretation and structured prompt formulation to machine-readable command generation, instrument execution, observation, and feedback. LLM: Large language model; BPM: beam position monitor.

Within this loop, incoming data support state assessment, optimization-task formulation, and evaluation of candidate control or forecasting strategies. Reinforcement-learning and autonomous-optimization methods provide a foundation for adaptive behavior, where policies improve through interaction with changing environments[29].

This approach is relevant to grid optimization, dynamic state estimation, integrated energy management, and resilient operation. Translation to real systems depends on reliable data pipelines, interpretable failures, operational safeguards, and clear accountability under uncertainty.

EARTH AND ENVIRONMENTAL INTELLIGENCE

Tool-augmented multimodal reasoning for Earth observation

Earth observation (EO) is inherently multimodal, combining optical, multispectral or hyperspectral imagery, SAR, elevation models, land-cover products, climate data, and contextual information. Complex EO tasks therefore benefit from tool-augmented reasoning rather than single-image classification alone[30,31]. Figure 8 presents a workflow coupling multimodal inputs with geospatial tools and iterative reasoning.

A generic EO agent follows a reason-act-observe cycle: it interprets the task, selects tools, processes intermediate outputs, and updates its state until an interpretation or quantitative result is supported by evidence. Tools include spectral analysis, geometry, retrieval, visualization, and AI models.

