Towards AI-Driven Science and Innovation: Paradigms, Pratices, Challenges, and Horizons

Xiaoyu Xiong[1], Hao Wang[1], Keming Wu[2], Zhenfei Yang[1], Hanjie Zhao[1], Hongxiang Wang[1]

Hao Liu[1], Deyi Xiong[1,3*]

[1] Tianjin University [2] Tsinghua University [3] Shenzhen Loop Area Institute

Abstract

Science and innovation advance in a mutually reinforcing cycle, yet progress today is increasingly constrained by information overload, fragmented expertise, slow experimentation, and the difficulty of recognizing genuine innovation. Recent advances in artificial intelligence, particularly the emergence of large language models, have made AI an increasingly important force in this landscape. We organize its role into two paradigms. The human-AI collaborative paradigm combines computational power with human expertise to enhance reasoning, creativity, and contextual judgment. The autonomous paradigm emphasizes efficiency, supporting large-scale knowledge mining, hypothesis generation, integrated experimental workflows and end-to-end science system. This survey provides a systematic account of AI contributions across the scientific pipeline, from problem formulation to experimentation, scientific dissemination, and application. Distinctively, we highlight the necessity of human-AI collaboration and the coupled evolution of science and innovation, two perspectives often overlooked in prior surveys. We systematically investigate persistent challenges in interpretability, evaluation, creativity, and the identification of genuine innovation, and discuss future directions for AI from a research accelerator to a genuine enabler of science and societal progress. A curated list of related papers has been publicly available at a GitHub repository.[1]

*Corresponding author.

[1] https://github.com/TJUNLP-xxy/Awesome-AI-Science-and-Innovation

Contents

1 Introduction

Science and innovation have long stood at the core of human progress, shaping both the understanding of the natural world and the transformation of society. From classical mechanics and electromagnetism to the discovery of DNA and the rise of modern computing, science has continually expanded the frontiers of knowledge, while innovation has provided new instruments, infrastructures, and practical avenues that have in turn stimulated further inquiry (Newton, 1833; Maxwell, 1865; Watson and Crick, 1953; Turing et al., 1936). While science primarily seeks to uncover the underlying principles and laws governing phenomena, innovation focuses on translating such understanding into tangible applications, technologies, or systems that alter human experience and capability. Rather than being sharply separated, the two evolve in a co-dependent manner: scientific advances often create the conceptual foundations upon which innovations flourish, whereas innovations frequently open unforeseen directions for scientific exploration. Yet, this reciprocal dynamic faces mounting challenges in the contemporary landscape, where the explosion of information (Bretscher, 2022; Hanson et al., 2024), the increasing fragmentation of expertise (Ananyin, 2024), and the lengthening cycles of experimentation have made the pursuit of science more demanding.

Against this backdrop, artificial intelligence (AI) catalyzes the science-innovation pipeline (Gil, 2017; Gridach et al., 2025) via two complementary paradigms: the *humanAI collaboration paradigm* and the *autonomous paradigm*. The *humanAI collaboration paradigm* integrates human expertise with AI systems (Gottweis et al., 2025; Passerini et al., 2025), where dialogue interfaces, feedback loops, and co-design platforms (Ni et al., 2024; Prasad et al., 2025) help refine hypotheses, connect fragmented literatures, and iteratively adapt experiment designs. In parallel, the autonomous paradigm operationalizes such autonomy and efficiency through large-scale literature mining (Glickman and Zhang, 2024), semantic modeling of scientific corpora (Beltagy et al., 2019), hypothesis generation (Yang et al., 2024a; Baek et al., 2024; Li et al., 2024b), and simulation-driven experimental design (Desai et al., 2025; Tian et al., 2021), thereby processing vast information, reducing trial-and-error, and expanding previously intractable problem spaces. Whats more, it increasingly advances toward end-to-end workflows that integrate multimodal data, standardize experimental protocols, and close the loop from data to hypothesis, experiment, and interpretation (Lu et al., 2024; Yamada et al., 2025; Yuan et al., 2025; Ruan et al., 2024). While autonomy alleviates cognitive and experimental burdens, collaboration supports creativity and contextual judgment. Together, these paradigms illustrate how AI can act both as a partner and as an accelerator, helping address the pressing bottlenecks of information overload, fragmented expertise, and extended experimental cycles in modern science. Figure 1 provides an overview of these two paradigms and depicts their respective trajectories of development.

Despite its transformative potential, AI for science and innovation still confronts profound obstacles. Scientific inquiry is inherently open-ended, requiring creativity that extends beyond routine optimization. At the same time, integrating fragmented knowledge into coherent theories calls for reasoning abilities that current models struggle to realize in a transparent manner. These epistemic limitations are compounded by practical concerns: many AI systems operate as black boxes, raising questions of interpretability and trust, while the lack of standardized benchmarks hampers rigorous evaluation and comparison. Overcoming these

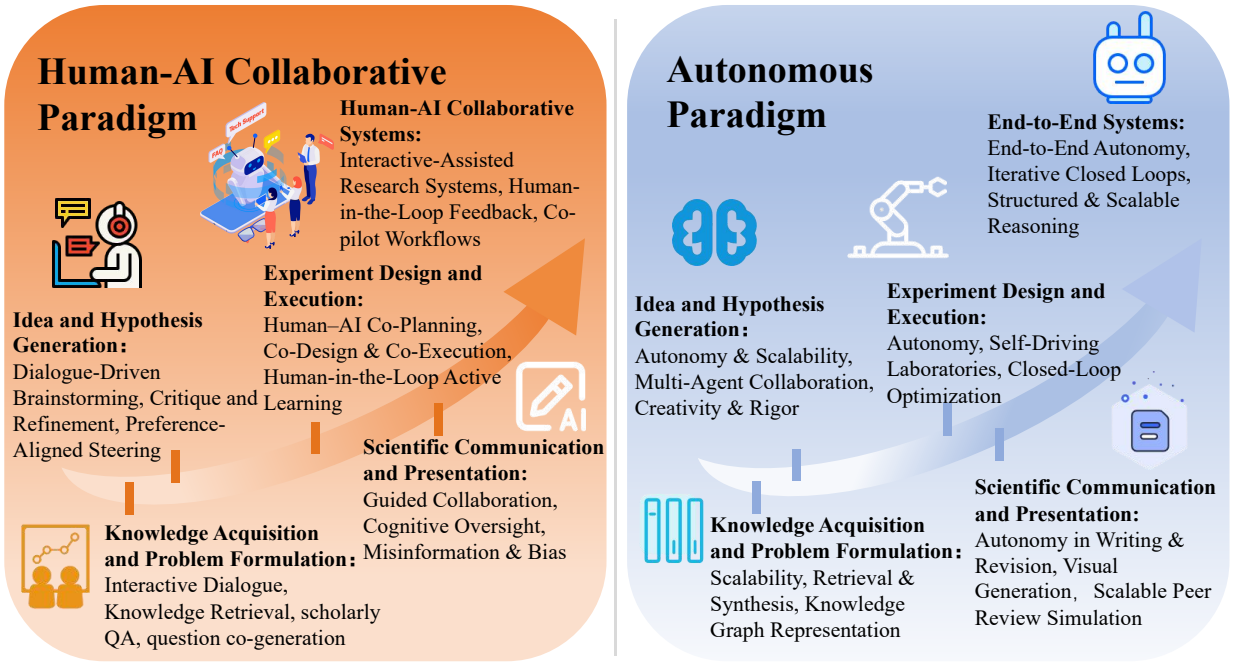

Figure 1: Human-AI Collaborative Paradigm vs. Autonomous Paradigm in AI-driven Science and Innovation, highlighting their keywords to knowledge acquisition and problem formulation, idea and hypothesis generation, experiment design and execution, communication of science, and science-integrated system.

intertwined challenges is crucial if AI is to evolve from a tool of acceleration into a genuine enabler of innovation.

In the past two years, a growing number of surveys have emerged on AI-driven science, reflecting the communitys effort to synthesize methods, frameworks, and applications in this rapidly evolving domain. Some works emphasize paradigm shifts, envisioning LLMs evolving from tools to autonomous scientific agents (Zhang et al., 2025d; Wei et al., 2025; Zheng et al., 2025); others highlight data foundations and system architectures, with a focus on Sci-LLMs and trustworthy scientific agents (Ren et al., 2025; Xie et al., 2025). In parallel, task-oriented surveys examine specific stages of the research workflow such as hypothesis generation, experimentation, and peer review (Eger et al., 2025; Luo et al., 2025; Alkan et al., 2025). Beyond these strands, another set of surveys adopts a broader disciplinary perspective, for instance, AI4Research (Chen et al., 2025d), which systematizes tasks and resources across scientific fields, or the Science of Science, which leverages AI to analyze and model the processes of research itself (Chen et al., 2025b). Building on these perspectives, our survey emphasizes two dimensions that have received comparatively less attention. The first is the collaborative paradigm: given that current LLMs remain largely opaque and imperfect, humanAI collaborative serves as a complementary approach, combining computational power with human judgment to jointly enhance reasoning, creativity, and decision-making, and thus remains an important area warranting continued attention. The second is the coupled evolution of science and innovation, which should not be treated in isolation. Scientific

discovery provides the conceptual foundations and insights that fuel innovative creation, while innovation opens new tools, methods, and applications that, in turn, guide further scientific exploration. Together, they form a mutually reinforcing spiral that continually drives both knowledge generation and practical impact.

Building on the intertwined evolution of science and innovation, the emerging role of AI, and the challenges that remain unresolved, this survey provides a systematic account of the field. We first contrast the human-AI collaborative and autonomous paradigms, within which we introduce a roadmap of science that traces the stages from problem formulation and hypothesis generation to experimental execution and theoretical communication, while detailing how AI intervenes at each point and how its role has evolved over time (Section 2). We then examine AI techniques across these stages of the science pipeline (Section 3), followed by a review of integrated systems aiming at end-to-end workflows (Section 4). Building on this, we highlight domain-specific innovations and emerging infrastructures (Sections 5, 6). Finally, we discuss challenges and risks (Section 7) and conclude with directions for future research (Section 9), offering a structured view toward the development of AI-driven science and innovation.

2 Paradigms of AI-Driven Science and Innovation: Human-AI Collaborative vs. Autonomous

In AI-driven science and innovation, most research efforts over recent years have focused on increasing autonomy in scientific and innovative workflows (Abramson et al., 2024; Szymanski et al., 2023b; Dai et al., 2024; Sheng et al., 2024; Gottweis et al., 2025; Yamada et al., 2025). Autonomous systems can operate at different levels, ranging from supporting specific stages of the scientific pipeline, such as knowledge acquisition and problem formulation (Wan et al., 2024; Ivanisenko et al., 2024; Susnjak et al., 2025), idea and hypothesis generation (Yang et al., 2024a; Ghafarollahi and Buehler, 2025; Yang et al., 2024b; Baek et al., 2024; Li et al., 2024b; Jansen et al., 2025), experimental design and execution (Shen et al., 2023b; Desai et al., 2025; Tian et al., 2021; Noh et al., 2024; Ruan et al., 2024), and scientific communication and presentation (Xu, 2025; Liang et al., 2025; Chen et al., 2025a; Wang et al., 2025d; Pang et al., 2025) to enabling end-to-end execution of the entire scientific process (Lu et al., 2024; Yamada et al., 2025; Yuan et al., 2025; Naumov et al., 2025; Team et al., 2025a). This paradigm offers clear advantages in efficiency and scalability, but it also faces several limitations. Its effectiveness depends on the availability and quality of input data (Gubbi et al., 2022; Abdel-Rehim et al., 2025; Li et al., 2025a), and it may struggle to incorporate nuanced domain knowledge. Moreover, autonomous systems often raise concerns regarding safety, interpretability, and reliability (Felderer and Ramler, 2021; Arrieta et al., 2020), and they may exhibit limited robustness when applied to novel or complex scenarios.

In contrast, human-AI collaborative approaches, where AI systems and human researchers interact in complementary ways across different stages of the scientific workflow, have received comparatively less attention to date. Such approaches can take many forms, including interactive hypothesis refinement (Ding et al., 2025; Abdel-Rehim et al., 2025; Li et al., 2025a), iterative experiment design (Andreasson et al., 2022; Adams et al., 2023; Nahal et al., 2024),

Table 1: AI roles across different stages of the scientific workflow under the collaborative and autonomous paradigms.

Stage	Collaborative Paradigm	Autonomous Paradigm
Knowledge Acquisition & Problem Formulation	Human-guided retrieval, AI-assisted synthesis and framing of research problems	Automated mining, summarization, comprehension, and direct problem formulation
Idea & Hypothesis Generation	AI-assisted brainstorming, iterative refinement, evidence-supported shaping	Autonomous generation, knowledge recombination, hypothesis formulation
Experiment Design & Execution	HumanAI co-design, interactive candidate selection, domain-guided refinement	Autonomous planning, robotic execution, active optimization
Scientific Communication & Presentation	Co-writing, interactive editing, and ethical oversight	Automated drafting, visualization, and review assistance
Integrated Systems for Scientific Discovery	Human-in-the-loop guidance, co-investigation, and feedback-driven refinement	End-to-end autonomous discovery, closed-loop experimentation, and adaptive optimization

guided data analysis (Mosqueira-Rey et al., 2023; Zhang et al., 2024; Pratiush et al., 2025; Dai et al., 2023), collaborative literature synthesis (van de Schoot et al., 2021; de Bruin et al., 2025; Callaghan and Müller-Hansen, 2020; Fok et al., 2025), and assisted scientific communication (Shen et al., 2023a; Gero et al., 2022; Chen et al., 2025a). By combining AI's computational capabilities with human contextual judgment, domain expertise, and creative insight, these systems offer opportunities to enhance the reliability, interpretability, efficiency, and overall quality of scientific processes, while addressing some of the limitations inherent in autonomous paradigms.

To more intuitively illustrate how AI intervenes across different stages of the scientific workflow, we summarize the two paradigms in Table 1 and further clarify this taxonomy in Figure 2, highlighting their distinct modes of contribution.

2.1 Four-Dimensional Comparative Analysis

Recognizing the importance and complementary roles of these two paradigms, this chapter presents a structured comparison and analysis along four key dimensions: input dependency, human feedback, AI closed-loop capability, and dominance in driving the research workflow. By framing the discussion in this way, we aim to provide researchers with insights into the trade-offs between different approaches and highlight directions for future development of AI in scientific discovery and innovation.

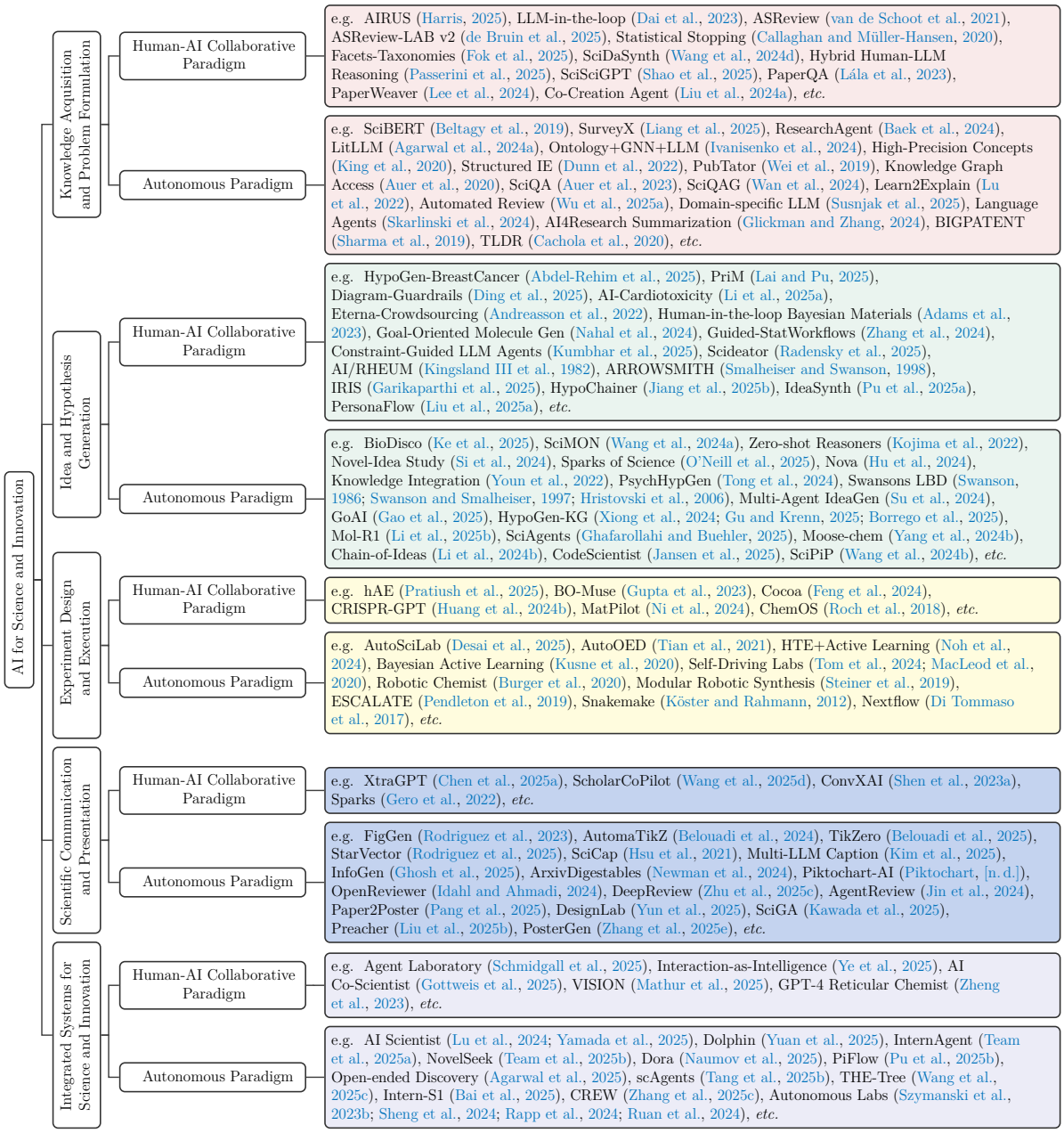

Figure 2: Overview of representative AI tools and systems for science and innovation. We organize existing research along four core stages of the scientific workflowknowledge acquisition and problem formulation, idea and hypothesis generation, experiment design and execution, and scientific communication and presentation, while distinguishing between humanAI collaborative paradigm and autonomous paradigm, which embed human expertise into the loop. Integrated end-to-end systems that unify multiple stages are also highlighted, illustrating the landscape of collaborative and autonomous in AI-driven science and innovation.

In AI-driven science and innovation, research workflows can be broadly organized into two paradigms: **human-AI collaborative systems** and **autonomous systems**. These differ

along four key dimensions*input dependency*, *human feedback*, *AI closed-loop capability*, and *dominance in the scientific process*each illustrated by representative examples.

Input dependency. In collaborative paradigms, human researchers actively guide the AI process by framing and contextualizing inputs. In CoQuest (Liu et al., 2024a), for example, experts refine AI-generated research questions by providing contextual information, and in SciDaSynth (Wang et al., 2024d), domain experts select and validate AI-generated data tables to ensure their relevance and accuracy. In contrast, autonomous systems allow AI to predominantly control the interpretation and utilization of input data, requiring minimal human intervention post-data preparation. For instance, AlphaFold predicts protein structures solely from sequence data using deep learning techniques (Jumper et al., 2021), and SCIPIP generates ideas by autonomously mining prior literature and data (Wang et al., 2024b).

Human feedback. Collaborative paradigms emphasize human agency in shaping the iterative process. BO-Muse integrates human-in-the-loop evaluation, allowing experts to guide parameter updates during Bayesian optimization (Gupta et al., 2023), while Scideator facilitates co-creation of research ideas by enabling experts to iteratively refine AI-generated concepts through interactive facet recombination (Radensky et al., 2025). In autonomous systems, AI independently drives iterative refinement processes with minimal human intervention. For instance, SciQAG autonomously generates and tests hypotheses without external guidance (Wan et al., 2024), and AlphaEvolve executes evolutionary simulations independently, refining algorithms through self-directed iterations (Novikov et al., 2025).

AI closed-loop capability. Collaborative systems typically maintain semi-autonomous loops in which human researchers intervene at critical junctures, guiding hypotheses, filtering results, or validating experimental outputs. For example, SciDaSynth allows experts to iteratively review and refine AI-generated data tables to ensure alignment with scientific goals and domain knowledge (Wang et al., 2024d), while coScientist positions AI as a virtual research partner whose hypotheses and experiment plans are continuously shaped through human oversight (Gottweis et al., 2025). Autonomous systems, in contrast, exhibit varying degrees of closed-loop operation. In some cases, AI governs the entire research cyclefrom hypothesis generation to experiment execution and analysisas in CAMEO (Kusne et al., 2020), which autonomously designs and tests experiments to discover novel materials, or A-Lab (Szymanski et al., 2023b), which automates synthesis and characterization workflows. In other cases, autonomy is localized to specific stages, such as hypothesis evaluation or data analysis, where AI iteratively refines outputs without requiring human intervention.

