From Automation to Autonomy: A Survey on Large Language Models in Scientific Discovery

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, Yangqiu Song
Department of Computer Science and Engineering, HKUST, Hong Kong SAR, China
https://github.com/HKUST-KnowComp/Awesome-LLM-Scientific-Discovery

Abstract

Large Language Models (LLMs) are catalyzing a paradigm shift in scientific discovery, evolving from task-specific automation tools into increasingly autonomous agents and fundamentally redefining research processes and human-AI collaboration. This survey systematically charts this burgeoning field, placing a central focus on the changing roles and escalating capabilities of LLMs in science. Through the lens of the scientific method, we introduce a foundational three-level taxonomy—*Tool, Analyst, and Scientist*—to delineate their escalating autonomy and evolving responsibilities within the research lifecycle. We further identify pivotal challenges and future research trajectories such as robotic automation, self-improvement, and ethical governance. Overall, this survey provides a conceptual architecture and strategic foresight to navigate and shape the future of AI-driven scientific discovery, fostering both rapid innovation and responsible advancement.

1 Introduction

The relentless advancement of Large Language Models (LLMs) has unlocked a suite of emergent abilities, such as planning (Huang et al., 2024b), complex reasoning (Huang and Chang, 2023), and instruction following (Qin et al., 2024). Moreover, integrating agentic workflows enables LLM-based systems to perform advanced functions, including web navigation (He et al., 2024), tool use (Qu et al., 2025), code execution (Jiang et al., 2024a), and data analytics (Sun et al., 2024). In scientific discovery, this convergence of advanced LLM capabilities and agentic functionalities is catalyzing a significant paradigm shift. This shift is poised not only to accelerate the research lifecycle but also to fundamentally alter the collaborative dynamics between human researchers and artificial intelligence in the pursuit of knowledge.

However, this rapid expansion of LLM applications and the ongoing paradigm shift in scientific discovery present notable challenges. The accelerated pace of LLM evolution and their deepening integration into complex research complicate systematic assessment, necessitating conceptual frameworks to structure current understanding and chart future directions. While existing surveys have provided valuable overviews of LLMs in various scientific domains (Zhang et al., 2024, 2025a) or have cataloged particular AI techniques for science (Luo et al., 2025; Reddy and Shojaee, 2025), they often focus on discipline-specific applications or a static snapshot of LLM capabilities. Consequently, existing reviews may overlook the crucial trend of increasing LLM autonomy and their evolving roles across the entire scientific method, leaving their comprehensive impact and trajectory towards greater independence underexplored.

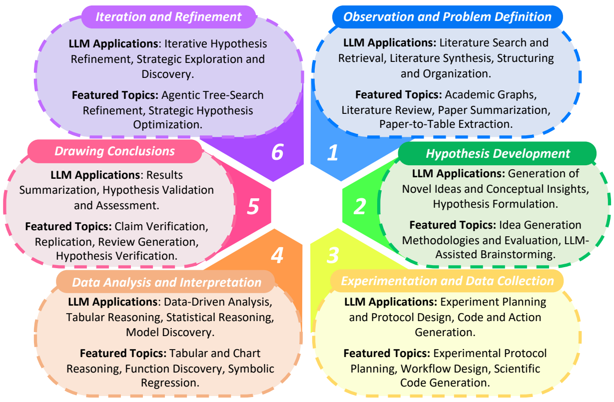

Figure 1: Stages of the scientific method with corresponding LLM applications and research topics.

To systematically chart this evolving landscape and address the identified gap, we anchor our analysis in the six stages (Figure 1) of the established scientific method (Popper, 1935; Kuhn, 1962): (1) observation and problem definition, (2) hypothesis development, (3) experimentation and data collection, (4) data analysis and interpretation, (5) drawing conclusions, and (6) iteration and refinement. Our examination of LLM applications across these

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 17733–17750
November 4-9, 2025 ©2025 Association for Computational Linguistics

Autonomy Levels	LLMs' Role	Human's Role	Task Scope	Agentic Workflow
Level 1 ***LLM as Tool***	Task Automation Tool	Task Allocation	Explicitly Defined	Simple & Static
Level 2 ***LLM as Analyst***	Data Modeling & Analytical Agent	Problem Definition & Output Validation	Goal-Oriented	Advanced
Level 3 ***LLM as Scientist***	Open Exploratory & Discovery Agent	Minimal Intervention	Open-Ended	Strategic & Iterative

Table 1: Three levels of autonomy in LLM-based scientific discovery.

stages reveals a significant trend: LLMs are progressing from performing discrete, task-oriented functions within a single stage to deployment in sophisticated, multi-stage agentic workflows. Notably, emerging research (Schmidgall et al., 2025; Yamada et al., 2025) now explores developing LLM-based systems capable of autonomously navigating nearly all these stages. To effectively capture and delineate this trajectory of increasing capability and independence, we introduce a foundational three-level taxonomy for LLM involvement in scientific discovery (Table 1): (i) ***LLM as Tool***, where models augment human researchers by performing specific, well-defined tasks under direct supervision; (ii) ***LLM as Analyst***, where models exhibit greater autonomy in processing complex information, conducting analyses, and offering insights with reduced human intervention; and (iii) ***LLM as Scientist***, representing a more advanced stage where LLM-based systems can autonomously conduct major research stages, from formulating hypotheses to interpreting results and suggesting new avenues of inquiry.

Building upon this taxonomic framework, we further identify critical gaps in the current research landscape and highlight pivotal challenges and future trajectories for the field, including: (1) fully autonomous discovery cycles for evolving scientific inquiry without human intervention; (2) robotic automation for interaction in the physical world for laboratory experimentation; (3) continuous self-improvement and adaptation from past research experiences; (4) transparency and interpretability of LLM-conducted research; and (5) ethical governance and societal alignment. Addressing these multifaceted challenges will be crucial for achieving a future where AI acts as a transformative partner in scientific exploration.

This survey focuses on LLM-based systems in scientific discovery, particularly their varying levels of autonomy. While acknowledging the broad impact of LLMs in science, we deliberately narrow our scope to exclude research on general-purpose scientific LLMs or LLMs for domain-specific scientific knowledge acquisition and reasoning, which are well covered in existing surveys (Zhang et al., 2024, 2025a). The remainder of this paper is organized as follows: Section 2 details our taxonomy and its interaction with the scientific method. Section 3 presents *LLM as Tool* applications, categorized by scientific method stages. Section 4 examines *LLM as Analyst* works by scientific domain, while Section 5 analyzes *LLM as Scientist* systems, focusing on their idea development and refinement strategies. Section 6 explores challenges and future directions.

2 Three Levels of Autonomy

Table 1 illustrates the three levels of autonomy in LLM-based scientific discovery with their associated features. In this section, we discuss their applications and characteristics in more detail.

LLM as Tool (Level 1). Level 1 represents the most foundational application of LLMs in scientific discovery. At this stage, LLMs function primarily as *tailored tools* under direct human supervision, designed to execute specific, well-defined tasks within a single stage of the scientific method. Their role is to augment human capabilities by automating or accelerating discrete activities such as literature summarization, drafting initial text for manuscripts, generating code snippets for data processing, or reformatting datasets. The autonomy of LLMs at this level is limited; they operate based on explicit human prompts and instructions, with outputs typically requiring human validation and integration into the broader research workflow. The primary goal is to enhance researcher efficiency and reduce routine task burdens.

LLM as Analyst (Level 2). In Level 2, LLMs exhibit a greater degree of autonomy and move beyond purely static, task-oriented applications. Here, LLMs function as *passive agents*, capable of more

