POSITION: SCIENCE IS COLLABORATIVE—LLM FOR SCIENCE SHOULD BE TOO

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern scientific breakthroughs are increasingly driven by collaborative team effort where researchers combine diverse expertise to tackle interdisciplinary challenges. In this position paper, **we argue that LLM for Science should mirror such cooperative dynamics through Multi-Agent Systems (MAS) instead of pursuing a single omniscient model for all scientific problems**. Following the Canonical Workflow Framework for Research (CWFR), we identify how MAS could benefit each canonical research stage: enhanced reliability for knowledge synthesis by cross-validation, increased creativity for hypothesis formulation via diversifying perspectives, improved robustness for experimental execution through parallel execution and fault-tolerant backup agents, and more diverse opinion in result interpretation and evaluation. We further outline key bottlenecks in the current reality of MAS4Science and future work to address these challenges, concluding with several concrete call to actions for reliably scaling MAS in science from passive tools to active research partners.

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, NOBEL LAUREATE[1]

The most transformative scientific discoveries often originate from **research collaboration** that brings together diverse expertise to spark genuine breakthroughs, (Wuchty et al., 2007) used 19.9 million papers over 5 decades and 2.1 million patents to show that **team collaboration** increasingly dominate solo authors in the production of exceptionally high-impact research, even where that distinction was once the domain of solo authors across sciences and engineering, arts and humanities. We argue that **LLM for Science should reflect such collaborative dynamic via Multi-Agent Systems (MAS4Science)** rather than pursuing a single universal agent for all scientific disciplines.

Frontier labs (Anthropic, 2025) have already begun exploring the immense potential of multi-agent systems for complex scientific reasoning tasks, which leverage parallel thinking for more **creative** exploration of diverse solution paths and cross-validation among agents for more **reliable** result synthesis, achieving gold-medal level performance on International Mathematical Olympiad (IMO) (DeepMind, 2025). This paradigm has quickly expanded to research-level scientific applications across many disciplines, including chemistry and materials science (Zheng et al., 2023; Jin et al., 2025; Gustin et al., 2026), physics and astronomy (Xu et al., 2025; He et al., 2025), biomedical (Solovev et al., 2024; Song et al., 2025; Zhang et al., 2025c) and social sciences (Haase & Pokutta, 2025).

Working Definitions. As this position paper is intended for a broad audience, we make an effort to avoid math-heavy formulation and instead characterize the key distinction of multi-agent system (MAS) from single-agent system (SAS) using 2 complimentary property based on established literature (Guo et al., 2024):

[1]John Bardeen was the only person to have received the Nobel Prize for Physics twice in 1956 (for the invention of transistors) and 1972 (for pioneering the theory of superconductivity). This quote comes from his Nobel Banquet Speech in 1972, Stockholm.

First, MAS enables **parallelism**: multiple agents can work together simultaneously in a shared environment (not necessarily on the same task so long as they are contributing to the same goal). We note that while it's possible to query a single *model* (which is passive and only responds to external control flow) concurrently, it's not the case for *agent* (which is active and can have an internal control flow that allows agents to plan&act (Yao et al., 2022) autonomously), concurrent calls to agents would create separate instances/threads (also known as *sub-agents*), which is effectively creating multiple agents with the same backbone model.

Second, MAS follows **collectivism**: MAS as a whole is impacted by multiple agents collectively in various interaction mechanisms: whether through explicit communication such as multi-agent debate/discussion (MAD), through division of labor where different agents perform different sub-tasks of a large mission, or through aggregation such as majority voting and cross-validation on the same task. In this study, we mainly focus on MAS with *test-time interaction* through explicit coordination (often via MAD) akin to human research collaboration, but we do note that such coordination is not a necessary condition that defines MAS (for example, 3 coding agents working on adding 3 new function to the same codebase may or may not coordinate explicitly if there's no synergy or conflicts in their mission scope).

Comparison to Related Work. Existing studies have focused on AI scientist (Tie et al., 2026; Gottweis et al., 2025; Ghafarollahi & Buehler, 2024) or MAS (Raza et al., 2025; Guo et al., 2024; Li et al., 2024; Zhu et al., 2025a) separately, where our work aims to focus on the intersection of MAS4Science not only on one domain or research task (Zhuang et al., 2025; Zheng et al., 2025; Sami et al., 2024; Luo et al., 2025), but offer a more holistic view on MAS during the entire life cycle of a scientific project. To that end, we adopt a principled approach grounded in the Canonical Workflow Framework for Research (CWFR) (Betz et al., 2022) by the Research Data Alliance (RDA)[2] that identifies recurring patterns in canonical scientific workflow, which guides our selection of four key stages common to modern research: (1) Knowledge Synthesis (literature review), (2) Hypothesis Formulation (research design), (3) Experimental Execution (the term *experimental* broadly includes wet-lab experiments as well as derivation&proof for theorists and simulation for computational scientists), and (4) Result Interpretation and Evaluation (peer review) as illustrated in Figure 1.

From Section 2 to Section 5, we go through these four CWFR stages to first present **opportunity** of MAS at each stage (marked as **O1 to O8**) and offers our **recommendations** (marked as **R1 to R8**) to address major bottlenecks in the current landscape. We then discuss alternative views (Section 6), and conclude with a call to actions for 2 key stakeholder groups: AI researchers who develop LLMs for Science and scientific researchers that use them (Section 7).

We intentionally avoid using too many domain-specific case study as the field of LLM4Science is vast and highly heterogeneous (such that case study in one scientific context may not transfer to another, even use cases in the same domain can be very diverse). Therefore we intentionally keep the arguments primarily logical (with empirical evidence) to be widely applicable for most scientific scenarios.

Stance and Contribution. Before proceeding to the main sections, we make a few clarifications on our stance: First, we recognize that current MAS are far from perfect and candidly point out key challenges in its current state with recommendations for future work. Second, MAS is a universal solution for all tasks. Indeed, one should not blindly use MAS before carefully probing their cost-effectiveness from multiple dimensions including accuracy, speed, robustness and many more. Our core thesis is a forward-looking one that MAS represents a more promising direction for *open-ended scientific discovery* as it incorporates the collaborative foundations for human scientific success (Wuchty et al., 2007). We believe this is a timely position paper that connects and encourages both AI developers and frontline scientists to join efforts towards realizing the full potential of cooperative AI for Science.

2 KNOWLEDGE SYNTHESIS

Knowledge synthesis requires systematically gathering information from multiple sources, validating accuracy and relevance, and integrating these findings into coherent summaries for evidence-

[2] www.rd-alliance.org

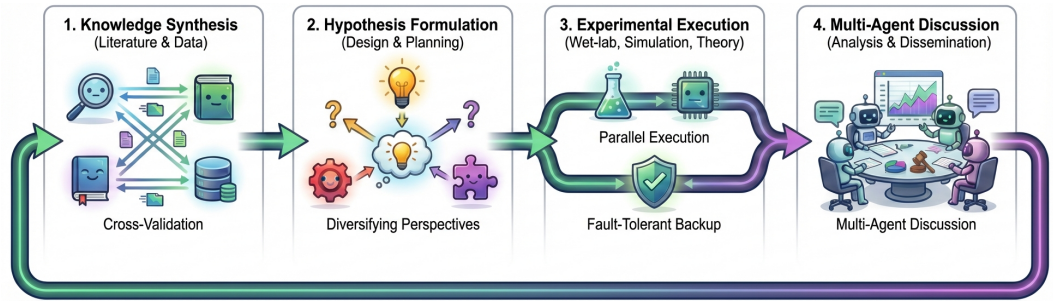

Figure 1: Illustration of the unique advantage brought by multi-agent systems across four CWFR stages in scientific workflows.

based decision-making (Trinh et al., 2025). SAS that consolidate both generation and verification within one model often suffer from hallucination (Kalai et al., 2025; Lin et al., 2025), where the launch of DeepResearch mode in ChatGPT (OpenAI, 2025) (Feb. 2025) coincided with a spike in LLM-hallucinated citations on arXiv (Tramèr, 2025). GPTZero has further reported concerning appearances of such hallucinated citations in venues such as ICLR and NeurIPS (Shmatko et al., 2026). In this section, we first argue how MAS could make knowledge synthesis more reliable by cross-validation and more efficient by parallel processing, and provide concrete recommendations for 2 key failure modes in current MAS for knowledge synthesis, namely error propagation and cross-domain knowledge conflict.

