Review

Collective Intelligence: On the Promise and Reality of Multi-Agent Systems for AI-Driven Scientific Discovery

Terry Jingchen Zhang * , Yongjin Yang , Yinya Huang , Sirui Lu , Bernhard Schölkopf , Zhijing Jin

Posted Date: 8 October 2025

doi: 10.20944/preprints202508.1640.v2

Keywords: AI for science; multi-agent systems; AI-driven scientific discovery

Review

Collective Intelligence: On the Promise and Reality of Multi-Agent Systems for AI-Driven Scientific Discovery

Terry Jingchen Zhang [1,*], Yongjin Yang [2], Yinya Huang [3], Sirui Lu [4], Bernhard Schölkopf [5] and Zhijing Jin [6]

[1] ETH Zurich, Switzerland
[2] University of Toronto, Canada
[3] ETH AI Center, Switzerland
[4] MPI of Quantum Optics, Germany
[5] MPI for Intelligent Systems, Germany
[6] MPI & University of Toronto
* Correspondence: zjingchen@ethz.ch

Abstract

Modern scientific progress is increasingly driven by collaborative endeavors that leverage specialized expertise and constructive peer critique. Multi-agent systems (MAS) offer a robust framework to emulate these collaborative dynamics inherent to human researcher teams by combining distributed information processing with discussion-driven validation, enabling collective intelligence that exceeds the capabilities of individual agents in addressing interdisciplinary challenges. We introduce an application-oriented taxonomy that maps canonical stages of the standard research workflow to both the promising potential and the current reality of MAS for scientific discovery, providing a coherent foundation for understanding, evaluating, and advancing autonomous MAS-powered AI co-scientists. We highlight the distinctive advantages of multi- over single-agent approaches, identify key bottlenecks limiting current deployments, and outline critical research frontiers to bridge the gap between potential and practice. We argue that MAS hold transformative promise to move beyond the role of assistive tools, evolving into autonomous co-scientists capable of parallel exploration of vast knowledge spaces and robust validation through diverse perspectives, thereby advancing open-ended scientific research alongside human investigators.

Keywords: AI for science; multi-agent systems; AI-driven scientific discovery

1. Introduction

> *"Science is a collaborative effort. The combined results of several people working together is often much more effective than an individual scientist working alone."*
>
> —John Bardeen[1]

Automating scientific discovery has evolved through technological epochs driven by advancing artificial intelligence reasoning capabilities. Pioneering systems like *Adam* [1] proposed closing hypothesis-experiment cycles through robot scientists, while deep learning breakthroughs produced landmark achievements including *AlphaFold* [2] for protein structure prediction and *AlphaProof* [3] for mathematical reasoning, drastically accelerating discovery across diverse domains by solving previously intractable problems with high accuracy and efficiency.

Building upon these specialized deep learning successes, the emergence of large language models (LLMs) has unlocked a more general form of scientific reasoning, enabling AI systems to integrate

[1] John Bardeen was the only person to have received the Nobel Prize in Physics twice, for inventing the transistors and the theory of superconductivity. https://www.nobelprize.org/prizes/physics/1972/bardeen/speech

knowledge across disciplines and engage in human-like discourse. This generalization capability has catalyzed a paradigm shift where AI systems evolved from assistive tools [4,5] toward autonomous agents [6–8] emulating independent researchers. These LLM-powered systems advance research across physics [9,10], biochemistry [11–14], causal inference [15], social sciences [16,17], and clinical diagnosis [18,19], demonstrating broad AI-driven scientific capabilities that integrate domain knowledge with adaptive reasoning to tackle multifaceted challenges.

Recent breakthroughs of *Grok-4-Heavy* [20] and *Gemini-DeepThink* [21] explored multi-agent schema [22,23] to mirror collective reasoning dynamic of human research teams, achieving leading performance on challenging benchmarks including the International Mathematical Olympiad[2] and Humanity's Last Exam [24]. This progress signals a promising transition from single-agent systems toward MAS architectures reflecting the collaborative intelligence underlying human scientific discovery, where emergent group dynamics enable superior problem-solving through division of labor and iterative refinement.

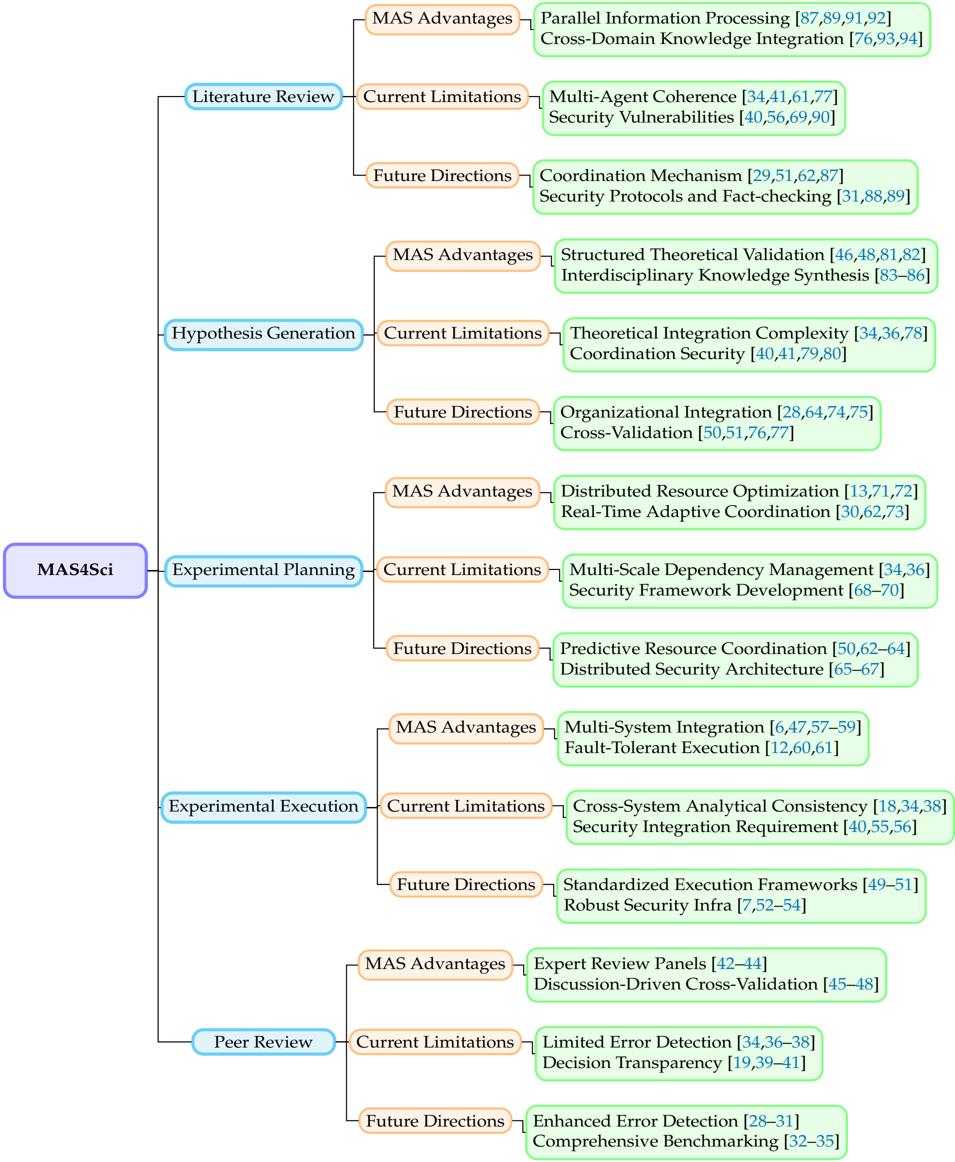
MAS4Sci
Literature Review
MAS Advantages
Parallel Information Processing [87,89,91,92]
Cross-Domain Knowledge Integration [76,93,94]
Current Limitations
Multi-Agent Coherence [34,41,61,77]
Security Vulnerabilities [40,56,69,90]
Future Directions
Coordination Mechanism [29,51,62,87]
Security Protocols and Fact-checking [31,88,89]
Hypothesis Generation
MAS Advantages
Structured Theoretical Validation [46,48,81,82]
Interdisciplinary Knowledge Synthesis [83–86]
Current Limitations
Theoretical Integration Complexity [34,36,78]
Coordination Security [40,41,79,80]
Future Directions
Organizational Integration [28,64,74,75]
Cross-Validation [50,51,76,77]
Experimental Planning
MAS Advantages
Distributed Resource Optimization [13,71,72]
Real-Time Adaptive Coordination [30,62,73]
Current Limitations
Multi-Scale Dependency Management [34,36]
Security Framework Development [68–70]
Future Directions
Predictive Resource Coordination [50,62–64]
Distributed Security Architecture [65–67]
Experimental Execution
MAS Advantages
Multi-System Integration [6,47,57–59]
Fault-Tolerant Execution [12,60,61]
Current Limitations
Cross-System Analytical Consistency [18,34,38]
Security Integration Requirement [40,55,56]
Future Directions
Standardized Execution Frameworks [49–51]
Robust Security Infra [7,52–54]
Peer Review
MAS Advantages
Expert Review Panels [42–44]
Discussion-Driven Cross-Validation [45–48]
Current Limitations
Limited Error Detection [34,36–38]
Decision Transparency [19,39–41]
Future Directions
Enhanced Error Detection [28–31]
Comprehensive Benchmarking [32–35]