Dominance in driving the process. Collaborative paradigms place humans in the primary decision-making role, retaining strategic leadership while leveraging AI for computational support, proposals, or efficiency gains. BO-Muse positions experts as the primary drivers of Bayesian optimization (Gupta et al., 2023), Scideator enables researchers to iteratively shape AI-generated ideas (Radensky et al., 2025), the MIT Jameel Clinics deep learningguided discovery of the antibiotic keeps final validation in human hands (Stokes et al., 2020), and the MatPilot system uses human feedback to guide AI-driven materials hypothesis generation and experimental design (Ni et al., 2024). In autonomous systems, by contrast, AI takes the lead in hypothesis initiation, experimental design, and iterative execution. For instance, Robot Scientist Adam autonomously formulates and tests hypothe-

ses in yeast genomics (King et al., 2009), and materials acceleration platforms like A-Lab (Szymanski et al., 2023b) perform closed-loop synthesis of compounds.

2.2 Evolution of Collaborative and Autonomous Paradigms

In the human-AI collaborative paradigm, advances can be traced along methodological, capability, and integration axes. Methodologically, the trajectory began with assistive information systems that supported researchers in literature retrieval, semantic search, and knowledge organization (Ammar et al., 2018; Wei et al., 2019; Auer et al., 2020; King et al., 2020), thereby helping scholars navigate rapidly expanding scientific corpora. This evolved into interactive assistants, such as PaperQA (Lála et al., 2023) or Scideator (Radensky et al., 2025), which go beyond retrieval to engage in hypothesis refinement, ideation, and cross-disciplinary synthesis, allowing researchers to test and expand their reasoning in dialogue with AI systems. In terms of capability, collaboration has advanced from providing static references to dynamically co-generating ideas, methods, and even manuscript content (Ding et al., 2025; Abdel-Rehim et al., 2025; Shen et al., 2023a; Gero et al., 2022), while remaining receptive to human critique and steering. At the level of integration, collaboration has evolved from early stand-alone digital tools for literature management and search (Mueen Ahmed and Dhubaib, 2011; Ammar et al., 2018), to dialog-based copilots embedded directly into research workflows (Shen et al., 2023a; Shao et al., 2024), and further toward hybrid platforms where human judgment and AI-generated insights are interleaved across multiple stages of discovery (Zheng et al., 2023; Zhang et al., 2025c). Collectively, these advances delineate a trajectory from assistive utilities to interactive copilots and collaborative research environments, in which AI extends human reasoning, amplifies creativity, and contributes to shaping new possibilities for scientific and technological innovation.

In the autonomous paradigm, progress can likewise be traced along methodological, capability, and integration axes. Methodologically, the trajectory began with symbolic expert systems such as Dendral (1960s) (Lindsay et al., 1993), which used rule-based reasoning to assist chemists in molecular identification, and advanced to the first robot scientists such as Adam (King et al., 2009) and Eve (Williams et al., 2015), which embodied closed-loop autonomy by coupling hypothesis formulation with robotic experimentation and data interpretation. The deep learning era then introduced powerful predictive models, with AlphaFold (Jumper et al., 2021) representing a landmark in leveraging learned molecular representations for protein structure prediction. More recently, autonomy has progressed toward agentic and multi-agent frameworks. Self-driving laboratories (MacLeod et al., 2020; Tom et al., 2024; Steiner et al., 2019; Burger et al., 2020) exemplify this trend by integrating hypothesis generation, experimental design, execution, and analysis into closed-loop systems. At the software layer, orchestration platforms such as ChemOS (Roch et al., 2018) and data-centric pipelines such as ESCALATE (Pendleton et al., 2019) provide the infrastructure to coordinate and manage autonomous experimentation, enabling these laboratories to operate with minimal human intervention. Capability-wise, autonomy has progressed from narrow, single-task applications (e.g., spectral interpretation (Goodacre, 2003) or protein structure prediction (Rohl et al., 2004)) to increasingly sophisticated multi-stage workflows that autonomously connect discovery, validation, and optimization (Rapp et al., 2024). On the level

of system integration, developments moved from isolated computational tools (Jain et al., [n. d.]; Jumper et al., 2021), through hybrid modelingsynthesis platforms (Roch et al., 2018; Pendleton et al., 2019; Kusne et al., 2020), to fully autonomous laboratories (King et al., 2009; Williams et al., 2015; MacLeod et al., 2020; Tom et al., 2024; Steiner et al., 2019; Burger et al., 2020; Rapp et al., 2024) that orchestrate hypothesis-driven inquiry via robotic execution and iterative feedback. Taken together, these advances chart a coherent trajectory from rule-based systems to agentic infrastructures that increasingly assume the central functions of scientific discovery, thereby accelerating the pace of research, enabling exploration at unprecedented scale, and opening new probabilities for scientific and technological innovation.

Comparative perspective The collaborative and autonomous paradigms represent distinct yet complementary orientations. Collaboration emphasizes interpretability, creativity, and alignment with human judgment, focusing on co-constructing ideas and exploring unconventional directions through humanAI interaction. Autonomy prioritizes scale, speed, and reproducibility by delegating tasks to AI-driven pipelines, while also increasingly incorporating generative capabilities that enable novel hypotheses and design strategies to emerge. Their trade-offs highlight different strengths and limitations, suggesting that both paradigms are necessary and should be chosen according to specific scientific objectives and constraints.

2.3 Summary

Overall, humanAI collaborative paradigms and autonomous represent two distinct routes in AI-driven science and innovation. Autonomy emphasizes efficiency, scalability, and self-directed execution but raises concerns of interpretability and robustness, while collaboration highlights human strategic leadership, contextual judgment, and creativity, with AI serving as an accelerator. By comparing them across input dependency, human feedback, closed-loop capability, and process dominance, we outline their complementary strengths and limitations. At the same time, both paradigms continue to evolve: autonomy is moving toward more generative and agentic forms, and collaboration is advancing toward more interactive and systemic integration. Together, they point to a future where autonomy and co-creativity jointly shape scientific and technological innovation.

3 AI Roles across Key Stages of the Science and Innovation Pipeline

3.1 Knowledge Acquisition and Problem Formulation

Knowledge acquisition and problem generation constitute the cornerstone of AI-driven science and innovation. Within the broader scientific process, genuine progress typically begins with two interdependent activities: first, the assimilation of existing knowledge, and second, the articulation of novel, meaningful, and testable research questions. This stage can thus be conceptualized as a *knowledge-to-question chain*, where the systematic collection, structuring, and interpretation of prior work create a fertile ground from which innovative hypotheses and problem statements can emerge. By bridging what is already known with what remains

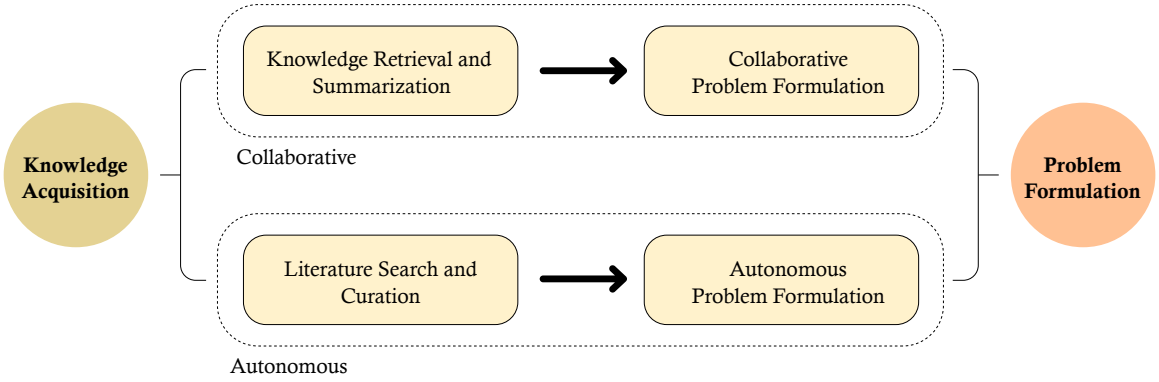

Figure 3: Structure of the two paradigms for knowledge acquisition and problem formulation. From scientific literature and data, the process splits into (i) a *collaborative* branch (humanAI knowledge structuring, interactive summarization, co-design) leading to co-created research questions, and (ii) an *autonomous* branch (RAG/LLM parsing, knowledge graphs, ontology-guided extraction) leading to autonomous problem/hypothesis formulation.

uncertain, this stage directly influences the trajectory of subsequent experimental design, validation, and application.

AI roles at this stage can be categorized into the *humanAI collaborative paradigm*, which highlights interactive co-creation between human and AI systems and foregrounds the centrality of human judgment, creativity, and domain-specific insights, and the *autonomous paradigm*, which emphasizes self-sustained knowledge extraction and problem formulation powered by AI. Figure 3 illustrates the overall structure of these two paradigms at this stage: beginning with scientific literature and data, the process diverges into (i) a collaborative path, where knowledge structuring and refinement occur in humanAI interactive cycles, leading to the co-creation of research questions; and (ii) an autonomous path, where retrieval-augmented formulation and large language models are applied to parse and represent knowledge, ultimately leading to autonomous problem formulation.

3.1.1 Collaborative Knowledge Acquisition and Problem Formulation

The collaborative paradigm emphasizes the synergistic integration of human with AI support. Rather than positioning autonomy as a replacement for researchers, this paradigm views AI as an active collaborator that amplifies human creativity, domain knowledge, and critical reasoning. The emphasis is thus placed on interactive processes in which humans and AI systems validate interpretations and iteratively refine research questions.

Within this paradigm, *knowledge retrieval and summarization* remain central but are reinterpreted through the lens of humanAI dialogue. Benchmark datasets such as BigPatent (Sharma et al., 2019) and TLDR (Cachola et al., 2020) have been used to train extreme summarization models that distill lengthy documents into concise, accessible insights. In addition, domain-specific pipelines for areas such as polymer science or biomedical research (Dunn et al., 2022), along with toolkits like LitLLM (Agarwal et al., 2024a), demon-

strate how customized summarization can make specialized knowledge more accessible to scientists.

Interactive frameworks exemplify this co-creative spirit. CoQuest (Liu et al., 2024a), for example, enables iterative exploration and refinement of research questions through dialogue between human users and AI agents. SciQA (Auer et al., 2023) serves as both a benchmark and an evaluative tool for question answering, emphasizing the integration of structured knowledge with human reasoning. Similarly, SciDaSynth (Wang et al., 2024d) demonstrates how automated extraction can be paired with human editing and validation to generate reliable, high-quality scientific syntheses. Even at the model-training level, domain-specific fine-tuning guided by expert feedback (Susnjak et al., 2025) shows how collaborative alignment can yield AI systems that are more faithful, accurate, and aligned with domain-specific expectations. Together, these examples illustrate how AI can act not merely as a data processor but as a genuine partner in knowledge creation.

3.1.2 Autonomous Knowledge Acquisition and Problem Formulation

In contrast, the autonomous paradigm leverages the capabilities of large language models , retrieval-augmented formulation, and knowledge graph to automatically extract, organize, and synthesize knowledge from massive scientific corpora. The primary advantages of this paradigm are its scalability and efficiency: by minimizing the need for constant human intervention, autonomous pipelines can rapidly process vast amounts of literature, uncover hidden connections, and systematically propose research questions.

A central task within autonomous knowledge acquisition is *literature search and curation*. For this, RAG-based methods have proven particularly effective. For example, PaperQA (Lála et al., 2023) introduces a generative agent that is capable of targeted question answering across large-scale corpora, while Paperweaver (Lee et al., 2024) contextualizes recommendations by integrating user-collected works, effectively transforming paper recommender systems into dynamic knowledge assistants. Emerging frameworks under the "AI Scientist" formalism (Lu et al., 2024) illustrate the potential of full autonomy across the pipelinefrom literature discovery and hypothesis generation to downstream evaluation.

Beyond retrieval and curation, autonomous systems are increasingly applied to construct benchmark resources and evaluate formulated problems. For instance, SciQAG (Wan et al., 2024) generates fine-grained science questionanswer datasets, while SciPIP (Wang et al., 2024b) and autonomous problem discovery frameworks (Yang et al., 2024a) integrate LLM prompting with real-time literature or web sources to propose novel research directions. On the representation side, platforms like the Open Research Knowledge Graph (ORKG) (Jaradeh et al., 2019) provide structured and queryable bases for scientific contributions. More advanced techniques, such as ontology-guided extraction and graph neural networks, further enrich these representations with semantic depth (Ivanisenko et al., 2024). Automated review generation using LLMs has also been explored (Wu et al., 2025a), offering new ways to synthesize research findings into coherent narratives at scale. Collectively, these advances underscore the growing capacity of autonomous systems to handle not only raw information but also the higher-level abstractions required for problem formulation.

3.1.3 Evaluation

The evaluation of knowledge acquisition and problem formulation presents unique challenges, as success depends not only on accuracy but also on novelty, utility, and interpretability. Reliable evaluation therefore requires a diverse set of resources, benchmarks, and access to scientific repositories.

Scientific literature is typically accessed through multiple types of repositories. By access policy, open-access repositories (e.g., PubMed Central, arXiv) provide unrestricted access, subscription-based repositories (e.g., ScienceDirect, SpringerLink) require paid access, and hybrid repositories (e.g., Taylor & Francis Online, Oxford Academic) offer mixed models. By content type, institutional repositories archive outputs of specific organizations (e.g., MIT DSpace, Harvard DASH), while preprint repositories (e.g., bioRxiv, chemRxiv) enable rapid, pre-review dissemination. Data repositories such as Dryad and Zenodo support reproducibility, while aggregator repositories like BASE and CORE (Knoth et al., 2023) broaden search coverage across sources. More recently, AI-powered platforms such as Elicit [2] and ORKG ASK [3] have introduced intelligent, multi-source querying, while tools like NotebookLM [4] enable customized, user-focused corpora exploration, and recommender systems such as Scholar Inbox deliver personalized literature alerts.

To evaluate knowledge acquisition specifically, literature-based discovery (LBD) benchmarks have emerged. Qi et al. (Qi et al., 2023), for instance, introduce a dataset constructed under strict temporal filtering, ensuring that systems are tested on their ability to identify insights unavailable in the training period. Kumar et al. (Kumar et al., 2024) develop a multi-disciplinary benchmark across Chemistry, Computer Science, Economics, Medicine, and Physics to test whether LLMs can propose genuinely novel research ideas.

Problem formulation, in turn, is typically evaluated through benchmark-driven tasks. SciQA (Auer et al., 2023) leverages the ORKG, which currently encodes nearly 170,000 resources spanning contributions from over 15,000 papers across more than 700 research fields. SciQAG (Wan et al., 2024) extends this work by automatically generating fine-grained QA datasets. Summarization-focused datasets such as BigPatent (Sharma et al., 2019) and TLDR (Cachola et al., 2020) further support evaluation of synthesis and question formulation tasks. Together, these resources establish a multi-layered evaluation ecosystem, enabling the systematic assessment of both autonomous and collaborative-augmented approaches in knowledge acquisition and problem formulation.

3.2 Idea and Hypothesis Generation

Idea and hypothesis generation constitutes a pivotal stage in the science pipeline, serving as the bridge from accumulated knowledge to novel inquiry. Although often conflated, ideas and hypotheses play distinct roles: ideas represent exploratory inspirations or problem framings that broaden the conceptual search space, whereas hypotheses embody structured, testable propositions that narrow this space into empirically verifiable claims (Nilsen et al.,

[2] https://elicit.com/solutions/search
[3] https://ask.orkg.org/
[4] https://notebooklm.google

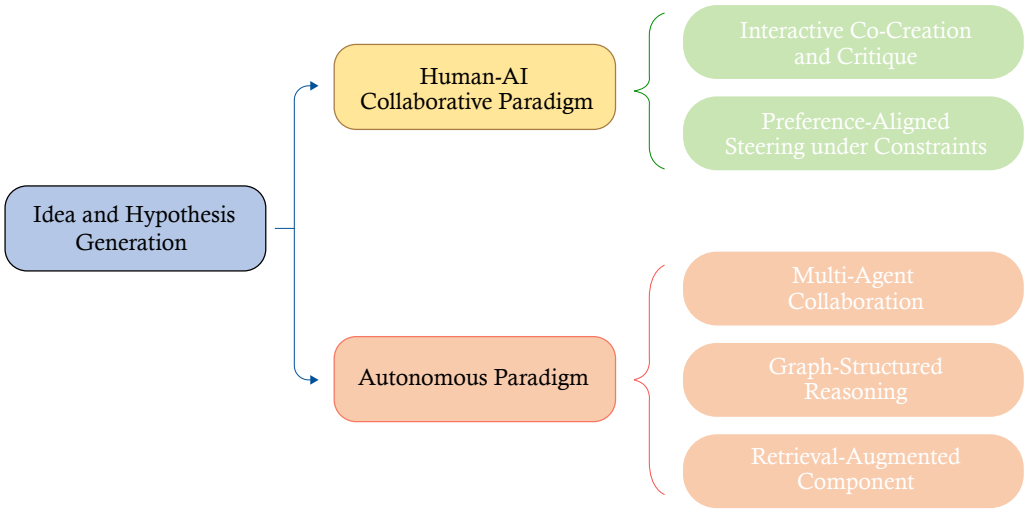

Figure 4: Paradigms of AI-supported idea and hypothesis generation. The humanAI collaborative paradigm focuses on interactive co-creation, critique, and preference-aligned steering under constraints, whereas the autonomous paradigm emphasizes autonomy and scalability through multi-agent collaboration, graph-structured reasoning, and retrieval-augmented component.

2020). This distinction is not merely semantic but methodological, since different forms of AI support are required for open-ended ideation versus hypothesis formalization. Recent developments reveal that technical approaches at this stage have coalesced into two paradigms, each reflecting a different locus of emphasis. The *collaborative* paradigm emphasizes amplifying human expertise and decision quality, embedding multi-round interactive feedback (Baek et al., 2024; Ding et al., 2025), preference and constraint-aware steering (Radensky et al., 2025; Kumbhar et al., 2025; Tong et al., 2024; Lai and Pu, 2025), and co-specification of variables and feasibility with downstream experimental planning (Gottweis et al., 2025), thereby aligning machine creativity with domain judgment. By contrast, the *autonomous* paradigm prioritizes autonomy and scalability, leveraging mechanisms such as LLM-based generation (Wang et al., 2024a; Yang et al., 2024a; Kojima et al., 2022; Si et al., 2024; O'Neill et al., 2025), multi-agent collaboration (Ke et al., 2025), graph-structured reasoning (Ghafarollahi and Buehler, 2025; Sybrandt et al., 2020), search and optimization-based self-refinement (Li et al., 2024b; Hu et al., 2024), and retrieval or knowledge-grounded augmentation to systematically expand and refine the hypothesis space (Youn et al., 2022) with minimal human intervention. Thus, while both paradigms address the same functional goal of idea and hypothesis generation, they diverge in methodological emphasis: collaboration seeks to strengthen human judgment through guided, iterative co-creation, whereas autonomy seeks to complete reasoning chains with minimal external input. These two paradigms and their representative technical components are summarized in Figure 4.

3.2.1 Collaborative Idea and Hypothesis Generation

Within the collaborative paradigm, idea and hypothesis generation is framed as a human-centered activity intended to expand researchers' capacity to navigate vast, fragmented knowledge spaces. These approaches stress tight coupling between domain expertise and machine support: systems assist by retrieving relevant literature and visualizing conceptual structures so that researchers can inspect, judge, and steer the exploration. Early implementations often rely on interactive citation-network browsers, ontology-based recommenders, and visual-analytics interfaces to guide user-driven investigation (Swanson and Smalheiser, 1997; Kingsland III et al., 1982; Smalheiser and Swanson, 1998). Although such tools broaden awareness and reduce cognitive burdens, their practical effectiveness depend heavily on the coverage and granularity of underlying datasets and on users' manual interpretation, which limits scalability across domains. The defining feature of this paradigm is that AI functions primarily as an intelligent assistant, while human judgment remains central to hypothesis evaluation, refinement, and selection.

