# *Can LLMs Reason Like Scientists?* A Survey on Hypothesis Generation

**Anonymous ACL submission**

## Abstract

*Can machines reason like scientists?* Scientific hypothesis generation—the process of formulating testable explanations for observed phenomena—remains the most critical bottleneck in accelerating scientific discovery. While recent advances in Large Language Models (LLMs) show promise for automating hypothesis generation, the field lacks a systematic understanding of their capabilities, limitations, and optimal application strategies. In this survey, we explore the emerging landscape of LLM-driven hypothesis generation. We present a structured taxonomy of current approaches, analyse domain-specific datasets and evaluation strategies, and discuss open challenges. We review 37 core LLM-based hypothesis/idea generation papers spanning diverse scientific domains from 2023 to 2025. Overall, our goal is to clarify the state of the art, motivate further interdisciplinary research, and provide practical guidance through a continuously updated GitHub[1] repository of relevant papers and resources.

## 1 Introduction

Large Language Models (LLMs) have been widely adopted across numerous natural language processing tasks, including information extraction (Wang et al., 2025), question answering (Kamalloo et al., 2023), summarisation (Ramprasad et al., 2024), and machine translation (Zhu et al., 2024). Building on their success in these tasks, recent research has begun to explore the potential of LLMs for more complex reasoning tasks, particularly scientific hypothesis generation, which requires creative, abductive reasoning rather than pattern recognition. From a philosophy of science perspective, a hypothesis is a tentative explanation or prediction about a phenomenon, formulated to allow for empirical testing and potential falsification (Popper, 1959). Hypothesis generation plays a central role in the scientific process, enabling researchers to propose testable ideas that may lead to discoveries. Traditionally, this process has relied on human intuition, expertise, and domain-specific knowledge.

However, as the volume of scientific literature grows exponentially, researchers are increasingly overwhelmed by the challenge of synthesising information between disciplines. This information overload creates cognitive bottlenecks that hinder the identification of novel insights and interdisciplinary connections. In this context, the question arises: *can LLMs assist in reasoning like scientists and help generate novel hypotheses?* This question has sparked growing interest in the research community. Since 2023, a rising number of studies have investigated the ability of LLMs to generate hypotheses in fields such as computational chemistry (Sprueill et al., 2024), biomedicine (Qi et al., 2024), astronomy (Ciucă et al., 2023) or even in mathematics (Romera-Paredes et al., 2024). Although hypothesis generation has been a long-standing topic of interest with early computational techniques (Karp, 1991; Voytek and Voytek, 2012), recent advances in LLMs have rapidly transformed the field. The pace of innovation in LLM-based approaches has accelerated so quickly that keeping up with emerging developments and challenges has become increasingly complex. One of the central challenges lies in evaluating the hypotheses generated by these models—a task that involves assessing their novelty, feasibility, and clarity. Crucially, it also raises another fundamental question: *to what extent can LLMs produce genuinely original ideas, rather than simply rephrasing or recombining existing knowledge?*

This paper aims to survey the current state of LLM-based hypothesis generation comprehensively. Our main contributions are as follows:

- We introduce a structured taxonomy of LLM-

[1] URL disclosed upon acceptance.

based approaches to hypothesis generation, capturing key modelling paradigms and design choices;

- We compile and analyse a curated list of domain-specific benchmarks and datasets used for evaluating hypothesis generation systems;
- We outline the current limitations and open challenges in the field, and propose directions for future research to strengthen and guide the development of this emerging area.

## 2 Traditional Hypothesis Generation

Before the advent of LLMs, researchers explored hypothesis generation through human-driven and computational methods. Although not exhaustive, this section outlines key pre-LLM approaches to contextualise current developments.

**Human-Centric Approaches** Historically, hypotheses emerged from expert intuition, collaborative reasoning, and domain-specific insights (Swanson, 1986a; Nonaka, 2009). Researchers identified trends or gaps through discourse and practical experience. However, this process was limited by cognitive biases (e.g., confirmation bias) and scalability issues, especially as scientific output increased.

**Literature-Based Discovery** Literature-Based Discovery (LBD) aimed to surface implicit links across publications to address these limits algorithmically. Swanson's foundational work (Swanson, 1986b) demonstrated that previously unlinked literature (e.g., fish oil and Raynaud's syndrome) could yield novel insights. Tools like `ARROWSMITH` (Smalheiser and Swanson, 1998) formalised this through the A-B-C model, identifying bridge terms across disconnected concepts. Later systems improved scalability and semantics: `MOLIERE` (Sybrandt et al., 2017) combined topic modeling (Blei et al., 2003) and phrase mining; `KnIT` (Spangler et al., 2014) used factual networks and information diffusion; and tools like `DiseaseConnect` (Liu et al., 2014) and `BrainSCANr` (Voytek and Voytek, 2012) relied on structured vocabularies such as MeSH (Lipscomb, 2000) to infer disease or gene associations.

**Summary** These early approaches established essential foundations for automated hypothesis generation by highlighting latent connections in literature. Yet they often relied on predefined structures, lacked generalisation across domains, and struggled with causal reasoning. Recent advances in large language models promise to overcome these limitations by enabling more flexible, context-aware, and scalable hypothesis generation, an evolution we explore in the next section.