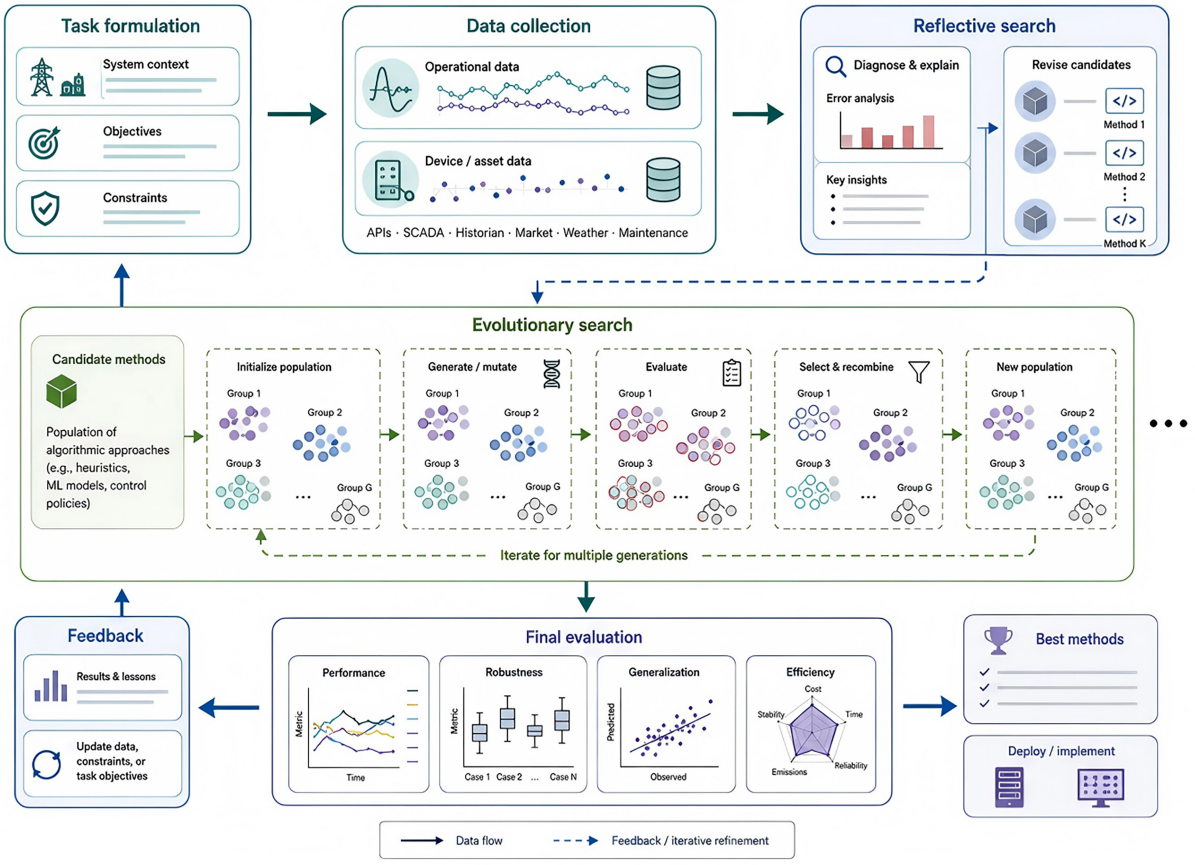


Figure 7. Energy-systems workflow linking task formulation, data collection, iterative search, evaluation, and feedback. ML: Machine learning; APIs: application programming interfaces; SCADA: supervisory control and data acquisition.

This approach brings agentic AI closer to practical geospatial analysis, where conclusions depend on layered and measurable evidence. Open challenges include sensor shift, missing metadata, uncertainty quantification, and reliable evaluation of tool use and reasoning trajectories[30].

EMBODIED INTELLIGENCE AND AUTONOMOUS SYSTEMS

Feedback-driven embodied intelligence

Embodied agents operate under contact mechanics, non-equilibrium conditions, and sensorimotor uncertainty, requiring tight coupling between high-level planning and low-level feedback. Vision-language-action models have demonstrated the potential to combine language-based reasoning with multimodal perception and robotic control, translating abstract instructions into executable physical behaviors[32-34]. Figure 9 highlights this coupling between planning, execution, and feedback.

In a representative pipeline, the system retrieves relevant skills, generates a high-level plan, and passes it to the trajectory planner and robot controller. Visual observations, position estimates, and force feedback return to the controller for online replanning and motion correction.

Such feedback-driven architectures are better suited to long-horizon manipulation than a single offline plan. Barriers remain in control latency, transfer across platforms, safe human-robot interaction, semantic grounding, and unified evaluation for open-ended physical tasks.