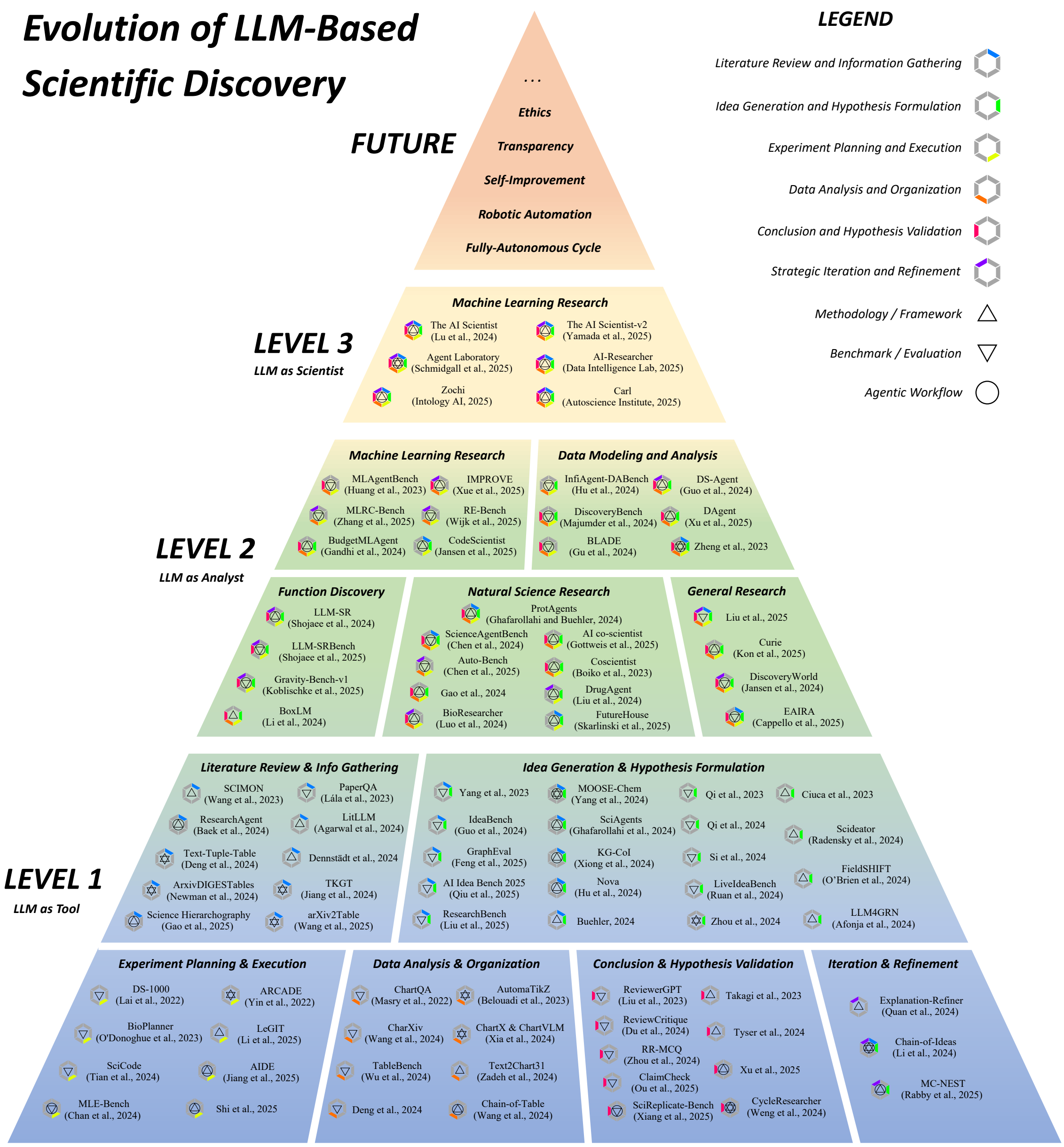

Figure 2: Taxonomy of research works in LLM-based scientific discovery with detailed categorization.

complex information processing, data modeling, and analytical reasoning with reduced human intervention for intermediate steps. While still operating within boundaries set by human researchers, these systems can independently manage sequences of tasks, such as analyzing experimental datasets to identify trends, interpreting outputs from complex simulations, or even performing iterative refinement of models. The human researcher typically defines the overall analytical goals, provides the necessary data, and critically evaluates the insights or interpretations generated by the LLM.

LLM as Scientist (Level 3). Level 3 applications signify a significant leap in autonomy, where LLM-based systems operate as *active agents* capable of orchestrating and navigating multiple stages of the scientific discovery process with considerable independence. These systems can demonstrate initiative in formulating hypotheses, planning and executing experiments, analyzing the resultant data, drawing preliminary conclusions, and potentially proposing subsequent research questions or avenues for exploration. LLM-based systems at this level can drive substantial portions of the research cycle, conducting scientific discovery with minimal human intervention.

Collectively, we present our full taxonomy with detailed categorization in Figure 2, which consolidates research works within our focused scope across all three levels of autonomy.

3 Level 1. LLM as Tool (Table A1)

In this section, we introduce *Level 1* research works in LLM-based scientific discovery, categorized by the stages in the scientific method they address.

3.1 Literature Review and Information Gathering

Literature Review Automatic literature search and retrieval is crucial for identifying research gaps and formulating research questions. Lála et al. (2023) first introduced the literature retrieval benchmark LitQA, featuring a RAG-based agent, PaperQA. LitLLM (Agarwal et al., 2024) further provided a comprehensive RAG-based toolkit for LLM-driven literature review. Taking this automation a step further, Wang et al. (2024c) demonstrated that large language models can automatically write entire survey papers. Dennstädt et al. (2024) directed their focus to the biomedical domain, highlighting the potential of LLMs in literature screening. Other approaches, such as SCIMON (Wang et al., 2024a) and ResearchAgent (Baek et al., 2025), have integrated active literature retrieval with the generation of research ideas. More recently, Gao et al. (2025) tackled the task of hierarchically organizing scientific literature through fine-grained paper abstraction. Nevertheless, several 'Deep Research' products (OpenAI, 2025; Google, 2025; xAI, 2025) have recently been released, featuring enhanced agentic workflows that support automated literature web search, organization, and report generation, thereby significantly accelerating traditional, human-intensive literature research processes.

Information Aggregation In parallel, several works have explored effective methods for aggregating information from scientific papers into tabular summaries. ArxivDIGESTables (Newman et al., 2024) investigated cross-literature table generation using LLMs, accompanied by an automated evaluation strategy. ArXiv2Table (Wang et al., 2025b) revised the evaluation protocol and provided a comprehensive benchmark. Methodologies such as Text-Tuple-Table (Deng et al., 2024b) and TKGT (Jiang et al., 2024b) have enhanced the quality of LLM-based table generation by incorporating tuple-based structures and graph modalities.

3.2 Idea Generation and Hypothesis Formulation

Idea Generation Numerous research efforts have focused on the automated generation of novel research ideas and hypotheses. In the general domain, benchmarks such as IdeaBench (Guo et al., 2024a) and LiveIdeaBench (Ruan et al., 2025) have evaluated the capability of LLMs to generate research ideas based on provided literature summaries. Concurrently, LLM-based agent frameworks, including Nova (Hu et al., 2024a), SciAgents (Ghafarollahi and Buehler, 2024b), and KG-CoI (Xiong et al., 2024), have been proposed to enhance idea generation through effective reasoning over academic knowledge graphs, iterative planning, and searching. More specific methodologies, such as employing dynamic control to guide the creative process, have also been introduced (Li et al., 2024c). Moreover, several exploratory studies have assessed the novelty and quality of LLM-generated ideas for AI research, underscoring the potential for automated idea generation when coupled with appropriate human guidance (Si et al., 2024; Feng et al., 2025; Qiu et al., 2025). Furthermore, many studies within natural science disciplines have investigated LLM-based idea generation in domain-specific contexts. For example, Ciucă et al. (2023) proposed adopting adversarial prompting for effective idea generation in astronomy. In biology, Buehler (2024) enhanced idea generation by integrating knowledge extraction and graph representations.

Hypothesis Formulation Building upon identified ideas, the design of testable scientific hypotheses has also been a significant focus. Qi et al. (2023) and Yang et al. (2024) examined the ability of LLMs to propose hypotheses, demonstrating their considerable capacity for generating novel yet valid hypotheses under open-ended constraints. Methodologies such as Scideator (Radensky et al., 2025) have been developed to investigate human-LLM collaboration to facilitate grounded idea and hypothesis generation. Other approaches have focused on ensuring the generated hypotheses are well-founded; for instance, HypER generates literature-grounded hypotheses with clear provenance (Vasu et al., 2025), while O'Neill et al. (2025) leverage structured data from scientific papers for the same purpose. Within natural science, benchmarks (Qi et al., 2024) and methods (O'Brien et al., 2024) have extended hypothesis generation into the biomedical domain. Meanwhile, MOOSE-Chem (Yang et al., 2025) offers a systematic evaluation benchmark and an agent framework specifically for hypothesis discovery in chemistry.

3.3 Experiment Planning and Execution

Experiment planning and execution constitute a crucial stage in LLM-based scientific discovery. While integral to advanced *Level 2* and *Level 3* agents, this subsection focuses on *Level 1* research, where LLMs serve as tools for experimental tasks.

Planning Regarding experiment planning, Li et al. (2025) discussed the effectiveness of incorporating LLMs into the design of causal discovery experiments. BioPlanner (O'Donoghue et al., 2023) introduced an automated evaluation framework for assessing LLMs in biological protocol planning. Furthermore, Shi et al. (2025) proposed a hierarchically encapsulated representation to complement LLMs in biological protocol design.

Execution For experiment execution, current research has primarily concentrated on code generation, particularly for artificial intelligence research, given the inherent compatibility of terminal interfaces with LLM experimental environments. Early code generation benchmarks, such as ARCADE (Yin et al., 2022) and DS-1000 (Lai et al., 2022), focused on data science tasks. Subsequent works, including MLE-Bench (Chan et al., 2025) and SciCode (Tian et al., 2024), incorporate more realistic scenarios, such as those encountered in machine learning engineering and natural science research, thereby presenting significant challenges for LLMs. To address these challenges, AIDE (Jiang et al., 2025) proposed enhancing complex code generation capabilities by adopting tree-search methodologies for code optimization.

3.4 Data Analysis and Organization

Tabular Data In this stage, LLMs assist the scientific workflow by automating processes related to data organization, presentation, and analysis. For data presented in tabular format, Chain-of-Table (Wang et al., 2024d) proposes a method to enhance tabular understanding by incorporating evolving tables within the reasoning chain of LLMs. Concurrently, Deng et al. (2024a) highlight the potential of integrating visual information to improve multimodal understanding, thereby aiding tabular comprehension. More recently, Wu et al. (2025) introduced TableBench, a comprehensive benchmark for table-based question answering under practical industrial scenarios.

Chart Data Beyond tabular data, charts represent another important format for organizing and storing information derived from experimental data. Early benchmarks, exemplified by ChartQA (Masry et al., 2022), examined the capabilities of vision transformers in chart-based question answering. Subsequent works, including CharXiv (Wang et al., 2024e) and ChartX (Xia et al., 2025), have expanded the scope of chart understanding scenarios by utilizing human-curated chart generation or by incorporating real-world chart data sourced from arXiv preprints. Regarding chart generation, AutomaTikZ (Belouadi et al., 2024) formulates the process as TikZ code generation from caption text and has demonstrated the efficacy of fine-tuning LLMs using open scientific figure data. More recently, Text2Chart31 (Zadeh et al., 2025) employed reinforcement learning with automated feedback to refine chart generation capabilities within the Matplotlib library.