More recent advances have sought to overcome these limitations by introducing adaptive and more interactive forms of machine support. The advent of large language models has been particularly transformative, as their ability to engage in dialogue-driven exchanges enables forms of collaboration that resemble real-time scientific brainstorming. On this foundation, two collaborative routes have become particularly salient. The first route, which is referred to as *interactive co-creation and critique*, emphasizes multi-turn dialogue that catalyzes idea generation and refinement. In this mode, the model functions as a conversational partner that helps researchers clarify intentions, challenge hidden assumptions, and propose controlled variants. Such exchanges facilitate iterative brainstorming and early-stage hypothesis shaping. By contrast, the second route, *preference-aligned steering under constraints*, directs ideas toward feasibility and rigor. Here, models incorporate explicit critiques, scoring guidelines, or real-world constraints such as ethical or resource considerations. This guidance helps researchers filter ideas, prioritize the most promising ones, and formalize them into testable hypotheses and experimental designs. Taken together, these routes use models not to replace human reasoning but to amplify human agency, expanding the search space, condensing relevant knowledge, and structuring creative exploration with greater efficiency and rigor.

Interactive Co-Creation and Critique. One prominent route leverages LLMs as real-time brainstorming partners. Through multi-turn dialogue, models can clarify researcher intent, surface hidden assumptions, and propose controlled variations of emerging ideas. This interactivity transforms idea generation from one-shot outputs into iterative co-creation, where scientists refine, critique, and reframe hypotheses with model support. By serving as both creative generator and critical interlocutor, such systems enable researchers to systematically explore alternative formulations and converge on more precise, testable hypotheses.

Several recent systems exemplify interactive co-creation in scientific hypothesis generation. HypoChainer (Jiang et al., 2025b) integrates human-in-the-loop visualization, knowledge graph traversal, and LLM suggestions across a three-stage process (exploration, hypothesis chaining, and validation prioritization), enabling experts to iteratively generate interpretable, evidence-grounded hypotheses. IRIS (Interactive Research Ideation System)

(Garikaparthi et al., 2025) combines granular feedback, provenance display, and adaptive exploration pacing to support researchers in refining hypotheses through guided, iterative interactions. IdeaSynth (Pu et al., 2025a) represents research ideas as editable graph nodes and leverages LLMs for literature-grounded suggestions, encouraging users to explore alternative formulations and elaborate ideas more fully. PersonaFlow (Liu et al., 2025a) employs configurable LLM-simulated expert personas from diverse disciplines, allowing researchers to receive structured critiques and multiple perspectives without additional cognitive load. Collectively, these systems illustrate that interactive co-creation transforms idea generation from a one-shot task into an iterative, researcher-driven process, expanding the ideational search space while enhancing the clarity, testability, and evidential grounding of proposed hypotheses.

Preference-Aligned Steering under Constraints. Recent advances illustrate how integrating researcher preferences and real-world constraints can structure hypothesis generation as a collaborative process. Rather than producing ideas in isolation, models are guided by human-defined desiderata, methodological boundaries, or ethical considerations, shaping the generative trajectory. For instance, Scideator (Radensky et al., 2025) combines LLM suggestions with researcher-selected research-paper facets to generate candidate ideas, which are then curated and expanded by humans. In materials science, goal- and constraint-guided agents produce candidate hypotheses that balance novelty with experimental feasibility, allowing researchers to iteratively select and refine promising directions. Automating Psychological Hypothesis Generation with AI (Tong et al., 2024) leverages causal knowledge graphs with LLMs to propose candidate psychological hypotheses, with researchers assessing validity and plausibility. PriM (Lai and Pu, 2025) employs multi-agent collaboration under principle-driven and experimental constraints, generating hypotheses that are both innovative and practically testable. Collectively, these systems show that preference- and constraint-aware steering does not restrict creativity; instead, it channels ideation into scientifically responsible and actionable directions, bridging the gap between imaginative exploration and feasible implementation.

3.2.2 Autonomous Idea and Hypothesis Generation

Within the autonomous paradigm, idea and hypothesis generation is approached with an emphasis on autonomy and scalability, seeking to minimize reliance on human intervention while systematically expanding the search space. Early efforts in this line primarily has leveraged text mining and literature-based discovery, extracting latent associations across scientific corpora to propose candidate ideas (Swanson, 1986). These systems typically cast the problem as link prediction or semantic network exploration (Hristovski et al., 2006; Henry and McInnes, 2017; Frijters et al., 2010), where hypothesis discovery is reformulated into identifying novel connections within graphs or co-occurrence structures. While effective at surfacing overlooked relationships, such methods are inherently constrained by the coverage, granularity, and static nature of their underlying knowledge bases, limiting their ability to support more dynamic and iterative hypothesis refinement.

The advent of large language models has reshaped autonomous idea and hypothesis generation by positioning them as dynamic processes of iterative exploration and refinement, rather

than static, one-shot pipelines. Building on this generative core, several promising technical routes have begun to emerge, each emphasizing different aspects of autonomous reasoning. Notable among them are *multi-agent orchestration*, which enhances diversity through structured role interaction, and *graph-structured reasoning*, which promotes coherence by embedding hypotheses within explicit relational structures. In addition, many systems incorporate auxiliary mechanisms such as retrieval-augmented generation to strengthen factual grounding and contextual support, ensuring that LLM outputs remain connected to external evidence and prior knowledge. While these approaches are still in an exploratory phase, they collectively illustrate how complementary techniques can be layered around LLMs to enhance both the creativity and rigor of autonomous hypothesis generation.

Multi-Agent Collaboration. A prominent line of research focuses on leveraging LLM-based agents configured in multi-role systems, where interaction among agents replaces single-pass generation. Instead of relying on a single model to perform the entire task, these approaches distribute idea and hypothesis generation across complementary roles. For example, some agents propose candidate ideas, others critique or refine them, and a coordinator guides the overall process. Through such iterative exchanges, the system broadens the search space of candidate hypotheses while simultaneously applying logical scrutiny and adaptive refinement. This paradigm mirrors the division of labor in human scientific communities, where dialogue and collaboration enhance both the diversity and the rigor of emerging hypotheses.

Several systems demonstrate how this paradigm can be instantiated. The AI Scientist (Lu et al., 2024) framework organizes agents into a pipeline covering idea generation, experimental design, and manuscript writing, with feedback loops across stages that reduce errors and encourage progressive refinement. VirSci (Su et al., 2024), developed by Shanghai AI Lab, emphasizes disciplinary diversity by simulating virtual researchers who engage in structured discussions, novelty evaluation, and iterative idea refinement; controlled experiments show that factors such as team size, expertise balance, and number of discussion rounds significantly affect the originality of generated outputs. PiFlow (Pu et al., 2025b) treats scientific discovery as a process of entropy reduction, where multiple agents guided by scientific principles (e.g., physical laws) collaboratively propose and refine hypotheses, achieving significant improvements in both efficiency and quality of idea generation. SciAgents (Ghafarollahi and Buehler, 2025) extends this approach to domain-specific discovery tasks in molecular science, coordinating specialized agents to achieve improvements in both novelty and reliability. General-purpose coordination frameworks such as CAMEL (Li et al., 2023) and AutoGen (Wu et al., 2023) further highlight the flexibility of this paradigm, providing reusable mechanisms for role assignment and structured communication (e.g., planner, solver, verifier) that can be adapted across scientific domains.

Collectively, these systems illustrate that multi-agent collaboration is not merely a technical variation but a principled strategy for balancing creativity with rigor. By embedding structured interaction and iterative refinement, they demonstrate that LLM-based research agents can achieve higher levels of diversity, robustness, and scalability in autonomous hypothesis generation.

Graph-Structured Reasoning. Another influential direction in autonomous idea and hypothesis generation is the use of graph-structured reasoning, which embeds scientific concepts and relationships within explicit network representations to guide LLM output. Instead of generating hypotheses in a freeform manner, these approaches situate the generation process within structured knowledge graphs, citation networks, or concept maps, thereby anchoring outputs in coherent relationships and reducing hallucination risk. The underlying methodological framework typically involves constructing a relational scaffoldsuch as nodes representing entities (e.g., genes, materials, phenomena) and edges denoting known interactionsthen prompting or conditioning LLMs to propose hypotheses that extend or bridge existing structures in logical, evidence-supported ways.

For instance, GoAI (Gao et al., 2025) constructs a knowledge graph to organize and relate research concepts, supporting the generation of novel scientific ideas. Similarly, KG-CoI (Xiong et al., 2024) introduces a knowledge-grounded chain-of-ideas framework, in which domain knowledge graphs constrain retrieval and reduce hallucinations during hypothesis formulation. In psychology, causal knowledge graphs built from large corpora have been combined with LLMs to generate hypotheses whose novelty and quality rival those produced by human doctoral scholars. SciAgents (Ghafarollahi and Buehler, 2025) integrates multi-agent orchestration with graph reasoning, where specialized agents traverse and update domain graphs to generate and refine hypotheses in molecular discovery. Other efforts follow a similar paradigm: Interesting Scientific Idea Generation builds personalized subgraphs for researchers to inspire novel ideas (Gu and Krenn, 2025), ResearchLink (Borrego et al., 2025) directly generates hypotheses over scholarly knowledge graphs, and BioDisco (Ke et al., 2025) combines biomedical graphs with literature mining to iteratively refine candidate hypotheses.

Collectively, these systems demonstrate that graph-structured reasoning is not merely a representational choice but a principled strategy for grounding creativity in structure. By embedding hypothesis generation in explicit relational scaffolds, they show that LLM-based systems can achieve higher levels of logical coherence, factual reliability, and domain relevance in autonomous scientific discovery.

Retrieval-Augmented Component. Building on the foundations of multi-agent orchestration and graph-structured reasoning, many recent systems integrate retrieval mechanisms to further improve the grounding and reliability of generated hypotheses. By sourcing relevant literature, datasets, or structured repositories, these components complement the diversity and coherence provided by the previous two approaches, ensuring that candidate hypotheses remain supported by external evidence. A notable example is SciPIP (Wang et al., 2024b), which introduces a retrieval-augmented generation (RAG) approach to enhance the novelty and feasibility of scientific idea generation. Unlike traditional methods that rely solely on keyword-based retrieval, SciPIP constructs a comprehensive literature database supporting semantic and citation-based retrieval. This is complemented by a dual-path idea generation framework that integrates both retrieved content and the internal knowledge of large language models. Through this integration, SciPIP effectively balances the originality and practicality of proposed research ideas, demonstrating its potential as a valuable tool for advancing scientific discovery.

In summary, autonomous approaches to hypothesis generation have evolved from static literature-based discovery to LLM-driven systems that incorporate multi-agent collaboration, graph reasoning, and retrieval augmentation. These methods greatly expand the search space and improve grounding, yet they still face clear limitations: errors can accumulate without human oversight, creativity remains difficult to measure, and context-specific judgment is often lacking.

3.2.3 Evaluation

The evaluation of AI-generated scientific ideas and hypotheses remains fragmented, largely due to the lack of standardized benchmarks. Existing practices draw on a combination of complementary strategies. The most established approach is expert review, where domain researchers assess outputs along dimensions such as novelty, feasibility, usefulness, clarity, and overall quality (Tong et al., 2024; Wang et al., 2024b;a). Alongside human judgments, there is a growing exploration of autonomous or semi-autonomous protocols. A prominent trend is the use of LLMs themselves as evaluators, typically through direct scoring, pairwise comparisons, or meta-review style setups with multiple reviewer agents to reduce variance; some studies further examine the consistency between LLM- and human-based ratings to gauge reliability (Si et al., 2024; Ghafarollahi and Buehler, 2025; Li et al., 2024b). Dedicated frameworks such as GraphEval have also emerged to improve automated idea evaluations, constructing a graph of idea components and leveraging lightweight reasoning or graph-based propagation to provide structured and explainable assessments (Feng et al., 2025a). In addition, several works propose objective bibliometric indicators to mitigate the inherent unreliability of LLM judgments. for instance, Virsci introduces quantifiable measures grounded in historical and contemporary similarity, citation influence, and overall novelty (Su et al., 2024). What's more, some completed science systems contribute further by validating feasibility through executable code or simulation-based testing (Jansen et al., 2025; Yamada et al., 2025; Gottweis et al., 2025).

Building on these evaluation approaches, several benchmark systems have been developed to provide more standardized and reproducible assessments of AI-generated research ideas. Among the most notable are:

- **IdeaBench** (Guo et al., 2025): This benchmark provides a standardized framework for evaluating large language models in scientific idea generation. It comprises both a dataset and an evaluation procedure. The dataset links target papers with their cited references, creating a literature-informed context in which LLMs generate research ideas. The evaluation follows a two-stage process: first, GPT-4o serves as a scalable proxy judge to rank generated and original ideas together according to user-specified criteria such as novelty or feasibility, enabling personalized assessment at scale; second, a normalized *Insight Score* is derived from the relative ranks to quantify model performance. This design supports systematic, fine-grained comparisons across models, while also revealing trade-offs between scalability and potential biases introduced by LLM-based evaluation.

- **AI Idea Bench 2025** (Qiu et al., 2025b): This benchmark presents a quantitative framework for evaluating research ideas generated by large language models within the AI domain. It comprises a curated dataset of 3,495 AI papers published after the cutoff of LLM training data, together with their associated "inspiration" papers, thereby mitigating the risk of knowledge leakage and grounding evaluation in realistic literature contexts. The evaluation methodology assesses idea quality along two complementary dimensions: (1) target-paper alignment, measuring how well generated ideas correspond to the ground-truth content of the target papers; and (2) reference-based evaluation, evaluating innovation and feasibility by comparing against broader reference literature. This dual-faceted design enables rigorous benchmarking of idea generation methods while addressing both fidelity and creative potential in LLM.

- **LiveIdeaBench** (Ruan et al., 2025): This benchmark evaluates LLMs' scientific idea generation under minimal context by using single-keyword prompts. Drawing on Guilford's theory of divergent thinking, it employs a dynamic panel of state-of-the-art models to judge generated ideas across four evaluation dimensions: originality, feasibility, fluency, and flexibility. The authors report extensive experiments with 20 leading models on 1,180 keywords spanning 18 scientific domains, and release a dataset, metrics, and a dynamic leaderboard for comparative evaluation. Results indicate that performance on LiveIdeaBench is not well predicted by conventional general-intelligence benchmarks, suggesting a partial dissociation between creative ideation capacity and standard measures of model intelligence.

Overall, these studies illustrate a multi-faceted yet still unsettled landscape for evaluating AI-generated scientific ideas. Human expert review remains indispensable for capturing contextual validity and nuanced judgment, while LLM-based evaluation offers scalability and flexibility, particularly when combined with structured frameworks. Objective measures provide complementary quantitative insights, and end-to-end systems further ensure feasibility through executable or simulated validation. Benchmarks reflect efforts to standardize and reproduce the evaluation of AI-generated scientific ideas, providing more consistent and comparable assessments across different generation methods and scientific domains. Despite these advances, establishing reliable and widely accepted evaluation protocols remains an open challenge. In particular, these developments underscore the need to carefully leverage LLM-based evaluation methods in ways that ensure reproducibility, reliability, interpretability, and alignment with scientific value.

3.3 Experiment Design and Execution

AI-driven science and innovation is not merely the process of proposing theories or conducting data-driven inference; it fundamentally hinges on how experiments are designed and executed. The design and execution of experiments constitute the critical bridge between abstract hypotheses and empirical validation, serving as the foundation upon which scientific knowledge is built. In this context, the integration of AI has brought about a transformative shift. By embedding machine learning algorithms, robotics, and LLMs into experimental workflows, two major paradigms can be identified in this emerging landscape: *collaborative*

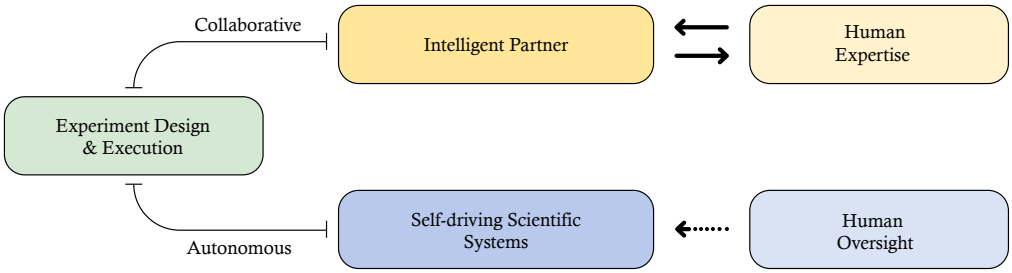

Figure 5: Paradigms of AI-driven experiment design and execution. The collaborative paradigm emphasizes human expertise through intelligent partnership, while the autonomous paradigm leverages self-driving scientific systems with human oversight.

experimenting and *autonomous experimenting*. The conceptual distinction between these two paradigms and their interaction with human expertise is illustrated in Fig. 5.

3.3.1 Collaborative Experiment Design and Execution

The collaborative paradigm centers on the essential role of human expertise in scientific inquiry, where creativity, intuition, and ethical oversight remain critical. Under this setting, AI systems provide suggestions, generate hypotheses, or design candidate experiments, while human researchers contribute domain-specific insights, assess feasibility, and ensure that experiments align with scientific and ethical standards.

HumanAI Co-Planning, Co-Design & Co-Execution. One promising direction in this paradigm is humanAI co-planning. For example, BO-Muse (Gupta et al., 2023) introduces a teaming framework where human experts provide high-level objectives while AI systems propose candidate experimental designs. Through iterative refinement, human and machine converge on effective strategies that balance novelty with feasibility. Human-in-the-loop active learning plays a central role in this process, as AI agents recommend potential experimental conditions while humans validate their relevance and correctness, ensuring that outcomes remain scientifically meaningful.

Beyond general frameworks, domain-specific collaborative systems have also emerged. Cocoa (Feng et al., 2024), for instance, explores the co-planning and co-execution of experimental tasks with AI agents, aiming to create seamless interfaces between autonomous systems and human researchers. In biotechnology, CRISPR-GPT (Huang et al., 2024b) exemplifies how LLM-based systems can augment the design of gene-editing experiments: researchers specify constraints and goals, while the AI proposes detailed experimental protocols and highlights potential outcomes. Such systems illustrate the potential of collaborative augmentation to preserve human creativity and interpretability, while still benefiting from the computational power and generative capacity of AI.

3.3.2 Autonomous Experiment Design and Execution

The autonomous paradigm focuses on constructing self-driving scientific systems that are capable of autonomously planning, executing, and optimizing experiments with minimal human oversight. These systems aim to establish a closed-loop workflow where hypotheses are generated, experiments are performed, results are analyzed, and new hypotheses are refinedall without continuous human intervention. Such a paradigm emphasizes efficiency, scalability, and systematic exploration, enabling scientists to accelerate discovery across vast and complex design spaces.

Self-Driving Laboratories. One active research direction is the development of self-driving laboratories. These platforms combine robotic execution with active learning algorithms to automatically explore experimental conditions and optimize outcomes. For example, AutoSciLab (Desai et al., 2025) presents an interpretable framework for autonomous experimentation, offering a structured approach to balancing exploration and exploitation in experimental design. Similarly, AutoOED (Tian et al., 2021) leverages adaptive search techniques to explore high-dimensional scientific variable spaces and efficiently identify optimal experimental conditions. By closing the loop between hypothesis and execution, these platforms significantly reduce the time and resources required for experimental iteration.

In addition to these general-purpose frameworks, domain-specific autonomous systems have demonstrated profound impact. AlphaFold (Jumper et al., 2021), for instance, has revolutionized protein structure prediction by applying deep learning to a problem that has challenged biologists for decades. Building upon this achievement, AlphaEvolve (Novikov et al., 2025) extends autonomy to protein design, showcasing how AI can not only analyze but also generate novel molecular structures. In materials science and chemistry, robotic laboratories combined with active learning have been applied to the discovery of optimal electrolyte formulations (Noh et al., 2024), while LLM-powered systems such as LLM-RDF (Ruan et al., 2024) provide end-to-end autonomous pipelines for chemical synthesis.

Taken together, the collaborative paradigms and autonomous represent complementary approaches to AI-driven experimental science. Collaborative approaches prioritize human interpretability, creativity, and ethical governance, whereas autonomous systems excel in scalability, reproducibility, and systematic exploration of complex design spaces.

3.3.3 Evaluation

Evaluating AI in experiment design and execution presents unique challenges. Unlike narrowly defined tasks such as classification or summarization, experimental design requires iterative decision-making, integration of diverse domain knowledge, and reasoning under uncertainty. Moreover, the success of an AI system in this context cannot be measured solely by accuracy or efficiency; creativity, adaptability, and alignment with scientific goals are equally essential. As a result, the evaluation of AI for experiment design has given rise to a new generation of benchmarks that attempt to capture these multifaceted requirements.

For instance, BoxingGym (Gandhi et al., 2025) introduces an environment-based benchmark that simulates the iterative nature of experimental reasoning, including hypothesis generation, experimental planning, and refinement. By embedding AI agents within controlled

yet dynamic environments, it allows researchers to evaluate how effectively systems can adaptively propose and test experiments.

At a broader scale, MLR-Bench (Chen et al., 2025c) aims to cover the entire research lifecycle, ranging from idea generation and proposal writing to experiment execution and paper drafting. Its dual evaluation framework, combining automated scoring with human expert review, emphasizes both the feasibility and the originality of AI-generated experimental designs.