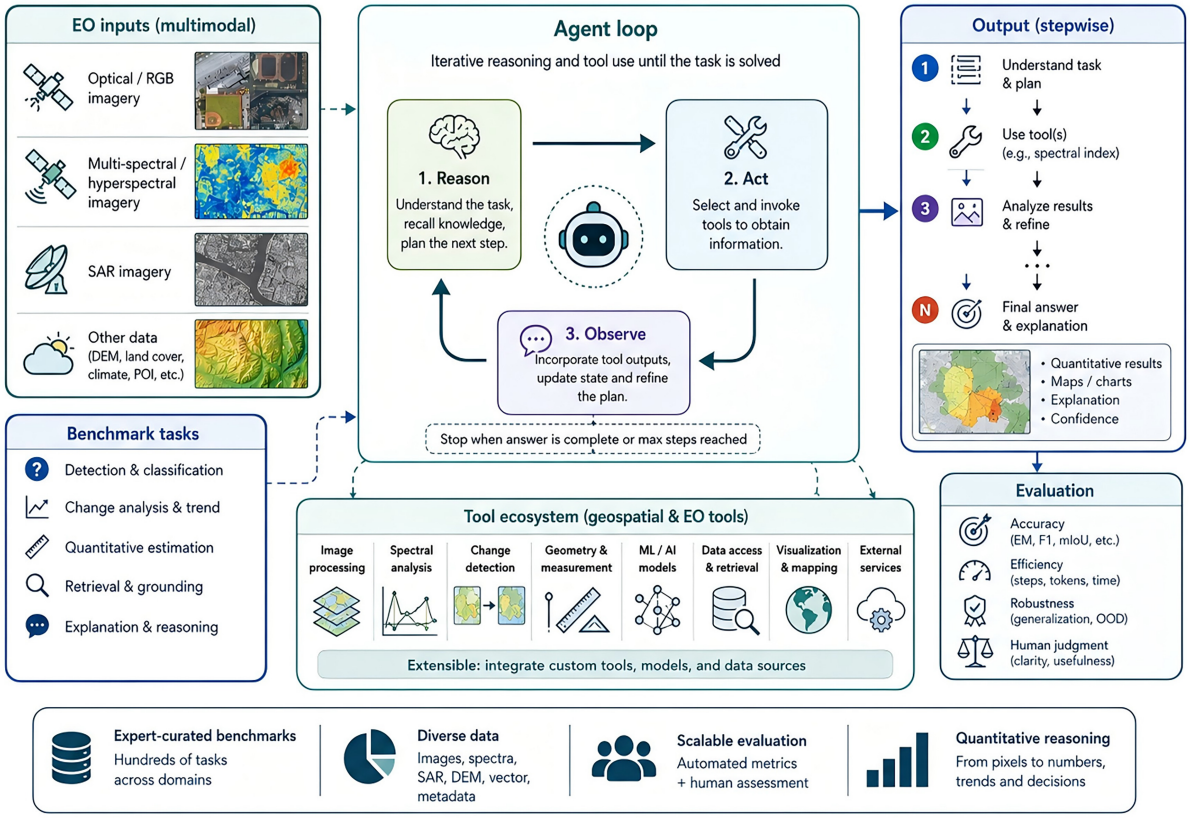


Figure 8. Earth-observation agent combining multimodal imagery, geospatial tools, iterative reasoning, and quantitative evaluation. EO: Earth observation; RGB: red, green, and blue; SAR: synthetic aperture radar; DEM: digital elevation model; POI: point of interest; ML: machine learning; AI: artificial intelligence; EM: exact match; F1: F1 score (harmonic mean of precision and recall); mIoU: mean intersection over union; OOD: out of distribution.

TRUST, SAFETY, AND ETHICAL GOVERNANCE

Full-stack security and governance for AI agents

Agents interacting with tools and physical systems face security risks beyond text generation. Vulnerabilities may arise across prompts, memory, planning, actions, tools, observations, and external environments, as summarized in Figure 10. Indirect prompt injection can compromise tool-integrated applications[35]. Language-model red-teaming supports adversarial evaluation[36], while Agent Security Bench assesses agent-specific attacks, including memory poisoning, and defenses[37].

This full-stack perspective shifts governance from post hoc content moderation toward engineering controls over interaction pathways. The central challenge is maintaining reliable evaluation as agent architectures and deployment environments evolve while balancing autonomy, usefulness, transparency, and security.

CONCLUDING PERSPECTIVES

Taken together, the ten representative advances reveal a broader transition from systems that primarily generate information toward agents that participate directly in scientific and societal workflows.