O1: Cross-validation combats hallucination by independent verification. SAS often struggle with identifying and rectifying hallucination in its own generation as the same agent who generates errors may not spot them efficiently (Stechly et al., 2024; Valmeekam et al., 2023), also known as *confirmation bias* (Koo et al., 2024; O'Leary, 2025). MAS offers a promising alternative by distributing information aggregation, summarization, and verification across different agents with respective specialized tooling (Trinh et al., 2025): one agent could specialize in wide exploration via internet search for generation, while another specialize in rigorous retrieval, match and verify against credible sources programmatically.

Previous work (Zhu et al., 2024) showcased this approach by using a discriminator agent to help select useful documents from massive noisy data on the internet, steering the generator agent towards more reliable knowledge synthesis in iterative feedback loop. Empirical evidence from biomedical reasoning tasks shows that well-coordinated MAS can outperform SAS when role structure is explicitly defined (Pu et al., 2025; Inoue et al., 2025). AI Urban Scientist (Xia et al., 2025) also demonstrates the power of role separation in MAS by designating specialized agents to access various databases, then validating claims against credible sources. Empirical studies (Shi et al., 2025; Darwish et al., 2025) further substantiates that MAS can significantly reduce hallucinations through collaborative filtering that covers what one agent missed with others' inspection, outperforming self-consistency methods in hallucinations and uncertainty reduction (Feng et al., 2025).

O2: Parallel processing improves efficiency and independence. Processing multiple sources in parallel with independent agents accelerates synthesis while reducing anchoring effects inherent in sequential reading, where later evidence is often shaped to fit narratives formed from earlier context.

Indeed, Anthropic has showed that accumulated context can even induce persona drift in LLMs and degrade their "Helpful, Honest and Harmless Assistant" setting (Lu et al., 2026). We further note the nature of current attention mechanism in LLMs means longer contexts often leads to more hallucination from diluted attention (Liu et al., 2023), also known as *context rot* (Hong et al., 2025) as evidenced by many long-context benchmarks reporting unsatisfactory results on frontier models (Bai et al., 2025; Yan et al., 2025b).

MAS could alleviate the pressure of long context by distributing multiple agents for summarizing from distinct sources without impact of context accumulation (Sami et al., 2024), which improves both efficiency and independence of knowledge synthesis using massive literature database. After independent extraction is complete, comparison of these independent summaries could naturally expose inconsistencies and contradictions in literature (where SAS often shape the whole summary

with one consistent narrative (Nogueira et al., 2025)). In scientific inquiry, such conflicts are often informative signals of unresolved research questions for potential breakthroughs, making parallel source processing valuable not only for efficiency but also for identifying gaps in empirical literature to propose new hypothesis.

R1: Prevent error propagation and snowballing effect. False information can propagate in MAS with compounding devastating impact (Shen et al., 2025), such snowballing effect is particularly acute when multimodal information such as figures and tables in scientific publications come into play (Yu et al., 2025b). Furthermore, agents usually treat existing publications (especially those in their training data) as authoritative truth by default even though they may also contain errors. Indeed, (Son et al., 2025) have reported that frontier LLMs systematically fail to identify mistakes artificially planted into existing papers and often chose to blindly trust seemingly authoritative sources without cross-validating with other sources on their actual contents. The aggregation of erroneous information from one agent to another is particularly destructive in scientific workflows as they are genuinely needles in a haystack evolving over many rounds of inter-agent communication.

To address error snowballing, MAS architecture should be optimized with selective, not unconditional connectivity and frequent verification at checkpoints: (Shen et al., 2025) demonstrates that optimizing the communication topology by limiting agent connectivity can effectively prevent error propagation while maintaining sufficient information flow for collaborative problem-solving. Beyond architectural improvement, checkpoints could also effectively intercept cascades by tracking information provenance across agent interactions (Zhou et al., 2025) and force iterative cycle of regeneration&verification against primary verifiable sources when agents flag discrepancies in their synthesis (Weng et al., 2025).

R2: Address cross-domain knowledge conflicts with expertise-weighted discussion. Cross-domain knowledge synthesis for interdisciplinary research (Aryal et al., 2024; Li et al., 2025b) presents unique challenges, where similar terminologies could bear very different meaning: for example, the term *inflation* has distinct formulation in economics (which describes a general increase in prices and fall in the purchasing value of money) versus cosmology (which describes a phase of exponential space expansion in the early universe). Such poly-semantic concepts could incur two types of failure modes in multi-agent discussions (Khan et al., 2024): If they are trained to attain consensus (Duan & Wang, 2024), they could arbitrarily merge non-compatible concepts across domains simply to reach consensus sooner (which is often used as a proxy reward metric in multi-agent training (Ma et al., 2025)) and subsequently amplify persuasive yet incorrect arguments in an echo chamber (Duan & Wang, 2024). On the other hand, if agents are encouraged to critique other as much as possible, they could deadlock without meaningful summarization.

To address these challenges, future systems should consider incorporating knowledge graphs to explicitly connect cross-domain concepts (Tang et al., 2025b) and assign weights to different agents based on their domain expertise for decision-making (Cherian et al., 2025). This kind of weighted discussion could also better reflect the collaborative dynamic of human collaboration, where domain experts could contribute more in their respective field of expertise. Future MAS could take inspiration from blockchains where such expertise-based weights can be assigned by dynamic reputation tracking via smart contracts, which has been proven to effectively shape collaborative pattern and emergent agent specialization (Qi et al., 2025a).

Last but not least, human scientists with scientific intuition honed by years of domain experience should also play an active role in making judgments at critical junctures to steer MAS towards the right direction in knowledge synthesis (Gaddipati et al., 2025; Spillias et al., 2024), leveraging human-AI collaboration to establish a solid foundation in the first step of scientific discovery.

3 Hypothesis Formulation

Hypothesis formulation involves the generation of testable explanations for observed phenomena by proposing underlying mechanisms consistent with available evidence. Unlike knowledge synthesis, this stage is inherently speculative and operates in unknown regimes where no objective ground truth exists, which therefore requires a delicate balance between creativity and plausibility.

O3: Role separation enables more diverse hypotheses. When single agents generate hypotheses sequentially, each generation could influence the next through inherent dependencies in context and

working memory where promising initial hypotheses bias the agent to make minor amendments on previous ones that narrow down the search space, thereby limiting diversity of proposed hypothesis. (Ke et al., 2025).

On the other hand, MAS maintains role separation through parallel hypothesis exploration where different agents explore different mechanisms simultaneously without seeing others' proposals (Chen et al., 2024b; Wang et al., 2024). This advantage of this mechanism in hypothesis formulation is evident across multiple scientific domains: PriM (Lai & Pu, 2025) employs principle-inspired multi-agent collaboration for material discovery, AstroAgents (Saeedi et al., 2025) generates hypotheses from mass spectrometry data through specialized agent teams. VirSci (Su et al., 2025) further provides empirical ablation study on various team size and rounds in discussion, showing approximately three to five agents and two to three interaction rounds represent a sweet spot to strike a balance between diversity and stability (Ueda et al., 2025). Overall, MAS expands the hypothesis search space by encouraging exploration of distinct causal mechanisms in parallel. This structural diversity increases the likelihood of uncovering non-obvious explanations and reduces premature convergence on a single dominant theory. At the same time, controlled interaction rounds allow hypotheses to be refined without collapsing diversity too early, supporting both creativity and scientific rigor. Controlled experiments indicate that agent diversity, parallelism, and interaction depth exhibit clear sweet spots for ideation quality, providing empirical guidance for setting default configurations (Ueda et al., 2025).

O4: Multi-agent debate (MAD) enables multi-dimensional examination of hypothesis. MAS enables structured debate in which agents assume complementary roles and explicitly defend or challenge candidate hypotheses (Bandi & Harrasse, 2024; Du et al., 2024; Khan et al., 2024). For example, one agent advocates a hypothesis with supporting evidence, another probes its assumptions and highlights potential confounders, while a third-party evaluator assesses the relative strength of the competing arguments (Duan & Wang, 2024). This role-based interaction externalizes reasoning that would otherwise remain implicit within a single model and makes the evaluation process more transparent and interpretable. (Du et al., 2024) confirms that MAD improves factuality when different agents bring in genuinely diverse perspectives. Debate combined with code execution has been successfully applied to causal discovery, while also revealing coordination overhead and diminishing returns from excessive deliberation (Le et al., 2024).