## 3 LLM-Based Hypothesis Generation

The emergence of LLMs enables new capabilities for scientific hypothesis generation that were barely possible with earlier methods. This section surveys key approaches, from simple prompting to complex autonomous systems, and highlights their potential and current limitations in supporting the scientific discovery process.

### 3.1 Direct Prompting

Initial efforts to use LLMs for scientific hypothesis generation relied on prompting-based approaches. In that section, we categorise these methods into three distinct types, which we describe in detail.

#### 3.1.1 Iterative Feedback

These studies follow a standard iterative loop: generating hypotheses, evaluating them (via tools, humans, or self-critique), and refining outputs based on feedback. Across domains, mathematics (Romera-Paredes et al., 2024), biomedicine (Abdel-Rehim et al., 2024), and the social sciences (Zhou et al., 2024), iterative prompting improved discovery: successful outputs were added back to the search space, experimental results helped refine drug hypotheses (with 3 of 4 new combinations showing synergy), and maintaining a "wrong example bank" boosted generalisation, even outperforming supervised baselines. Other approaches enhanced internal reasoning: Sprueill et al. (2023) used Monte Carlo Tree Search to refine LLM prompts for catalyst design without external data, while `Nova` (Hu et al., 2024) integrated planning, retrieval, and self-correction to generate more diverse and high-quality ideas—though gains plateaued after three iterations. Finally, Qiu et al. (2024) proposed a symbolic evaluation-refinement loop, showing LLMs can produce valid hypotheses across domains but struggle with internal consistency under noise.

#### 3.1.2 Search-Based and Combinatorial Exploration

This category treats hypothesis generation as a search problem over a vast knowledge space, using structured or combinatorial prompting to

guide exploration. CHEMREASONER (Sprueill et al., 2024) exemplifies this approach by coupling LLM-generated catalyst hypotheses with feedback from a GNN trained on quantum-chemical data. It iteratively generates and evaluates natural-language queries using adsorption-energy and reaction-barrier scores, refining prompts via a closed-loop mechanism. Without human intervention, it steers the search toward energetically favourable catalysts. It matches or outperforms expert-designed baselines in key reaction benchmarks—demonstrating how grounded, feedback-driven prompting can accelerate reliable scientific discovery.

#### 3.1.3 General Creativity and Idea Generation

These studies focus on broad idea generation, followed by human or AI-driven evaluation to surface the most novel or valuable outputs. Across domains, from product ideation (Girotra et al., 2023) to neuroscience and biology (O'Brien et al., 2024), and NLP research (Si et al., 2024), prompting large language models yields promising results. While average LLM outputs may be less novel or more homogeneous, the top 10% consistently outperform baselines, with AI-generated ideas up to seven times more likely to rank among the highest-quality. Methods like retrieval-augmented prompting, cross-domain transfer, and self-critique help surface diverse, plausible hypotheses. However, performance plateaus and limited self-evaluation capacities point to the need for external filtering. Creativity assessments such as the AUT (Haase and Hanel, 2023) found GPT-4's originality comparable to humans, and expert evaluations across domains (Park et al., 2023) confirmed that prompt-driven ideation can produce experimentally viable hypotheses. Still, concerns about factual errors, computational cost, and ethical oversight highlight the importance of grounding these systems through robust evaluation loops.

### 3.2 External Knowledge Integration

An emerging trend in LLM-based hypothesis generation is integrating structured external knowledge, such as academic graphs, ontologies, or curated corpora, to enhance factual consistency, novelty, and contextual relevance. This section outlines several approaches that operationalise this idea across domains and methods.

#### 3.2.1 Knowledge and Causal Graph–Based Augmentation

These studies enhance hypothesis generation by grounding LLMs in structured knowledge and causal reasoning. ResearchAgent (Baek et al., 2025) uses a multi-agent, iterative process to identify problems, propose methods, and design experiments, leveraging academic knowledge graphs and entity-centric stores. Human and automated evaluations found it produced more creative and relevant ideas than baselines, though scalability and hallucination remain challenges. In psychology, LLMCG (Tong et al., 2024) combined large-scale literature retrieval, GPT-4–based causal extraction, and link prediction to generate novel, conceptually rich hypotheses—outperforming both scholars and LLMs alone—though some causal links misaligned with expert judgment. To further improve factual grounding, KG-CoI (Xiong et al., 2024) integrated knowledge graphs into idea generation and hallucination detection, boosting consistency and accuracy. However, its performance depended heavily on the quality of the input KG and evaluation dataset.

#### 3.2.2 Literature-Based Inspiration

Wang et al. (2024a) introduced SCIMON, a framework that retrieves "inspirations" from the literature and iteratively optimises hypothesis generation for novelty. Applied to AI/NLP and biomedical domains, SCIMON outperformed baseline LLMs, although the generated ideas still lacked the depth and originality of expert-written papers.

### 3.3 Collaborative Multi-Agent Systems

Multi-agent systems built on LLMs have recently emerged as powerful tools for automating complex scientific workflows, including hypothesis generation. These systems distribute distinct roles, such as ideation, critique, validation, and planning, among specialised agents, enabling them to emulate the collaborative dynamics of real-world scientific teams. Their ability to engage in interactive dialogue, cooperative or adversarial, to refine ideas and improve reasoning (Wu et al., 2023) is a central strength.