Figure 1. An application-oriented taxonomy of Multi-Agent Systems for scientific discovery mapped to key stages of standard research workflow

[2] https://www.nature.com/articles/d41586-025-02343-x

1.1. Scope and Comparison to Other Surveys

Despite these advances, existing surveys [25–27] remain fragmented across different domains and isolated tasks such as paper review and experimental execution, lacking a holistic view of MAS potential in the complete research workflows. We aim to address this gap through a comprehensive analysis that details MAS advantages over single agents, confronts current limitations with key bottlenecks, and outline a roadmap of future directions to remediate these gaps towards transforming MAS from ideals into reliable co-scientists as research companion.

1.2. Paper Organization

We introduce a comprehensive application-oriented taxonomy structured around three core analytical dimensions: first, we examine the advantages of multi-agent versus single-agent systems across five key stages of the research workflow (Section 2); second, we analyze the current reality and fundamental bottlenecks limiting MAS deployment in scientific discovery (Section 3); and third, we outline strategic future work directions toward realizing the full potential of MAS for science (Section 4). Given the rapid evolution of MAS, we prioritize recent studies highlighting their unique advantages and challenges in scientific discovery.

2. Multi vs. Single Agent across Key Stages in Scientific Research Workflow

We present an analytical taxonomy to compare Multi-Agent Systems (MAS) against conventional single-agent approaches across 5 key stages of standard scientific workflow, including literature review, hypothesis generation, experimental planning, experimental execution, and peer review.

2.1. Literature Review

The transformation from sequential processing to emergent knowledge synthesis fundamentally alters how computational systems comprehend scientific literature, shifting from linear comprehension to multidimensional synthesis through distributed cognitive frameworks. The core innovation demonstrated in [27] lies not in parallel reading capabilities but in maintaining multiple simultaneous interpretive frameworks where each specialized agent processes literature through distinct conceptual lenses while contributing to unified understanding, effectively addressing the inherent working memory limitations of sequential processing where new information forces compression or discarding of older knowledge. Retrieval agents employing domain-specific strategies [95] demonstrate sophisticated semantic understanding by recognizing conceptual equivalences such as "nucleophilic substitution" and "SN2 reaction" as identical mechanisms, while fact-checking agents [89] construct evidence networks that reveal previously hidden contradictions and support patterns across disparate studies through systematic cross-validation. This difference enables knowledge graph construction [87] that discovers latent interdisciplinary connections existing not within individual papers but rather in the conceptual spaces between them, exemplified by agents linking bacterial quorum sensing to neural synchronization through shared mathematical frameworks of threshold-dependent collective behaviors. The computational biology applications [22] particularly exemplify this advantage where agents simultaneously analyze molecular simulations using physics-based models, gene regulatory networks through Boolean networks and differential equations, and phylogenetic patterns via sequence alignments, ultimately synthesizing insights about how molecular changes propagate through regulatory networks to produce evolutionary adaptations requiring multi-scale understanding. Coordination frameworks [76] preserve domain-specific nuance while facilitating cross-pollination through semantic bridges that create shared conceptual spaces, enabling meaningful comparison across fields without forcing artificial standardization. Knowledge-enhanced frameworks [91] achieve distributed bias mitigation where each agent's systematic errors become visible through disagreement patterns, approaching objectivity through diversity rather than attempting bias-free individual agents. The simulated expert discussions [92] reveal that MAS engage in dialectical processes where contradictions drive deeper investigation rather than defaulting to simple voting or averaging, producing richer understanding than consensus-seeking single agents, representing genuine interdisciplinary synthesis from earth science [93] to drug discovery [94] where emergent patterns arise from agent interactions rather than individual analysis.

2.2. Hypothesis Generation

Adversarial tension revolutionizes hypothesis generation by harnessing productive tension between competing perspectives through structured adversarial frameworks that systematically explore possibility spaces beyond what confirmation-biased single agents would naturally investigate. Debate-based systems [46] create computational analogs to the scientific method itself by assigning agents opposing stances and forcing each to strengthen positions through evidence and reasoning, uncovering arguments that would remain unexplored in single-agent generation where inherent confirmation bias leads to premature convergence. Adversarial validation frameworks [81] ensure hypotheses survive only through withstanding rigorous challenge from multiple perspectives, not simply attempting falsification but probing boundaries to explore under what conditions hypotheses hold, where they break down, and what modifications might extend their applicability to broader domains. Conditional effect evaluation [48] reveals how adversarial agents uncover edge cases, failure modes, and necessary modifications that proposing agents, influenced by confirmation bias toward their creations, would never investigate without external pressure. This process, formalized through cooperative reinforcement learning [82], produces hypotheses that are simultaneously more creative and more rigorous because agents learn not just from their successes but from successful challenges posed by adversaries, creating evolutionary pressure toward robust novelty. The transformative potential manifests most clearly in interdisciplinary synthesis where multi-agent code generation in physics [83] combines quantum field theory thinking in terms of particle interactions, general relativity conceptualizing spacetime curvature, and condensed matter physics focusing on collective phenomena through iterative refinement where each domain's constraints shape emerging hypotheses. Drug discovery applications demonstrate practical impact where agent swarms [85] transform traditionally linear pipelines into parallel iterative processes with structural biology agents identifying binding pockets while medicinal chemistry agents explore ligands, pharmacokinetic agents evaluate absorption profiles, and toxicology agents flag safety concerns simultaneously. Autonomous molecular design [86] reveals that MAS discover entirely new compound classes violating traditional drug-like rules yet proving effective precisely because they satisfy constraints emerging from multiple perspective intersections. Principle-aware frameworks [84] ensure creative exploration remains scientifically grounded through guided creativity where critic agents serve as theoretical consistency guardians, not suppressing creativity but ensuring creative leaps are conscious choices rather than errors, producing hypotheses both more innovative and more likely valid than either unconstrained brainstorming or rigid rule-following could achieve.