3.5 Conclusion and Hypothesis Validation

In the concluding stages of research, LLMs can provide feedback on, or verify, claims and conclusions derived from experiments.

Paper Review In this context, a significant focus of contemporary research involves investigating the utility of LLMs as reviewers for artificial intelligence papers. ReviewerGPT (Liu and Shah, 2023) initially explored the capability of LLMs to identify deliberately inserted errors within research papers, highlighting the necessity for more robust systems to conduct comprehensive reviews. Zhou et al. (2024a) further evaluated static LLMs in the context of reviewing real-world conference papers using a multiple-choice format. Du et al. (2024) conducted a comprehensive analysis of LLM review quality through extensive human studies and comparisons, revealing weaknesses in their ability to identify deficiencies. ClaimCheck (Ou et al., 2025) further investigated the capabilities of LLMs in critiquing research claims, demonstrating that this task remains challenging even for highly advanced models such as OpenAI's o1 (OpenAI, 2024). Beyond reviewing, other work has focused on the subsequent step of paper revision, with systems like XtraGPT enabling human-AI collaboration for controllable academic paper revisions (Chen et al., 2025a). Concurrently, research highlights the potential to address these limitations by incorporating multi-agent systems with specialized roles (Tyser et al., 2024; Xu et al., 2025a), through LLM alignment via reinforcement learning (Weng et al., 2025), or by employing novel frameworks like generative adversarial reviews (Bougie and Watanabe, 2024).

Hypothesis Validation Another important application at this stage is the automatic validation of hypotheses by LLMs. Takagi et al. (2023) demonstrated that LLMs possess considerable capabilities in automatically generating code to verify research hypotheses within simplified machine learning problems. Benchmarks such as SciReplicate-Bench (Xiang et al., 2025) and PaperBench (Starace et al., 2025) have further extended this concept to evaluating the replication of real-world research papers. A distinct but related line of inquiry explores predicting empirical AI research outcomes directly with language models, assessing whether LLMs can anticipate experimental results without full execution (Wen et al., 2025). Furthermore, Xu et al. (2025c) have navigated this domain into physics research, aiming to enhance the interpretability of the discovery process through the use of multi-agent workflows.

3.6 Iteration and Refinement

The iterative refinement of research hypotheses, as a distinct area of investigation, has received comparatively less attention in current research. Explanation-Refiner (Quan et al., 2024) employed theorem provers to verify and subsequently refine LLM-generated hypotheses. Chain-of-Idea (Li et al., 2024a) introduced an LLM-based agent framework designed to organize literature and develop research ideas by building upon or extending existing lines of inquiry. More recently, MC-NEST (Rabby et al., 2025) adopted Monte-Carlo Tree Search to iteratively verify and refine scientific hypotheses across multiple research domains.

4 Level 2: LLM as Analyst (Table A2)

In this section, we introduce *Level 2* research works in LLM-based scientific discovery, categorized according to their task nature and domains.

4.1 Machine Learning Research

Automated Machine Learning (AutoML) (Shen et al., 2024) endeavors to generate high-performing modeling configurations for a given task in a data-driven manner. With the advent of LLM-based agents, several studies have explored their application in the automated modeling of machine learning (ML) tasks. A suite of benchmarks has emerged to track progress in this area. MLAgentBench (Huang et al., 2024a) evaluates the capabilities of LLMs in designing and executing ML experiments, revealing that performance is often contingent upon task familiarity. Similarly, MLRC-Bench (Zhang et al., 2025b) and RE-Bench (Wijk et al., 2024) further probe the limits of these agents, assessing their ability to solve novel ML research challenges and comparing their R&D capabilities against human experts. MLGym (Nathani et al., 2025) offers valuable resource and benchmark for advancing these AI research agents.

To address the challenges posed by these benchmarks, various agentic frameworks have been proposed. The IMPROVE framework (Xue et al., 2025) highlighted the significance of iterative refinement mechanisms. CodeScientist (Jansen et al., 2025) incorporated an ML modeling agent with machine-generated ideas, while BudgetMLAgent (Gandhi et al., 2025) leveraged curated expert collaboration frameworks to achieve superior results with cost-effective models. More recent end-to-end systems like MLR-Copilot (Li et al., 2024d) and the multi-agent framework MLZero (Fang et al., 2025) aim for fully autonomous machine learning research and automation. Pushing the boundaries of automation even further, some work has explored the use of language models to directly propose LM architectures (Cheng et al., 2025a), moving beyond orchestration to direct model creation.

4.2 Data Modeling and Analysis

Automated data-driven analysis, encompassing statistical data modeling and hypothesis validation, represents a foundational application area for LLM-assisted scientific discovery. InfiAgent-DABench (Hu et al., 2024b) benchmarked the capabilities of LLMs in static code generation and execution for data analysis using CSV files. Subsequent benchmarks, such as BLADE (Gu et al., 2024), DiscoveryBench (Majumder et al., 2024), and DSBench (Jing et al., 2024), have improved evaluation robustness by incorporating real-world research papers and expert-curated analytics to assess how far agents are from human expert performance. These studies indicate that most LLMs struggle with complex data analytics tasks, even when operating within an agent framework (Zheng et al., 2023). To address these challenges, DS-Agent (Guo et al., 2024b) proposes to enhance LLM performance by incorporating a case-based reasoning method to improve domain knowledge acquisition. In a related effort, DAgent (Xu et al., 2025b) extended the application domain to querying relational databases and enabled report generation using results derived from decomposed sub-problems.

4.3 Function Discovery

Function discovery, which aims to identify the underlying equations from observational data of variables, has been significantly influenced by the advancement of AI-driven symbolic regression (SR) (Udrescu and Tegmark, 2020; Kamienny et al., 2022). To enhance this process, LLM-SR (Shojaee et al., 2025a) leveraged the prior domain knowledge of LLMs and incorporated feedback from clustered memory storage, while DrSR (Wang et al., 2025a) proposed a dual reasoning framework that utilizes both data and experience for scientific equation discovery. To systematically assess these capabilities, LLM-SRBench (Shojaee et al., 2025b) introduced a benchmark for evaluating LLMs as function discovery agents, which incorporates function transformations to mitigate data contamination. Furthermore, other studies have explored the capabilities of LLMs in discovering complex models within specific domains, such as Physics (Koblischke et al., 2025), Statistics (Li et al., 2024b), and automated neural scaling law discovery (Lin et al., 2025).

4.4 Natural Science Research

Research has largely focused on applying LLMs to autonomous research workflows for natural science discovery. Auto-Bench (Chen et al., 2025b) evaluated LLMs on chemistry and social science tasks based on causal graph discovery, revealing that LLMs perform effectively only when task complexity is highly limited. In contrast, ScienceAgentBench (Chen et al., 2025c) provided a multidisciplinary benchmark for LLMs operating within agent frameworks such as CodeAct (Wang et al., 2024b) and self-debug (Chen et al., 2023). This benchmark highlighted the necessity for tailored agent workflows for such explorative tasks.

In the biomedical domain, Gao et al. (2024) discussed potential applications of AI agents in brainstorming, experimental planning, and execution. BioResearcher (Luo et al., 2024) proposed an end-to-end framework for biomedical research involving dry lab experiments. DrugAgent (Liu et al., 2025b) adopted multi-agent collaboration to automate drug discovery. In chemistry, Coscientist (Boiko et al., 2023) incorporated the use of tools by LLMs to support semi-autonomous chemistry experiment design and execution. ProtAgents (Ghafarollahi and Buehler, 2024a) facilitated biochemistry discovery by building a multi-agent framework for automating protein design. Recent works, such as FutureHouse (Skarlinski et al., 2025) and AI Co-scientist (Gottweis et al., 2025), contributed to formulating demonstrably novel research hypotheses and proposals using multi-agent systems guided by predefined research goals.

4.5 General Research

Apart from specialized domain applications, some benchmarks have broadly evaluated diverse tasks from different stages of scientific discovery. DiscoveryWorld (Jansen et al., 2024) created a virtual environment for LLM agents to conduct simplified scientific exploration. In (Liu et al., 2025a), various application scenarios for AI agents in research were comprehensively discussed, supported by preliminary evaluation datasets. Similarly, CURIE (Kon et al., 2025) proposed a benchmark and an agentic framework for rigorous and automated scientific experimentation. While EAIRA (Cappello et al., 2025) focused on assessing the ability of LLMs to perform in a real-world research assistant role using various task formats.

5 Level 3. LLM as Scientist (Table A3)

Recently, several research efforts and commercial products have demonstrated prototypes of fully autonomous research within the artificial intelligence domain. These systems typically encompass a comprehensive workflow, from initial literature review to iterative refinement cycles where hypotheses or designs are progressively improved. A common feature is using an agent-based framework to autonomously produce research outputs, often culminating in draft research papers. This section will further compare these approaches, focusing on their methodologies for idea development and iterative refinement, as these aspects critically distinguish them from *Level 2* agents.