Other benchmarks have focused on narrower but foundational aspects. ResearchBench (Liu et al., 2025c), for example, targets the early phases of the experimental pipeline, such as inspiration retrieval, hypothesis construction, and ranking of candidate hypotheses across disciplines. Meanwhile, Scientist-Bench (Tang et al., 2025a) adopts a challenge-based format with both guided and autonomous tasks, assessing whether AI systems can independently propose and justify novel research directions.

Together, these initiatives represent an emerging ecosystem of evaluation tools for AI-driven experimentation. While still in their infancy, they are vital for establishing transparent, systematic, and reproducible standards to assess how AI contributes to the scientific process. Such benchmarks not only provide a basis for comparison but also help identify the strengths and limitations of different AI paradigms, paving the way for future improvements in both collaboration and autonomy.

3.4 Scientific Communication and Presentation

The landscape of scientific communication and presentation is undergoing a profound transformation with the advent of artificial intelligence. From the initial stages of drafting a manuscript to the final dissemination of research findings, AI is introducing new tools and paradigms that promise to enhance efficiency, accuracy, and reach. This report provides a detailed overview of AI's role in this domain, categorizing its applications across key stages of the scientific communication lifecycle and exploring the emerging ethical and collaborative considerations. To better illustrate this multi-stage landscape, Figure 6 presents a systematic framework of AI's role in scientific communication and presentation.

3.4.1 Collaborative Scientific Communication and Presentation

Although a substantial portion of current research centers on fully autonomous systems for scientific writing, visualization, and review, the emerging trajectory of innovation increasingly points toward the humanAI collaborative paradigm. Rather than fully replacing human roles, these systems are designed to enhance creativity, critical reasoning, and accountability through interactive and co-creative engagement at each stage of the communication pipeline.

Manuscript Co-Authoring and Interactive Drafting. AI is increasingly being integrated into the writing process as a collaborative partner. Frameworks such as XtraGPT (Chen et al., 2025a) and ScholarCopilot (Wang et al., 2025d) provide instruction-driven, revision-oriented workflows where human authors remain in control of argumentation and tone, while AI assists with structural reorganization and citation retrieval. Interactive systems like AIRUS (Harris, 2025) allow researchers to co-draft sections under a supervision

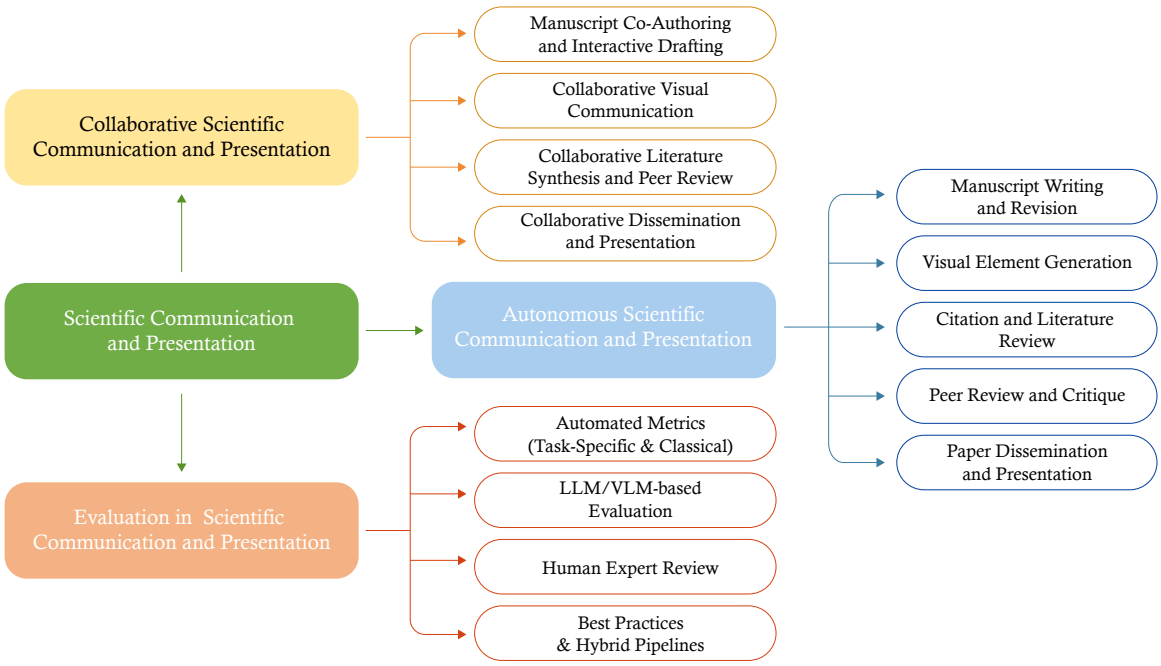

Figure 6: A systematic overview of AI in scientific communication and presentation. The workflow can be viewed from three major perspectives: (1) **HumanAI Collaboration Paradigms**, encompassing interactive assistance, guided collaboration, and cognitive oversight; and (2) **Autonomous Generation and Assistance**, including manuscript writing, visual element generation, citation and literature review, peer review, and dissemination; (3) **Ethical and Social Considerations**, highlighting authorship and integrity, misinformation and bias, over-reliance risks, and responsible use. This framework illustrates how AI reshapes the lifecycle of scientific communication from content creation to collaborative practices and ethical governance.

paradigm, where AI proposes hypotheses and textual components that humans validate or rejectbridging autonomy and authorship integrity.

Collaborative Visual Communication. In figure and graphic generation, the paradigm is shifting from automated rendering to co-design workflows. DesignLab (Yun et al., 2025) and PosterGen (Zhang et al., 2025e) employ iterative feedback loops between AI designers and human reviewers, mirroring professional visual design processes. Such human-in-the-loop design ensures both aesthetic quality and scientific accuracy, while maintaining interpretability and traceability of each revision step. Similarly, Paper2Poster (Pang et al., 2025) incorporates "visual-in-the-loop" interactions, where the user guides layout and content balance through conversational feedback.

Collaborative Literature Synthesis and Peer Review. AI systems are evolving into reflective collaborators in literature analysis and critique. AgentReview (Jin et al., 2024) and DeepReview (Zhu et al., 2025c) simulate reviewer discussions and allow humans to adjust the reasoning process, not merely the results. Such integration of cognitive oversight

enables experts to supervise evidence retrieval, argument structure, and critique generation. In citation and synthesis tasks, models like LitLLMs (Agarwal et al., 2024b) now include guided query refinement and user-verified document selection, embedding accountability within retrieval and reasoning loops.

Collaborative Dissemination and Presentation. The dissemination phasetraditionally human-drivenis now becoming a site of creative co-production. PosterGen (Zhang et al., 2025e) and Preacher (Liu et al., 2025b) exemplify multi-agent collaboration in transforming papers into posters and videos, but crucially, these frameworks incorporate human aesthetic judgments and scientific validation at each iteration. Through dialogue-based refinement and explainable reasoning, AI becomes a co-presenter that supports researchers in communicating ideas effectively across modalities.

From Autonomy to Co-Creation. Across these stages, a clear evolution emergesfrom autonomous autonomy to interactive, guided, and reflective collaboration. This shift embeds ethical and social governance directly within design: authorship accountability is ensured by human supervision; bias and misinformation risks are reduced via transparent reasoning loops; and over-reliance concerns are mitigated by shared cognitive responsibility. In this sense, collaboration is not a separate dimension but a reconfigured mode of autonomy, transforming scientific communication into a humanAI co-creative process.

3.4.2 Autonomous Scientific Communication and Presentation

AI-driven tools are automating various aspects of scientific communication, from generating text and visuals to performing critical reviews.

Manuscript Writing and Revision. The most immediate application of AI in scientific communication and presentation is in text generation and refinement. LLMs are being used to generate initial drafts, brainstorm ideas, and summarize complex concepts. Systems like Grammarly serve as AI writing partners that offer suggestions for polishing entire paragraphs, adjusting tone, and ensuring clarity. AI tools are particularly popular for basic tasks such as grammar checking (22% of usage) and improving readability (51%) (Xu, 2025).

More specialized systems, such as SurveyX (Liang et al., 2025), have been developed to automate the entire process of writing a survey paper, including generating outlines, refining content for fluency, and ensuring citation quality. Another tool, ScholarCopilot (Wang et al., 2025d), is designed to generate professional academic articles with accurate and contextually relevant citations, dynamically retrieving references and integrating them into the text.

Visual Element Generation. AI is revolutionizing the creation of visual assets, which are crucial for effective scientific communication.

- **Figures and Graphics:** Systems are being developed to generate scientific figures directly from text descriptions. For example, FigGen (Rodriguez et al., 2023) is a diffusion model for text-to-figure generation. For vector graphics, AutomaTikZ (Belouadi et al., 2024) and its successor TikZero (Belouadi et al., 2025) generate TikZ code from natural language descriptions, which can then be compiled into vector graphics. This approach reframes vectorization as a code generation task, allowing

models to leverage high-level graphics languages. StarVector (Rodriguez et al., 2025) extends this line of work by generating Scalable Vector Graphics (SVG) code from both images and text prompts, enabling the creation of icons, diagrams, and charts.

- **Captions:** The generation of high-quality figure captions is a critical task. SciCap (Hsu et al., 2021) proposes an end-to-end neural framework to automatically generate informative captions for scientific figures. Multi-LLM Collaborative Caption Generation (MLBCAP) (Kim et al., 2025) introduces a multi-agent framework where several LLMs collaborate to generate and refine captions, integrating both textual and visual information to produce contextually rich results.

- **Infographics and Tables:** AI can create complex infographics and structured tables to simplify data. Infogen (Ghosh et al., 2025) is a two-stage framework that generates complex statistical infographics from text-heavy documents. ArxivDIGESTables (Newman et al., 2024) addresses the challenge of creating literature review tables by automatically generating them from research papers. Tools like Piktochart AI (Piktochart, [n. d.]) can transform documents into visually engaging infographics, reports, and presentations in seconds.

Citation and Literature Review. Automating the labor-intensive process of managing citations and synthesizing literature is a key area of focus for AI. ScholarCopilot (Wang et al., 2025d) is a unified framework that enhances LLMs for academic writing by generating accurate and contextually relevant citations. The system dynamically determines when to retrieve references by generating a special token and then querying a citation database. LitLLMs (Agarwal et al., 2024b) explores using LLMs to assist with writing literature reviews by breaking down the task into two components: retrieving related works and generating the review text. The system uses a hybrid search strategy that combines keyword-based and document-embedding-based search to improve precision and recall. ArxivDIGESTables (Newman et al., 2024) offers a framework for automatically generating literature review tables, which provides a structured way to compare and contrast papers.

Peer Review and Critique. AI is also being developed to assist in the peer review process, providing rapid feedback and simulating review dynamics. OpenReviewer (Idahl and Ahmadi, 2024) is a specialized LLM fine-tuned on a large dataset of expert reviews to generate critical, structured reviews for machine learning and AI papers. It aims to provide authors with constructive feedback before submission to improve their manuscripts. DeepReview (Zhu et al., 2025c) is a multi-stage framework that emulates expert reviewers by incorporating structured analysis, literature retrieval, and evidence-based argumentation to address limitations like hallucinated reasoning and a lack of structured evaluation in existing systems. AgentReview (Jin et al., 2024) is a simulation framework that uses LLM agents to explore peer review dynamics, revealing insights into factors like reviewer biases and social influence on paper decisions.

Paper Dissemination and Presentation. AI is transforming static research papers into dynamic and visually engaging presentation materials. Paper2Poster (Pang et al., 2025) introduces a multi-agent framework that automatically generates academic posters from

full-length papers through a "visual-in-the-loop" process, where a Painter agent creates panels and a Commenter agent iteratively refines them. Building on this, PosterGen (Zhang et al., 2025e) explicitly models the workflow of professional designers with specialized agents for content parsing, layout planning, style design, and rendering, combined with a VLM-based rubric for evaluating aesthetics, producing presentation-ready posters with minimal human refinement. Preacher (Liu et al., 2025b) extends dissemination to video abstracts, employing a top-down summarization strategy and a Progressive Chain of Thought (P-CoT) for fine-grained planning, followed by bottom-up video synthesis to deliver coherent, stylistically rich multimedia presentations. Beyond posters and videos, DesignLab (Yun et al., 2025) automates slide creation through iterative interactions between a design reviewer and contributor LLM, mirroring human design workflows, while SciGA (Kawada et al., 2025) provides a dataset and benchmark for graphical abstract generation, supporting more effective visual communication in academic publishing. EvoPresent (Liu et al., 2025d) introduces a self-improving aesthetic agent for academic presentations, integrating storytelling, design, and delivery through reinforcement learningbased aesthetic feedback. PaperTalker (Zhu et al., 2025b) proposes a multi-agent framework for automatically generating academic presentation videos, combining slide design, narration, and talking-head rendering for coherent and engaging delivery.

3.4.3 Evaluation

The evaluation landscape for AI-driven scientific communication and presentation is currently heterogeneous: researchers combine expert human review, task-specific automatic metrics, and increasingly model-based (LLM/VLM) evaluators to capture different dimensions of quality. Several recent systems exemplify this mixed approach. SurveyX (Liang et al., 2025), for instance, pairs automated metrics, *Citation Recall*, *Citation Precision* and IoU (Insertion-over-Union for insertion/localization accuracy), with human evaluation using the same criteria; IoU quantifies overlap between predicted and true insertion intervals, while Recall and Precision measure retrieval coverage and correctness. XtraGPT (Chen et al., 2025a) introduces the Length-Controlled Win Rate (LC-Win Rate) to mitigate bias from differing response lengths and scales comparisons with an automatic judge (e.g. `alpaca_eval_gpt4_turbo_fn`), which has been reported to reach roughly 68.1% agreement with human annotators in prior work. ScholarCopilot (Wang et al., 2025d) separates evaluation into two primary axes, generation quality and retrieval accuracy, where generation quality is scored on five dimensions (Content Relevance, Logical Coherence, Academic Rigor, Information Completeness, Scholarly Innovation) and GPT-4o is used to rate model outputs against ground-truth texts on a 1,000-paper test set; retrieval accuracy is measured by Recall@k (k=1..10) in a masked-citation protocol: citations and downstream content are masked and the model must predict the citation from the preceding context. ScholarCopilot reports that using only the last sentence before a citation as the query avoids context leak and improves baseline comparability.

For visual and multimodal outputs the community likewise mixes classical metrics and human/VLM judgment. Figure-generation systems such as FigGen, AutomaTikZ, TikZero and StarVector typically report FID, IS, KID and CLIP-similarity to quantify perceptual quality and coarse semantic alignment (Rodriguez et al., 2023; Belouadi et al., 2024; 2025; Rodriguez

et al., 2025), while caption-generation frameworks like MLBCAP rely primarily on human evaluation to assess informativeness and contextual accuracy (Kim et al., 2025). Systems that generate more complex infographics (e.g. Infogen) introduce task-specific, custom metrics to capture statistical and semantic correctness beyond generic image scores (Ghosh et al., 2025). For end-user artifacts such as posters and slide decks, Paper2Poster and PosterGen combine VLM-as-judge rubrics (layout balance, readability, aesthetic coherence) with targeted human review to evaluate both design and semantic fidelity (Pang et al., 2025; Zhang et al., 2025e).

Each evaluation modality has clear strengths and limitations. Automated retrieval and perceptual metrics (Recall@k, FID/IS/KID, CLIP) are reproducible and scalable but do not capture higher-order scientific qualities such as novelty, argument coherence, or the numerical correctness of plotted data. LLM/VLM-based evaluators scale rubric-based scoring and pairwise comparisons but are sensitive to prompt design, model versioning, and internal biaseshence reported modelhuman agreement (e.g., 68.1% for one automatic judge) must be measured and disclosed. Human expert review remains indispensable for assessing contextual validity and scholarly merit but is costly and subject to inter-annotator variability. Practical experience from the cited systems suggests several concrete practices: (1) mask-based citation tests that use only the immediate preceding sentence as query reduce context leakage when measuring Recall@k; (2) explicitly controlling for response length (LC-Win Rate) prevents length-driven wins in pairwise comparisons; (3) combine perceptual image metrics with structured, autonomous checks for plot/axis/data consistency and supplement with VLM + human calibration for layout/design; and (4) always report evaluator prompts, model versions, sampling settings, and humanmodel agreement statistics to enable reproducibility and proper interpretation.

In practice a hybrid, multi-stage evaluation pipeline is recommended: using task-appropriate automatic metrics for large-scale filtering, applying calibrated LLM/VLM judges for scalable comparative scoring (with explicit length-control and prompt disclosure), and retaining targeted human expert adjudication for final quality assessment and failure analysis.

4 Integrated Systems for Science and Innovation

The previous section reviewed how AI techniques support individual stages of the science and innovation pipeline, from knowledge acquisition to scientific communication. In practice, these stages are increasingly assembled into *integrated systems* that connect them within a unified workflow to support the complete exploration of science. Consistent with the earlier discussion, such systems can likewise be viewed through the *human-AI collaborative paradigm* and the *autonomous paradigm*. Figure 7 illustrates how integrated systems for science and innovation can be organized under the two paradigms, together with their representative technical routes.

4.1 Human-AI Collaborative Science and Innovation Systems

Human-AI Collaborative Science and Innovation Systems position AI agents as "co-investigators" or intelligent assistants that support, rather than supplant, human expertise.

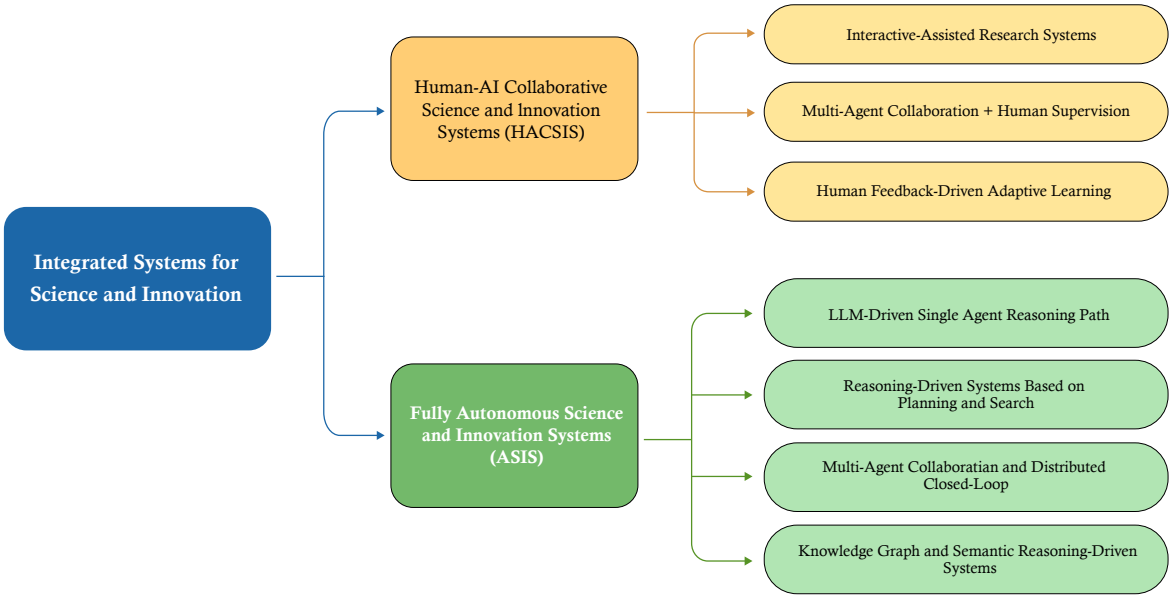

Figure 7: Illustration of the two major paradigms of AI-driven discovery and innovation systems: (1) Human-AI Collaborative Science and Innovation Systems (HACSIS), which emphasize interactive assistance, multi-agent collaboration under human supervision, and human feedbackdriven adaptive learning. This taxonomy highlights the transition from linear, LLM-centered pipelines toward distributed, knowledge-integrated, and human-in-the-loop systems; and (2) Autonomous Science and Innovation Systems (ASIS), which pursue end-to-end autonomy through approaches such as LLM-driven single-agent reasoning, search and planning-based systems, multi-agent collaborative frameworks, and knowledge graph-driven reasoning.

These systems prioritize human supervision and feedback, ensuring that scientific outputs remain not only technically sound but also aligned with ethical, contextual, and domain-specific considerations. In this way, HACSIS embody a vision of symbiotic collaboration, in which human creativity and critical reasoning are combined with AI's computational efficiency and scalability.