#### 3.3.1 Role-Based Multi-Agents

A growing line of work leverages multiple LLM-based agents with specialised roles to emulate collaborative scientific workflows.

Qi et al. (2023) pioneered this direction by introducing a biomedical benchmark with temporally split background–hypothesis pairs and systematically evaluating LLMs under zero-shot, few-shot, and fine-tuning settings. Their cooperative framework assigned structured roles—*Analyst*, *Engineer*, *Scientist*, and *Critic*—with agents coordinating via tool calls and chain-of-thought prompting. Outputs were assessed using four complementary metrics (novelty, relevance, significance, verifiability), and zero-shot hypotheses were later corroborated by real publications, demonstrating genuine generative generalisation.

Similarly, Qi et al. (2024) extended this setup to unseen biomedical datasets, highlighting the benefits of collaborative tool use and uncertainty modelling while noting persistent issues like hallucinations and external knowledge integration.

In a domain-specific adaptation, Ghafarollahi and Buehler (2024) introduced `SciAgents`, a multi-agent framework for materials biology. Combining LLMs, ontologies, and data retrieval tools enables agents such as an "*Ontologist*" and a "*Novelty Assistant*" to build structured scientific concept graphs and uncover non-obvious material properties.

Su et al. (2024) proposed *Virtual Scientists* (`VirSci`), which instantiates GPT-4-based agents with curated expertise profiles and simulates team dynamics through asynchronous collaboration cycles—spanning topic discussion, hypothesis drafting, novelty assessment, and abstract writing. This dual-layered interaction (internal critique + external consultation) significantly improved novelty and projected impact over single-agent baselines.

In biomedical research, Ghareeb et al. (2025) introduced Robin, a lab-in-the-loop system that iteratively proposes and tests hypotheses, bridging LLM reasoning with empirical validation.

Lastly, the AI Co-Scientist system by Gottweis et al. (2025), built on Gemini 2.0, features a "generate–debate–evolve" loop with dedicated agents for generation, reflection, and ranking, coordinated by a central supervisor. Its outputs, evaluated via the GPQA benchmark, correlated strongly with expert-rated quality and novelty, surpassing single-agent LLMs.

#### 3.3.2 Domain-Specific Scientific Agents

Some LLM-based systems are tailored to specific scientific domains or datasets, enabling focused hypothesis generation grounded in domain expertise.

In astrophysics, Ciucă et al. (2023) explored adversarial prompting by immersing GPT-4 in a corpus of 1,000 papers—agents engaged in critical exchanges, producing more robust and higher-quality hypotheses than non-adversarial setups.

In the domain of scientific QA and literature analysis, Skarlinski et al. (2024) introduced `PaperQA2`, an agentic LLM framework combining retrieval-augmented generation with modules for literature search, contradiction detection, and citation tracing. It achieved superhuman performance in tasks like scientific question answering and contradiction identification, though at a higher computational cost.

#### 3.3.3 Mixed-Initiative and Human-Centred Systems

These systems blend human input with LLM-driven generation to support interactive, creative scientific ideation. `Scideator` (Radensky et al., 2024) exemplifies this approach by combining LLM-based retrieval and in-context prompting to recombine paper "facets" (e.g., purpose, mechanism, evaluation) into novel hypotheses. A dedicated novelty checker flags overlaps and suggests refinements, forming a closed idea generation and validation loop. Wu et al. (2025) introduced `CollabLLM`, a framework that encourages proactive, multi-turn planning in LLMs. The system promotes more interactive and goal-directed behaviour by simulating future dialogue paths and optimising for a multiturn-aware reward signal. Empirical results across tasks such as document editing and code assistance show significant gains in accuracy, interactivity, and user satisfaction compared to standard LLMs.

#### 3.3.4 Autonomous Research Agents

These systems aim to automate the scientific discovery process—from hypothesis generation to experimentation and reporting, with minimal human intervention.

In the social sciences, `MOOSE` (Yang et al., 2024) chains LLM-powered modules into a sequential pipeline for open-domain discovery, incorporating past, present, and future feedback loops. It outperforms baselines in novelty and helpfulness, although its generalizability beyond the 50-paper dataset remains untested.

Pushing toward full automation, Lu et al. (2024) introduced `The AI Scientist`, a closed-loop system that handles ideation, code generation, experiment execution (via the Aider coding assistant),

result visualisation, LaTeX paper writing, and simulated peer review. Applied to three machine learning subfields, it generated low-cost, high-quality manuscripts, some of which surpassed acceptance thresholds at top venues. While promising, the system still faces challenges in robustness, hallucination control, and long-term autonomy.

In computational chemistry, `MOOSE-Chem` (Yang et al., 2025b) and `MOOSE-Chem2` (Yang et al., 2025a) organise hypothesis generation into inspiration, composition, and ranking phases. The latter reframes the task as a combinatorial optimisation problem using hierarchical search and finds that homogeneous agent ensembles outperform heterogeneous ones in output quality.