5.1 Idea Development

The genesis of research in *Level 3* systems involves transforming initial concepts into validated hypotheses, with distinct approaches to sourcing and vetting these ideas. Agent Laboratory (Schmidgall et al., 2025) predominantly conducts literature reviews based on human-defined research objectives. Moving towards greater autonomy, several systems initiate their process from broader human inputs, such as reference papers (Data Intelligence Lab, 2025; Autoscience, 2025) or general research domains (IntologyAI, 2025), subsequently exploring literature to autonomously identify gaps and formulate novel hypotheses. The AI Scientist (v1 (Lu et al., 2024) and v2 (Yamada et al., 2025)) showcases an even more generative approach: v1 brainstorms ideas from templates and past work, while v2 can generate diverse research proposals from abstract thematic prompts. Crucially, these systems employ diverse methods to evaluate their ideas prior to full implementation. AI Scientist-v1 uses self-assessed scores for interestingness, novelty, and feasibility, supplemented by external checks with Semantic Scholar. AI Scientist-v2 integrates literature review tools early in its idea formulation stage to assess novelty. This spectrum reveals a clear trend: while humans often initiate ideas, advanced systems can autonomously explore, generate, and validate the scientific merit and originality of research objectives before development.

5.2 Iterative Refinement

Iterative refinement within *Level 3* systems involves sophisticated feedback loops that enable not just incremental improvements but also fundamental reassessments of the research trajectory. A key differentiator lies in the primary source and nature of this high-level feedback. The AI Scientist (v1 and v2) incorporates highly automated internal review and refinement processes. It employs AI reviewers, LLM evaluators for experimental choices, and VLMs to critique figures, fostering a rich internal feedback loop for iterative development. In contrast, Zochi (IntologyAI, 2025) integrates human expertise for macro-level guidance, where feedback can trigger complete re-evaluations of hypotheses or designs. This allows it to act on critiques challenging the core research premise, even reverting to hypothesis regeneration if results are unsatisfactory. Overall, while automated self-correction is a common goal, the current landscape reveals a pragmatic blend: some systems focus on enhancing autonomous deep reflection, while others integrate human oversight for robust, high-level iterative refinement and strategic redirection.

6 Challenges and Future Directions

Throughout this survey, we have systematically reviewed the escalating roles of Large Language Models in scientific discovery, delineating their progression through distinct levels of autonomy and capability—from foundational assistants and analysts to increasingly autonomous scientific researchers. In particular, we have underscored the evolving methodologies, task complexities, and the nature of human-LLM interaction that define each stage of this maturation. Beyond reviewing these advancements and current applications, this section presents several significant challenges and outlines promising directions for future research, aiming to inspire further exploration into the development and responsible deployment of LLMs as transformative tools in scientific inquiry.

Fully-Autonomous Research Cycle While current *Level 3* systems can navigate multiple stages of the scientific method for a specific inquiry, they often operate within a single research instance or predefined topic. The scientific method, however, is inherently cyclical, characterized by continuous iteration, refinement, and the pursuit of evolving research questions. A significant future direction, therefore, is to develop LLM-based systems capable of engaging in a truly autonomous research cycle. This would entail not merely executing a given research task from start to finish, but possessing the foresight to discern the broader implications of their findings, proactively identify promising avenues for subsequent investigation, and strategically direct their efforts towards practical advancements that build upon previous work.

Robotic Automation A key barrier to fully autonomous scientific discovery in natural science is LLM agents' inability to conduct physical laboratory experiments. While adept in computational research, their application in fields requiring physical interaction remains limited. Integrating LLMs with robotic systems empowers them to translate their planning capabilities into direct experimental actions. Early works in LLM-robotic integration (Yoshikawa et al., 2023; Song et al., 2024; Darvish et al., 2025) already highlights this potential in chemical experimentation. Such automation is poised to significantly broaden LLM-based research, enabling end-to-end discovery in disciplines like chemistry and materials science, thereby advancing autonomous scientific exploration.

Transparency and Interpretability The *black-box* nature (or opacity) of advancing LLMs in science undermines scientific validation, trust, and the assimilation of AI-driven insights (Xu et al., 2025c). Addressing this opacity demands more than superficial Explainable AI (XAI) techniques (Ahadian and Guan, 2024). It necessitates a paradigm shift towards systems whose internal operations are inherently designed for verifiable reasoning and justifiable conclusions (Bengio et al., 2025). Consequently, the challenge is not just explaining outputs, but ensuring the AI's internal logic aligns with scientific principles and can reliably differentiate asserted claims from verifiable truths. This profound interpretability is vital for reliable and reproducible LLM-based scientific discovery.

Continuous Self-Improvement The iterative and evolving nature of scientific inquiry demands systems capable of learning from ongoing engagement, assimilating experimental outcomes, and adapting research strategies. Current research integrating continual learning with agent-based systems already highlights the potential for LLMs to adapt to new tasks or environments without catastrophic forgetting (Majumder et al., 2023; Kim et al., 2024). Within scientific discovery, a promising future direction is to incorporate online reinforcement learning frameworks (Carta et al., 2023). This integration promises to continuously enhance scientific agents' capabilities over their operational lifetime through successive discoveries, thereby advancing sustainable autonomous exploration.

Ethics and Societal Alignment As LLM-based systems gain independent reasoning and action capabilities, their potential for risks—ranging from amplified societal biases to deliberate misuse like generating harmful substances or challenging human control—becomes increasingly salient and complex (He et al., 2023; Ahadian and Guan, 2024; Bengio et al., 2025). With AI capabilities and societal norms in constant flux, alignment is consequently an imperative, continuous process demanding adaptive governance and evolving value systems (Li et al., 2024e). This requires embedding ethical constraints directly in scientific AI design frameworks, alongside vigilant oversight, to ensure advancements serve human well-being and the common good.

Limitations

This survey provides a systematic review of LLMs in scientific discovery, with a particular emphasis on the paradigm shift characterized by their escalating levels of autonomy. Our analysis and the selection of reviewed literature are therefore centered on works that illustrate this transition across the stages of the scientific method, categorized within our proposed three-level autonomy framework: LLM as Tool, LLM as Analyst, and LLM as Scientist.

Consequently, the scope of this survey has certain limitations. Firstly, we do not provide an exhaustive review of research focused on the development of general-purpose scientific LLMs for domain-specific reasoning or application. These areas, while crucial to the broader landscape of AI in science, are extensively covered in other existing surveys and fall outside our specific focus on the autonomy paradigm. Secondly, while we acknowledge the importance of fundamental LLM capabilities such as planning, code generation, and agentic decision-making, this survey does not delve deeply into orthogonal benchmarks or methodologies related to these general abilities. These exclusions were deliberate to maintain a focused exploration of the transformative roles and increasing independence of LLMs throughout the scientific research lifecycle.

Ethics Statement

Our paper presents a comprehensive survey of LLMs in scientific discovery, with a specific focus on their role transformation from task automation tools to autonomous agents. All research works reviewed in this survey are properly cited. To the best of our knowledge, the referenced materials are publicly accessible or available under licenses permitting their use for research review. We did not conduct additional dataset curation or human annotation work. Consequently, we believe that this paper does not raise any ethical concerns.

Acknowledgements

We thank all the anonymous reviewers and meta reviewers for their valuable comments. The authors of this paper were supported by the ITSP Platform Research Project (ITS/189/23FP) from ITC of Hong Kong, SAR, China, and the AoE (AoE/E-601/24-N), the RIF (R6021-20) and the GRF (16205322) from RGC of Hong Kong, SAR, China.

References

Tejumade Afonja, Ivaxi Sheth, Ruta Binkyte, Waqar Hanif, Thomas Ulas, Matthias Becker, and Mario Fritz. 2024. Llm4grn: Discovering causal gene regulatory networks with llms – evaluation through synthetic data generation. *Preprint*, arXiv:2410.15828.

Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H. Laradji, Krishnamurthy DJ Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. 2024. Litllm: A toolkit for scientific literature review. *Preprint*, arXiv:2402.01788.

Pegah Ahadian and Qiang Guan. 2024. Ai trustworthy challenges in drug discovery. In *Trustworthy Artificial Intelligence for Healthcare*, pages 1–12, Cham. Springer Nature Switzerland.

Autoscience. 2025. Meet carl: The first ai system to produce academically peer-reviewed research.

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.

Jonas Belouadi, Anne Lauscher, and Steffen Eger. 2024. Automatikz: Text-guided synthesis of scientific vector graphics with tikz. *Preprint*, arXiv:2310.00367.

Yoshua Bengio, Michael Cohen, Damiano Fornasiere, Joumana Ghosn, Pietro Greiner, Matt MacDermott, Sören Mindermann, Adam Oberman, Jesse Richardson, Oliver Richardson, Marc-Antoine Rondeau, Pierre-Luc St-Charles, and David Williams-King. 2025. Superintelligent agents pose catastrophic risks: Can scientist ai offer a safer path? *Preprint*, arXiv:2502.15657.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624:570–578.

Nicolas Bougie and Narimasa Watanabe. 2024. Generative adversarial reviews: When llms become the critic. *Preprint*, arXiv:2412.10415.

Markus J. Buehler. 2024. Accelerating scientific discovery with generative knowledge extraction, graph-based representation, and multimodal intelligent graph reasoning. *Preprint*, arXiv:2403.11996.

Franck Cappello, Sandeep Madireddy, Robert Underwood, Neil Getty, Nicholas Lee-Ping Chia, Nesar Ramachandra, Josh Nguyen, Murat Keceli, Tanwi Mallick, Zilinghan Li, Marieme Ngom, Chenhui Zhang, Angel Yanguas-Gil, Evan Antoniuk, Bhavya Kailkhura, Minyang Tian, Yufeng Du, Yuan-Sen Ting, Azton Wells, and 7 others. 2025. Eaira: Establishing a methodology for evaluating ai models as scientific research assistants. *Preprint*, arXiv:2502.20309.