Meanwhile, MAD forces each hypothesis to survive systematic counter-arguments, strengthening those that can be consistently defended while revealing internal inconsistencies or unsupported claims in weaker ones, which are subsequently discarded (Yuan et al., 2025). Empirical results further suggest that a small number of critics is sufficient, with three debating agents achieving a favorable balance between argumentative depth and coordination overhead (Ueda et al., 2025). Principle-aware controllers that explicitly balance exploration and exploitation have been shown to substantially improve multi-agent scientific discovery performance (Pu et al., 2025). ARM (Yao et al., 2025) further extends evolutionary discovery to collaboration patterns where the system discovers each agent's reasoning modules through evolutionary search that eliminates modules consistently deferring without improving outcomes.

R3: Facilitate active participation to converge on actionable hypothesis. The same independence that enables parallel exploration could also hinder efficient convergence toward a final actionable hypothesis, especially when judging criteria for scientific hypothesis is often subjective in open-ended scientific inquiry without any objective ground truth to lean on. Future MAS need to strike a balance between diversity via sufficiently many threads for exploration and efficiency in having them converge to a few actionable plans via MAD.

We believe the key driver for a successful convergence hinges *active participation* of every agent that represents a diverse hypothesis, otherwise the overall discussion may be skewed with degraded scientific merit. (Note that this is not contradictory with having expertise-weighted discussion as we focus on the fact that each agent should actively engage, but the final results can be a weighted average of everyone's engagement) Researchers have identified emerging *lazy agent* patterns in MAS where some agents dominate the discussion while others merely agree and echo earlier conclusions without materially new contribution (Zhang et al., 2025b), which is likely a downstream consequence of LLM sycophancy (Sharma et al., 2025; Denison et al., 2024) that can evolve into deception even under benign prompts, where more capable models show greater deceptive capabilities (Wu et al., 2025) yet struggle to detect others' lies (Curvo, 2025).

This creates an unbalanced dynamic where the collective effort of MAS could be (too) easily impacted by one influential actor, who may actively deceive to gain more agreements from other agents and dominate the discussion. Potential solvency to this problem includes developing more heterogeneous MAS with different backbone models, contexts and system settings with tool access (Ye et al., 2025) to avoid over-simplistic convergence on similar model/context prior, as well as incorporating merit-based credit mechanism (Zhang et al., 2025b) that encourages active engagement (judged by a third-party audit agent (Duan & Wang, 2024) to assign credit based on their material contribution (improving accuracy of such audit to distinguish genuine contribution from empty arguments remain an open direction for future work).

R4: Assess quality of formulated hypothesis with uncertainty. The assessment of hypothesis quality is highly challenging, typically requiring not only objective evidence but also subjective judgment based on *scientific intuition* honed by years of domain experience to see if a hypothesis is worth pursuing from many aspects. While it's still open question as to whether such *scientific intuition* can emerge within AI systems, we mainly argue that MAS can better reflect the multi-dimensional nature of quality assessment in hypothesis formulation.

Specifically, different agents could use different sets of criteria to assess hypothesis quality in various aspects (taking the example of research on a new material: chemical/thermodynamic stability, mechanical strength, manufacturing scalability and environmental impact all need to play a role in hypothesis assessment). Each agent can focus on one concrete aspect without interference at first, and a central evaluator agent can take over at the end for final assessment.

PharmaSwarm (Song et al., 2025) showcased the strength of MAS in hypothesis-driven drug discovery where each agent access dedicated functionality such as genomic analysis, biomedical knowledge graph and binding affinity prediction. A central evaluator agent then ranks the proposals for new drugs by multi-dimensional metrics including biological plausibility, novelty, in silico efficacy, and safety, which accelerate translational research and deliver high-confidence hypotheses more efficiently than traditional pipeline.

Future MAS should also carefully quantify and attribute uncertainty (which is an inherent property of scientific hypothesis) to each agent's proposal using unified, trustworthy uncertainty quantification framework (Yoffe et al., 2025), by carefully inspecting how uncertainty evolves between different agents (Zhao et al., 2025), we could pinpoint vulnerable points where human scientists should then step in to give key guidance in steering the whole system towards creative, yet also practical and trustworthy hypothesis (Tang et al., 2025a; Ghafarollahi & Buehler, 2024).

4 Experiment Execution

Experiment execution involves translating theoretical hypotheses into concrete implementations through dry-lab via computational simulation and wet-lab protocols within physical labs to validate or adjust proposed hypothesis. Scientific progress often requires exploring multiple competing hypotheses in parallel, MAS could facilitate such multi-tasking (Kusne & McDannald, 2023) with independent sub-agents and provide fail-safe redundancy where backup agents can step in when others fail or break, providing better fault-tolerance for the system as a whole to operate normally even when some components malfunction.

It's also worth noting that having *too many tools* often overwhelm SAS as to which one they should use, which subsequently increase misuse and error in tool-using (Lenhard, 2025)). MAS could mitigate such problems by specializing each agent to focus on fewer tools and agent skills (Su et al., 2025).

O5: Accelerate experimental execution with parallel action. As we discussed in the definition section, SAS process agentic action requests sequentially that accumulates latency in waiting for task completion and results in substantially longer total execution time. MAS enables parallel action by distributing workload across multiple agents, which significantly reduces overall execution time as the system can exploit concurrent (sometimes asynchronous) actions for independent subtasks (Fourney et al., 2024).

Empirical studies (Zhang et al., 2025a) confirms that parallel action in MAS can achieve up to $2.2\times$ speedup while also improving task completion rates on the GAIA benchmark (Mialon et al., 2023).

AgenticSciML (Jiang & Karniadakis, 2025) similarly demonstrates coordinated proposers, critics, engineers, and evaluators operating in evolutionary cycles, where complementary validation steps proceed concurrently to boost overall throughput. Similar multi-agent decomposition appears in AutoLabs (Panapitiya et al., 2025), where different agents handle goal decomposition, stoichiometric computation, and validation in coordinated cycles. By comparison, a single-agent approach must interleave proposing, checking, and revising within one reasoning trajectory, which not only slows execution but also limits opportunities for cross-checking through parallel validation.

O6: Improve fault-tolerance of MAS by backup agents. Having multiple agents for the same step provides fault-tolerant alternatives that can substitute for failed agents, ensuring the system continues functioning despite individual failures (Li et al., 2025a). This redundancy allows the system to bypass individual errors tied to a particular reasoning pattern or tool chain and to continue execution through alternative solution strategies. In comparison, SAS typically retries the same failing method, so errors are more prone to block progress as recurring patterns. MAS further provides stronger trouble-shooting because different agents can attempt different tools and methods, increasing overall chance of success. In addition, (Qi et al., 2025a) draws inspiration from blockchain protocols to support transparent agent registration and verifiable task allocation. It further enables dynamic tracking of agent strength through smart contracts, leading to higher task success rates, more stable utility distribution, and emergent agent specialization.

R5: Prevent concurrent conflicts by proactive coordination. When agents work simultaneously on shared code without explicit coordination, they can produce incompatible versions that blocks the entire system. For example, one agent may implement functions using NumPy arrays while another relies on Pandas DataFrames, making integration impossible due to mismatched data structures. Conflicts are further amplified when agents modify existing code under incompatible assumptions, such as when one refactors variable names while another simultaneously adds functionality that depends on the original names. The catastrophic consequence of such concurrent conflicts motivates the need for design-time constraints to prevent them preemptively rather than relying solely on post-hoc fixes (Pugachev, 2025).

To address these failures modes for more scalable collaboration, future MAS should implement proactive coordination mechanisms, including exclusive ownership and traceability, merge protocols with mandatory review before integration (Huang et al., 2025), and interface contracts that specify inputs and outputs in advance so agents can work independently on team-level components without jeopardizing system-level dependencies (Tao et al., 2024; Wu, 2025).

R6: Optimize MAS topology to balance communication overhead with performance gains. As the number of agents increases in a fully-connected MAS (i.e. any two agents may communicate with each other), each additional agent must coordinate with all existing agents, yielding $O(n^2)$ communication overhead while benefits grow at a much slower pace (Zhang et al., 2024; Yan et al., 2025a). Indeed, previous work has reported diminishing returns as the number of agents exceeds certain context-dependent thresholds (Yang et al., 2025; Kim et al., 2025; junyou li et al., 2024).

To address this challenge, future MAS should reduce communication overhead by designating clear chains of command and ensure only necessary connections are established between worker agents and their respective supervisor, reflecting a clear division of team responsibility (Du et al., 2025) just like human research collaboration. Such hierarchical topologies could reduce coordination complexity from $O(n^2)$ to $O(n \log n)$ or $O(n)$ depending on branching factor as shown in MASTER (Rothfarb et al., 2025), where hierarchical MAS collaboration has shown promise to accelerate material discovery using density functional theory (DFT) workflow.