Finally, the Chain-of-Ideas (CoI) framework by Li et al. (2024a) builds developmental chains from prior work to generate and refine future-facing hypotheses. It incorporates trend analysis and introduces the Idea Arena for evaluation, aligning well with human judgment. While CoI excels in novelty and significance, its experimental designs remain less feasible than those produced by human researchers.

**Summary** Together, these works trace the evolution from single-agent prompting to sophisticated, domain-specific multi-agent ecosystems that emulate key steps of the scientific process. They underscore the potential of mixed-initiative designs to enhance creativity, collaboration, and engagement—positioning LLMs as intelligent partners in hypothesis generation, while also highlighting persistent challenges such as cost, hallucination control, and domain generalisation in real-world scientific workflows.

### 3.4 Autonomous Scientific Discovery Systems

Recent efforts aim to automate the scientific discovery pipeline—spanning hypothesis generation, program synthesis, validation, and full-cycle autonomous agents—by integrating prompting, reasoning, execution, and evaluation.

Wang et al. (2024b) introduced a pipeline that enhances LLMs' inductive reasoning by translating natural language hypotheses into executable Python code, tested via feedback loops. Applied to datasets like ARC, SyGuS, and List Functions, it outperforms prompting-only baselines, though it remains computationally intensive and struggles with visual tasks and precise code generation. From verification to simulation, Ma et al. (2024) proposed the Scientific Generative Agent (SGA), which combines LLM-guided hypothesis generation with differentiable physical simulations in a bilevel optimisation loop. It achieves strong material and molecular design results but raises concerns about interpretability, differentiability, dependence, and computational demands. Toward full autonomy, Li et al. (2024c) presented `MLR-Copilot`, a multi-agent system with dedicated components for ideation, experimentation, and feedback. It improves throughput in ML research using GPT-4 and Claude, though its domain generalizability and need for oversight remain open issues. Similarly, Li et al. (2024b) developed `BoxLM`, where LMs iteratively propose and refine probabilistic programs using Box's Loop. While it achieves expert-level model discovery, it's limited to static datasets and lacks full critique automation. To benchmark such agents, Jansen et al. (2024) created `DISCOVERYWORLD`, a text-based environment with 120 tasks across scientific domains. While it simulates the discovery process, LLM-based agents like ReAct and Plan+Exec struggle compared to humans, and the environment's abstraction and computational cost present challenges. These works highlight progress in integrating LLMs with code and simulation environments for autonomous discovery. Yet, scalability, domain transfer, cost, and human oversight persist, positioning current systems as powerful augmentations rather than complete replacements for human researchers.

## 4 Hypothesis Validation Strategies

Evaluating systems for scientific hypothesis generation is a complex task. Unlike traditional NLP evaluation, hypothesis generation aims to produce novel, plausible, and testable scientific ideas—often in domains where ground truth is incomplete or non-existent. This open-endedness renders standard evaluation metrics insufficient and necessitates a multi-faceted approach combining human expertise, automated metrics, multi-modal integration, and domain-specific validation. In this section, we first review established methodologies before outlining promising directions for future research.

### 4.1 Human Expert Evaluation

Evaluations conducted by domain experts remain the most reliable method for assessing the relevance, originality, and scientific merit of machine-

generated hypotheses. Over time, these assessments have become more structured and methodologically rigorous. Recent protocols have involved large panels of experts from diverse academic backgrounds to evaluate hypotheses along dimensions such as clarity, innovation potential, and expected impact. Comparative studies have shown that, when supported by LLMs, researchers can generate more compelling and diverse ideas than with traditional search-based workflows. Such findings suggest that expert-in-the-loop systems not only support hypothesis refinement but can also enhance ideation itself.

In highly specialised fields such as biomedicine, structured evaluations have been designed to focus on clinical relevance and biological plausibility. Frameworks developed for this purpose often involve expert reviews centred on real-world applicability and potential translational impact. Some benchmark efforts have incorporated expert assessments across multiple research tasks, offering a broader view of how LLMs contribute to domain-specific scientific workflows.

**Blind Review and Pairwise Comparison** To reduce bias and ensure fair evaluation, blind review protocols are increasingly employed. Experts are unaware whether a human or an AI system has generated a hypothesis in these settings. This approach has revealed that, in many cases, AI-generated hypotheses can be as highly rated—or even surpass—those written by human researchers regarding novelty and scientific interest. Building on this principle, some recent evaluation strategies employ direct pairwise comparisons in tournament-style formats, where hypotheses compete against each other and are ranked based on expert preference. These structured comparison schemes offer a scalable and interpretable method for evaluating generative systems.

**Multi-Rater Reliability** One of the persistent challenges in expert-based evaluation is achieving consistency across annotators. Scientific hypothesis assessment often involves subjective judgment, leading to variability in ratings. Earlier studies have highlighted relatively low agreement levels among reviewers, emphasising the complexity of the task. However, newer frameworks are addressing this by introducing more formalised scoring rubrics, multiple rounds of review, and collaborative assessment protocols. These improvements have contributed to more stable and reproducible evaluation outcomes, reflecting a growing understanding of effectively integrating human judgment into validating AI-generated scientific content.