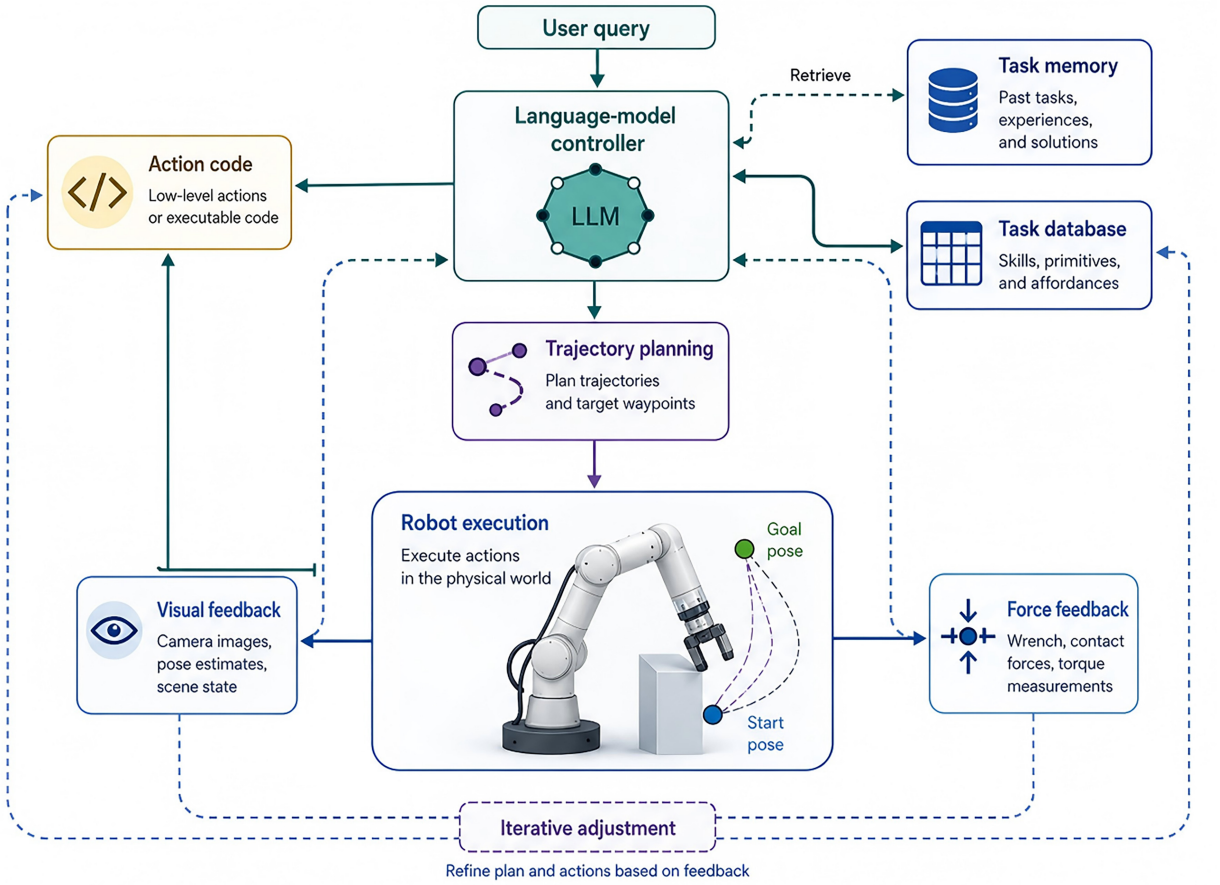


Figure 9. Embodied-agent workflow connecting language-model planning, task memory, robot execution, and visual and force feedback. LLM: Large language model.

Across these domains, successful agents combine long-horizon planning, domain-specific tools, multimodal information processing, memory and state tracking, and iterative feedback[1]. A particularly important feature is the emergence of closed-loop workflows in which hypothesis generation, tool execution, outcome evaluation, and strategy refinement are connected through interaction with external knowledge, computational models, experimental platforms, or physical environments[16,17].

Several challenges recur across application areas. Long-horizon autonomy remains vulnerable to accumulated planning and tool-use errors, while generalization, uncertainty quantification, standardized evaluation, and reproducibility remain unresolved. Transparency, human oversight, safety constraints, and rigorous validation therefore remain essential for high-impact applications[35-37].

Future development is likely to focus on more reliable integration of foundation models, specialized knowledge, scientific tools, and autonomous experimentation. Rather than replacing human expertise, next-generation agents are more plausibly viewed as collaborative systems that extend scientific reasoning and decision-making while preserving meaningful human supervision and accountability.