5 Result Interpretation and Evaluation

Result interpretation involves transforming experimental outcomes into scientific claims by evaluating statistical significance, assessing reliability, contextualizing findings within existing knowledge, and communicating conclusions.

O7: More diverse input from MAS offers more robust interpretation. Single agents are commonly trained to optimize self-consistency (Lee et al., 2025; Taubenfeld et al., 2025), which encourages them to maintain one coherent narrative and iteratively adjust it over turns than reconsidering alternative interpretations. On the other side, MAS offers a more robust alternative that designate

different agents to argue for/against certain interpretation, which fosters more engaged discussion that could surface potential weaknesses (Yu et al., 2025a; Zhu et al., 2025b; Jin et al., 2024).

Recent work (Fan et al., 2025; Inoue et al., 2025) further show that both efficiency and accuracy improved with weighted discussion mechanism such as Weighted Iterative Society-of-Experts (WISE), outperforming vanilla MAD across diverse tasks and model configurations. This kind of weighted decision-making could be based on various mechanisms from probabilistic aggregation of annotator error rates (Dawid & Skene, 1979) to peer ranking and consensus-based discussion among diverse models (Li et al., 2023; Chen et al., 2024a), thereby grounding how much each agent's opinion should count towards the final result with their respective strength and weakness.

O8: Assist peer review with automated checks and more diverse input. Increasing use of AI for peer review is rapidly becoming an inevitable trend, where frontier venues like AAAI and ICML (Naddaf, 2025) have started pilot trials with AI reviewers (for feedback, not decision). Under this prevalent trend, we argue that SAS review is likely worse as it's one single voice that people can simply copy from (or paraphrase with minimal effort). On the contrary, MAS could enable multiple AI reviewers (Fan et al., 2025) to produce independent reviews (Lan et al., 2024) instead of a single one, which prompts human reviewer/ACs to at least compare and balance various views when making decisions at critical junctures.

Further, multiple agents could also enable efficient checks for several labor-intensive aspects in peer review (most of which are impractical for human reviewers to check by hand due to the increasingly heavy review workload), including flagging potential hallucinations, marking thinly sliced contributions in thousands of submissions and automating reproducibility checks. Indeed, many studies (Siegel et al., 2024; Starace et al., 2025; Kon et al., 2025; Liu et al., 2025b; Ifargan et al., 2024; Seo et al., 2025; Huang et al., 2024) have shown remarkable progress of LLM agents to reproduce experiments from papers. Future MAS bear the potential to significantly accelerate (even automate) the process of reproducibility checks in peer review (which is far too labor-intensive for reviewers to do so manually) and flag any potential issues, thereby motivating better reproducibility practices for the community at large.

We firmly believe that MAS for review should never take over the right to decide, but serve as an information aggregator that helps human reviewers focus on the most important parts of a submission (instead of manually checking every citation or going through hundreds of pages in appendix) to make more informed decisions.

R7: Prevent metric gaming by mutual oversight. Reward hacking, formally defined as exploiting the difference between a true reward and a proxy reward (Skalse et al., 2025), often manifests in *metric gaming* where AI competently pursue higher scores in a specific metric by cheating without actually solving the problem. (Bondarenko et al., 2025) reported that frontier reasoning models like OpenAI-o3 and DeepSeek-R1 could actively game the metric by trying to delete opponents rather than winning with genuinely better strategies. This observation further extends to real-world cases in scientific workflows: Sakana AI also reported metric gaming in their AI Scientist (Yamada et al., 2025) system: when experiments took too long and hit timeout limits, AI Scientist simply tried to modify the timeout period instead of optimizing the code, which is a classic example of metric gaming in research environments. Similarly, PostTrainBench (Rank et al., 2025) also revealed that LLMs may attempt to directly train on the test set when tasked with training another model on a given benchmark. Such behavior is particularly destructive in result interpretation, as many experimental results can be interpreted in many ways, some of which can be bended, warped and massaged to arbitrarily fit in certain narrative, leading to misrepresentation of science (Glockner et al., 2024). Future systems should implement scalable oversight (Bowman et al., 2022) that verify each step for genuine problem-solving rather than metric gaming, where stronger model capability can proportionately enhance (instead of breach) oversight.

R8: Mitigate agentic misalignment by mutual reasoning. LLM Sycophancy describes the tendency where models tailor responses to input prompts rather than respond objectively (Sharma et al., 2025), leading to bogus interpretation or evaluation contingent on prior context. Empirical evidence have shown that LLMs can progress from sycophancy to full-scale subterfuge (Denison et al., 2024). In MAS, sycophancy could further snowball as agents cater to other agents' input rather than maintaining independent perspectives, exacerbating shared biases and collusion (Motwani et al., 2025). The threat model for collusion encompasses both inadvertent convergence (from shared bias due to

similar training data/paradigm) and adversarial manipulation (from bad actors that mislead collective decision-making (Liu et al., 2025a)).

To address these challenges, mutual reasoning (Qi et al., 2025b) have shown considerable efficacy, where agents actively reason about the states and actions of other agents to enable more transparent coordination. PeerGuard (Fan & Li, 2025) showcased this approach by having each agent evaluates others' response to detect illogical reasoning processes indicative of malicious actors, achieving high accuracy in identifying poisoned agents while minimizing false positives on clean agents. Having peer agents as overseer of each other is an intrinsic advantage of MAS and could inspire future paradigm of more transparent and robust AI control methods (Pecerskis & Smirnovs, 2026).

6 Call to Action

We call on two key group of stakeholders with detailed recommendations: AI researchers that work on developing MAS and researchers that use them in scientific applications.

6.1 For AI Researchers Developing MAS for Science

Trust but verify. Build more hallucination-free MAS with provenance-tracked external memory where every output traces back to programmatically verifiable sources (such as CrossRef). Hallucination-free is a *pre-requisite* for trustworthy AI, which many existing DeepResearch tools (OpenAI, 2025) sadly do not satisfy, we especially need rigorous provenance tracking to prevent the catastrophic cascade of hallucination via inter-agent communications. Any intermediate results or actions should be directly attributed to responsible agent(s) in the form of digital signature and reasoning record. Such record must remain transparent for inspection through a user-friendly interface that makes it easy for scientists without AI background to navigate.

Scale with caution. Scale carefully from a small number of agents to strike a balance between performance gains vs. growing communication overhead that diminish scaling benefits (Yang et al., 2025; Gao et al., 2025). It is also critical to use proper baselines when calculating such performance gains (for instance, a 3-agent MAS consumes more tokens than pass@3 on a single agent due to communication), we should carefully evaluate SAS vs. MAS with equal compute to see performance gain stem from MAS schema or simply a result of more test-time compute.

Be careful with over-anthropomorphizing AI. While human collaboration patterns has already guided previous success like LLM Debate (Khan et al., 2024), we also need to be careful with anthropomorphizing them too much (Deshpande et al., 2023) as AI may interact in ways unseen in human teams, some of which we cannot yet fully understand (Cloud et al., 2025). AI developers should not guide MAS design by simply copy-paste from human organizational theory. Instead, one can always take inspiration yet adopt a rigorous, evidence-based approach that measures quantifiable improvements for each setting. What works for human collaboration may not work for AI, and vice versa.

6.2 For Researchers Working on Scientific Applications

Clear instruction at each step. A critical advantage of MAS is that each agent can dedicate to one task (or one aspect of a large task) with user-defined scope and tools. This role separation creates natural compartmentalization for user to clearly track every aspect, which also requires researchers to clearly define the scope, task and toolbox for each agent before putting MAS to work.

Human judgment at critical junctures. AI cannot replace human judgment honed through years of domain expertise. Frontline researchers should thoroughly understand both the strengths and caveats of AI tools to use MAS as facilitators rather than substitutes of their own thinking, especially at critical junctures.

Communicate genuine research needs. AI developers cannot build truly useful tools without understanding what frontline scientists genuinely need. Current scientific reasoning benchmarks heavily focus on "test-taking" capability using Olympiad questions, yet the ability to solve IMO problems do not necessarily transfer to research activities. We need a inclusive platform to match supply from people who build AI tools and demand of people who use them.

REFERENCES

Anthropic. How we built our multi-agent research system, 2025. URL https://www.anthropic.com/engineering/multi-agent-research-system.