Interactive-Assisted Research Systems. The most prevalent form of HACSIS is the interactive assistance model, where humans initiate a research task and provide continuous, incremental guidance throughout the process. In this paradigm, AI functions as an execution engine, capable of automating repetitive or computationally intensive subtasks such as literature summarization, code generation, and data analysis, while humans remain "in the loop" to validate, refine, and redirect outputs. For example, the AIRUS (AI Research Under Supervision) (Harris, 2025) framework demonstrates a lightweight workflow in which researchers use LLMs to generate hypotheses and code snippets, subsequently transferring them into external environments such as Jupyter notebooks for execution. This "cut-and-paste" mode illustrates a minimal-overhead but effective approach to embedding AI within

existing research practices. Similarly, systems such as Agent Laboratory (Schmidgall et al., 2025) adopt a modular co-pilot mode, in which AI agents complete subtasks and present their outputs for human review. Researchers then provide "high-level notes" that inform the agent's subsequent steps, ensuring that each cycle of generation is shaped by human expertise. Cocoa (Feng et al., 2024) introduces interactive plans as a novel design pattern for humanAI collaborative paradigm. By enabling co-planning and co-execution, it allows researchers to flexibly delegate tasks while maintaining steerability, demonstrating effectiveness in real-world scientific research projects. Extending beyond software-based workflows, VISION (Mathur et al., 2025) adapts this paradigm to physical laboratories by providing natural language and voice interfaces for direct humaninstrument interaction, thereby lowering operational barriers and accelerating experimentation. Across these systems, human-in-the-loop feedback functions as a safeguard that both constrains the AI's exploratory tendencies and harnesses its productive capabilities.

Multi-Agent Collaboration with Human Supervision. A more advanced category of HACSIS combines the strengths of multi-agent architectures with explicit human oversight, resulting in hybrid systems where AI agents propose, debate, and refine hypotheses under the guidance of domain experts. In this setup, specialized agents perform distinct roles, e.g., such as idea generation, literature review, or experimental validation, while a human scientist provides strategic direction and evaluative feedback. One representative system, Towards an AI Co-Scientist (Gottweis et al., 2025), employs a "generate, debate, and evolve" paradigm in which multiple agents produce competing hypotheses that enter a tournament-style evaluation process. Human experts are integrated into this loop, contributing their own proposals, supplying domain-specific constraints, and adjudicating among AI-generated alternatives. The result is a tournament-evolution process where human judgment tempers AI-driven exploration. Similarly, systems such as CodeScientist (Jansen et al., 2025) offer hybrid modes that allow humans to intervene in code generation, idea filtering, or evaluation, making the discovery process more efficient and controllable than in fully autonomous execution. In these systems, expert oversight not only increases the robustness of outcomes but also ensures alignment with human-defined goals and ethical standards.

Human Feedback-Driven Adaptive Learning. Beyond direct supervision, some collaborative systems explore the deeper integration of human feedback into the AI's learning process itself. The aim is to align the AI's internal models with human values, preferences, and expertise, leading to more adaptive and natural collaboration over time. For instance, the Deep Cognition (Ye et al., 2025) framework introduces the notion of "cognitive supervision", wherein human researchers intervene strategically at critical junctures of the AI's reasoning. This is facilitated by interaction designs that render the AI's reasoning process transparent and interruptible, allowing humans to shape not only outputs but also the underlying logic. In addition to explicit feedback mechanisms, such as corrective notes or preference rankings, future directions point toward incorporating implicit feedback signals, including patterns of human editing and behavioral cues, to refine the AI's alignment with researcher intent. Another emerging theme is the development of shared mental models between human and AI collaborators. By cultivating a mutual understanding of task objectives, constraints, and context, these systems aim to achieve smoother coordination and anticipate each other's needs. Such alignment is seen as a crucial step toward enabling true

"co-PI" (Prasad et al., 2025) relationships, where human and AI partners engage in iterative cycles of hypothesis generation, evaluation, and refinement as epistemic peers.

Overall, HACSIS highlight the importance of human-centered design in the trajectory of science. By embedding mechanisms for interaction, oversight, and adaptive learning, these systems strike a balance between computational autonomy and human creativity, demonstrating that the most productive path toward scientific discovery may lie in carefully orchestrated human-AI symbiosis rather than in fully autonomous paradigm.

4.2 Autonomous Science and Innovation Systems

Whereas the collaborative paradigm seeks to strengthen human judgment through guided interaction, autonomous scientific systems (ASIS) aim to minimizeor even eliminatethe role of human involvement. Autonomous science and innovation systems are designed to complete the entire scientific research process, from hypothesis generation to experimental execution and even paper writing, with minimal human intervention. These systems usually use large language models as their core, breaking down the research process into a series of autonomous steps. Taking "The AI Scientist" as an example (Lu et al., 2024), the system can generate novel research ideas, write code, perform experiments, generate result graphs, and write entire papers. By continuously iterating this closed-loop process, the system continuously improves the research plan in multiple cycles, simulating the human scientific community's continuous exploration of groundbreaking ideas.

LLM-Driven Single-Agent Reasoning Path. This approach leverages a robust LLM to sequentially perform problem understanding, hypothesis generation, experimental execution, results analysis, and report writing, guided by carefully designed prompt engineering. For example, the Dolphin system generates novel ideas by integrating feedback from prior experiments and employs an *exception backtracking debugging* mechanism to automatically correct generated code (Yuan et al., 2025). NovelSeek further extends this paradigm by maintaining an embedding-based *Idea Bank* across multiple iterations, recording previous ideas to guide subsequent generations and ensure coherent, non-redundant exploration (Team et al., 2025b). Spacer exemplifies the creative potential of single-agent reasoning, autonomously producing factually grounded scientific concepts through extraction and refinement of high-potential keyword sets from large-scale literature (Lee et al., 2025). Complementing these systems, Intern-S1 represents a multimodal, Mixture-of-Experts (MoE) foundation model that integrates general scientific reasoning with domain expertise, enabling complex analysis across diverse scientific datasets and tasks, and achieving state-of-the-art performance in specialized scientific benchmarks (Bai et al., 2025). Collectively, these systems demonstrate the capacity of LLM-driven single-agent paths in ASIS to achieve end-to-end, coherent, and creative scientific discovery and innovation.

- Prompt Engineering and Chained Reasoning: The system guides the LLM through each step of its thinking through staged prompts. Dolphin first sorts relevant literature, then uses the LLM to generate ideas and feeds experimental results back into the prompts.

- Context/Memory Management: To maintain long-term memory, systems like NovelSeek (Team et al., 2025b) use embedding storage to track previously explored ideas, avoiding duplication of effort and enabling the discovery of more novel solutions.
- Code debugging and optimization: To address the instability of LLM-generated code, Dolphin (Yuan et al., 2025) uses backtracking debugging, which allows LLM to analyze compilation error logs and automatically repair the code to improve the success rate of experiments.

Reasoning-Driven Systems Based on Planning and Search. This technical route conceptualizes science and innovation as a large-scale combinatorial search problem, where the AI system must traverse vast hypothesis and solution spaces to identify promising directions. Unlike linear, single-path pipelines, these systems leverage planning algorithms and heuristic search strategies to systematically generate, evaluate, and refine hypotheses. At the core, tree-search methods represent one of the most influential approaches. For example, AI Scientist-v2 (Yamada et al., 2025) introduces a "progressive agentic tree-search methodology" in which agents iteratively formulate hypotheses, design experiments, and draft manuscripts, while an experiment manager guides the expansion and pruning of branches. Similarly, AIDE-ML (Jiang et al., 2025a) employs tree-search to autonomously generate, debug, and benchmark machine learning code, optimizing for user-defined metrics.

Beyond tree structures, other approaches enrich the search-based reasoning process. CodeScientist (Jansen et al., 2025) frames the ideation and experiment construction loop as a form of genetic algorithmdriven exploration, where combinations of research articles and code blocks are recombined and mutated to produce novel hypotheses. This genetic search perspective encourages diverse, cross-domain discoveries, enabling the system to move beyond narrow benchmark optimization toward open-ended scientific creativity.

Meanwhile, probabilistic search strategies such as Monte Carlo Tree Search (MCTS) extend reasoning to particularly complex spaces. DrugMCTS (Yang et al., 2025b) exemplifies this approach by integrating MCTS with a multi-agent collaboration framework for drug repurposing. Through iterative simulation and evaluation of candidate paths, the system identifies high-value therapeutic opportunities, demonstrating how MCTS can scale reasoning over highly uncertain biomedical landscapes. AutoDS presents an open-ended autonomous scientific discovery approach that leverages Bayesian surprise to guide hypothesis exploration. By integrating Monte Carlo Tree Search (MCTS) with progressive widening, the system systematically navigates complex hypothesis spaces, efficiently identifying surprising and novel findings across diverse domains (Agarwal et al., 2025). This exemplifies a reasoning-driven, planning-and-search-based ASIS framework.

Together, these planning- and search-driven systems highlight a reasoning-centric paradigm: AI agents act not only as passive predictors but as active explorers of scientific possibility spaces, systematically balancing breadth of exploration with depth of evaluation.

Multi-Agent Collaboration and Distributed Closed-Loop. Scientific discovery often requires diverse expertise and coordinated effort, a process that can be emulated through multi-agent systems in AI research. Instead of relying on a single monolithic model, multi-agent collaboration distributes tasks across multiple specialized agents with complementary

capabilities, thereby creating a closed-loop research workflow. Each agent assumes a distinct role, such as literature survey, hypothesis generation, experimental execution, or evaluation, and their interactions are orchestrated to achieve coherent progress toward a shared scientific goal.

From a technical perspective, role specialization and orchestration form the foundation of this approach. For instance, A representative example is the DORA AI Scientist (Naumov et al., 2025), which organizes hierarchical teams of domain-specific and generalist agents to perform tasks ranging from hypothesis generation and data analysis to automated report drafting. Operating under user guidance, DORA integrates specialized tools and open data repositories to streamline research workflows while leaving high-level discovery and oversight to human researchers. Moreover, a recent work, PiFlow (Pu et al., 2025b), extends multi-agent systems by framing scientific discovery as a principle-aware uncertainty reduction process. Instead of relying solely on workflow orchestration, PiFlow incorporates scientific laws as guiding constraints, leading to more systematic hypothesisevidence linking and significantly improved discovery efficiency across multiple domains. NovelSeek (Team et al., 2025b) introduces a unified multi-agent framework in which dedicated agents, such as the Survey Agent, Idea Innovation Agent, Code Review Agent, and Assessment Agent, operate under the coordination of an Orchestration Agent to ensure seamless collaboration. To address the increasing complexity of scientific tasks, hierarchical multi-agent architectures have also been proposed. A representative example is AgentOrchestra (Zhang et al., 2025b), which employs a central planning agent to decompose high-level objectives into sub-tasks and delegate them to specialized sub-agents. This hierarchical design improves task completion rate, adaptability, and scalability compared to flat multi-agent baselines. Effective communication and conflict resolution mechanisms further underpin these systems. Although not always explicitly formalized, shared memory structures and standardized communication protocols are commonly employed to synchronize information flow, mitigate conflicts, and maintain workflow consistency. scAgents is a fully autonomous multi-agent framework for end-to-end single-cell perturbation analysis(Tang et al., 2025b). Given only raw data and research objectives, the system autonomously generates model architectures and executable code for training and inference, achieving significant improvements over task-specific baselines. This demonstrates the capabilities of multi-agent collaboration and distributed closed-loop systems in ASIS.

Collectively, these multi-agent frameworks demonstrate how distributed, role-aware collaboration can approximate the structure and efficiency of human research teams, enabling more resilient and adaptive pathways to discovery.

Knowledge Graph and Semantic Reasoning-Driven Systems. Another line of research integrates the reasoning capabilities of large language models with structured representations such as knowledge graphs (KGs), thereby enabling more grounded and semantically coherent scientific discovery. Unlike free-form text generation, knowledge graphs provide explicit entityrelation structures that reduce factual hallucination, improve verifiability, and facilitate the identification of novel, logically consistent connections. By leveraging these structured representations, AI systems can reason more systematically and produce hypotheses that are both innovative and scientifically grounded.

The technical core of this paradigm lies in three complementary components. First, knowledge graph construction is achieved by using LLMs to extract entities and infer relations from diverse scientific corpora, thereby creating and enriching a dynamic, semantic-rich representation of the domain. Second, knowledge-constrained generation guides hypothesis formulation with reference to the constructed graph. For example, the AccelMat (Kumbhar et al., 2025) framework initiates hypothesis generation by coupling an input prompt with a knowledge graph, ensuring that new proposals are grounded in existing knowledge while highlighting unexplored gaps. Finally, chain-of-thought reasoning over structured graphs allows the system to perform transparent, step-by-step inference, linking each generated hypothesis to verifiable evidence in the graph. This integration of symbolic structure and neural reasoning not only enhances interpretability but also promotes the discovery of previously overlooked scientific relationships. THE-Tree constructs domain-specific evolution trees from scientific literature, leveraging LLM-based reasoning to propose, cite, and verify scientific advancements. By combining structured knowledge with semantic reasoning, it enhances hypothesis validation, causal inference, and predictive evaluation, exemplifying knowledge-graph-driven approaches within ASIS (Wang et al., 2025c). Complementing this, Mol-R1 focuses on explicit Long-CoT reasoning for molecular discovery, introducing a distillation-based reasoning dataset and iterative adaptation strategies to enhance explainability and reasoning depth in chemistry-oriented tasks (Li et al., 2025b). Together, these approaches demonstrate how structured representations and advanced reasoning models jointly push ASIS toward more interpretable and scientifically rigorous discoveries.

4.3 Evaluation

The evaluation of AI-driven scientific discovery and innovation systems remains a central yet unsettled challenge, reflecting the difficulty of assessing outputs that must be at once novel, feasible, interpretable, and scientifically valuable. Current practices combine a spectrum of complementary strategies, spanning human expert judgment, autonomous evaluation protocols, and emerging benchmark platforms.

Human-centered evaluation remains the most established approach. Domain experts assess system outputs along dimensions such as novelty, correctness, clarity, and potential impact. This form of evaluation is indispensable for capturing contextual validity and nuanced judgment, but it is limited by scalability and potential subjectivity. For instance, The AI Scientist (Lu et al., 2024) system simulates peer-review by having LLMs generate full research papers that are then scored by reviewer agents mimicking human referees. While efficient, this approach suffers from the well-known limitations of LLM judgment reliability and may fail to capture true scientific value. In a more pragmatic vein, The AI Scientist v2 (Yamada et al., 2025) validates its effectiveness by submitting system-generated manuscripts to real workshops, with acceptance decisions serving as the ultimate external validation. Such evaluations are costly but provide high-stakes evidence of scientific merit.

Autonomous evaluation protocols have been explored to address scalability. A common trend is the use of LLMs as evaluators, either through direct scoring, pairwise comparisons, or multi-agent meta-review to reduce variance. Systems such as DOLPHIN extend this principle by validating their generated methods on standard tasks, showing that the proposed ideas

can indeed yield effective results. NovelSeek (Team et al., 2025b) adopts a similar approach but scales to hundreds of tasks simultaneously, thereby testing robustness across diverse domains. Other systems move beyond evaluation by proxy: AgentOrchestra (Zhang et al., 2025b) benchmarks system performance on tasks with known ground-truth answers, while AI co-scientist (Gottweis et al., 2025) advances a step further by validating generated hypotheses through end-to-end wet-lab experiments in biomedical domains such as drug repurposing, discovery of therapeutic targets, and antimicrobial resistance studies. This rare but rigorous approach ensures feasibility, though it remains prohibitively resource-intensive.

Benchmark-driven evaluation represents a rapidly growing frontier, aiming to standardize and reproduce assessments across systems. A variety of domain-specific and cross-domain benchmarks have been introduced, including Matbench Discovery (Riebesell et al., 2023) for crystal stability prediction, LlaSMol (Yu et al., 2024) for chemistry-focused LLM evaluation, ProteinLMBench (Shen et al., 2024) for protein understanding, MicroVQA (Burgess et al., 2025) for microscopy-based reasoning, and EarthSE (Xu et al., 2025) for Earth science exploration. Similarly, PHYSICS (Feng et al., 2025b) benchmarks university-level problem solving, MLR-Bench (Chen et al., 2025c) and LMR-BENCH (Yan et al., 2025) test AI systems on open-ended machine learning research and reproduction of prior work, while RExBench (Edwards et al., 2025) probes coding agents' ability to autonomously implement research extensions. For multimodal and domain-intensive tasks, MSEarth (Zhao et al., 2025a), AstroVisBench (Joseph et al., 2025), and XLRS-Bench (Wang et al., 2025e) expand evaluation to remote sensing, astronomy, and ultra-high-resolution imagery, respectively. Beyond technical tasks, SciArena (Zhao et al., 2025b) provides an open evaluation platform for scientific literature tasks, while LimitGen-Human collects real human-written limitations to assess whether LLMs can identify critical flaws in research. Collectively, these benchmarks reflect attempts to bring consistency, comparability, and reproducibility into an evaluation landscape that remains fragmented.

Despite this progress, each evaluation approach carries trade-offs. Human review captures depth and contextual grounding but lacks scalability. LLM-based evaluation provides scalability and flexibility but raises concerns of reliability and circularity, especially if the same families of models act as both generators and evaluators. Objective benchmarks offer rigor and reproducibility but risk oversimplification, failing to capture the open-ended creativity central to scientific discovery and innovation. End-to-end experimental validation, while the gold standard, remains resource-demanding and infeasible for large-scale adoption.

Taken together, these practices illustrate a multi-layered evaluation ecosystem. Effective assessment of AI for science discovery and innovation systems will likely require hybrid strategies, integrating human expertise, autonomous evaluation, structured benchmarks, andwhere possibleexperimental validation. The key open challenge is to establish reliable, transparent, and widely accepted protocols that align system performance not only with computational correctness but also with genuine scientific progress.

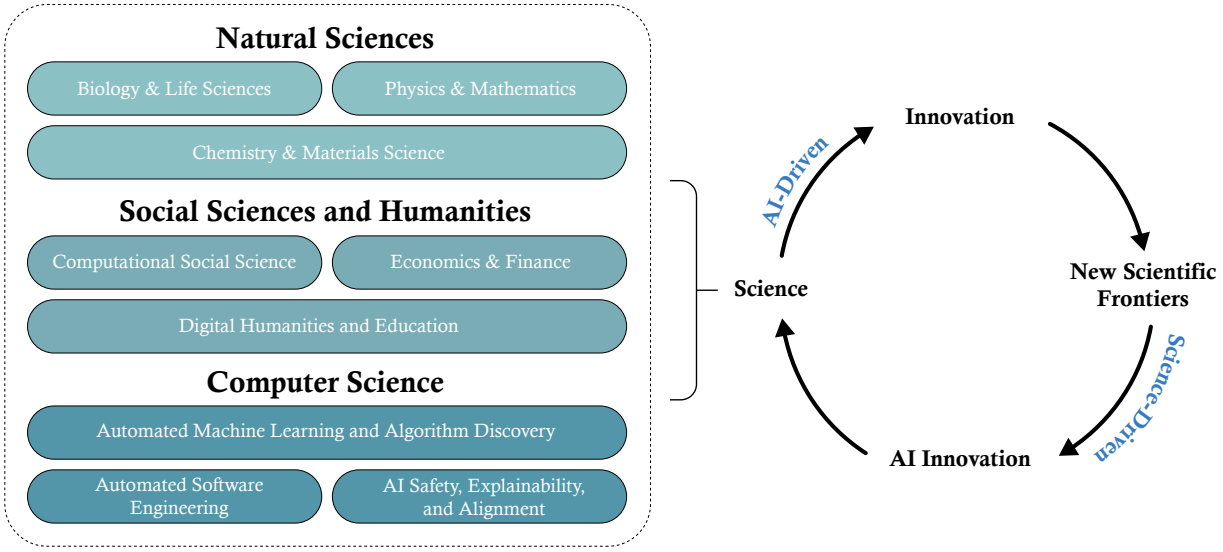

Figure 8: The symbiotic spiral of AI-driven scientific innovation. AI catalyzes breakthroughs across disciplines, while scientific advances provide new insights, data, and materials that, in turn, enhance AI capabilities. This self-reinforcing feedback loop between science, AI, and innovation accelerates discovery and expands the frontiers of human knowledge.

5 AI-Driven Innovations across Scientific Fields

AI has catalyzed an unprecedented era of scientific and technological innovation, fundamentally transforming how we conceptualize, approach, and solve humanity's most complex challenges. Beyond accelerating computation, AI now formulates hypotheses, designs experiments, and reveals structures that extend beyond human intuition. From protein folding (Jumper et al., 2021) to autonomous materials discovery (Szymanski et al., 2023a) and climate modeling (Bi et al., 2023), AI systems are reshaping both the scale and method of research. In this section, we examine how AI drives transformative innovation across three domains: Natural Sciences, Social Sciences and Humanities, and Computer Science. Together, these domains illustrate a transition toward machine-augmented science: an era in which AI acts not just as a tool, but as a creative collaborator in the process of discovery, forming a self-reinforcing scienceinnovation spiral in which advances in AI, scientific discovery, and scientific understanding continuously accelerate one another (Figure 8).