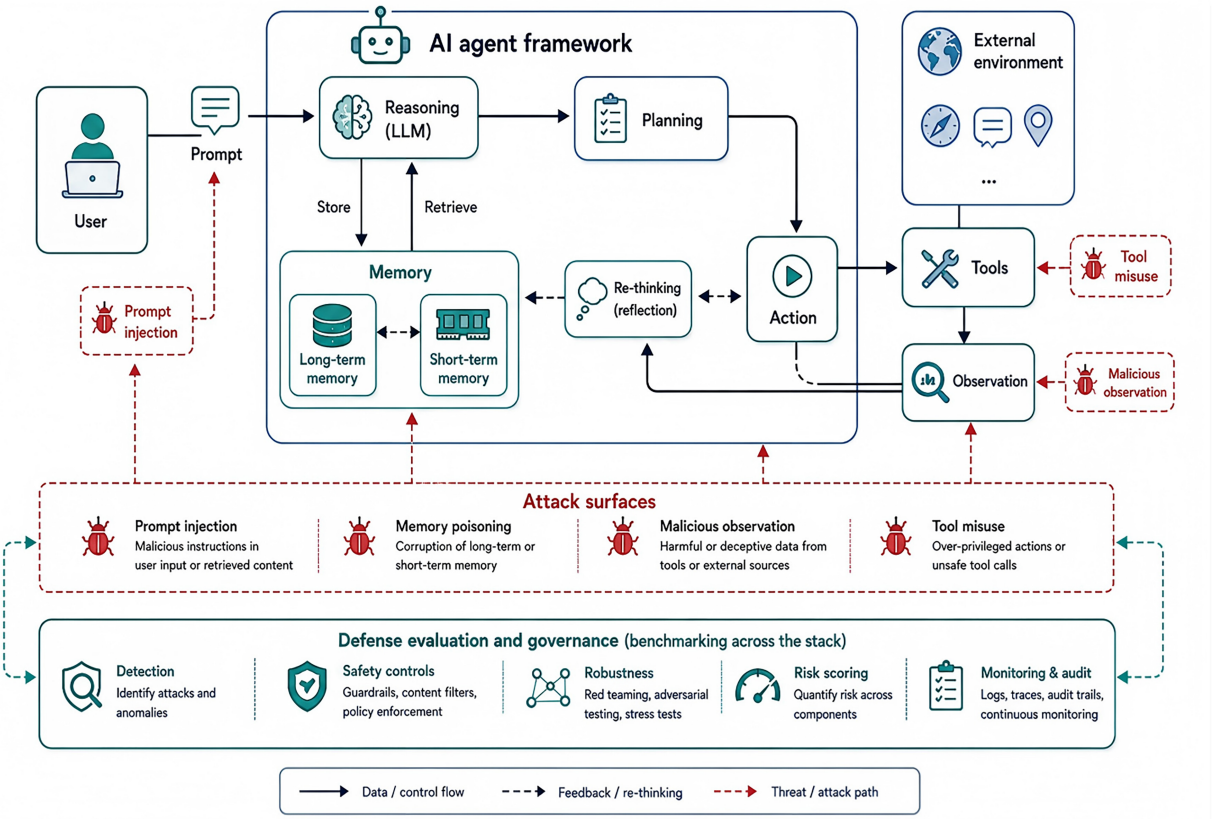


Figure 10. AI agent framework with four attack surfaces and corresponding defense governance. AI: Artificial intelligence; LLM: large language model.

DECLARATIONS

Acknowledgments

The author thanks Hanlin Zhuang from Suzhou MatSource Technology Co., Ltd. for helpful comments on the manuscript and technical assistance with the organization of literature resources, the development and refinement of schematic figures, and the preparation of supporting materials.

Authors' contributions

The author contributed solely to the article.

Availability of data and materials

Not applicable.

AI and AI-assisted tools statement

During the preparation of this manuscript, the AI tool ChatGPT (OpenAI, version GPT-5.5, released 2026-04-23) was used to assist in preparing some decorative icons and schematic elements in Figures 1-10, and GPT-6 Sol (released 2026-09-22) was used to assist in preparing decorative illustrations representing scientific and technological themes in the Graphical Abstract. These tools did not influence the study design, data collection, analysis, interpretation, or the scientific content of the work. The author takes full responsibility for the accuracy, integrity, and final content of the manuscript.

Financial support and sponsorship

This work was supported by Suzhou MatSource Technology Co., Ltd. (Suzhou, China) and the Gusu Laboratory of Materials (Suzhou, China), under grant number Y2501.

Conflicts of interest

Li, L. is an Editorial Board Member of *AI Agent* and is affiliated with Suzhou MatSource Technology Co., Ltd., which provided financial support for this work. Li, L. was not involved in any steps of the editorial process, including reviewers' selection, manuscript handling, or decision-making.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2026.