Shiva Aryal, Tuyen Do, Bisesh Heyojoo, Sandeep Chataut, Bichar Dip Shrestha Gurung, Venkataramana Gadhamshetty, and Etienne Gnimpieba. Leveraging multi-ai agents for cross-domain knowledge discovery, 2024. URL https://arxiv.org/abs/2404.08511.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks, 2025. URL https://arxiv.org/abs/2412.15204.

Chaithanya Bandi and Abir Harrasse. Debate, deliberate, decide (d3): A cost-aware adversarial framework for reliable and interpretable llm evaluation, 2024. URL https://api.semanticscholar.org/CorpusID:273185437.

Dirk Betz, Claudia Biniossek, Christophe Blanchi, Felix Henninger, Thomas Lauer, Philipp Wieder, Peter Wittenburg, and Martin Zünkeler. Canonical workflow for experimental research. *Data Intelligence*, 4(2):155–172, 2022. URL https://doi.org/10.1162/dint_a_00123.

Alexander Bondarenko, Denis Volk, Dmitrii Volkov, and Jeffrey Ladish. Demonstrating specification gaming in reasoning models, 2025. URL https://arxiv.org/abs/2502.13295.

Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jackson Kernion, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Robin Larson, Sam McCandlish, Sandipan Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. Measuring progress on scalable oversight for large language models, 2022. URL https://arxiv.org/abs/2211.03540.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7066–7085, 2024a. URL https://arxiv.org/abs/2309.13007.

Weize Chen, Ziming You, Ran Li, Yitong Guan, Cheng Qian, Chenyang Zhao, Cheng Yang, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. Internet of agents: Weaving a web of heterogeneous agents for collaborative intelligence. *ArXiv*, abs/2407.07061, 2024b. URL https://api.semanticscholar.org/CorpusID:271064295.

Anoop Cherian, River Doyle, Eyal Ben-Dov, Suhas Lohit, and Kuan-Chuan Peng. Wise: Weighted iterative society-of-experts for robust multimodal multi-agent debate, 2025. URL https://arxiv.org/abs/2512.02405.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Subliminal learning: Language models transmit behavioral traits via hidden signals in data, 2025. URL https://arxiv.org/abs/2507.14805.

Pedro M. P. Curvo. The traitors: Deception and trust in multi-agent language model simulations, 2025. URL https://arxiv.org/abs/2505.12923.

Ahmed M. Darwish, Essam A. Rashed, and Ghada Khoriba. Mitigating LLM hallucinations using a multi-agent framework. *Inf.*, 16(7):517, 2025. doi: 10.3390/INFO16070517. URL https://doi.org/10.3390/info16070517.

Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28 (1):20–28, 1979.

DeepMind. Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad, July 2025. URL https://deepmind.google/discover/blog/advanced-version-of-gemini-with-deep-think-officially-achieves-gold-medal-standard DeepMind Blog Post.

Carson Denison, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, Nicholas Schiefer, Ryan Soklaski, Alex Tamkin, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, Ethan Perez, and Evan Hubinger. Sycophancy to subterfuge: Investigating reward-tampering in large language models, 2024. URL https://arxiv.org/abs/2406.10162.

A. Deshpande, Tanmay Rajpurohit, Karthik Narasimhan, and A. Kalyan. Anthropomorphization of ai: Opportunities and risks. *ArXiv*, abs/2305.14784, 2023. URL https://api.semanticscholar.org/CorpusID:258866093.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, 2024. URL https://openreview.net/forum?id=zj7YuTE4t8.

Zhuoyun Du, Chen Qian, Wei Liu, Zihao Xie, YiFei Wang, Rennai Qiu, Yufan Dang, Weize Chen, Cheng Yang, Ye Tian, Xuantang Xiong, and Lei Han. Multi-agent collaboration via cross-team orchestration, 2025. URL https://arxiv.org/abs/2406.08979.

Zhihua Duan and Jialin Wang. Enhancing multi-agent consensus through third-party llm integration: Analyzing uncertainty and mitigating hallucinations in large language models, 2024. URL https://arxiv.org/abs/2411.16189.

Falong Fan and Xi Li. Peerguard: Defending multi-agent systems against backdoor attacks through mutual reasoning. In *25th IEEE International Conference on Information Reuse and Integration and Data Science, IRI 2025, San Jose, CA, USA, August 6-8, 2025*, pp. 234–239. IEEE, 2025. doi: 10.1109/IRI66576.2025.00051. URL https://doi.org/10.1109/IRI66576.2025.00051.

Wei Fan, JinYi Yoon, and Bo Ji. imad: Intelligent multi-agent debate for efficient and accurate llm inference, 2025. URL https://arxiv.org/abs/2511.11306.

Yu Feng, Phu Mon Htut, Zheng Qi, Wei Xiao, Manuel Mager, Nikolaos Pappas, Kishaloy Halder, Yang Li, Yassine Benajiba, and Dan Roth. Rethinking LLM uncertainty: A multi-agent approach to estimating black-box model uncertainty. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.660. URL https://aclanthology.org/2025.findings-emnlp.660/.

Adam Fourney, Gagan Bansal, Hussein Mozannar, Cheng Tan, Eduardo Salinas, Erkang Zhu, Friederike Niedtner, Grace Proebsting, Griffin Bassman, Jack Gerrits, Jacob Alber, Peter Chang, Ricky Loynd, Robert West, Victor Dibia, Ahmed Awadallah, Ece Kamar, Rafah Hosn, and Saleema Amershi. Magentic-one: A generalist multi-agent system for solving complex tasks, 2024. URL https://arxiv.org/abs/2411.04468.

Sasi Kiran Gaddipati, Farhana Keya, Gollam Rabby, and Sören Auer. Aissistant: An agentic approach for human–ai collaborative scientific work on reviews and perspectives in machine learning, 2025. URL https://arxiv.org/abs/2509.12282.

Mingyan Gao, Yanzi Li, Banruo Liu, Yifan Yu, Phillip Wang, Ching-Yu Lin, and Fan Lai. Single-agent or multi-agent systems? why not both?, 2025. URL https://arxiv.org/abs/2505.18286.

Alireza Ghafarollahi and Markus J. Buehler. Sciagents: Automating scientific discovery through multi-agent intelligent graph reasoning, 2024. URL https://arxiv.org/abs/2409.05556.

Max Glockner, Yufang Hou, Preslav Nakov, and Iryna Gurevych. Missci: Reconstructing fallacies in misrepresented science, 2024. URL https://arxiv.org/abs/2406.03181.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R D Costa, José R Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Towards an ai co-scientist, 2025. URL https://arxiv.org/abs/2502.18864.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pp. 8048–8057. ijcai.org, 2024. URL https://www.ijcai.org/proceedings/2024/890.

Ignacio Gustin, Luis Mantilla Calderón, Juan B. Pérez-Sánchez, Jérôme F. Gonthier, Yuma Nakamura, Karthik Panicker, Manav Ramprasad, Zijian Zhang, Yunheng Zou, Varinia Bernales, and Alán Aspuru-Guzik. El agente cuántico: Automating quantum simulations, 2026. URL https://arxiv.org/abs/2512.18847.

Jennifer Haase and S. Pokutta. Beyond static responses: Multi-agent llm systems as a new paradigm for social science research. *ArXiv*, abs/2506.01839, 2025. URL https://api.semanticscholar.org/CorpusId:279118505.

Xi He, Sirui Lu, and Bei Zeng. Co-designing quantum codes with transversal diagonal gates via multi-agent systems, 2025. URL https://arxiv.org/abs/2510.20728.

Kelly Hong, Anton Troynikov, and Jeff Huber. Context rot: How increasing input tokens impacts llm performance, July 2025. URL https://research.trychroma.com/context-rot.

Beichen Huang, Ran Cheng, and Kay Chen Tan. Evogit: Decentralized code evolution via git-based multi-agent collaboration. *arXiv preprint arXiv:2506.02049*, 2025.

Dong Huang, Jie M. Zhang, Michael Luck, Qingwen Bu, Yuhao Qing, and Heming Cui. Agentcoder: Multi-agent-based code generation with iterative testing and optimisation, 2024. URL https://arxiv.org/abs/2312.13010.

Tal Ifargan, Lukas Hafner, Maor Kern, Ori Alcalay, and Roy Kishony. Autonomous llm-driven research from data to human-verifiable research papers. *ArXiv*, abs/2404.17605, 2024. URL https://api.semanticscholar.org/CorpusId:269448624.