5.1 AI for Natural Science Research

In natural sciences, AI is increasingly reshaping how experiments are conceived and knowledge is generated. AI systems are now applied to automate synthesis in chemistry and materials research, accelerate data-driven discovery in physics, and assist in uncovering compact relationships in complex datasets. For example, autonomous laboratories can independently plan and execute solid-state synthesis experiments (Szymanski et al., 2023a); symbolic regression methods such as AI Feynman recover interpretable physical equations

Table 2: AI innovations transforming natural science research.

Scientific Field	Revolutionary Breakthroughs	Enabling AI Methodologies	Representative Systems and Impact
Biology and Life Sciences	Protein structure prediction; multi-omics integration; autonomous hypothesis generation; explainable single-cell modeling; etc.	Deep learning (Transformers, GNNs); multimodal fusion; generative LLMs; explainable AI.	AlphaFold revolution (Jumper et al., 2021); MIDAS for multimodal omics integration (He et al., 2024); DeepOmics for variant discovery (Heo et al., 2024); Lomics for pathway generation (Wong et al., 2024); scExGraph for tumor microenvironment interpretation (Du et al., 2025).
Chemistry and Materials Science	Autonomous materials discovery; inverse design; retrosynthesis autonomy; closed-loop self-driving laboratories; etc.	GNNs; Bayesian optimization; active learning; LLMs; generative diffusion models.	Autonomous lab frameworks (Tom et al., 2024; Lo et al., 2024); Bayesian optimization for materials design (Chitturi et al., 2024); RetroExplainer for interpretable synthesis planning (Wang et al., 2023a); BatGPT-Chem for LLM-assisted retrosynthesis (Yang et al., 2024c); inverse design via transfer learning (Buterez et al., 2024).
Physics and Mathematics	Physics-informed simulation; autonomous law discovery; neural-symbolic theorem proving; etc.	Physics-Informed Neural Networks (PINNs); Fourier Neural Operators; symbolic regression; transformer-based discovery; neuralsymbolic reasoning.	PINNs for physics-constrained learning (Raissi et al., 2019); operator learning for multi-scale systems (Wong et al., 2025); symbolic regression and AI-Newton rediscovering physical laws (Fang et al., 2025a); AlphaGeometry achieving Olympiad-level reasoning (Trinh et al., 2024).

from data (Udrescu and Tegmark, 2020); and deep-learning-based surrogates for simulation enable efficient modeling of molecular and quantum systems (Noé et al., 2020). These developments illustrate a broader methodological shift: AI is no longer a passive analytical tool but an active collaborator that extends the reach of scientific exploration. Table 2 provides a comprehensive overview of revolutionary AI innovations transforming nature science.

5.1.1 Biology & Life Sciences: Accelerating the Decoding of Life's Mysteries

Life sciences have become a frontier of radical AI-driven innovation, where breakthrough discoveries emerge at an accelerating pace, fundamentally transforming our understanding of life itself (Luo et al., 2024). AI has moved beyond merely accelerating research to becoming a catalyst for innovation. It now enables scientists to explore new biological mechanisms, discover novel therapies, and design life systems with unprecedented precision.

Genomics and Protein Science: Revolutionary Breakthroughs in Life's Fundamental Code. AI contributes to progress in biology and life sciences by supporting tasks such as data integration, mechanism exploration, pathway prediction, and experimental autonomy. For instance, in pan-cancer genomics, models like MIDAS (He et al., 2024) can help integrate multimodal omics data, including genome, transcriptome, and proteome information, to produce unified feature matrices for analysis on large datasets. Approaches using regularized barycentric mapping (Zhu et al., 2025a) may assist in capturing correlations among mutations, gene expression, and protein phosphorylation, potentially aiding detection of mutation sites like those in EGFR for lung cancer. DeepOmics (Heo et al., 2024) supports end-to-end multi-omics integration to facilitate variant discovery in cancer research. Additionally, unsupervised learning methods, such as those in DeepProfile (Qiu et al., 2025a), can help identify transcriptomic signatures across various human cancers.

Mechanistic Exploration and Disease Modeling. In exploring disease mechanisms, multimodal attention fusion networks (Chauhan et al., 2020) integrate data from sources like metabolomics, genomics, imaging, and clinical records to identify potential pathways in conditions such as diabetic nephropathy. For Alzheimer's disease, AI-based analyses (Jiao et al., 2024; Ying et al., 2021; Mishra and Li, 2020) combine multimodal data to extract biomarkers, diagnose conditions, and reveal gene-phenotype associations, including synergistic effects on hippocampal atrophy and tau protein levels. In lung cancer research, multi-omics graph models such as MOLUNGN integrate mRNA, miRNA, methylation, and clinical data into a Graph Attention Network (GAT) framework to infer stage-specific progression biomarkers (Zhang et al., 2025a).

Pathway Design and Experimental Autonomy. For proposing biosynthetic or metabolic pathways, tools like Lomics leverage fine-tuned large language models to generate custom pathway and gene set hypotheses from omics data, reducing the search space of candidate routes (Wong et al., 2024). On the experimental side, explainable graph neural networks such as scExGraph have been developed to interpret tumor microenvironment components from single-cell sequencing data, enabling feedback and quality control in protocol design for pathology and cell-level assays (Du et al., 2025).

5.1.2 Chemistry & Materials Science: Intelligently Creating New Matter

Chemistry and materials science are undergoing a paradigm shift in which AI navigates vast chemical and structural spaces to propose molecules and materials with previously unattainable properties (Cheng et al., 2025; Tom et al., 2024). These developments have given rise to autonomous research ecosystems, self-driving laboratories (SDLs) and AI chemists, that

tightly couple robotics, high-throughput experimentation, and machine learning to compress the designsynthesischaracterization cycle.

Autonomous Materials Discovery and Design. Inverse-design and generative approaches now allow researchers to specify target properties and have models propose candidate structures that satisfy those constraints, effectively inverting the traditional search paradigm (Buterez et al., 2024). Closed-loop SDLs implement this pattern in the lab by autonomously proposing experiments, executing them, and iterating on model predictions; recent reviews and demonstrations show these systems dramatically accelerate discovery while improving data quality and reproducibility (Tom et al., 2024; Lo et al., 2024). Bayesian optimization and active learning strategies remain central to efficient exploration of high-dimensional composition and process spaces (Chitturi et al., 2024).

Autonomous Chemical Synthesis and Reaction Planning. Retrosynthesis and route planning have been transformed by data-driven neural methods: interpretable assembly-based models, iterative string-editing approaches, and LLM-augmented retrosynthesis systems now generate plausible, multi-step synthetic routes and rank alternatives for experimental validation (Wang et al., 2023a; Han et al., 2024; Yang et al., 2024c). Integrating these planners with robotic execution and closed-loop analytics is making end-to-end autonomous synthesis increasingly feasible in both medicinal chemistry and materials synthesis workflows (Tom et al., 2024).

5.1.3 Physics & Mathematics: Unveiling Fundamental Laws and Abstract Structures

Physics and mathematics are experiencing a substantive shift as AI transitions from computational assistant to active theorist, autonomously proposing hypotheses and aiding in proof construction (Trinh et al., 2024; Makke and Chawla, 2024). Modern AI systems are increasingly capable of uncovering fundamental physical relationships and mathematical structures that have been difficult to detect through conventional analysis.

Simulation of Complex Physical Systems. Simulating nonlinear dynamical phenomena, from turbulent fluids to interacting quantum systems, remains computationally intensive and algorithmically challenging. Physics-Informed Neural Networks (PINNs) address this challenge by embedding governing differential equations directly into the network loss, thereby biasing learning toward physically consistent solutions while fitting observational data (Raissi et al., 2019; Toscano et al., 2024). Recent reviews and comparative studies highlight issues of PINN scalability, optimization difficulties for stiff systems, and multi-scale generalization, and propose hybrid numericalML strategies and operator-learning (e.g., Fourier Neural Operators) as promising directions to improve stability and efficiency (Wong et al., 2025).

Autonomous Discovery of Physical Laws and Mathematical Conjectures. Symbolic regression, sparse identification (SINDy) and modern transformer-based approaches have advanced the autonomous discovery of governing equations from data, enabling the recovery of interpretable symbolic laws under noise and limited sampling (Makke and Chawla, 2024; Mower and Bou-Ammar, 2025). Recent systems combine foundation models, inductive priors, and symmetry constraints to produce more robust and physically meaningful

candidates for differential equations and conservation laws (Mower and Bou-Ammar, 2025; Yang et al., 2025a). As a concrete demonstration, the AI-Newton system has been shown to rediscover Newtonian mechanics laws (including Newton's second law and energy conservation) from noisy experiments without explicit prior physics knowledge (Fang et al., 2025a). In mathematics, neuralsymbolic systems such as AlphaGeometry (and follow-on systems) have achieved near-human (Olympiad-level) performance on challenging geometry problems by combining massive synthetic theorem datasets, neural search, and symbolic proof modules (Trinh et al., 2024).

These developments point to an emergent research ecology in which data-driven discovery tools accelerate hypothesis generation and formal reasoning, while domain experts continue to validate, formalize, and place new results into scientific and mathematical context.

5.2 AI for Social Sciences and Humanities

Artificial intelligence is transforming how scholars investigate societies, cultures, and histories by shifting research from mainly descriptive accounts toward computationally grounded explanations (Xu et al., 2024). This transformation unfolds along three overlapping paradigms: (1) computational modeling, which captures system-level social dynamics and emergent patterns at scale (Gurcan, 2024); (2) interpretive augmentation, where AI tools structure, extend, and amplify humanistic close reading and qualitative interpretation (Hitch, 2024); and (3) cognitive simulation, in which agentic and cognitive models reproduce emergent social behaviors to test hypotheses and develop theory (Jiang et al., 2024). Compared with traditional statistical or ethnographic methods, contemporary AI approaches make it feasible to extract mechanisms from multimodal, unstructured sources, such as text, images, and interaction traces, and to run high-resolution simulations that probe causal processes and generative explanations (Chapinal-Heras and Díaz-Sánchez, 2023).

5.2.1 Computational Social Science: Modeling Society at Scale

Computational social science has evolved into a data-intensive and theory-driven field that integrates large-scale behavioral datasets with computational modeling to uncover the mechanisms of social structure and change. As noted by Xu et al. (Xu et al., 2024), AI-driven analytics now enable researchers to bridge the gap between micro-level actions and macro-level social patterns.The integration of diverse data streams through graph neural networks and NLP models provides unprecedented analytical capabilities for understanding social complexity.

Analyzing Public Opinion and Social Dynamics. Recent advances in natural language processing, such as transformer-based sentiment and opinion modeling frameworks (Jahin et al., 2024), have allowed scholars to examine how opinions evolve and narratives spread through online ecosystems.For example, Ghafouri et al. (Ghafouri et al., 2024) uses Transformer models to measure how tightly users cluster around homogeneous views and how discourse diversity shrinks within these clusters. This method demonstrates how AI can generate analytical metrics of echo chambers, shifting our understanding of how discourse polarization emerges in large-scale digital publics.

Agent-Based Social Simulation. As Gürcan (Gurcan, 2024) argues, agent-based models (ABMs) are essential for exploring how complex societal patterns emerge from individual interactions. The recent integration of large language models into ABMs , as shown by Gao et al. (Gao et al., 2024) , endows simulated agents with human-like reasoning, communication, and adaptive behavior. These LLM-augmented simulations make it possible to test theories about norm emergence, cooperation, and cultural evolution in silico, providing a new methodological bridge between qualitative theorizing and quantitative experimentation.

Network Science and Community Detection. Human societies can also be represented as relational graphs. Graph Neural Networks (GNNs), as increasingly explored in methodological studies (Leeney and McConville, 2023; Liu et al., 2024b), have become central to the analysis of relational structures. GNN-based approaches enable community detection, link prediction, and influencer identification in large-scale datasets, from citation networks to social media graphs, revealing latent subcommunities that shape social cohesion, information diffusion, and group identity.

Together, these developments illustrate a convergence of data analytics, agentic simulation, and network science, a methodological synthesis that redefines how social scientists study human collectives at computational scale.

5.2.2 Economics and Finance: Discovering Patterns in Markets

Economics and finance have long been data-intensive fields, and AI is providing new tools for modeling market behavior, managing risk, and informing economic policy (Xu et al., 2024).

Algorithmic Trading and Market Prediction. Financial markets generate vast streams of time-series data, and recent work shows that hybrid deep learning models combining LSTM, Transformer architectures, or modified ensembles can capture non-linear dynamics and improve forecasting accuracy (Oukhouya et al., 2025). For instance, "Higher Order Transformers" exploit multimodal time-series data (including prices and social media signals) to enhance stock movement prediction (Omranpour et al., 2024), while "Cross-Modal Temporal Fusion" demonstrates improved forecasts by integrating macroeconomic indicators with news and price data (Pei et al., 2025).

Risk Management and Credit Scoring. Accurately assessing financial risk has benefited greatly from AI: machine learning and non-traditional data sources outperform traditional models during periods of stress in credit scoring applications in China (Gambacorta et al., 2024). Moreover, a novel model "LLM-FP-CatBoost" integrates large language models to process narrative data and address class imbalance in fintech lending, improving default prediction performance (Xia et al., 2025).

Economic Modeling and Policy Analysis. AI is also being applied to macroeconomic growth modeling; for example, empirical work using ARDL and non-linear ARDL models finds that AI (as measured via various proxies) has a positive long-term impact on economic growth in European countries (Kalai et al., 2024). These methods allow economists to simulate the effects of different government policies, sectoral shocks, or technology adoption in ways that go beyond simplified theoretical models, producing more empirically-grounded and policy-relevant insights.

5.2.3 Digital Humanities and Education: Quantifying Culture and Learning

Digital humanities combine computation with traditional humanities disciplines, while AI in education is increasingly shaping personalized, adaptive, and inclusive learning environments (Peláez-Sánchez et al., 2024; Yan et al., 2023).

Quantitative Analysis of Culture and History. Recent advances in remote sensing and computer vision enable scholars to detect archaeological patterns from aerial and satellite imagery at scale; for example, the zero-shot experiments with visual foundation models have demonstrated performance comparable to specialized methods in spotting archaeological features without fine-tuning (Landauer and Klassen, 2025). In addition, large open datasets such as Enki support supervised learning for automatic recognition of ancient settlement tells and invisible landscape features across geographies (Rosati et al., 2025).

AI in Education. Large language models are being rigorously evaluated for their effects on student learning outcomes: a recent meta-analysis of over one hundred experimental studies shows that LLM-assisted tutoring yields strong improvements in qualification outcomes, though effects on socialisation and subjectification are more variable depending on intervention design (Huang et al., 2025a). Moreover, multilingual LLM-based tutoring has been shown to significantly enhance learning when feedback is given in the student's native language, especially for low-resource languages in mathematics contexts (Tonga et al., 2025). Systematic reviews also highlight practical and ethical challenges, including issues of reproducibility, bias, and academic integrity, that must be addressed to ensure the responsible use of AI in humanities and education (Yan et al., 2023).

5.3 AI for Computer Science & AI: A Reflexive Turn

The most revolutionary and philosophically profound innovation in AI lies in its capacity for self-improvement: systems that can design better systems, thus generating a feedback loop of accelerating enhancement and reshaping not only computer science but the nature of intelligence itself. This reflexive transformation can be conceptualized as a multilayered ecosystem, in which AI not only develops new algorithms and architectures but also supports its own research, deployment, and governance. Figure 9 illustrates this self-improving cycle, spanning technical foundations, applied systems, safety mechanisms, and societal feedback.

5.3.1 Autonomous Machine Learning and Algorithm Discovery: AI Designing AI

A major research frontier in AI involves automating architecture and algorithm design. AutoML seeks to partially offload the manual burden of model selection and hyperparameter tuning by creating systems that assist, and eventually supersede, human designers (He et al., 2021; Baratchi et al., 2024). This self-improving cycle represents a paradigm shift in how AI systems are developed and optimized.

Neural Architecture Search (NAS). Neural Architecture Search explores vast structural search spaces to identify architectures that outperform or match human-crafted designs under resource or performance constraints. Modern surveys emphasize advances such as weight-sharing, differentiable NAS, and transformer-aware architecture search (Baratchi et al., 2024). For example, recent work in contrastive architecture comparators and zero-

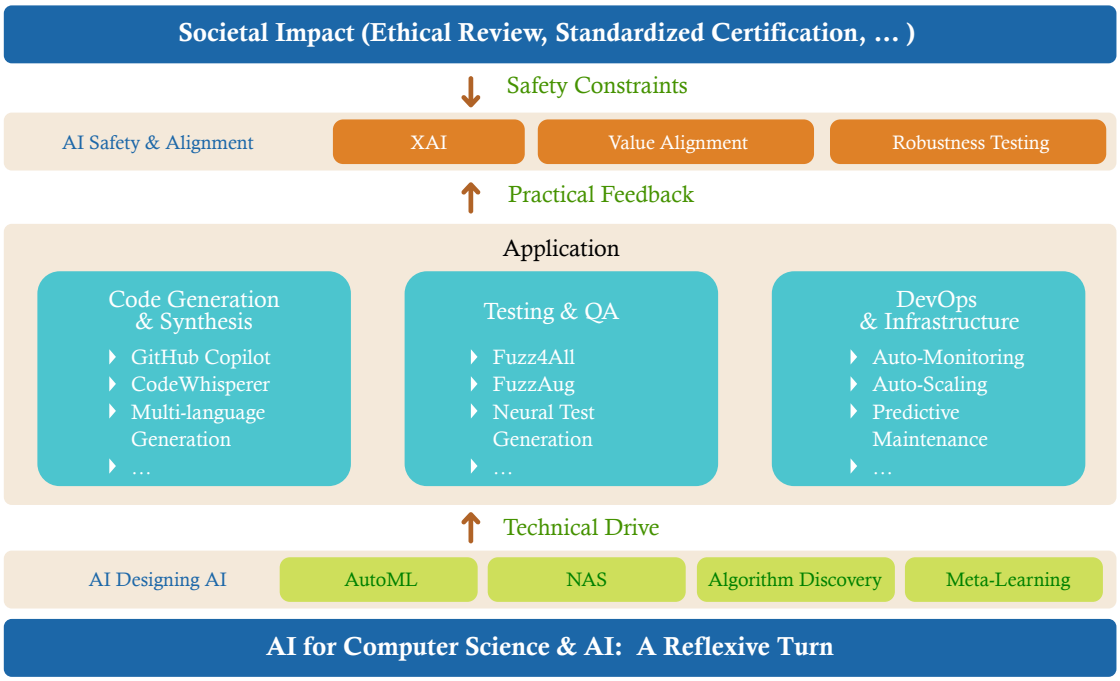

Figure 9: The Reflexive Ecosystem of AI in Computer Science. AI systems increasingly participate in a self-reinforcing cycle of advancement through autonomous design, practical applications, and feedback integrating safety, alignment, and societal governance.

shot performance prediction helps NAS avoid costly full evaluations of candidate architectures (Wang and Zhu, 2024).

Autonomous Algorithm Discovery. Beyond architecture, AI systems are now discovering entirely new algorithms. DeepMind's AlphaDev is a flagship example: by formulating sorting as a game and using deep reinforcement learning, AlphaDev discovered sorting and hashing routines that outperform prior human designs, for instance, new hashing functions (Mankowitz et al., 2023). This kind of work shows the power of letting AI systems push the boundaries of foundational computation.

Meta-Learning and Few-Shot Adaptation. To make AutoML systems more flexible, meta-learning and few-shot adaptation techniques aim to let models generalize rapidly to new tasks with little data. But real-world challenges such as task distribution shift, unstable fine-tuning, and overfitting remain significant hurdles. Recent surveys of meta-learning methods indicate that while many methods excel on benchmark suites, their robustness in open, heterogeneous domains is less understood (Peng, 2020). Some works now frame algorithm selection and parameterization as meta-learning problems in their own right, treating AutoML as algorithm selection under uncertainty (Garouani, 2025).

5.3.2 Autonomous Software Engineering: Transforming Code Creation and Maintenance

The principles of AI are being broadly applied across the entire software development lifecycle, from initial specification through deployment and ongoing maintenance (Müller et al., 2022; Ahmed et al., 2025).

Intelligent Code Generation and Synthesis. Platforms like GitHub Copilot, Amazon CodeWhisperer, and advanced LLM-based code models are reshaping how software is created: developers increasingly describe specifications in natural or formal language, and AI synthesizes multi-language, multi-framework implementations that often require minimal revision (Wang et al., 2025a). Recent work shows that combining AI code generation with verification strategies (e.g. consistency checking or context-aware retriever models) enhances correctness compared to raw generation alone (Keshri et al., 2025).