Thomas Carta, Clément Romac, Thomas Wolf, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. 2023. Grounding large language models in interactive environments with online reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 3676–3713. PMLR.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng, and Aleksander Mądry. 2025. Mle-bench: Evaluating machine learning agents on machine learning engineering. *Preprint*, arXiv:2410.07095.

Aonan Chen, Zhaoning Yang, Ruochen Li, Jiachen Liu, Muntasir Chowdhury, Moontae Lee, Dong-Yu Du, and Tao Yu. 2025a. Xtragpt: Context-aware and controllable academic paper revision via human-ai collaboration. *Preprint*, arXiv:2505.11336.

Tingting Chen, Srinivas Anumasa, Beibei Lin, Vedant Shah, Anirudh Goyal, and Dianbo Liu. 2025b. Autobench: An automated benchmark for scientific discovery in llms. *Preprint*, arXiv:2502.15224.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023. Teaching large language models to self-debug. *Preprint*, arXiv:2304.05128.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. 2025c. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. *Preprint*, arXiv:2410.05080.

Junyan Cheng, Peter Clark, and Kyle Richardson. 2025a. Language modeling by language models. *Preprint*, arXiv:2506.20249.

Ming-Chang Cheng, Yixuan Wang, Yuchi Wang, Yair Schiff, Kevin Bello, J. Zico Kolter, Yacine Jernite, William Yang Wang, Daniel Fried, Tian-Li Yu, and Andrew Gordon Wilson. 2025b. Language modeling by language models. *Preprint*, arXiv:2506.20249.

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.

Kourosh Darvish, Marta Skreta, Yuchi Zhao, Naruki Yoshikawa, Sagnik Som, Miroslav Bogdanovic, Yang Cao, Han Hao, Haoping Xu, Alán Aspuru-Guzik, Animesh Garg, and Florian Shkurti. 2025. Organa: A robotic assistant for automated chemistry experimentation and characterization. *Preprint*, arXiv:2401.06949.

Data Intelligence Lab. 2025. Ai-researcher: Fully-automated scientific discovery with llm agents. `https://github.com/HKUDS/AI-Researcher`.

Naihao Deng, Zhenjie Sun, Ruiqi He, Aman Sikka, Yulong Chen, Lin Ma, Yue Zhang, and Rada Mihalcea. 2024a. Tables as texts or images: Evaluating the table reasoning ability of llms and mllms. *Preprint*, arXiv:2402.12424.

Zheye Deng, Chunkit Chan, Weiqi Wang, Yuxi Sun, Wei Fan, Tianshi Zheng, Yauwai Yim, and Yangqiu Song. 2024b. Text-tuple-table: Towards information integration in text-to-table generation via global tuple extraction. *Preprint*, arXiv:2404.14215.

Fabio Dennstädt, Johannes Zink, Paul Martin Putora, Janna Hastings, and Nikola Cihoric. 2024. Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain. *Systematic Reviews*, 13:158.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, and 21 others. 2024. Llms assist nlp researchers: Critique paper (meta-)reviewing. *Preprint*, arXiv:2406.16253.

Haoyang Fang, Boran Han, Nick Erickson, Xiyuan Zhang, Su Zhou, Anirudh Dagar, Jiani Zhang, Ali Caner Turkmen, Cuixiong Hu, Huzefa Rangwala, Ying Nian Wu, Bernie Wang, and George Karypis. 2025. Mlzero: A multi-agent system for end-to-end machine learning automation. *Preprint*, arXiv:2505.13941.

Tao Feng, Yihang Sun, and Jiaxuan You. 2025. Grapheval: A lightweight graph-based llm framework for idea evaluation. *Preprint*, arXiv:2503.12600.

Shubham Gandhi, Manasi Patwardhan, Lovekesh Vig, and Gautam Shroff. 2025. Budgetmlagent: A cost-effective llm multi-agent system for automating machine learning tasks. *Preprint*, arXiv:2411.07464.

Muhan Gao, Jash Shah, Weiqi Wang, and Daniel Khashabi. 2025. Science hierarchography: Hierarchical organization of science literature. *Preprint*, arXiv:2504.13834.

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. 2024. Empowering biomedical discovery with ai agents. *Cell*, 187(22):6125–6151.

A. Ghafarollahi and M. J. Buehler. 2024a. Protagents: Protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Preprint*, arXiv:2402.04268.

Alireza Ghafarollahi and Markus J. Buehler. 2024b. Sciagents: Automating scientific discovery through multi-agent intelligent graph reasoning. *Preprint*, arXiv:2409.05556.

Google. 2025. Gemini deep research - your personal research assistant.

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.

Ken Gu, Ruoxi Shang, Ruien Jiang, Keying Kuang, Richard-John Lin, Donghe Lyu, Yue Mao, Youran Pan, Teng Wu, Jiaqian Yu, Yikun Zhang, Tianmai M. Zhang, Lanyi Zhu, Mike A. Merrill, Jeffrey Heer, and Tim Althoff. 2024. Blade: Benchmarking language model agents for data-driven science. *Preprint*, arXiv:2408.09667.

Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Eric Xie, Stefan Bekiranov, and Aidong Zhang. 2024a. Ideabench: Benchmarking large language models for research idea generation. *Preprint*, arXiv:2411.02429.

Siyuan Guo, Cheng Deng, Ying Wen, Hechang Chen, Yi Chang, and Jun Wang. 2024b. Ds-agent: Automated data science by empowering large language models with case-based reasoning. *Preprint*, arXiv:2402.17453.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. 2024. Webvoyager: Building an end-to-end web agent with large multimodal models. *Preprint*, arXiv:2401.13919.

Jiyan He, Weitao Feng, Yaosen Min, Jingwei Yi, Kunsheng Tang, Shuai Li, Jie Zhang, Kejiang Chen, Wenbo Zhou, Xing Xie, Weiming Zhang, Nenghai Yu, and Shuxin Zheng. 2023. Control risk for potential misuse of artificial intelligence in science. *Preprint*, arXiv:2312.06632.

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

Xueyu Hu, Ziyu Zhao, Shuang Wei, Ziwei Chai, Qianli Ma, Guoyin Wang, Xuwu Wang, Jing Su, Jingjing Xu, Ming Zhu, Yao Cheng, Jianbo Yuan, Jiwei Li, Kun Kuang, Yang Yang, Hongxia Yang, and Fei Wu. 2024b. Infiagent-dabench: Evaluating agents on data analysis tasks. *Preprint*, arXiv:2401.05507.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2024a. Mlagentbench: Evaluating language agents on machine learning experimentation. *Preprint*, arXiv:2310.03302.

Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. 2024b. Understanding the planning of llm agents: A survey. *Preprint*, arXiv:2402.02716.

IntologyAI. 2025. Zochi technical report: The first artificial scientist.

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.

Peter Jansen, Oyvind Tafjord, Marissa Radensky, Pao Siangliulue, Tom Hope, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Daniel S. Weld, and Peter Clark. 2025. Codescientist: End-to-end semi-automated scientific discovery with code-based experimentation. *Preprint*, arXiv:2503.22708.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2024a. A survey on large language models for code generation. *Preprint*, arXiv:2406.00515.

Peiwen Jiang, Xinbo Lin, Zibo Zhao, Ruhui Ma, Yvonne Jie Chen, and Jinhua Cheng. 2024b. TKGT: Redefinition and a new way of text-to-table tasks based on real world demands and knowledge graphs augmented LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16112–16126, Miami, Florida, USA. Association for Computational Linguistics.

Zhengyao Jiang, Dominik Schmidt, Dhruv Srikanth, Dixing Xu, Ian Kaplan, Deniss Jacenko, and Yuxiang Wu. 2025. Aide: Ai-driven exploration in the space of code. *Preprint*, arXiv:2502.13138.

Liqiang Jing, Zhehui Huang, Xiaoyang Wang, Wenlin Yao, Wenhao Yu, Kaixin Ma, Hongming Zhang, Xinya Du, and Dong Yu. 2024. Dsbench: How far are data science agents from becoming data science experts? *Preprint*, arXiv:2409.07703.

Pierre-Alexandre Kamienny, Stéphane d'Ascoli, Guillaume Lample, and François Charton. 2022. End-to-end symbolic regression with transformers. *Preprint*, arXiv:2204.10532.

Byeonghwi Kim, Minhyuk Seo, and Jonghyun Choi. 2024. Online continual learning for interactive instruction following agents. *Preprint*, arXiv:2403.07548.

Nolan Koblischke, Hyunseok Jang, Kristen Menou, and Mohamad Ali-Dib. 2025. Gravity-bench-v1: A benchmark on gravitational physics discovery for agents. *Preprint*, arXiv:2501.18411.

Patrick Tser Jern Kon, Jiachen Liu, Qingyun Ding, Yuxin Qiu, Zhaoning Yang, Yufan Huang, Jer-Shannassa, Moontae Lee, Muntasir Chowdhury, and Aonan Chen. 2025. Curie: Toward rigorous and automated scientific experimentation with ai agents. *Preprint*, arXiv:2502.16069.

Thomas Samuel Kuhn. 1962. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago.

Yuhang Lai, Chengxi Li, Yiming Wang, Tianyi Zhang, Ruiqi Zhong, Luke Zettlemoyer, Scott Wen tau Yih, Daniel Fried, Sida Wang, and Tao Yu. 2022. Ds-1000: A natural and reliable benchmark for data science code generation. *Preprint*, arXiv:2211.11501.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. 2023. Paperqa: Retrieval-augmented generative agent for scientific research.