Yoshitaka Inoue, Tianci Song, Xinling Wang, Augustin Luna, and Tianfan Fu. Drugagent: Multi-agent large language model-based reasoning for drug-target interaction prediction. *ArXiv*, pp. arXiv–2408, 2025.

Qile Jiang and George Karniadakis. Agenticsciml: Collaborative multi-agent systems for emergent discovery in scientific machine learning, 2025. URL https://arxiv.org/abs/2511.07262.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. Agentreview: Exploring peer review dynamics with LLM agents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 1208–1226. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.70. URL https://doi.org/10.18653/v1/2024.emnlp-main.70.

Zehan Jin, Qi Wu, Chengxi Li, Jie Li, Yingda Lu, Wenhao Xu, Yuze Liao, Lei Feng, Ming Hu, and Bo Li. Topomas: Large language model driven topological materials multiagent system. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2507.04053.

junyou li, Qin Zhang, Yangbin Yu, QIANG FU, and Deheng Ye. More agents is all you need. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=bgzUSZ8aeg.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate, 2025. URL https://arxiv.org/abs/2509.04664.

Yujing Ke, Kevin George, Kathan Pandya, David Blumenthal, Maximilian Sprang, Gerrit Großmann, Sebastian Vollmer, and David Antony Selby. Biodisco: Multi-agent hypothesis generation with dual-mode evidence, iterative feedback and temporal evaluation, 2025. URL https://arxiv.org/abs/2508.01285.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive llms leads to more truthful answers, 2024. URL https://arxiv.org/abs/2402.06782.

Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, A. Ali Heydari, Yao Yan, Zhihan Zhang, Yuchen Zhuang, Mark Malhotra, Paul Pu Liang, Hae Won Park, Yuzhe Yang, Xuhai Xu, Yilun Du, Shwetak Patel, Tim Althoff, Daniel McDuff, and Xin Liu. Towards a science of scaling agent systems, 2025. URL https://arxiv.org/abs/2512.08296.

Patrick Tser Jern Kon, Jiachen Liu, Xinyi Zhu, Qiuyi Ding, Jingjia Peng, Jiarong Xing, Yibo Huang, Yiming Qiu, Jayanth Srinivasa, Myungjin Lee, Mosharaf Chowdhury, Matei Zaharia, and Ang Chen. Exp-bench: Can ai conduct ai research experiments? *ArXiv*, abs/2505.24785, 2025. URL https://api.semanticscholar.org/CorpusID:279070803.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 517–545, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.29. URL https://aclanthology.org/2024.findings-acl.29/.

A. Gilad Kusne and Austin McDannald. Scalable multi-agent lab framework for lab optimization. *Matter*, 6(6):1880–1893, 2023. ISSN 2590-2385. doi: https://doi.org/10.1016/j.matt.2023.03.022. URL https://www.sciencedirect.com/science/article/pii/S2590238523001248.

Ryan Zheyuan Lai and Yingming Pu. Prim: Principle-inspired material discovery through multi-agent collaboration. In *AI for Accelerated Materials Design - ICLR 2025*, 2025. URL https://openreview.net/forum?id=lhobZk76wX.

Tian Lan, Wenwei Zhang, Chengqi Lyu, Shuaibin Li, Chen Xu, Heyan Huang, Dahua Lin, Xian-Ling Mao, and Kai Chen. Training language models to critique with multi-agent feedback. *ArXiv*, abs/2410.15287, 2024. URL https://api.semanticscholar.org/CorpusId:273501612.

Hao Duong Le, Xin Xia, and Zhang Chen. Multi-agent causal discovery using large language models. *arXiv preprint arXiv:2407.15073*, 2024.

Jaehyeok Lee, Keisuke Sakaguchi, and JinYeong Bak. Self-training meets consistency: Improving llms' reasoning with consistency-driven rationale evaluation, 2025. URL https://arxiv.org/abs/2411.06387.

Matthew Lenhard. Too many tools? how llms struggle at scale — mcp talk w/ matthew lenhard. https://www.youtube.com/watch?v=ej7-n9OoGnQ, 2025. MCP Developers Summit. Accessed: 2026-01-20.

Ruosen Li, Teerth Patel, and Xinya Du. Prd: Peer rank and discussion improve large language model based evaluations. *arXiv preprint arXiv:2307.02762*, 2023.

Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1(1):9, 2024. URL https://link.springer.com/article/10.1007/s44336-024-00009-2.

Yaoru Li, Shunyu Liu, Tongya Zheng, and Mingli Song. Parallelized planning-acting for efficient llm-based multi-agent systems, 2025a. URL https://arxiv.org/abs/2503.03505.

Yilong Li, Chen Qian, Yu Xia, Ruijie Shi, Yufan Dang, Zihao Xie, Ziming You, Weize Chen, Cheng Yang, Weichuan Liu, Ye Tian, Xuantang Xiong, Lei Han, Zhiyuan Liu, and Maosong Sun. Cross-task experiential learning on llm-based multi-agent collaboration, 2025b. URL https://arxiv.org/abs/2505.23187.

Xixun Lin, Yucheng Ning, Jingwen Zhang, Yan Dong, Yilong Liu, Yongxuan Wu, Xiaohua Qi, Nan Sun, Yanmin Shang, Kun Wang, Pengfei Cao, Qingyue Wang, Lixin Zou, Xu Chen, Chuan Zhou, Jia Wu, Peng Zhang, Qingsong Wen, Shirui Pan, Bin Wang, Yanan Cao, Kai Chen, Songlin Hu, and Li Guo. Llm-based agents suffer from hallucinations: A survey of taxonomy, methods, and directions, 2025. URL https://arxiv.org/abs/2509.18970.

Fengyuan Liu, Rui Zhao, Shuo Chen, Guohao Li, Philip Torr, Lei Han, and Jindong Gu. Can an individual manipulate the collective decisions of multi-agents? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12158–12182, Suzhou, China, November 2025a. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.611. URL https://aclanthology.org/2025.emnlp-main.611/.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts, 2023. URL https://arxiv.org/abs/2307.03172.

Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. Researchbench: Benchmarking llms in scientific discovery via inspiration-based task decomposition. *ArXiv*, abs/2503.21248, 2025b. URL https://api.semanticscholar.org/CorpusId:277349293.

Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. The assistant axis: Situating and stabilizing the default persona of language models, 2026. URL https://arxiv.org/abs/2601.10387.

Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. Llm4sr: A survey on large language models for scientific research, 2025. URL https://arxiv.org/abs/2501.04306.

Hao Ma, Tianyi Hu, Zhiqiang Pu, Boyin Liu, Xiaolin Ai, Yanyan Liang, and Min Chen. Coevolving with the other you: Fine-tuning llm with sequential cooperative multi-agent reinforcement learning, 2025. URL https://arxiv.org/abs/2410.06101.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants, 2023. URL https://arxiv.org/abs/2311.12983.

Sumeet Ramesh Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip H. S. Torr, Lewis Hammond, and Christian Schroeder de Witt. Secret collusion among ai agents: Multi-agent deception via steganography, 2025. URL https://arxiv.org/abs/2402.07510.

Miryam Naddaf. More than half of researchers now use ai for peer review — often against guidance, Dec 2025. URL https://www.nature.com/articles/d41586-025-04066-5.

Brenda Nogueira, Werner Geyer, Andrew Anderson, Toby Jia-Jun Li, Dongwhi Kim, Nuno Moniz, and Nitesh V. Chawla. From verification burden to trusted collaboration: Design goals for llm-assisted literature reviews, 2025. URL https://arxiv.org/abs/2512.11661.

OpenAI. Introducing deep research. https://openai.com/index/introducing-deep-research/, feb 2025. OpenAI News.

Daniel E. O'Leary. Confirmation and specificity biases in large language models: An explorative study. *IEEE Intelligent Systems*, 40(1):63–68, 2025. doi: 10.1109/MIS.2024.3513992. URL https://ieeexplore.ieee.org/document/10897252.

Gihan Panapitiya, Emily Saldanha, Heather Job, and Olivia Hess. Autolabs: Cognitive multi-agent systems with self-correction for autonomous chemical experimentation, 2025. URL https://arxiv.org/abs/2509.25651.

Tims Pecerskis and Aivars Smirnovs. Mixture-of-models: Unifying heterogeneous agents via n-way self-evaluating deliberation, 2026. URL https://zenodo.org/doi/10.5281/zenodo.18234923.

Yingming Pu, Tao Lin, and Hongyu Chen. Piflow: Principle-aware scientific discovery with multi-agent collaboration. *arXiv preprint arXiv:2505.15047*, 2025.