Autonomous Testing and Quality Assurance. AI is pushing forward test autonomy via novel techniques like LLM-powered fuzzing and neural test generation. For example, the Fuzz4All system uses LLMs to generate diverse inputs across many languages and has uncovered real bugs in compilers and runtimes (Xia et al., 2024). Another line of work, Data Augmentation by Fuzzing (FuzzAug), combines fuzzing and neural test generation to expand training data and improve branch coverage in semantic test generation models (He et al., 2025).

Intelligent Bug Detection and Repair. Recent advances in automatic program repair (APR) show that retrieval-augmented patch generation systems such as RAP-Gen achieve higher repair rates on benchmarks like Defects4J and TFix by leveraging external bug-fix repositories alongside generative models (Wang et al., 2023b). Another approach, InferFix, combines transformer-based models with static analyzers to repair both security and performance bugs, achieving strong top-1 fix accuracy on Java and `C#`repositories (Jin et al., 2023).

DevOps and Infrastructure Autonomy. On deployment and operations, research such as Enhancing Code Consistency in AI Research demonstrates how retrieval-augmented models can verify alignment between algorithm descriptions and code implementations, a step toward autonomous monitoring and anomaly detection in pipelines (Keshri et al., 2025). Emerging systems also incorporate predictive analytics and auto-scaling policies guided by usage traces to proactively maintain performance.

5.3.3 AI Safety, Explainability, and Alignment: Ensuring Responsible AI Development

The quest to ensure safe and beneficial AI has spurred rapid methodological and governance advances; contemporary safety research now spans empirical robustness testing, formal verification, and socio-technical evaluation frameworks that together seek to constrain harms produced by increasingly capable systems (Makin, 2024; Fang et al., 2025b). In a reflexive turn, AI systems themselves are being used to generate safety artifacts, from synthetic preference data to autonomous adversarial testing, creating meta-cognitive pipelines that both stress-test and improve model behaviour (Mei et al., 2023).

AI Safety and Robustness. Research in robustness has moved beyond single-shot adversarial examples toward continuous, distributional stress-testing and antifragile perspectives that treat model evaluation as an evolving process rather than a fixed benchmark (Jin and Lee, 2025; Makin, 2024). Work on containment, capability control, and formal guarantees is proceeding in parallel with empirical red-team evaluations, reflecting the multi-layered nature of modern safety engineering (Fang et al., 2025b).

Explainable AI and Interpretability. Explainable AI (XAI) methods, from local attribution techniques such as LIME and SHAP to concept-level and causal explanation tools, are now judged not only by fidelity metrics but by their utility for stakeholders (developers, auditors, and affected users) and their robustness to model choice and feature collinearity (Salih et al., 2025; Mersha et al., 2024). Recent work also explores how large language models can act as explanation mediators, translating opaque model reasoning into human-readable rationales while introducing new evaluation challenges for faithfulness and trustworthiness (Bilal et al., 2025).

AI Alignment and Value Learning. Alignment research concentrates on methods that align model behavior with human values, notably reinforcement learning from human feedback (RLHF) and variants such as reinforcement learning from AI feedback (RLAIF) and constitutional AI; however, recent critical analyses emphasize conceptual limits and sociotechnical trade-offs in these approaches, cautioning that RLHF alone is not a panacea for value alignment (Lindström et al., 2024; Lambert, 2025). Complementary lines of work study value learning, scalable oversight, and incentive-aware training regimes intended to close the gap between short-term behaviour shaping and long-term value-sensitive deployment (Fang et al., 2025b).

Governance and Evaluation Frameworks. Finally, there is growing convergence on the need for interoperable governance structures, standardized benchmarks for safety, independent model audits, and certification pathways, coupled with socio-technical impact assessment to ensure that deployment choices align with public interest and regulatory expectations (Makin, 2024).

6 Platforms and Toolchains

The ecosystem of platforms and toolchains supports the modern, AI-driven scientific workflow. To systematically elucidate the role of these tools throughout the research lifecycle, discussion follows the intrinsic logical sequence of scientific inquiry, which is visually summarized in Figure 10: from the initial phase of knowledge acquisition and hypothesis generation, through the intermediate stages of experimental design, execution, and data analysis, to the final communication and presentation of scientific findings. This section aims to present readers with a clear technological landscape of science and innovation. The tools and platforms for the four research areas are summarized in Table 3.

Table 3: Summary of tools and platforms

Stages	**Category**	**Tools/Platforms**
Knowledge Acquisition and Problem Formulation	Comprehensive Search and Intelligent Analysis	Elicit, Perplexity, Consensus, Semantic Scholar, Scite.ai, ScholarGPS, Iris.ai
	Citation Network Visualization and Knowledge Graph Construction	Connected Papers, Research Rabbit, Litmaps, CiteSpace, Sci2 Tool
	Reference Management and Personal Knowledge Systems	Zotero, Mendeley, EndNote, Obsidian, Roam Research
	Deep Reading and Interactive Inquiry	NotebookLM, ChatPDF, Enago Read, PaperQA2, Scholarcy
Idea and Hypothesis Generation	Dialogue-Based Brainstorming with Generative AI	ChatGPT, Claude, DeepSeek, Elicit, ResearchKick, SciPIP
	Connecting Knowledge to Spark Innovation	Obsidian, Roam Research
	Visual and Collaborative Concept Structuring	Whimsical AI, Miro AI, NotebookLM, Research Flow
Experiment Design and Execution	Standardized Design and Sharing of Experimental Protocols	Protocols.io, Synthace, Strateos, Weights & Biases, Llama Factory
	Electronic Lab Notebooks and Data Provenance	LabArchives, SciNote, eLabFTW
	Data Processing, Computational Execution, and Statistical Analysis	RStudio, Jupyter Notebooks, Google Colab, Snakemake, Nextflow, GraphPad Prism, Origin, MATLAB, etc.
	Access to Key Resources and Virtual Simulation	Open Science, GitHub, GitLab, Zenodo, Figshare, Dryad, Consensus, ScholarGPS, Elicit, Labster, etc.
Scientific Communication and Presentation	The Modern Manuscript Workflow	Overleaf, Google Docs, ChatGPT, Claude, DeepSeek, Scispace, HyperWrite, OmniThink
	Language Polishing and Stylistic Proofreading	Grammarly, DeepL Write, Wordtune, Quillbot, Writefull
	Citation Management and Formatting Standards	Zotero, Mendeley, EndNote, Overleaf, Cite This For Me, MassiveRef
	Scientific Illustration and Data Visualization	Matplotlib, GraphPad Prism, Origin, Tableau, BioRender, Inkscape, CiteSpace, Vizcom, Microsoft Copilot, etc.

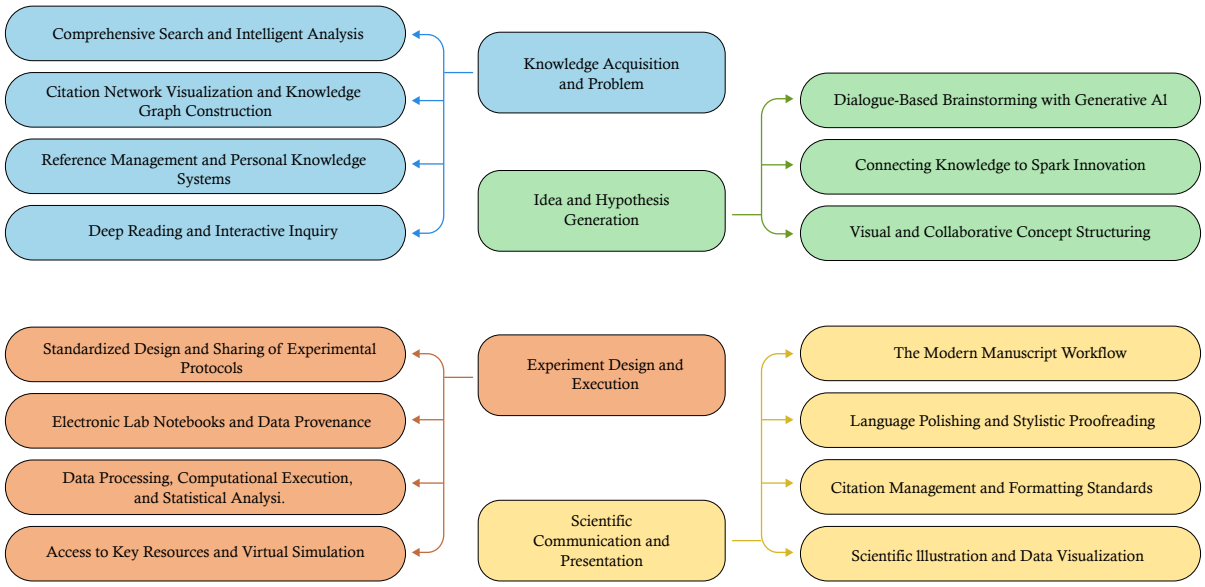

Figure 10: The Ecosystem of Tools and Platforms for the Modern Research Lifecycle.

6.1 Knowledge Acquisition and Problem Formulation

Comprehensive Search and Intelligent Analysis. A paradigm shift is underway in academic research, as traditional keyword-based search methodologies are progressively being superseded by more intelligent and responsive frameworks. The new generation of integrated search and analysis platforms, powered by the fusion of Natural Language Processing and Large Language Models, has facilitated a crucial leap from mere information retrieval to sophisticated knowledge synthesis a transition from search to solution. Platforms such as Elicit [5], Perplexity [6], and Consensus [7], empower researchers to pose complex questions in natural language. Fundamentally, these platforms leverage techniques like Retrieval-Augmented Generation, combining the vast knowledge encoded in LLMs with real-time, targeted information retrieval from scientific literature databases to ensure accuracy and traceability. In response, these tools distill and summarize pertinent findings directly from a vast corpus of literature, delivering answers that are cogently synthesized and meticulously cited.

Building on this foundation, another class of tools has emerged, offering deeper meta-analytical capabilities. Semantic Scholar [8], for example, employs AI to auto-generate concise Too Long; Didn't Read summaries and extract key figures from publications, significantly reducing the cognitive load associated with preliminary literature screening. In a pioneer-

[5] https://elicit.org
[6] https://www.perplexity.ai/
[7] https://consensus.app
[8] https://www.semanticscholar.org

ing move, Scite.ai [9] (Nicholson et al., 2021) analyzes the context of citations, classifying them as supporting, contrasting, or merely mentioning. This functionality provides a critical basis for assessing the academic impact and points of contention surrounding a specific study. Complementing these are platforms like ScholarGPS [10], which focuses on macro-level analyses of researcher profiles and institutional landscapes, and Iris.ai [11], which offers a semi-autonomous solution for conducting large-scale systematic reviews. Together, these tools constitute a multidimensional ecosystem for intelligent literature analysis.

Citation Network Visualization and Knowledge Graph Construction. To truly comprehend the architecture and evolution of a research field, scholars must move beyond the confines of individual publications to discern the underlying intellectual currents and connections. Visualization tools for citation networks provide powerful support for this endeavor. Platforms such as Connected Papers [12], Research Rabbit [13], and Litmaps [14] automatically generate graphical maps of a given paper's citation network, intuitively displaying its scholarly lineage. This helps researchers to rapidly identify foundational works, pivotal connecting literature, and emerging research frontiers.

For more specialized applications in scientometrics and knowledge graph analysis, sophisticated software like CiteSpace [15] and Sci2 Tool [16] offers a suite of more complex and powerful functionalities. These programs can perform co-occurrence, clustering, and temporal analyses on extensive datasets of scholarly literature. By doing so, they quantitatively unveil the knowledge structure, evolutionary pathways, and future trends of a given domain, providing a data-driven cartography of the field.

Reference Management and Personal Knowledge Systems. The efficient acquisition of knowledge is inextricably linked to the systematic management of literature resources. Reference management software, epitomized by tools like Zotero [17], Mendeley [18], and EndNote [19], has become a standard component of the modern research workflow. By enabling the seamless capture, storage, organization, and cross-device synchronization of bibliographic information, as well as autonomous in-text citation and bibliography generation, these programs form the cornerstone of a researcher's personal academic library.

More recently, driven by the increasing demand for deeper knowledge internalization and innovation, Personal Knowledge Management tools based on bi-directional linking, such as Obsidian [20] and Roam Research [21], have garnered significant attention within the academic community. The core advantage of these platforms lies in their ability to transcend the

[9] https://scite.ai
[10] https://scholargps.com/
[11] https://iris.ai
[12] https://www.connectedpapers.com
[13] https://www.researchrabbit.ai
[14] https://www.litmaps.com
[15] http://cluster.cis.drexel.edu/~cchen/citespace/
[16] https://sci2.cns.iu.edu/user/index.php
[17] https://www.zotero.org
[18] https://www.mendeley.com
[19] https://endnote.com
[20] https://obsidian.md/
[21] https://roamresearch.com/

limitations of traditional, linear note-taking. By establishing a networked web of connections between discrete knowledge nodes, they create an ideal environment for researchers to deeply integrate and cross-pollinate ideas from disparate sources, thereby fostering the generation of novel insights and the construction of a unique, personalized knowledge architecture.

Deep Reading and Interactive Inquiry. Addressing the challenge posed by the sheer information density of individual research articles, a new suite of AI-driven deep-reading tools has emerged to enhance the efficiency of textual comprehension. Platforms such as Scholarcy [22], NotebookLM [23], ChatPDF [24], and Enago Read [25] allow users to upload PDF documents and engage with them in a conversational format. Researchers can direct the AI to automatically generate summary flashcards, summarize specific sections, elucidate complex concepts, extract key data points, or compare arguments presented in different parts of the text. This interactive reading model transforms a static document into a dynamic knowledge dialogue, dramatically accelerating the process of digesting and internalizing complex scholarly material. Concurrently, more specialized question-answering systems like PaperQA2 [26] (Skarlinski et al., 2024) are being developed with a dedicated focus on providing high-precision answers derived from academic texts, further refining the accuracy of information retrieval.

6.2 Idea and Hypothesis Generation

Dialogue-Based Brainstorming with Generative AI. The advent of generative Large Language Models has introduced a novel interactive paradigm for hypothesis generation: dialogue-based brainstorming. Alongside premier general-purpose models such as ChatGPT [27], Claude [28], and DeepSeek [29], more focused academic platforms like SciPIP [30] are emerging. Collectively, these tools can serve as indefatigable academic dialogue partners. By engaging these models in Socratic-style inquiry, multi-perspective debates, and simulated thought experiments, researchers can rigorously probe the boundaries of a research question, uncover latent relationships between variables, and extend the applicability of theoretical frameworks. This interactive modality is highly conducive to divergent thinking.

Building upon these general-purpose models, specialized research platforms are now offering a deeper integration of literature analysis and ideation. Tools like Elicit and ResearchKick [31], for example, not only perform knowledge acquisition tasks but also feature explicit idea generation functionalities. Based on their analysis of existing research gaps in the literature, they can proactively propose under-explored research questions or testable scientific hypotheses, thereby bridging the gap between data-driven heuristics and the researcher's creative intuition.

[22] https://www.scholarcy.com
[23] https://notebooklm.google.com
[24] https://www.chatpdf.com
[25] https://www.read.enago.com/
[26] https://github.com/inproxima/paperqa
[27] https://chatgpt.com/
[28] https://claude.ai
[29] https://www.deepseek.com/
[30] https://github.com/cheerss/SciPIP
[31] https://www.researchkick.com/

Connecting Knowledge to Spark Innovation. A significant wellspring of scientific innovation lies in the novel recombination and connection of existing knowledge. Traditional linear modes of note-taking and thinking, however, have inherent limitations in fostering such cross-domain associations. To address this, Personal Knowledge Management tools like Obsidian and Roam Research have introduced core mechanisms of bi-directional linking and knowledge graphs. Within these platforms, every piece of knowledge a researcher records be it a note on a paper or a nascent experimental idea becomes an independent node that can be linked non-linearly to any other node. By navigating and exploring this networked knowledge structure, researchers can readily discover unexpected associations between disparate concepts. Such serendipitous encounters are precisely the catalyst required for generating breakthrough hypotheses. Similarly, by creating a conversational environment within a user-curated repository of documents, tools like NotebookLM also promote the deep internalization and synthesis of a specific knowledge domain, thereby sparking new insights.

Visual and Collaborative Concept Structuring. The process of translating abstract ideas into a structured, visualized format is a necessary step in the transition from the divergent to the convergent phase of hypothesis generation. Mind mapping and collaborative whiteboard tools play a pivotal role in this stage. Whimsical AI [32] deeply integrates AI into the creation of mind maps and flowcharts. A research team need only input a core research topic, and the AI can automatically generate a mind map complete with main branches, sub-questions, and key nodes. As a leader in collaborative whiteboards, Miro AI [33] offers equally powerful built-in AI features. During team brainstorming sessions, its AI can cluster scattered virtual sticky notes and distill key themes in real-time, rapidly identifying core concepts and potential connections within the discussion. Furthermore, more specialized tools like Research Flow [34] are designed specifically for creating visual research plans, using AI to structure the entire workflow from literature review to experimental design.

6.3 Experiment Design and Execution

Standardized Design and Sharing of Experimental Protocols. Experimental reproducibility begins with a protocol that is clear, detailed, and standardized. In response to the reproducibility crisis, open platforms such as Protocols [35] have emerged to provide a central repository for creating, sharing, peer-reviewing, and version-controlling experimental protocols. This not only accelerates the dissemination of knowledge but also enables other researchers to replicate and validate experiments with greater precision.

This principle extends robustly into the digital realm of computational science . Here, platforms like Weights & Biases [36] serve as essential MLOps tools for tracking every component of a machine learning experiment from hyperparameters and code versions to datasets and model weights ensuring computational results are fully traceable and reproducible. Furthermore, open-source frameworks like Llama Factory [37] standardize complex procedures such as

[32] https://whimsical.com/ai
[33] https://miro.com/ai/
[34] https://ponder.ing/
[35] https://www.protocols.io/
[36] https://wandb.ai
[37] https://github.com/hiyouga/LLaMA-Factory

fine-tuning large language models, providing a consistent methodology that allows different teams to replicate training runs with high fidelity.

For specific domains, Synthace [38] and Strateos [39] are cutting-edge platforms representative of Computer-Aided Biology. These platforms are not merely tools for autonomy but are early instantiations of the Self-Driving Lab paradigm. They represent a fundamental shift towards closing the loop of scientific discovery, where AI agents can autonomously design experiments based on prior knowledge, execute them via robotic hardware, analyze the results, and formulate new hypotheses in a fully integrated, iterative cycle. They combine laboratory autonomy with AI software to create Self-Driving Labs capable of autonomously planning and executing experiments.

Electronic Lab Notebooks and Data Provenance. Electronic Lab Notebooks (ELNs) are rapidly supplanting traditional paper notebooks to become the standard practice for ensuring data integrity, traceability, and collaborative efficiency. Benchling [40] has gained significant traction in the life sciences due to its deep integration of an ELN with molecular biology tools and a sample management system. In parallel, general-purpose ELNs such as LabArchives [41] and SciNote [42] provide cross-disciplinary solutions, supporting the embedding of diverse data formats, version control, and electronic signatures compliant with regulatory standards like FDA 21 CFR Part 11. For laboratories with budgetary constraints, open-source solutions like eLabFTW [43] offer a highly customizable alternative. The core value of an ELN lies in its creation of an immutable, timestamped, and easily searchable digital research log, which provides a robust technical foundation for scientific integrity.

Data Processing, Computational Execution, and Statistical Analysis. The processing and analysis of experimental data form the crucial bridge between raw observations and scientific conclusions. For computationally intensive research, interactive development environments have become the norm. Platforms such as RStudio [44] (for the R language) and Jupyter Notebooks [45] or Google Colab [46] (primarily for Python) employ a Literate Programming paradigm. This allows researchers to weave together code, computational outputs, data visualizations, and narrative text into a single document, which vastly enhances the transparency and reproducibility of the analytical workflow. While these notebook environments provide the interactive substrate for research, the core of AI-driven scientific modeling is powered by underlying deep learning frameworks such as PyTorch, TensorFlow, and JAX. These libraries provide the essential building blocks for creating, training, and deploying sophisticated models for scientific tasks within the notebook interface. For more complex procedures, particularly in fields like bioinformatics, Workflow Management Systems such as

[38] https://www.synthace.com/
[39] https://strateos.com/
[40] https://www.benchling.com/
[41] https://www.labarchives.com/
[42] https://www.scinote.net/
[43] https://www.elabftw.net/
[44] https://www.rstudio.com/
[45] https://jupyter.org
[46] https://colab.research.google.com

Snakemake [47] (Köster and Rahmann, 2012) or Nextflow [48] (Di Tommaso et al., 2017) enable the autonomy, scaling, and deployment of multi-step analysis pipelines.