Junyi Li, Yongqiang Chen, Chenxi Liu, Qianyi Cai, Tongliang Liu, Bo Han, Kun Zhang, and Hui Xiong. 2025. Can large language models help experimental design for causal discovery? *Preprint*, arXiv:2503.01139.

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, Liqiang Jing, and Xinya Du. 2024c. Learning to generate research idea with dynamic control. *Preprint*, arXiv:2412.14626.

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

Shimin Li, Tianxiang Sun, Qinyuan Cheng, and Xipeng Qiu. 2024e. Agent alignment in evolving social norms. *Preprint*, arXiv:2401.04620.

Haowei Lin, Yacine Jernite, H. T. Kung, and Andrew Gordon Wilson. 2025. Evosld: Automated neural scaling law discovery with large language models. *Preprint*, arXiv:2507.21184.

Chengwei Liu, Chong Wang, Jiayue Cao, Jingquan Ge, Kun Wang, Lvye Zhang, Ming-Ming Cheng, Penghai Zhao, Tianlin Li, Xiaojun Jia, Xiang Li, Xinfeng Li, Yang Liu, Yebo Feng, Yihao Huang, Yijia Xu, Yuqiang Sun, Zhenhong Zhou, and Zhengzi Xu. 2025a. A vision for auto research with llm agents. *Preprint*, arXiv:2504.18765.

Ryan Liu and Nihar B. Shah. 2023. Reviewergpt? an exploratory study on using large language models for paper reviewing. *Preprint*, arXiv:2306.00622.

Sizhe Liu, Yizhou Lu, Siyu Chen, Xiyang Hu, Jieyu Zhao, Yingzhou Lu, and Yue Zhao. 2025b. Drugagent: Automating ai-aided drug discovery programming through llm multi-agent collaboration. *Preprint*, arXiv:2411.15692.

Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. 2025c. Researchbench: Benchmarking llms in scientific discovery via inspiration-based task decomposition. *Preprint*, arXiv:2503.21248.

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. *Preprint*, arXiv:2408.06292.

Yi Luo, Linghang Shi, Yihao Li, Aobo Zhuang, Yeyun Gong, Ling Liu, and Chen Lin. 2024. From intention to implementation: Automating biomedical research via llms. *Preprint*, arXiv:2412.09429.

Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025. Llm4sr: A survey on large language models for scientific research. *Preprint*, arXiv:2501.04306.

Bodhisattwa Prasad Majumder, Bhavana Dalvi Mishra, Peter Jansen, Oyvind Tafjord, Niket Tandon, Li Zhang, Chris Callison-Burch, and Peter Clark. 2023. Clin: A continually learning language agent for rapid task adaptation and generalization. *Preprint*, arXiv:2310.10134.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. 2024. Discoverybench: Towards data-driven discovery with large language models. *Preprint*, arXiv:2407.01725.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *Preprint*, arXiv:2203.10244.

Deepak Nathani, Lovish Madaan, Nicholas Roberts, Nikolay Bashlykov, Ajay Menon, Vincent Moens, Amar Budhiraja, Despoina Magka, Vladislav Vorotilov, Gaurav Chaurasia, Dieuwke Hupkes, Ricardo Silveira Cabral, Tatiana Shavrina, Jakob Foerster, Yoram Bachrach, William Yang Wang, and Roberta Raileanu. 2025. Mlgym: A new framework and benchmark for advancing ai research agents. *Preprint*, arXiv:2502.14499.

Benjamin Newman, Yoonjoo Lee, Aakanksha Naik, Pao Siangliulue, Raymond Fok, Juho Kim, Daniel S. Weld, Joseph Chee Chang, and Kyle Lo. 2024. Arxivdigestables: Synthesizing scientific literature into tables using language models. *Preprint*, arXiv:2410.22360.

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.

Odhran O'Donoghue, Aleksandar Shtedritski, John Ginger, Ralph Abboud, Ali Essa Ghareeb, Justin Booth, and Samuel G Rodriques. 2023. Bioplanner: Automatic evaluation of llms on protocol planning in biology. *Preprint*, arXiv:2310.10632.

Charles O'Neill, Tirthankar Ghosal, Roberta Răileanu, Mike Walmsley, Thang Bui, Kevin Schawinski, and Ioana Ciucă. 2025. Sparks of science: Hypothesis generation using structured paper data. *Preprint*, arXiv:2504.12976.

OpenAI. 2024. Introducing openai o1 preview.

OpenAI. 2025. Introducing deep research.

Jiefu Ou, William Gantt Walden, Kate Sanders, Zhengping Jiang, Kaiser Sun, Jeffrey Cheng, William Jurayj, Miriam Wanner, Shaobo Liang, Candice Morgan, Seunghoon Han, Weiqi Wang, Chandler May, Hannah Recknor, Daniel Khashabi, and Benjamin Van Durme. 2025. Claimcheck: How grounded are llm critiques of scientific papers? *Preprint*, arXiv:2503.21717.

Karl R. Popper. 1935. *The Logic of Scientific Discovery*. Routledge, London, England.

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.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. Infobench: Evaluating instruction following ability in large language models. *Preprint*, arXiv:2401.03601.

Yansheng Qiu, Haoquan Zhang, Zhaopan Xu, Ming Li, Diping Song, Zheng Wang, and Kaipeng Zhang. 2025. Ai idea bench 2025: Ai research idea generation benchmark. *Preprint*, arXiv:2504.14191.

Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-rong Wen. 2025. Tool learning with large language models: a survey. *Frontiers of Computer Science*, 19(8).

Xin Quan, Marco Valentino, Louise A. Dennis, and André Freitas. 2024. Verification and refinement of natural language explanations through llm-symbolic theorem proving. *Preprint*, arXiv:2405.01379.

Gollam Rabby, Diyana Muhammed, Prasenjit Mitra, and Sören Auer. 2025. Iterative hypothesis generation for scientific discovery with monte carlo nash equilibrium self-refining trees. *Preprint*, arXiv:2503.19309.

Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S. Weld. 2025. Scideator: Human-llm scientific idea generation grounded in research-paper facet recombination. *Preprint*, arXiv:2409.14634.

Chandan K Reddy and Parshin Shojaee. 2025. Towards scientific discovery with generative ai: Progress, opportunities, and challenges. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):28601–28609.

Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, and Hao Sun. 2025. Liveideabench: Evaluating llms' divergent thinking for scientific idea generation with minimal context. *Preprint*, arXiv:2412.17596.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. 2025. Agent laboratory: Using llm agents as research assistants. *Preprint*, arXiv:2501.04227.

Zhenqian Shen, Yongqi Zhang, Lanning Wei, Huan Zhao, and Quanming Yao. 2024. Automated machine learning: From principles to practices. *Preprint*, arXiv:1810.13306.

Yu-Zhe Shi, Mingchen Liu, Fanxu Meng, Qiao Xu, Zhangqian Bi, Kun He, Lecheng Ruan, and Qining Wang. 2025. Hierarchically encapsulated representation for protocol design in self-driving labs. *Preprint*, arXiv:2504.03810.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K Reddy. 2025a. Llm-sr: Scientific equation discovery via programming with large language models. *Preprint*, arXiv:2404.18400.

Parshin Shojaee, Ngoc-Hieu Nguyen, Kazem Meidani, Amir Barati Farimani, Khoa D Doan, and Chandan K Reddy. 2025b. Llm-srbench: A new benchmark for scientific equation discovery with large language models. *Preprint*, arXiv:2504.10415.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *Preprint*, arXiv:2409.04109.

Michael Skarlinski, Tyler Nadolski, James Braza, Remo Storni, Mayk Caldas, Ludovico Mitchener, Michaela Hinks, Andrew White, and Sam Rodriques. 2025. Futurehouse platform: Superintelligent ai agents for scientific discovery.

Tao Song, Man Luo, Linjiang Chen, Yan Huang, Qing Zhu, Daobin Liu, Baicheng Zhang, Gang Zou, Fei Zhang, Weiwei Shang, Jun Jiang, and Yi Luo. 2024. A multi-agent-driven robotic ai chemist enabling autonomous chemical research on demand. *ChemRxiv*. Preprint.

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. 2025. Paperbench: Evaluating ai's ability to replicate ai research. *Preprint*, arXiv:2504.01848.

Maojun Sun, Ruijian Han, Binyan Jiang, Houduo Qi, Defeng Sun, Yancheng Yuan, and Jian Huang. 2024. A survey on large language model-based agents for statistics and data science. *Preprint*, arXiv:2412.14222.

Shiro Takagi, Ryutaro Yamauchi, and Wataru Kumagai. 2023. Towards autonomous hypothesis verification via language models with minimal guidance. *Preprint*, arXiv:2311.09706.

Minyang Tian, Luyu Gao, Shizhuo Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, Shengyan Liu, Di Luo, Yutao Ma, Hao Tong, Kha Trinh, Chenyu Tian, Zihan Wang, Bohao Wu, Yanyu Xiong, and 11 others. 2024. Scicode: A research coding benchmark curated by scientists. *Preprint*, arXiv:2407.13168.

Keith Tyser, Ben Segev, Gaston Longhitano, Xin-Yu Zhang, Zachary Meeks, Jason Lee, Uday Garg, Nicholas Belsten, Avi Shporer, Madeleine Udell, Dov Te'eni, and Iddo Drori. 2024. Ai-driven review systems: Evaluating llms in scalable and bias-aware academic reviews. *Preprint*, arXiv:2408.10365.

Silviu-Marian Udrescu and Max Tegmark. 2020. Ai feynman: a physics-inspired method for symbolic regression. *Preprint*, arXiv:1905.11481.