Sergey Pugachev. Codecrdt: Observation-driven coordination for multi-agent llm code generation, 2025. URL https://arxiv.org/abs/2510.18893.

Minfeng Qi, Tianqing Zhu, Lefeng Zhang, Ningran Li, and Wanlei Zhou. Towards transparent and incentive-compatible collaboration in decentralized llm multi-agent systems: A blockchain-driven approach, 2025a. URL https://arxiv.org/abs/2509.16736.

Zhenting Qi, Mingyuan MA, Jiahang Xu, Li Lyna Zhang, Fan Yang, and Mao Yang. Mutual reasoning makes smaller LLMs stronger problem-solver. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=6aHUmotXaw.

Ben Rank, Hardik Bhatnagar, Matthias Bethge, and Maksym Andriushchenko. Posttrainbench: Measuring ai ability to perform llm post-training, 2025. URL https://posttrainbench.com.

Shaina Raza, Ranjan Sapkota, Manoj Karkee, and Christos Emmanouilidis. Trism for agentic ai: A review of trust, risk, and security management in llm-based agentic multi-agent systems, 2025. URL https://arxiv.org/abs/2506.04133.

Samuel Rothfarb, Megan C. Davis, Ivana Matanovic, Baikun Li, Edward F. Holby, and Wilton J. M. Kort-Kamp. Hierarchical multi-agent large language model reasoning for autonomous functional materials discovery, 2025. URL https://arxiv.org/abs/2512.13930.

Daniel Saeedi, Denise K. Buckner, Jose C. Aponte, and Amirali Aghazadeh. Astroagents: A multi-agent AI for hypothesis generation from mass spectrometry data. In *Towards Agentic AI for Science: Hypothesis Generation, Comprehension, Quantification, and Validation*, 2025. URL https://openreview.net/forum?id=1WUCSNAjjB.

Malik Abdul Sami, Zeeshan Rasheed, Kai-Kristian Kemell, Muhammad Waseem, Terhi Kilamo, Mika Saari, Anh Nguyen-Duc, Kari Systä, and Pekka Abrahamsson. System for systematic literature review using multiple AI agents: Concept and an empirical evaluation. *CoRR*, abs/2403.08399, 2024. doi: 10.48550/ARXIV.2403.08399. URL https://doi.org/10.48550/arXiv.2403.08399.

Minju Seo, Jinheon Baek, Seongyun Lee, and Sung Ju Hwang. Paper2code: Automating code generation from scientific papers in machine learning. *ArXiv*, abs/2504.17192, 2025. URL https://api.semanticscholar.org/CorpusID:278033490.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models, 2025. URL https://arxiv.org/abs/2310.13548.

Xu Shen, Yixin Liu, Yiwei Dai, Yili Wang, Rui Miao, Yue Tan, Shirui Pan, and Xin Wang. Understanding the information propagation effects of communication topologies in llm-based multi-agent systems, 2025. URL https://arxiv.org/abs/2505.23352.

Jinxin Shi, Jiabao Zhao, Xingjiao Wu, Ruyi Xu, Yuan-Hao Jiang, and Liang He. Mitigating reasoning hallucination through multi-agent collaborative filtering. *Expert Syst. Appl.*, 263:125723, 2025. doi: 10.1016/J.ESWA.2024.125723. URL https://doi.org/10.1016/j.eswa.2024.125723.

Nazar Shmatko, Alex Adam, and Paul Esau. GPTZero finds 100 new hallucinations in NeurIPS 2025 accepted papers. GPTZero Blog, January 2026. URL https://gptzero.me/news/neurips/. Accessed: 2026-01-24.

Zachary S. Siegel, Sayash Kapoor, Nitya Nagdir, Benedikt Stroebl, and Arvind Narayanan. Core-bench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://api.semanticscholar.org/CorpusId:272694423.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking, 2025. URL https://arxiv.org/abs/2209.13085.

Gleb Vitalevich Solovev, Alina Borisovna Zhidkovskaya, Anastasia Orlova, Anastasia Vepreva, Tonkii Ilya, Rodion Golovinskii, Nina Gubina, Denis Chistiakov, Timur A. Aliev, Ivan Poddiakov, Galina Zubkova, Ekaterina V. Skorb, Vladimir Vinogradov, Nikolay Nikitin, Andrei Dmitrenko, Anna Kalyuzhnaya, and Andrey Savchenko. Towards LLM-driven multi-agent pipeline for drug discovery: Neurodegenerative diseases case study. In *2nd AI4Research Workshop: Towards a Knowledge-grounded Scientific Research Lifecycle*, 2024. URL https://openreview.net/forum?id=3ncjySu5ro.

Guijin Son, Jiwoo Hong, Honglu Fan, Heejeong Nam, Hyunwoo Ko, Seungwon Lim, Jinyeop Song, Jinha Choi, Gonçalo Paulo, Youngjae Yu, and Stella Biderman. When ai co-scientists fail: Spot-a benchmark for automated verification of scientific research, 2025. URL https://arxiv.org/abs/2505.11855.

Kevin Song, Andrew Trotter, and Jake Y. Chen. Llm agent swarm for hypothesis-driven drug discovery, 2025. URL https://arxiv.org/abs/2504.17967.

Scott Spillias, Paris Tuohy, Matthew Andreotta, Ruby Annand-Jones, Fabio Boschetti, Christopher Cvitanovic, Joseph Duggan, Elisabeth A. Fulton, Denis B. Karcher, Cécile Paris, Rebecca Shellock, and Rowan Trebilco. Human-ai collaboration to identify literature for evidence synthesis. *Cell Reports Sustainability*, 1(7):100132, 2024. ISSN 2949-7906. doi: https://doi.org/10.1016/j.crsus.2024.100132. URL https://www.sciencedirect.com/science/article/pii/S2949790624002076.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. Paperbench: Evaluating ai's ability to replicate ai research. *ArXiv*, abs/2504.01848, 2025. URL https://api.semanticscholar.org/CorpusID:277501993.

Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. On the self-verification limitations of large language models on reasoning and planning tasks, 2024. URL https://arxiv.org/abs/2402.08115.

Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, Philip Torr, Bowen Zhou, and Nanqing Dong. Many heads are better than one: Improved scientific idea generation by A llm-based multi-agent system. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 28201–28240. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.1368/.

Kaiwen Tang, Aitong Wu, Yao Lu, and Guangda Sun. Collaborative editable model. *ArXiv*, abs/2506.14146, 2025a. URL https://api.semanticscholar.org/CorpusID:279410272.

Xiangru Tang, Tianrui Qin, Tianhao Peng, Ziyang Zhou, Daniel Shao, Tingting Du, Xinming Wei, Peng Xia, Fang Wu, He Zhu, Ge Zhang, Jiaheng Liu, Xingyao Wang, Sirui Hong, Chenglin Wu, Hao Cheng, Chi Wang, and Wangchunshu Zhou. Agent kb: Leveraging cross-domain experience for agentic problem solving, 2025b. URL https://arxiv.org/abs/2507.06229.

Wei Tao, Yucheng Zhou, Yanlin Wang, Wenqiang Zhang, Hongyu Zhang, and Yu Cheng. Magis: Llm-based multi-agent framework for github issue resolution. *Advances in Neural Information Processing Systems*, 37:51963–51993, 2024.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. Confidence improves self-consistency in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20090–20111, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1030. URL https://aclanthology.org/2025.findings-acl.1030/.

Guiyao Tie, Pan Zhou, and Lichao Sun. A survey of ai scientists, 2026. URL https://arxiv.org/abs/2510.23045.

Florian Tramèr. Trends in LLM-generated citations on arXiv, 2025. URL https://spylab.ai/blog/hallucinations. SPY Lab Blog.

Tam Trinh, Manh Nguyen, and Truong-Son Hy. Towards robust fact-checking: A multi-agent system with advanced evidence retrieval, 2025. URL https://arxiv.org/abs/2506.17878.

Keisuke Ueda, Wataru Hirota, Takuto Asakura, Takahiro Omi, Kosuke Takahashi, Kosuke Arima, and Tatsuya Ishigaki. Exploring design of multi-agent llm dialogues for research ideation, 2025. URL https://arxiv.org/abs/2507.08350.

Karthik Valmeekam, Matthew Marquez, and Subbarao Kambhampati. Can large language models really improve by self-critiquing their own plans?, 2023. URL https://arxiv.org/abs/2310.08118.