### 4.2 Automated Evaluation

**Text-based Relevance** Initial efforts to evaluate LLM outputs relied heavily on surface-level metrics such as BLEU and ROUGE, which measure word overlap between generated and reference hypotheses. However, such metrics often fall short of capturing an idea's semantic depth and scientific value. As a result, more sophisticated evaluation tools have been developed that incorporate semantic precision and recall and hybrid scores that combine symbolic and neural representations. These allow for a more meaningful assessment of whether a hypothesis is contextually appropriate and scientifically relevant. Some benchmarks now include domain-specific metrics tailored to the complexity and requirements of particular research tasks, such as code execution or model reproducibility.

**Model-Based Metrics** Recent evaluation frameworks have increasingly turned to large language models as evaluators of generated hypotheses. When fine-tuned or provided with structured prompts, these models can approximate human-level assessments across dimensions such as plausibility, novelty, and relevance. Some systems now rely on LLMs to score hypotheses using composite metrics that account for internal coherence and broader scientific context. For instance, measures have been developed to quantify how dissimilar a proposed idea is from past knowledge and how closely it aligns with emerging literature trends, thus reflecting historical uniqueness and prospective impact.

**Novelty Assessment** Measuring novelty remains one of the central goals in hypothesis evaluation. Automated approaches have evolved to estimate the originality of ideas by analysing their semantic distance from existing publications. This often involves embedding-based comparisons using pre-trained scientific language models combined with ranking strategies that assess the rarity or innovation of proposed connections. Some systems build structured citation graphs or ideation chains to contextualise a hypothesis within a broader intellectual lineage, enabling more informed judgments about its uniqueness.

**Domain-Specific Evaluation** Evaluation strategies tailored to specific scientific fields are increasingly recognised as essential due to the varied standards of evidence, feasibility, and validation across disciplines. In biomedical research, hypothesis evaluation often relies on alignment with curated clinical databases or known gene-disease associations, enabling automated cross-referencing against structured biomedical knowledge. In the chemical sciences, evaluation protocols typically focus on structural validity and chemical plausibility, incorporating molecular simulation or synthesis pathway prediction techniques. Astronomy and astrophysics present unique challenges, where hypothesis evaluation may involve the integration of large-scale observational datasets or comparing generated hypotheses with complex knowledge graphs. On the other hand, social science domains prioritise theoretical grounding and temporal context, often requiring evaluation of whether a hypothesis is consistent with existing paradigms or predictive of future trends. These domain-specific practices underscore the importance of aligning evaluation methodologies with disciplinary norms, highlighting the need for adaptable frameworks to accommodate modern science's epistemological diversity.

### 4.3 Domain-Specific Benchmarks

We present a curated set of benchmarks organised by scientific domain to facilitate comparison, focusing on their tasks, evaluation metrics, and design principles.

#### 4.3.1 Computational Chemistry

Recent efforts have produced benchmarks to evaluate LLMs' reasoning and hypothesis generation in chemistry. `BioFuelQR` (Sprueill et al., 2023) includes complex reasoning questions on catalysis, with a set of 20 queries targeting $CO_2$ conversion. It was later extended with the `CO2-Fuel` subset (Sprueill et al., 2024), emphasizing structured catalyst discovery. `TOMATO-Chem` (Yang et al., 2025b) features 51 high-impact post-2024 chemistry papers annotated by PhD-level students, spanning subfields such as Polymer, Organic, Inorganic, and Analytical Chemistry. A forthcoming extension promises more fine-grained hypothesis annotations and safeguards against training data contamination(Yang et al., 2025a).

#### 4.3.2 Biomedicine and Computational Biology

Qi et al. (2023, 2024) introduced a benchmark designed to test LLMs' zero-shot generalisation in biomedical hypothesis generation. It uses temporally split literature (pre- vs. post-2023) to prevent data leakage. It evaluates outputs with standard metrics (BLEU, ROUGE) and four custom criteria—Novelty, Relevance, Significance, and Verifiability—assessed by humans and GPT-4. Results demonstrate strong alignment between human and model judgments, highlighting areas for improvement, such as automated extraction bias, limited tool integration, and the need for richer evaluation protocols.

#### 4.3.3 Social Sciences

The `TOMATO` benchmark (Yang et al., 2024) evaluates LLMs on generating novel and valid hypotheses from open-domain web corpora in the social sciences. It includes 50 annotated papers and emphasises ideas new to humanity rather than commonsense reasoning. Evaluation focuses on validity, novelty, and helpfulness, using a mix of expert and GPT-4 assessments. While it offers a realistic, high-quality dataset and rigorous evaluation, its limited size and disciplinary scope may constrain generalizability to broader or interdisciplinary scientific contexts.

## 5 Challenges and Future Research Directions

This section identifies the challenges of leveraging LLMs for scientific hypothesis generation and validation. Building on the recommendations and limitations highlighted across the literature, we propose a set of structured future research directions for the community.

### 5.1 Challenges

**Creativity** There is an ongoing debate about whether LLMs exhibit genuine creativity or recombine existing knowledge. While its outputs may resemble human free association, the underlying mechanisms remain fundamentally derivative.