2.3. Experimental Planning

Adaptive coordination through distributed intelligence revolutionizes experimental planning by reconceptualizing experiments as complex adaptive systems requiring continuous coordination rather than predetermined schedules that inevitably fail when confronted with real-world uncertainties. The MULTITASK framework [71] treats laboratory resources as ecosystems where agents engage in continuous negotiation creating market-based optimization, naturally balancing competing demands through local interactions rather than central planning requiring perfect information. Chemical coordination systems [13] demonstrate how agents managing different resources trade priorities based on real-time progress where synthesis reactions completing faster immediately release equipment for other experiments while unexpected cell growth patterns trigger resource reallocation requests, handling intricate interdependencies through agents understanding local constraints and negotiating solutions. The innovation lies in anticipatory adaptation through parallel contingency modeling where astrophysical applications using OpenMAS [4,72] show agents maintaining dozens of parallel plans optimized for different scenarios including weather uncertainties, equipment failures, and changing priorities, enabling smooth transitions to pre-optimized alternatives rather than reactive troubleshooting. This proactive resilience transforms planning from deterministic scheduling into adaptive strategies maintaining readiness for multiple futures simultaneously. Parallel planning systems [73] extend this through game-theoretic modeling where agents simulate thousands of what-if scenarios using probabilistic models of failures, delays, and unexpected results to identify robust strategies performing well across contingencies. Multi-agent reinforcement learning [62] adds temporal depth with agents learning from collective experience to recognize subtle patterns predicting complications and proactively adjusting schedules. Fault detection MAS [30] identify subtle patterns presaging problems hours or days before causing failure, including gradual temperature drifts suggesting equipment degradation, unusual consumption patterns indicating contamination, and unexpected correlations revealing hidden dependencies, enabling preemptive intervention maintaining experimental continuity. This predictive capability transforms experimental planning from deterministic scheduling into adaptive strategies where emergent coordination from hundreds of agent interactions produces patterns no central planner could anticipate, including natural equipment utilization balancing, predictive reagent ordering, and automatic safety protocol emergence.

2.4. Experimental Execution

Robustness through redundant intelligence emerges from overlapping agent capabilities creating system-level robustness exceeding individual reliability through multiple independent perspectives providing different views of the same experimental process. El Agente's quantum chemistry framework [6] exemplifies this through memory structures where strategic agents understand scientific objectives, tactical agents translate these into computational strategies selecting appropriate theory levels and convergence criteria, and operational agents handle actual execution. When encountering convergence failures, the system doesn't simply retry but engages in multi-level error handling where operational agents communicate with tactical agents about potential causes, tactical agents recognize common problems and suggest alternatives, and strategic agents evaluate whether alternatives serve objectives, demonstrating automated scientific reasoning rather than mechanical procedure following. DeepMind's agent constellation [57,96–100] shows seamless integration across abstraction levels from low-level code optimization to high-level strategic planning where insights at one level immediately inform decisions at others. Self-reflective frameworks [58] add metacognitive capabilities where agents monitor their performance and develop understanding of why certain approaches fail, progressively improving at predicting which methods work for new problems based on accumulated experience. Paper-to-code systems [47,59] bridge theory-practice gaps by translating paper descriptions into executable code while understanding underlying physics, recognizing numerical stability requirements, and generating efficient implementations, even questioning whether papers' descriptions are complete or contain implicit assumptions when implementations don't produce expected results. Biomedical applications [12,60] coordinate robotic instrumentation while making scientific judgments distinguishing technical failures requiring repetition from biological variations providing scientific insight, achieving better reproducibility than human researchers not by eliminating variation but by documenting and controlling for it more systematically. Knowledge conflict resolution [61] ensures that when agents disagree about conflicting experimental modalities, the system engages in structured deliberation revealing whether conflicts arise from artifacts, resolution limitations, or different aspects of complex phenomena, creating experimental systems exhibiting judgment approaching human researchers while maintaining perfect reproducibility.

2.5. Peer Review

Collective intelligence through structured deliberation transforms peer review into processes that capture multi-perspective evaluation while eliminating human inconsistencies and biases through systematic assessment protocols. AgentReview [42] carefully constructs reviewing ecosystems with agents exhibiting varying commitment levels, expertise domains, and cognitive styles calibrated to mirror actual review panels while eliminating dysfunctions such as personal grudges or institutional biases, preventing individual agents from having disproportionate influence through sophisticated aggregation mechanisms. The system demonstrates how social influence theory creates rational convergence based on evidence rather than conformity where agents with stronger expertise in specific areas have appropriately greater influence and positions adjust based on evidence quality rather than social pressure. Deep review systems [43] engage in structured argumentation through distinct phases analyzing claims and evidence, debating methodological soundness and theoretical novelty, synthesizing arguments considering how strengths and weaknesses balance, and engaging in meta-review examining evaluations for bias and completeness. This structured approach produces reviews that are not only accurate but explainable with clear reasoning paths from evidence to assessment, addressing the black-box criticism often leveled at automated systems. Frameworks [44] adapt evaluation depth based on paper characteristics where straightforward empirical studies receive rapid methodology-focused review while theoretical breakthroughs get deep multi-perspective analysis and interdisciplinary works receive evaluation from agents spanning relevant fields. Resource allocation based on potential impact ensures revolutionary claims get extra scrutiny while incremental advances receive efficient evaluation, with breakthrough ideas that might be dismissed by conservative reviewers getting careful consideration from innovation-specialized agents. CycleResearcher [45] adds iterative refinement where reviews generate author responses triggering re-evaluation cycles that progressively improve both review and paper quality, creating virtuous cycles where the review process actively improves research rather than merely gatekeeping. This creates genuine scientific dialogue where papers don't just get accepted or rejected but actively improve through the process, establishing MAS as superior quality assurance systems enhancing scientific knowledge reliability through systematic multi-perspective validation.

3. Current Reality and Key Bottlenecks

3.1. Literature Review

Distributed knowledge processing could bring semantic fragmentation among different agents rather than integrated understanding, which is a core challenge in MAS knowledge synthesis. Research reveals that specialized agents operate with fundamentally incompatible conceptual frameworks leading to systematic misinterpretation [33] where biology agents understand "energy" as ATP and metabolic processes while physics agents think of conserved quantities governed by Hamiltonians, reflecting different ways of explaining reality with highly diverse theory. Knowledge conflicts [61] could cascade through networks as agents build increasingly specialized representations with exponentially deteriorating communication ability, creating isolated knowledge islands defeating the integrated understanding purpose. Security vulnerabilities through adversarial injection [41,80] exploit democratic discussion mechanisms where controlling just 15% to 20% of agents through subtle bias injection including consistent interpretation emphasis, research area downplaying, and false controversy creation can significantly shift system conclusions without detection. Current mitigation strategies each compromise core advantages where standardizing ontologies reduces diversity enabling creative insights, verification protocols add computational overhead negating efficiency gains, and restricting agent diversity eliminates multiple perspectives making MAS valuable.