Rosni Vasu, Chandrayee Basu, Bhavana Dalvi Mishra, Cristina Sarasua, Peter Clark, and Abraham Bernstein. 2025. HypER: Literature-grounded hypothesis generation and distillation with provenance. *Preprint*, arXiv:2506.12937.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024a. Scimon: Scientific inspiration machines optimized for novelty. *Preprint*, arXiv:2305.14259.

Runxiang Wang, Boxiao Wang, Kai Li, Yifan Zhang, and Jian Cheng. 2025a. Drsr: Llm based scientific equation discovery with dual reasoning from data and experience. *Preprint*, arXiv:2506.04282.

Weiqi Wang, Jiefu Ou, Yangqiu Song, Benjamin Van Durme, and Daniel Khashabi. 2025b. Can llms generate tabular summaries of science papers? rethinking the evaluation protocol. *Preprint*, arXiv:2504.10284.

Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. 2024b. Executable code actions elicit better llm agents. *Preprint*, arXiv:2402.01030.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Linyi Yang, Jindong Wang, Cunxiang Wang, Kaijie Zhu, Yiqiao Jin, and Xing Xie. 2024c. Autosurvey: Large language models can automatically write surveys. *Preprint*, arXiv:2406.10252.

Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. 2024d. Chain-of-table: Evolving tables in the reasoning chain for table understanding. *Preprint*, arXiv:2401.04398.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024e. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Preprint*, arXiv:2406.18521.

Yutai Wen, Yifan Jiang, Zitao Li, Whytnee Wade, J. D. Zamfirescu-Pereira, Xinyun Chen, Yuxin Wen, Yiming Yang, Graham Neubig, Jacob Andreas, and Daniel S. Weld. 2025. Predicting empirical ai research outcomes with language models. *Preprint*, arXiv:2506.00794.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2025. Cycleresearcher: Improving automated research via automated review. *Preprint*, arXiv:2411.00816.

Hjalmar Wijk, Tao Lin, Joel Becker, Sami Jawhar, Neev Parikh, Thomas Broadley, Lawrence Chan, Michael Chen, Josh Clymer, Jai Dhyani, Elena Ericheva, Katharyn Garcia, Brian Goodrich, Nikola Jurkovic, Holden Karnofsky, Megan Kinniment, Aron Lajko, Seraphina Nix, Lucas Sato, and 4 others. 2024. Rebench: Evaluating frontier ai rd capabilities of language model agents against human experts. *Preprint*, arXiv:2411.15114.

Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xinrun Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, Guanglin Niu, Tongliang Li, and Zhoujun Li. 2025. Tablebench: A comprehensive and complex benchmark for table question answering. *Preprint*, arXiv:2408.09174.

xAI. 2025. Grok 3 beta – the age of reasoning agents.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Peng Ye, Min Dou, Botian Shi, Junchi Yan, and Yu Qiao. 2025. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *Preprint*, arXiv:2402.12185.

Yanzheng Xiang, Hanqi Yan, Shuyin Ouyang, Lin Gui, and Yulan He. 2025. Scireplicate-bench: Benchmarking llms in agent-driven algorithmic reproduction from research papers. *Preprint*, arXiv:2504.00255.

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.

Baixuan Xu, Chunyang Li, Weiqi Wang, Wei Fan, Tianshi Zheng, Haochen Shi, Tao Fan, Yangqiu Song, and Qiang Yang. 2025a. Towards multi-agent reasoning systems for collaborative expertise delegation: An exploratory design study. *Preprint*, arXiv:2505.07313.

Wenyi Xu, Yuren Mao, Xiaolu Zhang, Chao Zhang, Xuemei Dong, Mengfei Zhang, and Yunjun Gao. 2025b. Dagent: A relational database-driven data analysis report generation agent. *Preprint*, arXiv:2503.13269.

Yinggan Xu, Hana Kimlee, Yijia Xiao, and Di Luo. 2025c. Advancing ai-scientist understanding: Making llm think like a physicist with interpretable reasoning. *Preprint*, arXiv:2504.01911.

Eric Xue, Zeyi Huang, Yuyang Ji, and Haohan Wang. 2025. Improve: Iterative model pipeline refinement and optimization leveraging llm agents. *Preprint*, arXiv:2502.18530.

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

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

Pengcheng Yin, Wen-Ding Li, Kefan Xiao, Abhishek Rao, Yeming Wen, Kensen Shi, Joshua Howland, Paige Bailey, Michele Catasta, Henryk Michalewski, Alex Polozov, and Charles Sutton. 2022. Natural language to code generation in interactive data science notebooks. *Preprint*, arXiv:2212.09248.

Naruki Yoshikawa, Marta Skreta, Kourosh Darvish, Sebastian Arellano-Rubach, Zhi Ji, Lasse Bjørn Kristensen, Andrew Zou Li, Yuchi Zhao, Haoping Xu, Artur Kuramshin, Alán Aspuru-Guzik, Florian Shkurti, and Animesh Garg. 2023. Large language models for chemistry robotics. *Autonomous Robots*, 47:1057–1086.

Fatemeh Pesaran Zadeh, Juyeon Kim, Jin-Hwa Kim, and Gunhee Kim. 2025. Text2chart31: Instruction tuning for chart generation with automatic feedback. *Preprint*, arXiv:2410.04064.

Qiang Zhang, Keyan Ding, Tianwen Lv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Mengyao Zhang, Jinlu Zhang, Jiyu Cui, Renjun Xu, Hongyang Chen, Xiaohui Fan, and 2 others. 2025a. Scientific large language models: A survey on biological & chemical domains. *ACM Comput. Surv.*, 57(6).

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024. A comprehensive survey of scientific large language models and their applications in scientific discovery. *Preprint*, arXiv:2406.10833.

Yunxiang Zhang, Muhammad Khalifa, Shitanshu Bhushan, Grant D. Murphy, Lajanugen Logeswaran, Jaekyeom Kim, Moontae Lee, Honglak Lee, and Lu Wang. 2025b. Mlrc-bench: Can language agents solve machine learning research challenges? *Preprint*, arXiv:2504.09702.

Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh T. N. Nguyen, Lauren T. May, Geoffrey I. Webb, and Shirui Pan. 2023. Large language models for scientific synthesis, inference and explanation. *Preprint*, arXiv:2310.07984.

Ruiyang Zhou, Lu Chen, and Kai Yu. 2024a. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351, Torino, Italia. ELRA and ICCL.