**Hallucinations and Factual Accuracy** LLMs frequently produce plausible but incorrect information hallucinations that pose serious risks in scientific contexts demanding precision. Such errors are especially problematic when mistaken for novel insights and can be amplified in multi-agent systems or combined with unreliable web content.

**Limited Novelty and Idea Diversity** LLMs often favour statistically likely outputs, resulting in repetitive and conservative suggestions. This bias can limit epistemic risk-taking and idea diversity without specialised prompting or intervention, potentially stifling innovation if broadly integrated into scientific workflows.

### 5.2 Future Research Directions

**Enhancing Novelty and Diversity** Applying data augmentation and debiasing techniques to diversify input datasets can help models generate hypotheses beyond traditional paradigms. Increasing uncertainty levels, for example, through collaborative multi-agent approaches, can diversify candidate generation and potentially enhance zero-shot hypothesis generation capabilities. Incorporating Dynamic Knowledge Graphs (DKGs) can allow systems to adapt to evolving datasets and uncover time-sensitive patterns, capturing trends and insights that static systems often miss.

**Feasibility and Practicality** Integrating multimodal data, including experimental results and sensor outputs, can improve feasibility assessments by grounding hypotheses in empirical evidence and domain-specific constraints. Developing foundation models with physical interaction capabilities, such as integrating robotic platforms in automated laboratories, can bridge the gap between theoretical predictions and experimental validation, enabling real-time feedback and refinement. Interdisciplinary Collaboration: Fostering partnerships between computational and experimental researchers can ground hypotheses in practical constraints and leverage diverse expertise.

**Human-AI Collaboration** Existing benchmarks focus on isolated model outputs rather than joint outcomes from human-AI teams. This approach fails to fully capture the limitations and capabilities of AI as a scientific collaborator, particularly in terms of the roles, expectations, and workflows envisioned by human researchers (Shao et al., 2025). Unlike chess, where AI systems have taught grandmasters novel strategies and reshaped expert practice (Schut et al., 2023), scientific discovery lacks equivalent frameworks for studying and evaluating human-AI collaboration. To move forward, evaluation protocols must include humans in the loop as active participants, not merely as annotators or evaluators. Automated metrics alone cannot account for scientific research's complex, iterative, and often exploratory nature. Human-AI systems require tailored evaluation strategies that reflect this interdependence. In particular, we must address three key challenges: (1) how AI systems complement or augment human expertise at various stages of the scientific process, (2) the system's responsiveness and adaptability to domain-specific guidance and constraints, and (3) the system's robustness in adversarial or ambiguous settings, such as resistance to user deception or misalignment. Despite increasing interest in collaborative intelligence, current evaluations of human-AI scientific workflows remain limited. Recent work has begun mapping what AI systems can and cannot do, and what researchers want them to do (Shao et al., 2025). However, there is still little methodological guidance for evaluating such systems in end-to-end scientific settings. In contrast to well-defined games like chess—where human-AI collaboration can be measured through novel move generation or improved win rates—we lack analogous metrics or interactive setups in science. Developing robust human-AI evaluation protocols would enable systems design that empowers researchers to explore novel directions rather than merely automate existing workflows. These protocols should be co-designed with domain experts and tested longitudinally. Empirical studies tracking real-world research outcomes from human-AI collaborations could yield actionable insights into system design, training strategies, and deployment best practices. Ultimately, embracing this interactive perspective will shift the focus from isolated performance to collaborative potential.

## 6 Conclusion

We reviewed LLM-based hypothesis generation to identify four major dimensions: direct prompting, multi-agent systems, external knowledge integration, and autonomous discovery systems. Our analysis reveals significant progress, with LLMs capable of generating hypotheses that experts judge as novel and plausible, with some achieving experimental validation. Multi-agent frameworks show particular promise in modelling the collaborative nature of the scientific process, while knowledge-augmented approaches help ground outputs in factual information. However, key challenges persist, including generating creative ideas, persistent hallucination, limited diversity in generated ideas, and difficulty evaluating scientific novelty and impact.

## 7 Limitations

While our survey offers a comprehensive overview of LLM-based hypothesis generation, it has several limitations. First, the fast-evolving nature of the field means our taxonomy and evaluation may quickly become outdated, despite efforts to curate recent and relevant works. Second, our study primarily focuses on English-language and high-resource domains (e.g., biomedicine, chemistry, and machine learning), which limits the generalizability of our insights to underrepresented disciplines or low-resource settings. Finally, this work adopts a technology-centric perspective. It does not sufficiently address the socio-technical and ethical implications of deploying LLMs as scientific collaborators, such as research reproducibility or bias amplification.

## References


Abbi Abdel-Rehim, Hector Zenil, Oghenejokpeme Orhobor, Marie Fisher, Ross J. Collins, Elizabeth Bourne, Gareth W. Fearnley, Emma Tate, Holly X. Smith, Larisa N. Soldatova, and Ross D. King. 2024. Scientific Hypothesis Generation by a Large Language Model: Laboratory Validation in Breast Cancer Treatment. *arXiv e-prints*, arXiv:2405.12258.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2025. Researchagent: Iterative research idea generation over scientific literature with large language models. *Preprint*, arXiv:2404.07738.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Ioana Ciucă, Yuan-Sen Ting, Sandor Kruk, and Kartheik Iyer. 2023. Harnessing the power of adversarial prompting and large language models for robust hypothesis generation in astronomy. *Preprint*, arXiv:2306.11648.