3.2. Hypothesis Generation

MAS discussion is also exposed to malicious agent injection attack that compromise scientific integrity through sophisticated manipulation exploiting the very openness to novel ideas that makes multi-agent systems valuable for breakthrough discovery. Adversarial agents achieve influence through bias injection shaping exploration spaces [40], selective evidence presentation, and reasonable-sounding objections to alternatives without explicit lying but through strategic emphasis gradually shifting collective attention. Previous study has demonstrated that controlling just 10% to 15% of agents can systematically guide hypothesis generation toward hidden agendas [66] through small biases compounding over multiple refinement rounds where each suggestion appears reasonable but collectively steers toward predetermined conclusions. This creates tension between measures preventing adversarial manipulation also could suppress legitimate creative exploration pushing boundaries.

3.3. Experimental Planning

Coordination challenges reveal integration difficulties where overhead exceeds distributed optimization benefits [29]. Each agent's local optimization creates global impact without predicting its actual impact. The lack of theoretical unification frameworks [78] also presents challenge where agents could operate with very different settings and contexts. MULTITASK attempts reveal that sophisticated negotiation protocols struggle to balance local-global coherence [71], producing theoretically optimal but practically unworkable plans requiring perfect execution where any deviation causes cascading failures throughout tightly coupled plans. The system spends more time replanning than executing, creating planning paralysis where theoretical optimization prevents practical progress particularly in exploratory research where outcome uncertainty means incorrect planning assumptions cause complete experimental schedule collapse. This particularly affects exploratory research where the very nature of discovery means unexpected results should trigger adaptation, but current systems treat deviations as failures.

3.4. Experimental Execution

While MAS enable parallel execution and control, the very same mechanism also obscures tractability of real-time decision-making, tracing reasoning with dozens of agents discuss together becomes exponentially more complicated [34] with each agent's decisions influencing others through complex feedback loops creating emergent behaviors unattributable to specific components. Error detection capabilities degrade as domain expertise becomes diluted across networks [37,38] where no individual agent has sufficient global understanding for systemic error identification. Security integration attempting data integrity adds complexity making real-time adaptation impossible [40,55] where every decision requires cross-validation through consensus mechanisms taking longer time. This creates an inevitable paradox where MAS can execute complex protocols but their complexity in turn complicates interpretable and efficient execution.

3.5. Peer Review

Consensus manipulation through opaque networks exposes how consensus mechanisms enabling collective evaluation create unprecedented manipulation vulnerabilities where multi-agent contributions through complex discussions produce untraceable decisions unattributable to evidence. Sophisticated manipulation shapes discussion patterns through doubt injection, concern amplification, and false consensus creation [19] exploiting social patterns where adversarial agents introduce skepticism through innocent questions planting doubt seeds, amplify minor concerns through strategic repetition, and coordinate seemingly independent similar concerns making isolated issues appear widespread. Social influence mechanisms designed to improve consistency become manipulation vectors where influential agents' early opinions create artificial consensus reflecting social patterns rather than scientific merit. Error propagation through multi-stage processes amplifies rather than corrects mistakes where initial methodology assessment errors become accepted facts influencing theoretical evaluation and ultimately accept/reject decisions, creating cascade failures in quality control. Verification attempts create rigid structures eliminating democratic advantages, producing reviews less reliable than single experts while requiring more computational resources, leaving the promise of achieving collective wisdom unrealized. Current systems potentially introduce new bias and manipulation forms threatening scientific quality control integrity more severely than traditional peer review problems they aimed to solve.

4. Future Work Towards MAS4Science

4.1. Literature Review

Resolving semantic fragmentation requires evolving knowledge graphs that maintain domain precision while enabling translation through learned bridges rather than predetermined mappings [28,87]. The key insight is treating semantic bridges as first-class learning entities that accumulate evidence about successful conceptual translations—for instance, when mathematical frameworks like barrier penetration probability successfully unify quantum tunneling and enzyme catalysis. Through co-evolutionary learning [74,82] and cross-task generalization [64], agents develop intermediate representations that preserve essential meaning including mathematical relationships, causal structures, and predictive power. This creates a form of semantic federalism where autonomous domains maintain conceptual sovereignty while participating in knowledge exchange through negotiated semantic treaties that emerge from successful collaboration patterns [51,76]. The result is living knowledge frameworks that adapt to scientific progress while maintaining coherence—achieving robust cross-domain understanding without sacrificing the specialized depth that makes domain expertise valuable.

4.2. Hypothesis Generation

Secure hypothesis generation requires cryptographic consensus preserving creative freedom while verifying reasoning validity without constraining exploration through zero-knowledge proofs enabling verification of good-faith scientific reasoning without revealing specific content. Peer-guard protocols [31] demonstrate agents proving logical rule following, evidence respect, and consistency maintenance without exposing ideas to theft or manipulation, creating trust without transparency that traditionally enables both collaboration and vulnerability. Multi-layered fact-checking [89] implements progressive verification with creative sandboxes allowing wild speculation in early stages with minimal constraints, graduating to rigorous evaluation as ideas mature from speculation to hypothesis. Differential privacy [65] enables sharing statistical insights about hypothesis spaces where agents communicate promising directions without revealing specific targets, allowing exploration adjustment avoiding redundancy without enabling intellectual property theft. Behavioral analysis identifies manipulation through reasoning patterns [66] detecting characteristic adversarial argumentation structures including overconfidence without evidence, contradicting data dismissal, and circular reasoning regardless of specific hypotheses promoted. Deadlock prevention uses game theory ensuring genuine contribution incentives [67] structuring rewards so agents benefit more from collaborative discovery than competitor sabotage, aligning individual and collective interests. Privacy-preserving reinforcement learning [55] enables collective learning without accessing individual contributions where successful strategies benefit all agents without revealing developers or specific hypotheses. Blockchain-inspired consensus creates immutable hypothesis evolution records [69] through cryptographic signing and linking preventing post-hoc manipulation while allowing legitimate building. Future systems must implement graduated transparency with time-delayed revelation where hypotheses remain encrypted during initial development, get partially revealed for collaborative refinement, and become fully transparent after key experiments.

4.3. Experimental Planning

Solving planning challenges requires compositional frameworks treating experiments as modular assemblies with defined interfaces. Parallel simulation [62,101] decomposes experiments into independent modules with specified inputs/outputs, transforming global optimization into local coordination. Virtual laboratories [63] model experimental physics including sample stability and calibration timing before committing resources. Agent-oriented planning [102] maintains probabilistic outcome models identifying critical uncertainties requiring early resolution. Reinforcement learning [71] enables agents to learn dependency structures: sequential versus parallel execution, catastrophic versus recoverable failures, and bottleneck versus excess resources. Modular interfaces with formal specifications enable independent development satisfying contracts. Violations require only local negotiation between affected modules rather than global replanning. This enables experimental marketplaces where distributed capabilities coordinate through standardized interfaces, creating virtual research organizations tackling problems beyond single group capabilities.

4.4. Experimental Execution

Future system must address the paradox where MAS complexity prevents the very transparency needed for scientific validation—demands living documentation that functions simultaneously as human-readable explanation and machine-executable programs [47,58]. Future MAS powered by multimodal language models should capture scientific intent beyond mechanical procedure, describing experimental aims, assumptions, and constraints while allowing implementation flexibility. These specifications must evolve through fault analysis [30], where execution failures serve as RL signal to continuously refine MAS. This approach transforms the transparency-interoperability trade-off into a manageable design choice where different domains implement appropriate transparency levels while maintaining system-wide scientific integrity [12]. Future systems must develop domain-specific languages that preserve essential scientific requirements while supporting gradual refinement from high-level intent to executable protocol.