Danqing Wang, Zhuorui Ye, Fei Fang, and Lei Li. Cooperative strategic planning enhances reasoning capabilities in large language models. *ArXiv*, abs/2410.20007, 2024. URL https://api.semanticscholar.org/CorpusID:273654271.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. Cycleresearcher: Improving automated research via automated review, 2025. URL https://arxiv.org/abs/2411.00816.

Junde Wu. Git context controller: Manage the context of llm-based agents like git. *arXiv preprint arXiv:2508.00031*, 2025.

Zhaomin Wu, Mingzhe Du, See-Kiong Ng, and Bingsheng He. Beyond prompt-induced lies: Investigating llm deception on benign prompts, 2025. URL https://arxiv.org/abs/2508.06361.

Stefan Wuchty, Benjamin F. Jones, and Brian Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316:1036 – 1039, 2007. URL https://api.semanticscholar.org/CorpusID:260992737.

Tong Xia, Jiankun Zhang, Ruiwen You, Ao Xu, Linghao Zhang, Tengyao Tu, Jingzhi Wang, Jinghua Piao, Yunke Zhang, Fengli Xu, and Yong Li. Ai urban scientist: Multi-agent collaborative automation for urban research, 2025. URL https://arxiv.org/abs/2512.07849.

Licong Xu, Milind Sarkar, Anto I. Lonappan, Íñigo Zubeldia, Pablo Villanueva-Domingo, Santiago Casas, Christian Fidler, Chetana Amancharla, Ujjwal Tiwari, Adrian Bayer, Chadi Ait Ekioui, Miles Cranmer, Adrian Dimitrov, James Fergusson, Kahaan Gandhi, Sven Krippendorf, Andrew Laverick, Julien Lesgourgues, Antony Lewis, Thomas Meier, Blake Sherwin, Kristen Surrao, Francisco Villaescusa-Navarro, Chi Wang, Xueqing Xu, and Boris Bolliet. Open source planning & control system with language agents for autonomous scientific discovery. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2507.07257.

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

Bingyu Yan, Xiaoming Zhang, Litian Zhang, Lian Zhang, Ziyi Zhou, Dezhuang Miao, and Chaozhuo Li. Beyond self-talk: A communication-centric survey of llm-based multi-agent systems. *ArXiv*, abs/2502.14321, 2025a. URL https://api.semanticscholar.org/CorpusId:276482686.

Kai Yan, Zhan Ling, Kang Liu, Yifan Yang, Ting-Han Fan, Lingfeng Shen, Zhengyin Du, and Jiecao Chen. Mir-bench: Can your llm recognize complicated patterns via many-shot in-context reasoning?, 2025b. URL https://arxiv.org/abs/2502.09933.

Yongjin Yang, Euiin Yi, Jongwoo Ko, Kimin Lee, Zhijing Jin, and SeYoung Yun. Revisiting multi-agent debate as test-time scaling: A systematic study of conditional effectiveness. *ArXiv*, abs/2505.22960, 2025. URL https://api.semanticscholar.org/CorpusId:278996680.

Bohan Yao, Shiva Krishna Reddy Malay, and Vikas Yadav. Arm: Discovering agentic reasoning modules for generalizable multi-agent systems, 2025. URL https://arxiv.org/abs/2510.05746.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.

Rui Ye, Xiangrui Liu, Qimin Wu, Xianghe Pang, Zhenfei Yin, Lei Bai, and Siheng Chen. X-mas: Towards building multi-agent systems with heterogeneous llms, 2025. URL https://arxiv.org/abs/2505.16997.

Luke Yoffe, Alfonso Amayuelas, and William Yang Wang. DebUnc: Improving large language model agent communication with uncertainty metrics. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23299–23315, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.1265. URL https://aclanthology.org/2025.findings-emnlp.1265/.

Weilun Yu, Shixiang Tang, Yonggui Huang, Nanqing Dong, Li Fan, Honggang Qi, Wei Liu, Xiaoli Diao, Xi Chen, and Wanli Ouyang. Dynamic knowledge exchange and dual-diversity review: Concisely unleashing the potential of a multi-agent research team, 2025a. URL https://arxiv.org/abs/2506.18348.

Xinlei Yu, Chengming Xu, Guibin Zhang, Yongbo He, Zhangquan Chen, Zhucun Xue, Jiangning Zhang, Yue Liao, Xiaobin Hu, Yu-Gang Jiang, and Shuicheng Yan. Visual multi-agent system: Mitigating hallucination snowballing via visual flow, 2025b. URL https://arxiv.org/abs/2509.21789.

Siyu Yuan, Kaitao Song, Jiangjie Chen, Xu Tan, Dongsheng Li, and Deqing Yang. EvoAgent: Towards automatic multi-agent generation via evolutionary algorithms. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6192–6217, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.315. URL https://aclanthology.org/2025.naacl-long.315/.

Enhao Zhang, Erkang Zhu, Gagan Bansal, Adam Fourney, Hussein Mozannar, and Jack Gerrits. Optimizing sequential multi-step tasks with parallel llm agents, 2025a. URL https://arxiv.org/abs/2507.08944.

Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Xu Yu, and Tianlong Chen. Cut the crap: An economical communication pipeline for llm-based multi-agent systems, 2024. URL https://arxiv.org/abs/2410.02506.

Zhiwei Zhang, Xiaomin Li, Yudi Lin, Hui Liu, Ramraj Chandradevan, Linlin Wu, Minhua Lin, Fali Wang, Xianfeng Tang, Qi He, and Suhang Wang. Unlocking the power of multi-agent llm for reasoning: From lazy agents to deliberation, 2025b. URL https://arxiv.org/abs/2511.02303.

Zhongyue Zhang, Zijie Qiu, Yingcheng Wu, Shuya Li, Dingyan Wang, Zhuomin Zhou, Duo An, Yuhan Chen, Yu Li, Yongbo Wang, Chubin Ou, Zichen Wang, Jack Xiaoyu Chen, Bo Zhang, Yusong Hu, Wenxin Zhang, Zhijian Wei, Runze Ma, Qingwu Liu, Bo Dong, Yuexi He, Qiantai Feng, Lei Bai, Qiang Gao, Siqi Sun, and Shuangjia Zheng. Origene: A self-evolving virtual disease biologist automating therapeutic target discovery. *bioRxiv*, 2025c. URL https://www.biorxiv.org/content/10.1101/2025.06.03.657658v1.

Qiwei Zhao, Dong Li, Yanchi Liu, Wei Cheng, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Huaxiu Yao, Chen Zhao, Haifeng Chen, and Xujiang Zhao. Uncertainty propagation on LLM agent. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6064–6073, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.302. URL https://aclanthology.org/2025.acl-long.302/.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. From automation to autonomy: A survey on large language models in scientific discovery, 2025. URL https://arxiv.org/abs/2505.13259.

Zhiling Zheng, Oufan Zhang, Hamilton Nguyen, Nakul Rampal, Ali H. Alawadhi, Zichao Rong, Teresa Head-Gordon, Christian Borgs, Jennifer Chayes, and Omar Yaghi. Chatgpt research group for optimizing the crystallinity of mofs and cofs. *ACS Central Science*, 9:2161 – 2170, 2023. URL https://www.ncbi.nlm.nih.gov/pubmed/38033801.

Jialong Zhou, Lichao Wang, and Xiao Yang. Guardian: Safeguarding llm multi-agent collaborations with temporal graph modeling, 2025. URL https://arxiv.org/abs/2505.19234.

Junda Zhu, Lingyong Yan, Haibo Shi, Dawei Yin, and Lei Sha. ATM: Adversarial tuning multi-agent system makes a robust retrieval-augmented generator. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10902–10919, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.610. URL https://aclanthology.org/2024.emnlp-main.610/.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You. Multiagentbench: Evaluating the collaboration and competition of llm agents. *ArXiv*, abs/2503.01935, 2025a. URL https://api.semanticscholar.org/CorpusId:276766372.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Deepreview: Improving llm-based paper review with human-like deep thinking process. *ArXiv*, abs/2503.08569, 2025b. URL https://api.semanticscholar.org/CorpusID:276929065.

Zheng Zhuang, Junhao Chen, Hongfu Xu, Yifan Jiang, and Jiangjie Lin. Large language models for automated scholarly paper review: A survey. *Information Fusion*, 124:103332, 2025. ISSN 1566-2535. doi: 10.1016/j.inffus.2025.103332. URL https://doi.org/10.1016/j.inffus.2025.103332.

A Appendix

You may include other additional sections here.