Alireza Ghafarollahi and Markus J. Buehler. 2024. SciAgents: Automating scientific discovery through multi-agent intelligent graph reasoning. *arXiv e-prints*, arXiv:2409.05556.

Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J. Szostkiewicz, Jon M. Laurent, Muhammed T. Razzak, Andrew D. White, Michaela M. Hinks, and Samuel G. Rodriques. 2025. Robin: A multi-agent system for automating scientific discovery. *Preprint*, arXiv:2505.13400.

Karan Girotra, Lennart Meincke, Christian Terwiesch, and Karl Ulrich. 2023. Ideas are dimes a dozen: Large language models for idea generation in innovation. *SSRN Electronic Journal*.

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, and 15 others. 2025. Towards an ai co-scientist. *Preprint*, arXiv:2502.18864.

Jennifer Haase and Paul H.P. Hanel. 2023. Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity. *Journal of Creativity*, 33(3):100066.

Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. 2024. Nova: An iterative planning and search approach to enhance novelty and diversity of llm generated ideas. *Preprint*, arXiv:2410.14255.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. 2024. Discoveryworld: A virtual environment for developing and evaluating automated scientific discovery agents. *Preprint*, arXiv:2406.06769.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.

Peter D. Karp. 1991. Artificial intelligence methods for theory representation and hypothesis formation. *Bioinformatics*, 7(3):301–308.

Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, Deli Zhao, Yu Rong, Tian Feng, and Lidong Bing. 2024a. Chain of ideas: Revolutionizing research via novel idea development with llm agents. *Preprint*, arXiv:2410.13185.

Michael Y. Li, Emily B. Fox, and Noah D. Goodman. 2024b. Automated statistical model discovery with language models. *Preprint*, arXiv:2402.17879.

Ruochen Li, Teerth Patel, Qingyun Wang, and Xinya Du. 2024c. Mlr-copilot: Autonomous machine learning research based on large language models agents. *Preprint*, arXiv:2408.14033.

C.E. Lipscomb. 2000. Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association*, 88(3):265.

Chun-Chi Liu, Yu-Ting Tseng, Wenyuan Li, Chia-Yu Wu, Ilya Mayzus, Andrey Rzhetsky, Fengzhu Sun, Michael Waterman, Jeremy J. W. Chen, Preet M. Chaudhary, Joseph Loscalzo, Edward Crandall, and Xianghong Jasmine Zhou. 2014. Diseaseconnect: a

comprehensive web server for mechanism-based disease–disease connections. *Nucleic Acids Research*, 42(W1):W137–W146.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. *arXiv e-prints*, arXiv:2408.06292.

Pingchuan Ma, Tsun-Hsuan Wang, Minghao Guo, Zhiqing Sun, Joshua B. Tenenbaum, Daniela Rus, Chuang Gan, and Wojciech Matusik. 2024. Llm and simulation as bilevel optimizers: A new paradigm to advance physical scientific discovery. *Preprint*, arXiv:2405.09783.

Ikujiro Nonaka. 2009. The knowledge-creating company. In *The economic impact of knowledge*, pages 175–187. Routledge.

Thomas O'Brien, Joel Stremmel, Léo Pio-Lopez, Patrick McMillen, Cody Rasmussen-Ivey, and Michael Levin. 2024. Machine learning for hypothesis generation in biology and medicine: exploring the latent space of neuroscience and developmental bioelectricity. *Digital Discovery*, 3:249–263.

Yang Jeong Park, Daniel Kaplan, Zhichu Ren, Chia-Wei Hsu, Changhao Li, Haowei Xu, Sipei Li, and Ju Li. 2023. Can chatgpt be used to generate scientific hypotheses? *Preprint*, arXiv:2304.12208.

Karl R. Popper. 1959. *The Logic of Scientific Discovery*. Routledge, London.

Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. 2023. Large Language Models are Zero Shot Hypothesis Proposers. *arXiv e-prints*, arXiv:2311.05965.

Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. 2023. Large language models are zero shot hypothesis proposers. *Preprint*, arXiv:2311.05965.

Biqing Qi, Kaiyan Zhang, Kai Tian, Haoxiang Li, Zhang-Ren Chen, Sihang Zeng, Ermo Hua, Hu Jinfang, and Bowen Zhou. 2024. Large language models as biomedical hypothesis generators: A comprehensive evaluation. *Preprint*, arXiv:2407.08940.

Linlu Qiu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, and Xiang Ren. 2024. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. *Preprint*, arXiv:2310.08559.

Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S. Weld. 2024. Scideator: Human-LLM Scientific Idea Generation Grounded in Research-Paper Facet Recombination. *arXiv e-prints*, arXiv:2409.14634.

Sanjana Ramprasad, Elisa Ferracane, and Zachary Lipton. 2024. Analyzing LLM behavior in dialogue summarization: Unveiling circumstantial hallucination trends. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12549–12561, Bangkok, Thailand. Association for Computational Linguistics.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. 2024. Mathematical discoveries from program search with large language models. *Nat.*, 625(7995):468–475.