4.5. Peer Review

Future MAS should offer traceable review through argumentation graphs making reasoning explicit and verifiable. MAS decompose reviews into atomic evaluations with edges representing logical relationships. Reproducibility frameworks [35,103] cross-validate research through independent implementation attempts. Principle-aware evaluation [84] enables explicit assessment conditions acknowledging context dependencies. Human-AI collaboration [104,105] shows transparency through confidence indicators and limitation acknowledgment. CycleResearcher's iterative refinement [45] creates scientific dialogue where authors address concerns and papers improve through back-and-forth. Future systems should justify decisions with detailed reasoning, incorporate uncertainty quantification, create valuable review artifacts contributing knowledge regardless of acceptance, transforming peer review into collaborative refinement.

5. Vision and Conclusions

The evolution from single-agent to multi-agent systems for scientific discovery represents a foundational paradigm shift from AI as only tools to collaborative intelligence networks that take inspirations from how human researchers attempt to expand knowledge frontiers into the unknowns. As these systems mature from current promising prototypes into actually functional research ecosystems, they promise to unlock capabilities that transcend current single-agent limitations of memory and context window to perform open-ended reasoning through discussion-driven mutual reasoning and parallel exploration of vast state spaces. The vision extends beyond mere automation toward enabling new forms of human-AI collaborative inquiry where AI agents serve as autonomous co-scientists with refreshing perspective alongside human researchers.

6. Limitations

This survey provides an overview of the current reality and future prospects of MAS for scientific discovery (MAS4Science). However, certain limitations in the scope and methodology of this paper warrant acknowledgment.

6.1. References and Methods

Due to page limits, this survey may not capture all relevant literature, particularly given the rapid evolution of both multi-agent and AI4Science research. We focus on frontier works published between 2023 and 2025 in leading venues, including conferences such as *ACL, ICLR, ICML, and NeurIPS, and journals like Nature, Science, and IEEE Transactions. Ongoing efforts will monitor and incorporate emerging studies to ensure the survey remains current.

6.2. Empirical Conclusions

Our analysis and proposed directions rely on empirical evaluations of existing MAS frameworks, which may not fully capture the field's macroscopic dynamics. The rapid pace of advancements risks outdating certain insights, and our perspective may miss niche or emerging subfields. We commit to periodically updating our assessments to reflect the latest developments and broader viewpoints.

6.3. Ethical Considerations

Despite our best effort to responsibly synthesize the current reality and future prospects at the intersection of AI4Science and Multi-Agent Research, ethical challenges remain as the selection of literature may inadvertently favor more prominent works, and may potentially overlook contributions from underrepresented research communities. We aim to uphold rigorous standards in citation practices and advocate for transparent, inclusive research to mitigate these risks.

References

1. King, R.; Whelan, K.E.; Jones, F.; Reiser, P.G.K.; Bryant, C.H.; Muggleton, S.; Kell, D.; Oliver, S. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature* **2004**, *427*, 247–252.
2. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.
3. DeepMind, G. AI achieves silver-medal standard solving International Mathematical Olympiad problems, 2024.
4. Xu, X.; Bolliet, B.; Dimitrov, A.; Laverick, A.; Villaescusa-Navarro, F.; Xu, L.; Íñigo Zubeldia. Evaluating Retrieval-Augmented Generation Agents for Autonomous Scientific Discovery in Astrophysics. *arXiv preprint* **2025**, [2507.07155].
5. Wang, H.; Fu, T.; Du, Y.; Gao, W.; Huang, K.; Liu, Z.; Chandak, P.; Liu, S.; Katwyk, P.V.; Deac, A.; et al. Scientific discovery in the age of artificial intelligence. *Nature* **2023**, *620*, 47–60.
6. Zou, Y.; Cheng, A.H.; Aldossary, A.; Bai, J.; Leong, S.X.; Campos-Gonzalez-Angulo, J.A.; Choi, C.; Ser, C.T.; Tom, G.; Wang, A.; et al. El Agente: An Autonomous Agent for Quantum Chemistry, 2025, [arXiv:cs.AI/2505.02484].
7. Lu, C.; Lu, C.; Lange, R.T.; Foerster, J.; Clune, J.; Ha, D. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery, 2024, [arXiv:cs.AI/2408.06292].
8. Gottweis, J.; Weng, W.H.; Daryin, A.; Tu, T.; Palepu, A.; Sirkovic, P.; Myaskovsky, A.; Weissenberger, F.; Rong, K.; Tanno, R.; et al. Towards an AI co-scientist, 2025, [arXiv:cs.AI/2502.18864].
9. Sarkar, M.; Bolliet, B.; Dimitrov, A.; Laverick, A.; Villaescusa-Navarro, F.; Xu, L.; Íñigo Zubeldia. Multi-Agent System for Cosmological Parameter Analysis. *arXiv preprint* **2024**, [arXiv:astro-ph.CO/2412.00431].
10. Lu, S.; Jin, Z.; Zhang, T.J.; Kos, P.; Cirac, J.I.; Schölkopf, B. Can Theoretical Physics Research Benefit from Language Agents?, 2025, [arXiv:cs.CL/2506.06214].
11. Jin, R.; Zhang, Z.; Wang, M.; Cong, L. STELLA: Self-Evolving LLM Agent for Biomedical Research. *arXiv preprint* **2025**, [2507.02004].
12. Gao, S.; Fang, A.; Lu, Y.; Fuxin, L.; Shao, D.; Zhu, Y.; Zou, C.; Schneider, J.; Chen, L.; Liu, C.; et al. Empowering biomedical discovery with AI agents. *Cell* **2024**, *187*, 6125–6151. https://doi.org/10.1016/j.cell.2024.09.022.
13. Boiko, D.A.; MacKnight, R.; Kline, B.; Gomes, G. Autonomous chemical research with large language models. *Nature* **2023**, *624*, 570–578. https://doi.org/10.1038/s41586-023-06792-0.
14. Bran, A.M.; Cox, S.; Schilter, O.; Baldassari, C.; White, A.D.; Schwaller, P. Augmenting large language models with chemistry tools. *Nature Machine Intelligence* **2024**, *6*, 525–535. Published 08 May 2024, https://doi.org/10.1038/s42256-024-00832-8.
15. Verma, V.; Acharya, S.; Simko, S.; Bhardwaj, D.; Haghighat, A.; Sachan, M.; Janzing, D.; Schölkopf, B.; Jin, Z. Causal AI Scientist: Facilitating Causal Data Science with Large Language Models. *Manuscript Under Review* **2025**.
16. Haase, J.; Pokutta, S. Beyond Static Responses: Multi-Agent LLM Systems as a New Paradigm for Social Science Research. *ArXiv* **2025**, *abs/2506.01839*.
17. Parkes, D.C.; Wellman, M.P. Economic reasoning and artificial intelligence. *Science* **2015**, *349*, 267 – 272.