REFERENCES

1. Xi, Z.; Chen, W.; Guo, X.; et al. The rise and potential of large language model based agents: a survey. *Sci. China. Inf. Sci.* **2025**, *68*, 121101. DOI
2. Yao, S.; Zhao, J.; Yu, D.; et al. ReAct: synergizing reasoning and acting in language models. *arXiv* **2023**, arXiv:2210.03629. Available online: https://doi.org/10.48550/arXiv.2210.03629 (accessed 2 November 2026).
3. Schick, T.; Dwivedi-Yu, J.; Dessi, R.; et al. Toolformer: language models can teach themselves to use tools. In NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, LA, USA; Curran Associates Inc.: Red Hook, NY, USA, 2023; pp 68539-51. https://dl.acm.org/doi/10.5555/3666122.3669119. (accessed 2026-11-02).
4. Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; Bernstein, M. S. Generative agents: interactive simulacra of human behavior. In *UIST '23: The 36th Annual ACM Symposium on User Interface Software and Technology*, San Francisco CA USA; Association for Computing Machinery: New York, NY, USA, 2023; pp 1-22. DOI
5. Wang, L.; Ma, C.; Feng, X.; et al. A survey on large language model based autonomous agents. *Front. Comput. Sci.* **2024**, *18*, 186345. DOI
6. Webb, T.; Mondal, S. S.; Momennejad, I. A brain-inspired agentic architecture to improve planning with LLMs. *Nat. Commun.* **2025**, *16*, 8633. DOI PubMed PMC
7. Huang, X.; Liu, W.; Chen, X.; et al. Understanding the planning of LLM agents: a survey. *arXiv* **2024**, arXiv:2402.02716. Available online: https://doi.org/10.48550/arXiv.2402.02716 (accessed 2 November 2026).
8. Wang, G.; Xie, Y.; Jiang, Y.; et al. Voyager: an open-ended embodied agent with large language models. *arXiv* **2023**, arXiv:2305.16291. Available online: https://doi.org/10.48550/arXiv.2305.16291 (accessed 2 November 2026).
9. Trinh, T. H.; Wu, Y.; Le, Q. V.; He, H.; Luong, T. Solving olympiad geometry without human demonstrations. *Nature* **2024**, *625*, 476-82. DOI PubMed PMC
10. Polu, S.; Sutskever, I. Generative language modeling for automated theorem proving. *arXiv* **2020**, arXiv:2009.03393. Available online: https://doi.org/10.48550/arXiv.2009.03393 (accessed 2 November 2026).
11. Zheng, K.; Han, J. M.; Polu, S. MiniF2F: a cross-system benchmark for formal olympiad-level mathematics. In *International Conference on Learning Representations (ICLR)*, 2022. https://openreview.net/forum?id=9ZPegFuFTFv. (accessed 2026-11-02).
12. Bran, A. M.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; Schwaller, P. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **2024**, *6*, 525-35. DOI PubMed PMC
13. Boiko, D. A.; MacKnight, R.; Kline, B.; Gomes, G. Autonomous chemical research with large language models. *Nature* **2023**, *624*, 570-8. DOI PubMed PMC
14. Burger, B.; Maffettone, P. M.; Gusev, V. V.; et al. A mobile robotic chemist. *Nature* **2020**, *583*, 237-41. DOI PubMed
15. Zhang, D.; Li, H. Digital Catalysis Platform (DigCat): a gateway to big data and AI-powered innovations in catalysis. *ChemRxiv* **2026**, Epub ahead of print. DOI
16. Zhang, D.; Jia, X.; Liu, H.; et al. Cloud synthesis: a global closed-loop feedback powered by autonomous AI-driven catalyst design agent. *AI. Agent.* **2025**, *1*, 2. DOI
17. Zhang, D.; Jia, X.; Hung, T. B.; et al. "DIVE" into hydrogen storage materials discovery with AI agents. *arXiv* **2025**, arXiv:2508.13251. Available online: https://doi.org/10.48550/arXiv.2508.13251 (accessed 2 November 2026).
18. Wang, H.; Li, K.; Ramsay, S.; Fehlis, Y.; Kim, E.; Hattrick-Simpers, J. Evaluating the performance and robustness of LLMs in materials science Q&A and property predictions. *Digit. Discov.* **2025**, *4*, 1612-24. DOI