Lisa Schut, Nenad Tomasev, Tom McGrath, Demis Hassabis, Ulrich Paquet, and Been Kim. 2023. Bridging the human-ai knowledge gap: Concept discovery and transfer in alphazero. *Preprint*, arXiv:2310.16410.

Yijia Shao, Humishka Zope, Yucheng Jiang, Jiaxin Pei, David Nguyen, Erik Brynjolfsson, and Diyi Yang. 2025. Future of Work with AI Agents: Auditing Automation and Augmentation Potential across the U.S. Workforce. *arXiv e-prints*, arXiv:2506.06576.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. *arXiv e-prints*, arXiv:2409.04109.

Michael D. Skarlinski, Sam Cox, Jon M. Laurent, James D. Braza, Michaela Hinks, Michael J. Hammerling, Manvitha Ponnapati, Samuel G. Rodriques, and Andrew D. White. 2024. Language agents achieve superhuman synthesis of scientific knowledge. *Preprint*, arXiv:2409.13740.

Neil R Smalheiser and Don R Swanson. 1998. Using arrowsmith: a computer-assisted approach to formulating and assessing scientific hypotheses. *Computer Methods and Programs in Biomedicine*, 57(3):149–153.

Scott Spangler, Angela D. Wilkins, Benjamin J. Bachman, Meena Nagarajan, Tajhal Dayaram, Peter Haas, Sam Regenbogen, Curtis R. Pickering, Austin Comer, Jeffrey N. Myers, Ioana Stanoi, Linda Kato, Ana Lelescu, Jacques J. Labrie, Neha Parikh, Andreas Martin Lisewski, Lawrence Donehower, Ying Chen, and Olivier Lichtarge. 2014. Automated hypothesis generation based on mining scientific literature. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 1877–1886, New York, NY, USA. Association for Computing Machinery.

Henry W. Sprueill, Carl Edwards, Khushbu Agarwal, Mariefel V. Olarte, Udishnu Sanyal, Conrad Johnston, Hongbin Liu, Heng Ji, and Sutanay Choudhury. 2024. Chemreasoner: Heuristic search over a large language model's knowledge space using quantum-chemical feedback. *Preprint*, arXiv:2402.10980.

Henry W. Sprueill, Carl Edwards, Khushbu Agarwal, Mariefel V. Olarte, Udishnu Sanyal, Conrad Johnston, Hongbin Liu, Heng Ji, and Sutanay Choudhury. 2024. ChemReasoner: Heuristic Search over a Large Language Model's Knowledge Space using Quantum-Chemical Feedback. *arXiv e-prints*, arXiv:2402.10980.

Henry W. Sprueill, Carl Edwards, Mariefel V. Olarte, Udishnu Sanyal, Heng Ji, and Sutanay Choudhury. 2023. Monte carlo thought search: Large language model querying for complex scientific reasoning in catalyst design. *Preprint*, arXiv:2310.14420.

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. 2024. Many Heads Are Better Than One: Improved Scientific Idea Generation by A LLM-Based Multi-Agent System. *arXiv e-prints*, arXiv:2410.09403.

Don R Swanson. 1986a. Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 30(1):7–18.

Don R Swanson. 1986b. Undiscovered public knowledge. *The Library Quarterly*, 56(2):103–118.

Justin Sybrandt, Michael Shtutman, and Ilya Safro. 2017. Moliere: Automatic biomedical hypothesis generation system. *Preprint*, arXiv:1702.06176.

Song Tong, Kai Mao, Zhen Huang, Yukun Zhao, and Kaiping Peng. 2024. Automating psychological hypothesis generation with ai: when large language models meet causal graph. *Humanities and Social Sciences Communications*, 11(1).

Jessica B. Voytek and Bradley Voytek. 2012. Automated cognome construction and semi-automated hypothesis generation. *Journal of Neuroscience Methods*, 208(1):92–100.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024a. Scimon: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 279–299. Association for Computational Linguistics.

Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D. Goodman. 2024b. Hypothesis search: Inductive reasoning with language models. *Preprint*, arXiv:2309.05660.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named entity recognition via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. *Preprint*, arXiv:2308.08155.

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. 2025. CollabLLM: From Passive Responders to Active Collaborators. *arXiv e-prints*, arXiv:2502.00640.

Guangzhi Xiong, Eric Xie, Amir Hassan Shariatmadari, Sikun Guo, Stefan Bekiranov, and Aidong Zhang. 2024. Improving scientific hypothesis generation with knowledge grounded large language models. *Preprint*, arXiv:2411.02382.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024. Large language models for automated open-domain scientific hypotheses discovery. *Preprint*, arXiv:2309.02726.

Zonglin Yang, Wanhao Liu, Ben Gao, Yujie Liu, Wei Li, Tong Xie, Lidong Bing, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. 2025a. Moose-chem2: Exploring llm limits in fine-grained scientific hypothesis discovery via hierarchical search. *Preprint*, arXiv:2505.19209.

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. 2025b. Moosechem: Large language models for rediscovering unseen chemistry scientific hypotheses. *Preprint*, arXiv:2410.07076.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. Hypothesis generation with large language models. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, page 117–139. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.