18. Chen, X.; Yi, H.; You, M.; Liu, W.Z.; Wang, L.; Li, H.; Zhao, Y. Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ Digital Medicine* **2025**, *8*, 65. https://doi.org/10.1038/s41746-025-01550-0.
19. Xiao, L.; Zhang, X.; Chen, J.X.; Hong, S. ArgMed-Agents: Explainable Clinical Decision Reasoning with LLM Disscusion via Argumentation Schemes. *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* **2024**, pp. 5486–5493.
20. xAI. Introducing Grok-4. https://x.ai/news/grok-4, 2025.
21. DeepMind. Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad, 2025. DeepMind Blog Post.
22. Ghafarollahi, A.; Buehler, M.J. SciAgents: Automating scientific discovery through multi-agent intelligent graph reasoning. *arXiv preprint* **2024**, [arXiv:cs.AI/2409.05556].
23. Ghareeb, A.E.; Chang, B.; Mitchener, L.; Yiu, A.; Warner, C.; Riley, P.; Krstic, G.; Yosinski, J. Robin: A Multi-Agent System for Automating Scientific Discovery. *arXiv preprint* **2025**, [2505.13400].
24. Phan, L.; Gatti, A.; Han, Z.; Li, N.; Hu, J.; Zhang, H.; Zhang, C.B.C.; Shaaban, M.; Ling, J.; Shi, S.; et al. Humanity's Last Exam, 2025, [arXiv:cs.LG/2501.14249].
25. Luo, Z.; Yang, Z.; Xu, Z.; Yang, W.; Du, X. LLM4SR: A Survey on Large Language Models for Scientific Research. *arXiv preprint arXiv:2501.04306* **2025**.
26. Zheng, T.; Deng, Z.; Tsang, H.T.; Wang, W.; Bai, J.; Wang, Z.; Song, Y. From Automation to Autonomy: A Survey on Large Language Models in Scientific Discovery. *arXiv preprint arXiv:2505.13259* **2025**.
27. Zhuang, Z.; Chen, J.; Xu, H.; Jiang, Y.; Lin, J. Large language models for automated scholarly paper review: A survey. *Information Fusion* **2025**, *124*, 103332. https://doi.org/10.1016/j.inffus.2025.103332.
28. Borghoff, U.M.; Bottoni, P.; Pareschi, R. An Organizational Theory for Multi-Agent Interactions Integrating Human Agents, LLMs, and Specialized AI. *Discover Computing* **2025**.
29. Yan, B.; Zhang, X.; Zhang, L.; Zhang, L.; Zhou, Z.; Miao, D.; Li, C. Beyond Self-Talk: A Communication-Centric Survey of LLM-Based Multi-Agent Systems. *ArXiv* **2025**, *abs/2502.14321*.
30. Khalili, M.; Zhang, X.; Cao, Y. Multi-Agent Systems for Model-based Fault Diagnosis. *IFAC-PapersOnLine* **2017**, *50*, 1211–1216. https://doi.org/10.1016/j.ifacol.2017.08.347.
31. Fan, F.; Li, X. PeerGuard: Defending Multi-Agent Systems Against Backdoor Attacks Through Mutual Reasoning. *ArXiv* **2025**, *abs/2505.11642*.
32. Chen, H.; Xiong, M.; Lu, Y.; Han, W.; Deng, A.; He, Y.; Wu, J.; Li, Y.; Liu, Y.; Hooi, B. MLR-Bench: Evaluating AI Agents on Open-Ended Machine Learning Research. *ArXiv* **2025**, *abs/2505.19955*.
33. Liu, Y.; Yang, Z.; Xie, T.; Ni, J.; Gao, B.; Li, Y.; Tang, S.; Ouyang, W.; Cambria, E.; Zhou, D. ResearchBench: Benchmarking LLMs in Scientific Discovery via Inspiration-Based Task Decomposition. *ArXiv* **2025**, *abs/2503.21248*.
34. Kon, P.T.J.; Liu, J.; Zhu, X.; Ding, Q.; Peng, J.; Xing, J.; Huang, Y.; Qiu, Y.; Srinivasa, J.; Lee, M.; et al. EXP-Bench: Can AI Conduct AI Research Experiments? *ArXiv* **2025**, *abs/2505.24785*.
35. Siegel, Z.S.; Kapoor, S.; Nagdir, N.; Stroebl, B.; Narayanan, A. CORE-Bench: Fostering the Credibility of Published Research Through a Computational Reproducibility Agent Benchmark. *Trans. Mach. Learn. Res.* **2024**, *2024*.
36. Son, G.; Hong, J.; Fan, H.; Nam, H.; Ko, H.; Lim, S.; Song, J.; Choi, J.; Paulo, G.; Yu, Y.; et al. When AI Co-Scientists Fail: SPOT-a Benchmark for Automated Verification of Scientific Research. *ArXiv* **2025**, *abs/2505.11855*.
37. L'ala, J.; O'Donoghue, O.; Shtedritski, A.; Cox, S.; Rodriques, S.G.; White, A.D. PaperQA: Retrieval-Augmented Generative Agent for Scientific Research. *ArXiv* **2023**, *abs/2312.07559*.
38. Starace, G.; Jaffe, O.; Sherburn, D.; Aung, J.; Chan, J.S.; Maksin, L.; Dias, R.; Mays, E.; Kinsella, B.; Thompson, W.; et al. PaperBench: Evaluating AI's Ability to Replicate AI Research. *ArXiv* **2025**, *abs/2504.01848*.
39. Li, G.; Hammoud, H.; Itani, H.; Khizbullin, D.; Ghanem, B. CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. In Proceedings of the Neural Information Processing Systems, 2023.
40. Zheng, C.; Cao, Y.; Dong, X.; He, T. Demonstrations of Integrity Attacks in Multi-Agent Systems. *ArXiv* **2025**, *abs/2506.04572*.
41. Amayuelas, A.; Yang, X.; Antoniades, A.; Hua, W.; Pan, L.; Wang, W. MultiAgent Collaboration Attack: Investigating Adversarial Attacks in Large Language Model Collaborations via Debate. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2024.