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

A Summary Tables of LLMs in Scientific Discovery

Research Works	Science Domain	Task Nature	Framework Methodology	Evaluation Benchmark	Agentic Workflow
		Literature Search and Info Aggregation			
LitLLM (Agarwal et al., 2024)	General	Literature	✓	✗	✗
Science Hierarchography (Gao et al., 2025)	General	Literature	✓	✗	✓
Dennstädt et al. (2024)	Biomedicine	Literature	✓	✗	✗
SCIMON (Wang et al., 2024a)	General	Literature, Idea Generation	✓	✗	✗
ResearchAgent (Baek et al., 2025)	General	Literature, Idea Generation	✓	✗	✓
Text-Tuple-Table (Deng et al., 2024b)	General	Text2Table	✓	✓	✗
TKGT (Jiang et al., 2024b)	General	Text2Table	✓	✓	✗
ArxivDIGESTables (Newman et al., 2024)	General	Literature, Text2Table	✓	✓	✗
arXiv2Table (Wang et al., 2025b)	General	Literature, Text2Table	✓	✓	✗
PaperQA & LitQA (Lála et al., 2023)	General	Literature	✓	✓	✗
AutoSurvey (Wang et al., 2024c)	General	Literature	✓	✗	✓
		Idea Generation and Hypothesis Formulation			
Si et al. (2024)	Artificial Intelligence	Idea Generation	✗	✓	✗
LiveIdeaBench (Ruan et al., 2025)	General	Idea Generation	✗	✓	✗
Nova (Hu et al., 2024a)	General	Literature, Idea Generation	✓	✗	✓
IdeaBench (Guo et al., 2024a)	General	Literature, Idea Generation	✗	✓	✗
GraphEval (Feng et al., 2025)	Artificial Intelligence	Literature, Idea Generation	✗	✓	✗
AI Idea Bench 2025 (Qiu et al., 2025)	Artificial Intelligence	Literature, Idea Generation	✗	✓	✗
Buehler (2024)	Biology	Literature, Idea Generation	✓	✗	✗
SciAgents (Ghafarollahi and Buehler, 2024b)	General	Literature, Idea / Hypothesis Generation	✓	✗	✓
MOOSE-Chem (Yang et al., 2025)	Chemistry	Literature, Idea / Hypothesis Generation	✓	✓	✓
Yang et al. (2024)	General	Literature, Idea / Hypothesis Generation	✗	✓	✗
ResearchBench (Liu et al., 2025c)	General	Literature, Idea / Hypothesis Generation	✗	✓	✗
KG-CoI (Xiong et al., 2024)	General	Literature, Idea / Hypothesis Generation	✓	✗	✓
HypER (Vasu et al., 2025)	General	Literature, Idea / Hypothesis Generation	✓	✗	✓
O'Neill et al. (2025)	General	Literature, Idea / Hypothesis Generation	✓	✗	✓
Ciucă et al. (2023)	Astronomy	Hypothesis Generation	✓	✗	✗
O'Brien et al. (2024)	Biomedicine	Hypothesis Generation	✓	✗	✗
LLM4GRN (Afonja et al., 2024)	Biology	Hypothesis Generation	✓	✗	✗
Zhou et al. (2024b)	General	Hypothesis Generation	✓	✓	✗
Qi et al. (2023)	General	Hypothesis Generation	✗	✓	✗
Qi et al. (2024)	Biomedicine	Hypothesis Generation	✗	✓	✗
Scideator (Radensky et al., 2025)	General	Idea / Hypothesis Generation	✓	✗	✗
Li et al. (2024c)	General	Idea / Hypothesis Generation	✓	✗	✗
		Experiment Planning and Execution			
Li et al. (2025)	General	Planning	✓	✗	✗
Shi et al. (2025)	Biology	Planning	✓	✗	✓
BioPlanner (O'Donoghue et al., 2023)	Biology	Planning	✗	✓	✗
ARCADE (Yin et al., 2022)	Artificial Intelligence	Code Generation	✓	✓	✗
AIDE (Jiang et al., 2025)	Artificial Intelligence	Code Generation	✓	✗	✓
SciCode (Tian et al., 2024)	Artificial Intelligence	Code Generation	✗	✓	✗
DS-1000 (Lai et al., 2022)	Artificial Intelligence	Code Generation	✗	✓	✗
MLE-Bench (Chan et al., 2025)	Artificial Intelligence	Code Generation	✗	✓	✓
		Data Analysis and Organization			
AutomaTikZ (Belouadi et al., 2024)	General	Text2Chart	✓	✓	✗
Text2Chart31 (Zadeh et al., 2025)	General	Text2Chart	✓	✗	✗
ChartX & ChartVLM (Xia et al., 2025)	General	Chart Reasoning	✓	✓	✗
CharXiv (Wang et al., 2024e)	General	Chart Reasoning	✗	✓	✗
ChartQA (Masry et al., 2022)	General	Chart Reasoning	✗	✓	✗
Chain-of-Table (Wang et al., 2024d)	General	Tabular Reasoning	✓	✗	✓
TableBench (Wu et al., 2025)	General	Tabular Reasoning	✗	✓	✗
Deng et al. (2024a)	General	Tabular Reasoning	✗	✓	✗
		Conclusion and Hypothesis Validation			
Tyser et al. (2024)	General	Review	✓	✗	✗
ClaimCheck (Ou et al., 2025)	Artificial Intelligence	Review	✗	✓	✗
Du et al. (2024)	Artificial Intelligence	Review	✗	✓	✗
Zhou et al. (2024a)	Artificial Intelligence	Review	✗	✓	✗
ReviewerGPT (Liu and Shah, 2023)	Artificial Intelligence	Review	✗	✓	✗
Bougie and Watanabe (2024)	General	Review	✓	✗	✓
CycleResearcher (Weng et al., 2025)	Artificial Intelligence	Review	✓	✓	✓
Takagi et al. (2023)	General	Hypothesis Validation	✓	✗	✗
Wen et al. (2025)	Artificial Intelligence	Hypothesis Validation	✓	✗	✗
PaperBench (Starace et al., 2025)	Artificial Intelligence	Hypothesis Validation	✗	✓	✓
SciReplicate-Bench (Xiang et al., 2025)	Artificial Intelligence	Hypothesis Validation	✗	✓	✓
Xu et al. (2025c)	Physics	Hypothesis Validation	✓	✗	✓
		Iteration and Refinement			
Quan et al. (2024)	General	Refinement	✓	✗	✗
MC-NEST (Rabby et al., 2025)	General	Hypothesis Generation, Refinement	✓	✗	✓
Chain of Ideas (Li et al., 2024a)	Artificial Intelligence	Idea Generation, Refinement	✓	✓	✓

Table A1: Comparison and classification of *Level 1* research works in LLM-based scientific discovery.

Research Works	Science Domain	Methodology Framework	Benchmark Evaluation	Scientific Method Stages					
				Obs.	*Hyp.*	*Exp.*	*Ana.*	*Con.*	*Ref.*
			Machine Learning Research						
CodeScientist (Jansen et al., 2025)	Artificial Intelligence	✓	✗	✓	✓	✓	✗	✗	✗
BudgetMLAgent (Gandhi et al., 2025)	Artificial Intelligence	✓	✗	✗	✓	✓	✓	✓	✗
IMPROVE (Xue et al., 2025)	Artificial Intelligence	✓	✗	✗	✗	✓	✓	✓	✓
MLAgentBench (Huang et al., 2024a)	Artificial Intelligence	✗	✓	✗	✗	✓	✓	✓	✗
MLR-Copilot (Li et al., 2024d)	Artificial Intelligence	✓	✗	✗	✗	✓	✓	✗	✓
MLRC-Bench (Zhang et al., 2025b)	Artificial Intelligence	✗	✓	✗	✗	✓	✓	✗	✓
RE-Bench (Wijk et al., 2024)	Artificial Intelligence	✗	✓	✗	✗	✓	✓	✗	✓
MLZero (Fang et al., 2025)	Artificial Intelligence	✓	✓	✗	✗	✓	✓	✗	✓
Genesys (Cheng et al., 2025b)	Artificial Intelligence	✓	✗	✗	✗	✓	✗	✗	✓
MLGym (Nathani et al., 2025)	Artificial Intelligence	✓	✓	✗	✗	✓	✓	✗	✓
			Data Modeling and Analysis						
DAgent (Xu et al., 2025b)	Data Science	✓	✗	✗	✓	✓	✓	✓	✗
DS-Agent (Guo et al., 2024b)	Data Science	✓	✗	✗	✓	✓	✓	✓	✓
InfiAgent-DABench (Hu et al., 2024b)	Data Science	✗	✓	✗	✓	✓	✓	✗	✗
BLADE (Gu et al., 2024)	Data Science	✗	✓	✗	✗	✓	✓	✓	✗
DiscoveryBench (Majumder et al., 2024)	Data Science	✗	✓	✗	✓	✓	✓	✓	✗
DSBench (Jing et al., 2024)	Data Science	✗	✓	✗	✗	✓	✓	✓	✗
Zheng et al. (2023)	General	✓	✓	✓	✓	✓	✓	✓	✗
			Function Discovery						
BoxLM (Li et al., 2024b)	Statistics	✓	✗	✗	✓	✓	✓	✓	✗
LLM-SR (Shojaee et al., 2025a)	General	✓	✗	✗	✓	✓	✗	✓	✓
LLM-SRBench (Shojaee et al., 2025b)	General	✗	✓	✗	✓	✓	✗	✓	✓
Gravity-Bench-v1 (Koblischke et al., 2025)	Physics	✗	✓	✗	✓	✓	✗	✓	✓
DrSR (Wang et al., 2025a)	General	✓	✗	✗	✓	✓	✓	✓	✓
EvoSLD (Lin et al., 2025)	Artificial Intelligence	✓	✗	✗	✓	✓	✓	✓	✓
			Natural Science Research						
Coscientist (Boiko et al., 2023)	Chemistry	✓	✗	✗	✓	✓	✓	✓	✗
Gao et al. (2024)	Biomedicine	✓	✗	✗	✓	✓	✓	✓	✗
BioResearcher (Luo et al., 2024)	Biomedicine	✓	✗	✗	✗	✓	✓	✓	✓
DrugAgent (Liu et al., 2025b)	Biomedicine	✓	✗	✗	✓	✓	✗	✗	✓
FutureHouse (Skarlinski et al., 2025)	Chemistry, Biology	✓	✗	✓	✓	✓	✗	✗	✗
ScienceAgentBench (Chen et al., 2025c)	Chemistry, Biology	✗	✓	✓	✓	✓	✓	✓	✗
ProtAgents (Ghafarollahi and Buehler, 2024a)	Chemistry, Biology	✓	✗	✓	✓	✓	✓	✓	✗
Auto-Bench (Chen et al., 2025b)	General	✗	✓	✗	✗	✓	✓	✗	✓
AI co-scientist (Gottweis et al., 2025)	General	✓	✗	✗	✓	✓	✓	✓	✗
			General Research						
DiscoveryWorld (Jansen et al., 2024)	General	✗	✓	✗	✓	✓	✓	✓	✓
Liu et al. (2025a)	General	✗	✓	✓	✓	✓	✓	✓	✓
Curie (Kon et al., 2025)	General	✓	✗	✗	✗	✓	✓	✓	✗
EAIRA (Cappello et al., 2025)	General	✗	✓	✓	✓	✓	✓	✓	✗

Table A2: Comparison and classification of *Level 2* research works in LLM-based scientific discovery.

Research Works	Science Domain	Methodology Framework	Benchmark Evaluation	Featured Functionality	Open-Sourced?
Agent Laboratory (Schmidgall et al., 2025)	Artificial Intelligence	✓	✓	literature review, experimentation, report writing, iterative research with human feedback loops.	✓
The AI Scientist (Lu et al., 2024)	Artificial Intelligence	✓	✗	idea generation, code generation, experiment execution, research paper writing.	✓
The AI Scientist-v2 (Yamada et al., 2025)	Artificial Intelligence	✓	✗	idea generation, code generation, experiment execution, research paper writing, with agentic tree-search and feedbacks.	✓
AI-Researcher (Data Intelligence Lab, 2025)	Artificial Intelligence	✓	✗	literature review, data analysis, report generation.	✓
Zochi (IntologyAI, 2025)	Artificial Intelligence	✓	✗	customizable workflows for data collection, analysis, and decision-making.	✓
Carl (Autoscience, 2025)	Artificial Intelligence	✓	✗	hypothesis generation, experiment design, data analysis, and manuscript writing.	✗

Table A3: Comparison and classification of *Level 3* research works in LLM-based scientific discovery.