19. Merchant, A.; Batzner, S.; Schoenholz, S. S.; Aykol, M.; Cheon, G.; Cubuk, E. D. Scaling deep learning for materials discovery. *Nature* **2023**, *624*, 80-5. DOI PubMed PMC
20. Jain, A.; Ong, S. P.; Hautier, G.; et al. Commentary: the Materials Project: a materials genome approach to accelerating materials innovation. *APL. Mater.* **2013**, *1*, 011002. DOI
21. Abramson, J.; Adler, J.; Dunger, J.; et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, *630*, 493-500. DOI PubMed PMC
22. Lin, Z.; Akin, H.; Rao, R.; et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379*, 1123-30. DOI PubMed
23. Jumper, J.; Evans, R.; Pritzel, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583-9. DOI PubMed PMC
24. Huang, K.; Zhang, S.; Wang, H.; et al. Biomni: a general-purpose biomedical AI agent. *bioRxiv* **2025**, Epub ahead of print. DOI PubMed PMC
25. Roohani, Y.; Lee, A.; Huang, Q.; et al. BioDiscoveryAgent: an AI agent for designing genetic perturbation experiments. In *International Conference on Learning Representations (ICLR)*, 2025. https://proceedings.iclr.cc/paper_files/paper/2025/hash/4252dc945318330290000f85dc5fac792-Abstract-Conference.html. (accessed 2026-11-02).
26. Gridach, M.; Nanavati, J.; Mack, C.; El Abidine, K. Z.; Mendes, L. Agentic AI for scientific discovery: a survey of progress, challenges, and future directions. In *ICLR 2025 Workshops: AgenticAI*, 2025. https://openreview.net/forum?id=TyCYakX9BD. (accessed 2026-11-02).
27. Vriza, A.; Prince, M. H.; Zhou, T.; Chan, H.; Cherukara, M. J. Operating advanced scientific instruments with AI agents that learn on the job. *npj. Comput. Mater.* **2026**, *12*, 160. DOI
28. Rolnick, D.; Donti, P. L.; Kaack, L. H.; et al. Tackling climate change with machine learning. *ACM. Comput. Surv.* **2022**, *55*, 1-96. DOI
29. Vázquez-canteli, J. R.; Nagy, Z. Reinforcement learning for demand response: a review of algorithms and modeling techniques. *Appl. Energy.* **2019**, *235*, 1072-89. DOI
30. Kao, C. H.; Zhao, W.; Revankar, S.; et al. Towards LLM agents for Earth observation. In *ICML 2025 Workshop on TerraBytes: Towards Global Datasets and Models for Earth Observation*, 2025. https://arxiv.org/abs/2504.12110v2. (accessed 2026-11-02).
31. Feng, P.; Lv, Z.; Ye, J.; et al. Earth-Agent: unlocking the full landscape of Earth observation with agents. *arXiv* **2026**, arXiv:2509.23141. Available online: https://doi.org/10.48550/arXiv.2509.23141 (accessed 2 November 2026).
32. Zitkovich, B.; Yu, T.; Xu, S.; et al. RT-2: vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the 7th Conference on Robot Learning*, 2023; Vol. 229, pp 2165-83. https://proceedings.mlr.press/v229/zitkovich23a.html. (accessed 2026-11-02).
33. Kim, M. J.; Pertsch, K.; Karamcheti, S.; et al. OpenVLA: an open-source vision-language-action model. *arXiv* **2024**, arXiv:2406.09246. Available online: https://doi.org/10.48550/arXiv.2406.09246 (accessed 2 November 2026).
34. Black, K.; Brown, N.; Driess, D.; et al. π0: a vision-language-action flow model for general robot control. *arXiv* **2026**, arXiv:2410.24164. Available online: https://doi.org/10.48550/arXiv.2410.24164 (accessed 2 November 2026).
35. Greshake, K.; Abdelnabi, S.; Mishra, S.; Endres, C.; Holz, T.; Fritz, M. Not what you've signed up for: compromising real-world LLM-integrated applications with indirect prompt injection. *arXiv* **2023**, arXiv:2302.12173. Available online: https://doi.org/10.48550/arXiv.2302.12173 (accessed 2 November 2026).
36. Perez, E.; Huang, S.; Song, F.; et al. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, December, 2022; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp 3419-48. DOI
37. Zhang, H.; Huang, J.; Mei, K.; et al. Agent Security Bench (ASB): formalizing and benchmarking attacks and defenses in LLM-based agents. In *International Conference on Learning Representations (ICLR)*, 2025. https://openreview.net/forum?id=V4y0CpX4hK. (accessed 2026-11-02).

Disclaimer/Publisher's Note: All statements, opinions, and data contained in this publication are solely those of the individual author(s) and contributor(s) and do not necessarily reflect those of OAE and/or the editor(s). OAE and/or the editor(s) disclaim any responsibility for harm to persons or property resulting from the use of any ideas, methods, instructions, or products mentioned in the content.

© The Author(s) 2026. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.