42. Jin, Y.; Zhao, Q.; Wang, Y.; Chen, H.; Zhu, K.; Xiao, Y.; Wang, J. AgentReview: Exploring Peer Review Dynamics with LLM Agents. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024; Al-Onaizan, Y.; Bansal, M.; Chen, Y., Eds. Association for Computational Linguistics, 2024, pp. 1208–1226. https://doi.org/10.18653/V1/2024.EMNLP-MAIN.70.
43. Zhu, M.; Weng, Y.; Yang, L.; Zhang, Y. DeepReview: Improving LLM-based Paper Review with Human-like Deep Thinking Process. *ArXiv* **2025**, *abs/2503.08569*.
44. Yu, W.; Tang, S.; Huang, Y.; Dong, N.; Fan, L.; Qi, H.; Guo, C. Dynamic Knowledge Exchange and Dual-Diversity Review: Concisely Unleashing the Potential of a Multi-Agent Research Team. *arXiv preprint arXiv:2506.18348* **2025**.
45. Weng, Y.; Zhu, M.; Bao, G.; Zhang, H.; Wang, J.; Zhang, Y.; Yang, L. CycleResearcher: Improving Automated Research via Automated Review. *ArXiv* **2024**, *abs/2411.00816*.
46. Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J.B.; Mordatch, I. Improving Factuality and Reasoning in Language Models through Multiagent Debate. In Proceedings of the Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024, 2024.
47. Seo, M.; Baek, J.; Lee, S.; Hwang, S.J. Paper2Code: Automating Code Generation from Scientific Papers in Machine Learning. *ArXiv* **2025**, *abs/2504.17192*.
48. Yang, Y.; Yi, E.; Ko, J.; Lee, K.; Jin, Z.; Yun, S. Revisiting Multi-Agent Debate as Test-Time Scaling: A Systematic Study of Conditional Effectiveness. *ArXiv* **2025**, *abs/2505.22960*.
49. Perera, R.; Basnayake, A.; Wickramasinghe, M. Auto-scaling LLM-based multi-agent systems through dynamic integration of agents. *Frontiers in AI* **2025**.
50. Tang, X.; Qin, T.; Peng, T.; Zhou, Z.; Shao, D.; Du, T.; Wei, X.; Xia, P.; Wu, F.; Zhu, H.; et al. Agent KB: Leveraging Cross-Domain Experience for Agentic Problem Solving, 2025, [arXiv:cs.CL/2507.06229].
51. Surabhi, P.S.M.; Mudireddy, D.R.; Tao, J. ThinkTank: A Framework for Generalizing Domain-Specific AI Agent Systems into Universal Collaborative Intelligence Platforms. *ArXiv* **2025**, *abs/2506.02931*.
52. Ifargan, T.; Hafner, L.; Kern, M.; Alcalay, O.; Kishony, R. Autonomous LLM-driven research from data to human-verifiable research papers. *ArXiv* **2024**, *abs/2404.17605*.
53. Schmidgall, S.; Moor, M. AgentRxiv: Towards Collaborative Autonomous Research. *ArXiv* **2025**, *abs/2503.18102*.
54. Zhang, Z.; Qiu, Z.; Wu, Y.; Li, S.; Wang, D.; Zhou, Z.; An, D. OriGene: A Self-Evolving Virtual Disease Biologist Automating Therapeutic Target Discovery. *bioRxiv* **2025**.
55. Mukherjee, A.; Kumar, P.; Yang, B.; Chandran, N.; Gupta, D. Privacy Preserving Multi-Agent Reinforcement Learning in Supply Chains. *ArXiv* **2023**, *abs/2312.05686*.
56. Shanmugarasa, Y.; Ding, M.; Chamikara, M.; Rakotoarivelo, T. SoK: The Privacy Paradox of Large Language Models: Advancements, Privacy Risks, and Mitigation. *ArXiv* **2025**, *abs/2506.12699*.
57. Li, Y.; Choi, D.; Chung, J.; Kushman, N.; Schrittwieser, J.; Leblond, R.; Eccles, T.; Keeling, J.; Gimeno, F.; Dal Lago, A.; et al. Competition-level code generation with AlphaCode. *Science* **2022**, *378*, 1092–1097.
58. Pan, R.; Zhang, H.; Liu, C. CodeCoR: An LLM-Based Self-Reflective Multi-Agent Framework for Code Generation. *ArXiv* **2025**, *abs/2501.07811*.
59. Lin, Z.; Shen, Y.; Cai, Q.; Sun, H.; Zhou, J.; Xiao, M. AutoP2C: An LLM-Based Agent Framework for Code Repository Generation from Multimodal Content in Academic Papers. *ArXiv* **2025**, *abs/2504.20115*.
60. Dobbins, N.J.; Xiong, C.; Lan, K.; Yetisgen-Yildiz, M. Large Language Model-Based Agents for Automated Research Reproducibility: An Exploratory Study in Alzheimer's Disease. *ArXiv* **2025**, *abs/2505.23852*.
61. Ju, T.; Wang, B.; Fei, H.; Lee, M.L.; Hsu, W.; Li, Y.; Wang, Q.; Cheng, P.; Wu, Z.; Zhang, Z.; et al. Investigating the Adaptive Robustness with Knowledge Conflicts in LLM-based Multi-Agent Systems. *ArXiv* **2025**, *abs/2502.15153*.
62. Azadeh, R. Advances in Multi-Agent Reinforcement Learning: Persistent Autonomy and Robot Learning Lab Report 2024. *arXiv preprint arXiv:2412.21088* **2024**.
63. Swanson, K.; Wu, W.; Bulaong, N.L.; Pak, J.E.; Zou, J. The Virtual Lab: AI Agents Design New SARS-CoV-2 Nanobodies with Experimental Validation. *bioRxiv* **2024**. Preprint, https://doi.org/10.1101/2024.11.11.623004.
64. Li, Y.; Qian, C.; Xia, Y.; Shi, R.; Dang, Y.; Xie, Z.; You, Z.; Chen, W.; Yang, C.; Liu, W.; et al. Cross-Task Experiential Learning on LLM-based Multi-Agent Collaboration. *ArXiv* **2025**, *abs/2505.23187*.

65. Szymanski, N.; Rendy, B.; Fei, Y.; Kumar, R.E.; He, T.; Milsted, D.; McDermott, M.J.; Gallant, M.C.; Cubuk, E.D.; Merchant, A.; et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* **2023**, *624*, 86 – 91.
66. Jin, Z.; Wu, Q.; Li, C.; Li, J.; Lu, Y.; Xu, W.; Liao, Y.; Feng, L.; Hu, M.; Li, B. TopoMAS: Large Language Model Driven Topological Materials Multiagent System. *arXiv preprint* **2025**, [2507.04053].
67. Wölflein, G.; Ferber, D.; Truhn, D.; Arandjelovi'c, O.; Kather, J. LLM Agents Making Agent Tools. *ArXiv* **2025**, *abs/2502.11705*.
68. Seo, S.; Kim, J.; Shin, M.; Suh, B. LLMDR: LLM-Driven Deadlock Detection and Resolution in Multi-Agent Pathfinding. *ArXiv* **2025**, *abs/2503.00717*.
69. de Witt, C.S. Open Challenges in Multi-Agent Security: Towards Secure Systems of Interacting AI Agents. *ArXiv* **2025**, *abs/2505.02077*.
70. Sun, L.; Yang, Y.; Duan, Q.; Shi, Y.; Lyu, C.; Chang, Y.C.; Lin, C.T.; Shen, Y. Multi-Agent Coordination across Diverse Applications: A Survey, 2025, [arXiv:cs.MA/2502.14743].
71. Kusne, A.G.; McDannald, A. Scalable multi-agent lab framework for lab optimization. *Matter* **2023**. https://doi.org/10.1016/j.matt.2023.05.025.
72. Xu, L.; Sarkar, M.; Lonappan, A.I.; Íñigo Zubeldia.; Villanueva-Domingo, P.; Casas, S.; Fidler, C.; Amancharla, C.; Tiwari, U.; Bayer, A.; et al. Open Source Planning & Control System with Language Agents for Autonomous Scientific Discovery. *arXiv preprint* **2025**, [2507.07257].
73. Li, Y.; Liu, S.; Zheng, T.; Song, M. Parallelized Planning-Acting for Efficient LLM-based Multi-Agent Systems. *ArXiv* **2025**, *abs/2503.03505*.
74. Park, C.; Han, S.; Guo, X.; Ozdaglar, A.; Zhang, K.; Kim, J.K. MAPoRL: Multi-Agent Post-Co-Training for Collaborative Large Language Models with Reinforcement Learning. *ArXiv* **2025**, *abs/2502.18439*.
75. Lan, T.; Zhang, W.; Lyu, C.; Li, S.; Xu, C.; Huang, H.; Lin, D.; Mao, X.L.; Chen, K. Training Language Models to Critique With Multi-agent Feedback. *ArXiv* **2024**, *abs/2410.15287*.
76. Du, Z.; Qian, C.; Liu, W.; Xie, Z.; Wang, Y.; Qiu, R.; Dang, Y.; Chen, W.; Yang, C.; Tian, Y.; et al. Multi-Agent Collaboration via Cross-Team Orchestration. *arXiv* **2024**, [arXiv:cs.CL/2406.08979]. Accepted to Findings of ACL 2025.
77. Zhu, K.; Du, H.; Hong, Z.; Yang, X.; Guo, S.; Wang, Z.; Wang, Z.; Qian, C.; Tang, X.; Ji, H.; et al. MultiAgentBench: Evaluating the Collaboration and Competition of LLM agents. *ArXiv* **2025**, *abs/2503.01935*.
78. Soldatova, L.N.; Rzhetsky, A. Representation of research hypotheses. *Journal of Biomedical Semantics* **2011**, *2*, S9 – S9.
79. Huang, Y.; Chen, Y.; Zhang, H.; Li, K.; Fang, M. Deep Research Agents: A Systematic Examination And Roadmap. *arXiv preprint arXiv:2506.18096* **2025**.
80. Tran, K.T.; Dao, D.; Nguyen, M.D.; Pham, Q.V.; O'Sullivan, B.; Nguyen, H.D. Multi-Agent Collaboration Mechanisms: A Survey of LLMs. *ArXiv* **2025**, *abs/2501.06322*.
81. Bandi, C.; Harrasse, A. Adversarial Multi-Agent Evaluation of Large Language Models through Iterative Debates. *ArXiv* **2024**, *abs/2410.04663*.
82. Pu, Z.; Ma, H.; Hu, T.; Chen, M.; Liu, B.; Liang, Y.; Ai, X. Coevolving with the Other You: Fine-Tuning LLM with Sequential Cooperative Multi-Agent Reinforcement Learning. *ArXiv* **2024**, *abs/2410.06101*.
83. Chun, J.; Chen, Q.; Li, J.; Ahmed, I. Is Multi-Agent Debate (MAD) the Silver Bullet? An Empirical Analysis of MAD in Code Summarization and Translation. *ArXiv* **2025**, *abs/2503.12029*.
84. Pu, Y.; Lin, T.; Chen, H. PiFlow: Principle-aware Scientific Discovery with Multi-Agent Collaboration. *ArXiv* **2025**, *abs/2505.15047*.
85. Song, K.; Trotter, A.; Chen, J.Y. LLM Agent Swarm for Hypothesis-Driven Drug Discovery. *ArXiv* **2025**, *abs/2504.17967*.
86. Koscher, B.A.; Canty, R.B.; McDonald, M.A.; Greenman, K.P.; McGill, C.J.; Bilodeau, C.L.; Jin, W.; Wu, H.; Vermeire, F.H.; Jin, B.; et al. Autonomous, multiproperty-driven molecular discovery: From predictions to measurements and back. *Science* **2023**, *382*.
87. Chen, X.; Che, M.; et al. An automated construction method of 3D knowledge graph based on multi-agent systems in virtual geographic scene. *International Journal of Digital Earth* **2024**, *17*, 2449185. https://doi.org/10.1080/17538947.2024.2449185.
88. Al-Neaimi, A.; Qatawneh, S.; Saiyd, N.A. Conducting Verification And Validation Of Multi- Agent Systems, 2012, [arXiv:cs.SE/1210.3640].
89. Nguyen, T.P.; Razniewski, S.; Weikum, G. Towards Robust Fact-Checking: A Multi-Agent System with Advanced Evidence Retrieval. *arXiv preprint* **2025**, [arXiv:cs.CL/2506.17878].

90. Ferrag, M.A.; Tihanyi, N.; Hamouda, D.; Maglaras, L. From Prompt Injections to Protocol Exploits: Threats in LLM-Powered AI Agents Workflows. *arXiv preprint arXiv:2506.23260* **2025**.
91. Wang, H.; Du, X.; Yu, W.; Chen, Q.; Zhu, K.; Chu, Z.; Yan, L.; Guan, Y. Learning to break: Knowledge-enhanced reasoning in multi-agent debate system. *Neurocomputing* **2023**, *618*, 129063.
92. Li, Z.; Chang, Y.; Le, X. Simulating Expert Discussions with Multi-agent for Enhanced Scientific Problem Solving. *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)* **2024**.
93. Pantiukhin, D.; Shapkin, B.; Kuznetsov, I.; Jost, A.A.; Koldunov, N. Accelerating Earth Science Discovery via Multi-Agent LLM Systems. *ArXiv* **2025**, *abs/2503.05854*.
94. Solovev, G.V.; Zhidkovskaya, A.B.; Orlova, A.; Vepreva, A.; Tonkii, I.; Golovinskii, R.; Gubina, N.; Chistiakov, D.; Aliev, T.A.; Poddiakov, I.; et al. Towards LLM-Driven Multi-Agent Pipeline for Drug Discovery: Neurodegenerative Diseases Case Study. In Proceedings of the OpenReview Preprint, 2024.
95. Sami, M.A.; Rasheed, Z.; Kemell, K.; Waseem, M.; Kilamo, T.; Saari, M.; Nguyen-Duc, A.; Systä, K.; Abrahamsson, P. System for systematic literature review using multiple AI agents: Concept and an empirical evaluation. *CoRR* **2024**, *abs/2403.08399*, [2403.08399]. https://doi.org/10.48550/ARXIV.2403.08399.
96. Mankowitz, D.J.; Michi, A.; Zhernov, A.; Gelada, M.; Selvi, M.; Paduraru, C.; Leurent, E.; Iqbal, S.; Lespiau, J.B.; Ahern, A.; et al. Faster sorting algorithms discovered using deep reinforcement learning. *Nature* **2023**, *618*, 257–263.
97. Trinh, T.H.; Wu, Y.; Le, Q.V.; He, H.; Luong, T. Solving olympiad geometry without human demonstrations. *Nature* **2024**, *625*, 476–482.
98. Fawzi, A.; Balog, M.; Huang, A.; Hubert, T.; Romera-Paredes, B.; Barekatain, M.; Novikov, A.; Ruiz, F.J.; Schrittwieser, J.; Swirszcz, G.; et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature* **2022**, *610*, 47–53.
99. Vinyals, O.; Babuschkin, I.; Czarnecki, W.M.; Mathieu, M.; Dudzik, A.; Chung, J.; Choi, D.H.; Powell, R.; Ewalds, T.; Georgiev, P.; et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* **2019**, *575*, 350–354.
100. Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. Mastering the game of Go without human knowledge. *Nature* **2017**, *550*, 354–359. https://doi.org/10.1038/nature24270.
101. Fukuda, M.; Gordon, C.; Mert, U.; Sell, M. MASS: A Parallelizing Library for Multi-Agent Spatial Simulation. In Proceedings of the Proceedings of the 2013 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation (PADS). ACM, 2013, pp. 161–170. https://doi.org/10.1145/2486092.2486120.
102. Li, A.; Chen, Y.; Lin, Y.; Li, W.; Ding, J.; Liu, J. Agent-Oriented Planning in Multi-Agent Systems. *arXiv preprint* **2024**, [arXiv:cs.AI/2410.02189].
103. Xiang, Y.; Yan, H.; Ouyang, S.; Gui, L.; He, Y. SciReplicate-Bench: Benchmarking LLMs in Agent-driven Algorithmic Reproduction from Research Papers. *ArXiv* **2025**, *abs/2504.00255*.
104. Tang, K.; Wu, A.; Lu, Y.; Sun, G. Collaborative Editable Model. *ArXiv* **2025**, *abs/2506.14146*.
105. Chen, N.; HuiKai, A.L.; Wu, J.; Hou, J.; Zhang, Z.; Wang, Q.; Wang, X.; He, B. XtraGPT: LLMs for Human-AI Collaboration on Controllable Academic Paper Revision. *ArXiv* **2025**, *abs/2505.11336*.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.