At the level of statistical analysis, a range of specialized software provides powerful support for researchers across different disciplines. GraphPad Prism [49] is widely used in the biomedical sciences for statistical analysis and publication quality graphing. SPSS [50] and Stata [51] remain mainstays for research in the social sciences and economics, while Origin [52] and MATLAB [53] play a vital role in data analysis and numerical computation within the physical sciences and engineering.

Access to Key Resources and Virtual Simulation. The execution of modern research is highly dependent on access to digital resources, including data and code. In alignment with the principles of Open Science, GitHub [54] and GitLab [55] have become the de facto standards for code sharing and version control, while general-purpose data repositories such as Zenodo [56], Figshare [57], and Dryad [58] provide reliable platforms for the public archiving of research data. It is noteworthy that a new trend is emerging where literature search tools are actively linking to these resources; for instance, platforms like Consensus and ScholarGPS now incorporate direct links to relevant code repositories, and Elicit can provide links to datasets within its analytical results.

Furthermore, virtual simulation technologies are providing an important supplement, and in some cases an alternative, to physical experimentation. Platforms like Labster [59] offer a vast library of highly interactive virtual laboratories, which are used for both educational purposes and the risk-free rehearsal of complex experimental techniques. In engineering and the physical sciences, multiphysics simulation software such as COMSOL Multiphysics [60] allows researchers to build high-fidelity digital twin models within a computer. These models enable the execution of in silico experiments that would otherwise be prohibitively expensive or physically impossible to perform.

6.4 Scientific Communication and Presentation

The Modern Manuscript Workflow. The composition of a scholarly paper has evolved from a solitary endeavor into a highly collaborative process, a shift underpinned by online collaborative platforms. For disciplines in the mathematical, physical, and engineering sci-

[47] https://snakemake.github.io/
[48] https://www.nextflow.io/
[49] https://www.graphpad.com/
[50] https://www.ibm.com/spss
[51] https://www.stata.com/
[52] https://www.originlab.com/
[53] https://www.mathworks.com/products/matlab.html
[54] https://github.com/
[55] https://about.gitlab.com/
[56] https://zenodo.org/
[57] https://figshare.com/
[58] https://datadryad.org/
[59] https://www.labster.com/
[60] https://www.comsol.com/

ences where LaTeX is the standard, Overleaf [61] has emerged as the de facto industry-leading online collaborative LaTeX editor, prized for its powerful formula handling, extensive templates, and robust version control. For researchers who prefer rich-text editing, platforms like Authorea [62], which are designed specifically for scientific writing, as well as the more general-purpose Google Docs [63], also provide powerful real-time co-authoring and commenting capabilities.

Concurrently, generative artificial intelligence is reshaping the drafting and structuring of manuscripts, acting as a writing partner for researchers. General-purpose Large Language Models such as ChatGPT, Claude, and DeepSeek can assist with brainstorming, outlining, drafting specific sections of a paper (e.g., Introduction, Methods), and performing structural reorganizations of existing text. In parallel, more vertical AI writing tools like Scispace [64] and HyperWrite [65] offer functions optimized for specific academic writing tasks, such as generating abstracts or drafting literature reviews.

Furthermore, emerging platforms like OmniThink [66] aim to unify these functionalities, creating an integrated workspace where AI-assisted brainstorming, literature analysis, and collaborative drafting occur within a single, seamless environment, thereby significantly enhancing composition efficiency.

Language Polishing and Stylistic Proofreading. The precision and clarity of language are the lifeline of a high-quality academic paper. AI-driven language enhancement tools provide powerful support for ensuring manuscript quality, a resource of particular importance for non-native English-speaking researchers. Grammarly [67] functions as a comprehensive English writing assistant, offering a full spectrum of recommendations ranging from basic spelling and grammar correction to advanced suggestions on style, clarity, and conciseness.

Tools like DeepL Write [68], Wordtune [69], and especially Writefull [70] is trained exclusively on a corpus of scientific papers excel at holistic sentence rewriting and stylistic optimization, providing multiple phrasing alternatives to help authors achieve a more idiomatical and scholarly tone. Quillbot [71] is distinguished by its powerful paraphrasing capabilities, which are invaluable for diversifying sentence structure and avoiding redundancy or inappropriate textual repetition while preserving the original meaning.

Citation Management and Formatting Standards. Meticulous and accurate reference management is a fundamental requirement of academic integrity. The manual compilation of bibliographies is not only inefficient but also highly susceptible to error.

[61] https://www.overleaf.com
[62] https://www.authorea.com/
[63] https://docs.google.com
[64] https://scispace.com/
[65] https://www.hyperwriteai.com/
[66] https://github.com/zjunlp/OmniThink
[67] https://www.grammarly.com
[68] https://www.deepl.com/write
[69] https://www.wordtune.com/
[70] https://www.writefull.com/
[71] https://quillbot.com/

Professional reference management software, exemplified by Zotero, Mendeley, and EndNote, allows researchers to build, organize, and maintain extensive personal libraries. For quick, on-the-fly citations, web-based generators like Cite This For Me [72] offer a streamlined solution. Furthermore, specialized tools such as MassiveRef [73] can assist in batch processing or verifying large lists of references.

Through seamless integration with writing platforms like Overleaf (via BibTeX), these tools enable the automatic insertion of in-text citations and the one-click generation of a complete reference list at the end of the manuscript. Researchers can effortlessly switch between thousands of different academic journal citation styles, ensuring that all references are standardized and accurate.

Scientific Illustration and Data Visualization. As the adage goes, a picture is worth a thousand words, and high-quality visualization is critical for the effective communication of complex data and scientific concepts. The tools in this domain can be broadly divided into two categories:

Data-Driven Plots and Charts: For data visualizations that require a high degree of customization and reproducibility, the best practice is programmatic generation using coding languages. Leading examples include the Matplotlib [74] and Seaborn [75] libraries in Python and the ggplot2 [76] package in R. For users who prefer a graphical user interface, software such as GraphPad Prism (biomedical sciences) and Origin (physical sciences and engineering) provide powerful integrated data analysis and plotting functionalities. Furthermore, business intelligence tools like Tableau [77] are adept at handling the exploratory visualization of large-scale datasets.

The advent of Large Language Models is now introducing a new, conversational paradigm. Tools like Microsoft Copilot [78]'s charting features and other LLM-based chart generators allow researchers to create complex visualizations simply by describing them in natural language. This AI-driven approach automates chart selection, data mapping, and styling, significantly lowering the barrier to producing insightful data plots.

Schematics and Scientific Illustrations: For creating non-data-driven schematics, models, and flowcharts, BioRender [79] has become an invaluable resource, offering a vast library of specialized icons. Its efficiency is being further enhanced by AI features like Smart Drawing, which can interpret a user's rough sketch and automatically convert it into a clean, standardized icon. For rapid conceptualization, AI platforms like Vizcom [80] can transform simple sketches or text prompts into polished, high-fidelity visuals, making it ideal for exploring early-stage ideas.

[72] https://www.citethisforme.com/
[73] https://www.bibcit.com/zh/massiveref
[74] https://matplotlib.org/
[75] https://seaborn.pydata.org/
[76] https://ggplot2.tidyverse.org/
[77] https://www.tableau.com/
[78] https://copilot.microsoft.com
[79] https://www.biorender.com/
[80] https://www.vizcom.com/

For work demanding maximum creative freedom and precision, vector graphics software remains the preferred choice, including Adobe Illustrator [81] and Inkscape [82]. It is also worth noting that outputs from analytical tools like Connected Papers or CiteSpace can serve as compelling, data-driven illustrations of the intellectual landscape of a research field.

6.5 Limitations and Future Directions

This section provides a comprehensive survey of the AI platforms and toolchains that underpin the modern scientific workflow. This burgeoning ecosystem exhibits considerable vitality and holds immense promise, with its development already catalyzing paradigm shifts in domains such as knowledge acquisition. However, a critical assessment reveals that its maturity is highly stratified and that significant challenges and gaps persist, particularly concerning system integration.

The Stratification of Maturity: From Infrastructure to the Frontier. The maturity of current platforms is not monolithic; rather, it is distinctly stratified across three tiers.

1. The Mature Foundational Layer. A select group of tools has become so deeply embedded in daily research practices that they now constitute an indispensable infrastructure. This tier includes reference management software such as Zotero and Mendeley, which are now standard components of the modern research workflow. It also encompasses collaborative writing platforms like Overleaf, which have emerged as the de facto standard in many scientific disciplines.

2. The Rapidly Adopted, Domain-Specific Layer. A second category of platforms is witnessing rapid adoption within particular fields, establishing new standards for domain-specific research. Prominent examples include electronic lab notebooks like Benchling, which offer deep integration of molecular biology tools for the life sciences, and workflow management systems such as Snakemake and Nextflow, which are becoming essential for automating complex pipelines in bioinformatics and related data-intensive fields.

3. The Nascent Frontier Layer. Representing the cutting edge of AI for Science, a third tier of tools possesses transformative potential but remains in the early stages of adoption. This includes self-driving laboratories, such as those developed by Synthace and Strateos, which are capable of autonomously planning and executing experiments. It also features next-generation knowledge synthesis platforms like Elicit, which can distill and summarize salient findings directly from vast bodies of scientific literature.

System Fragmentation and the Lack of Integration. Despite this rich diversity of tools, the most formidable challenge facing the ecosystem described in this section is its profound fragmentation. A multidimensional landscape of specialized tools, each highly optimized for a specific task within the research lifecycle. Yet, there is little evidence of seamless integration between them. Although an emerging trend sees some literature discovery platforms beginning to link to external code and data repositories, this practice is far from standard. Consequently, the prevailing research workflow involves a series of disjointed steps, compelling researchers to manually transfer data and insights between disparate plat-

[81] https://www.adobe.com/products/illustrator.html
[82] https://inkscape.org/

forms. This fundamental lack of interoperability represents a critical barrier to realizing truly autonomous, end-to-end scientific workflows.

Future Directions. This fragmentation points to critical gaps in the current toolchain, highlighting promising avenues for future development.

1. The Absence of Closed-Loop Autonomous Discovery. While self-driving laboratories mark a significant advance in experimental autonomy, the broader AI for Science framework remains heavily reliant on human-in-the-loop supervision. A crucial missing component is a platform capable of not only executing experiments but also independently interpreting the results, contextualizing them within the broader scientific landscape, and intelligently formulating the next most impactful line of inquiry. Without this capability, the cycle of autonomous scientific discovery remains incomplete.

2. The Systemic Neglect of Negative Results. The vast majority of platforms and toolchains discussed are overwhelmingly oriented towards the successful execution and dissemination of research, culminating in peer-reviewed publications. Conspicuously absent is a systemic framework for documenting, sharing, and analyzing failed experiments or null results. The development of such a system is paramount, as it would dramatically enhance the overall efficiency of scientific progress by preventing the redundant pursuit of unfruitful research avenues.

7 Challenge and Risks

In the era of AI-driven science and innovation, the integration of foundation models and autonomous agents into research practice has introduced both unprecedented opportunities and profound risks. While these systems promise to accelerate scientific discovery and expand the frontiers of knowledge, they simultaneously expose fundamental challenges concerning ethics and safety, limitations of current foundation models, and the uncertain implications of emergent behaviors. The structure of these challenges is illustrated in Figure 11. This chapter reviews these challenges across three dimensions: (1) ethics and safety, focusing on academic integrity, reliability, and governance gaps; (2) foundation model limitations, including constraints in dynamic knowledge integration, reasoning, multilinguality, and multimodality; and (3) emergent risks, arising from creativity, evolution, and multi-agent collaboration. Taken together, these perspectives highlight that realizing the full potential of AI-driven science and innovation requires not only technical breakthroughs but also comprehensive governance frameworks to ensure rigor, inclusivity, and trustworthiness in future research.

7.1 Ethics and Safety

As AI-driven science systems evolve from auxiliary tools to autonomous research agents, the ethical and safety landscape of scientific research is undergoing fundamental transformation. While these systems possess powerful scientific discovery capabilities, their increasing autonomy simultaneously exposes serious ethical blind spots: they are unable to make ethical judgments about the societal impacts of their research work, nor do they self-regulate based on potential risks that their research findings might trigger (Ferrara, 2024; González-Sendino

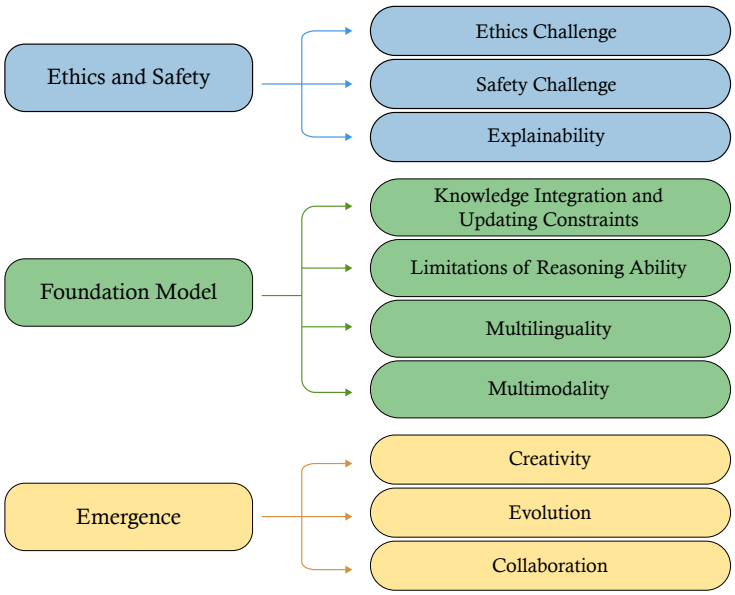

Figure 11: Overview of major challenges in AI-driven science and innovation across three dimensions: ethics and safety, foundation model limitations, and emergent risks.

et al., 2024; Hagendorff, 2024). These issues not only threaten the integrity foundation of scientific research, but may also have profound impacts on social justice and human welfare, urgently requiring coordinated responses from academia, industry, and policymakers.

7.1.1 Ethics Challenge

AI tools are already permeating academic workflows. An estimated 6.5%16.9% of reviews at AI conferences are affected by LLM usage, raising concerns about the reliability of quality control (Ye et al., 2024). Lin (2025b) exposed critical vulnerabilities in LLM-based reviewing: authors can embed hidden prompts in PDF submissions, using tiny or white-font text invisible to humans but readable by LLMs, thereby biasing reviews and undermining their independence. Peer reviews face recurring weaknesses, including lazy thinking, where feedback is shallow or poorly justified (Purkayastha et al., 2025), and excessive politeness, where softened language reduces critical evaluation (Bharti et al., 2024). Together, these issues threaten the fairness, clarity, and reliability of peer review. Beyond the review process, AI-generated content introduces broader risks for scientific progress. One example is the productionprogress paradox: although the volume of scientific publications has grown exponentially, genuine advancement appears to be slowing, a trend that widespread AI adoption may further intensify (Narayanan and Kapoor, 2024). Citation reliability adds another critical concern, as only 17.75% of LLM-generated DOIs are valid (Sinha et al., 2025). Such deficiencies risk accelerating a potential plagiarism singularity, where excessive AI-generated text undermines originality and amplifies ethical and copyright challenges (Schmidt and Meir, 2024).

In response, the global academic community is beginning to explore governance mechanisms, such as transparency guidelines for AI applications and mandatory labeling of AI-generated content (Bengio et al., 2024). Yet, the pace of technological development far outstrips regulatory adaptation, leaving current frameworks unable to mitigate emerging risks such as AI manipulation and content inflation. These dynamics underscore the urgent need for comprehensive ethical guidelines, standardized review procedures, and robust technical safeguards to ensure that AI-driven systems can accelerate scientific progress while preserving the integrity, reliability, and trustworthiness of academic research.

7.1.2 Safety Challenge

Safety concerns in AI-driven science and innovation are no less pressing than ethical ones. The MedHallu benchmark shows that even state-of-the-art systems such as GPT-4o and Llama-3.1 fail to detect more than one-third of medical errors in hallucination detection tasks (Pandit et al., 2025). Subsequent evaluations further suggest that domain-specific medical models may be even more vulnerable than general-purpose counterparts, displaying recurrent failures in three areas: visual misinterpretation, knowledge deficiency, and contextual misalignment (Chang et al., 2025). Compounding these risks, the rapid pace of technological development far outstrips the creation of adequate safeguards, leaving existing regulatory frameworks unable to address emerging threats such as large-scale hallucinations and manipulative misuse (Sahoo et al., 2024).

Against this backdrop, the research community is beginning to build a multi-dimensional framework for risk prevention and control. On the technical front, efforts include advanced hallucination detection systems (Chakraborty et al., 2025), integrated AI security platforms (Buhl et al., 2025), and fairness-aware training pipelines that embed safety constraints directly into model development (Kulkarni et al., 2025; Qi et al., 2024). At the educational level, initiatives emphasize strengthening researchers' literacy in AI ethics and safety, equipping them to critically evaluate AI-generated outputs and identify potential harms (Lin, 2025a). Yet a central challenge remains: safety research must balance reliability with scientific explorationoverly stringent safeguards risk suppressing innovation, whereas insufficient protections may amplify harm. This tension underscores the need for adaptive governance mechanisms that evolve in parallel with technological progress.

7.1.3 Explainability

The ultimate goal of science is to understand the world, and fully harnessing machine learning for discovery requires interpretable models (Wetzel et al., 2025). Within AI-driven science and innovation, interpretability issues are not merely technical transparency problems, but fundamental concerns regarding the authenticity and reliability of scientific discoveries. Sublime (2024) notes that current machine learning methods have forgotten the basic statistical principle that "correlation does not imply causation" often producing flawed causal models similar to astrology. Furthermore, empirical research by Lin (2025a) demonstrates that powerful black-box models often sacrifice interpretability, yet this trade-off relationship is not strictly monotonic, with interpretable models potentially outperforming black-box models in certain situations. Therefore, ensuring that AI outputs represent genuine breakthrough

discoveries rather than simple recombinations of existing data, and achieving traceability in scientific discovery, urgently requires the establishment of standardized interpretability frameworks.

In light of these concerns, there is an urgent need for a comprehensive governance system. Key priorities include: (1) developing emergency monitoring and early-warning mechanisms to track AI usage and risks in scientific research in real-time, with the ability to suspend high-risk applications until adequate safety measures are in place(Cheng and Zhang, 2025); (2)restructuring academic evaluation by introducing manipulation-resistant review processes and implementing rigorous safety certification standards, particularly in sensitive domains such as medical AI (Shin et al., 2025); (3) establishing a global coordination body for AI-driven science and innovation to formulate international ethical guidelines and accountability frameworks, mandating transparency in generation processes and disclosure of associated risks (Wu et al., 2025a). Proactive implementation of these measures can help ensure that AI accelerates scientific progress while safeguarding research integrity and the credibility of scholarly evaluation.

7.2 Foundation Model

Foundation models have become critical infrastructure for AI-driven science and innovation, yet they continue to show domain-specific limitations: difficulty with dynamic knowledge integration and updating, insufficient causal and logical reasoning, imbalanced multilingual representation that exacerbates global inequities, and constrained multimodal understanding of heterogeneous scientific data (Xie et al., 2025; Huang et al., 2025b; Khan et al., 2025). These issues undermine the reliability of model-enabled discovery and pose risks to the rigor and inclusivity of future research.

7.2.1 Knowledge Integration and Updating Constraints

AI-driven scientific inquiry leverages massive and complex, cross-disciplinary datasets. Current large language models struggle to comprehend and integrate such information (Wu et al., 2024; 2025b). A central bottleneck is the limited context window, which restricts long-document processing and cross-source synthesis (Wu et al., 2024). The result can be incomplete concept extraction and improper clustering, degrading the quality of scientific knowledge integration. Scientific knowledge also evolves rapidly. New findings add entities, reshape relationships, and occasionally overturn established frameworks (Jiang et al., 2023). By contrast, large language models are trained on static corpora and cannot readily update internal knowledge. Empirical assessments confirm that LLMs systematically lag behind the scientific literature, with coverage eroding as fields progressraising the critical risk of outdated claims, missed paradigms, and misleading conclusions (Wang et al., 2025b).

To mitigate these challenges, Retrieval-Augmented Generation (RAG) methods have been proposed. By accessing relevant external content, RAG can improve factual accuracy and broaden multi-document coverage(Agarwal et al., 2024b; Ali et al., 2024). However, its effectiveness depends on retrieval quality and does not fully overcome constraints in context length or knowledge updating. Overcoming these barriers will require innovations in dynamic

context management, hierarchical memory architectures, and adaptive retrieval strategies capable of scaling with the complexity of scientific knowledge.

7.2.2 Limitations of Reasoning Ability

Beyond constraints in context processing, reasoning remains a central limitation of current foundation models. Built on autoregressive, correlation-driven learning, these systems lack intrinsic mechanisms for causal reasoning, and as a result, when confronted with tasks requiring causal inference or strict logical consistency, they can produce uncertain or misleading conclusions (Wu et al., 2025b). Benchmark evidence further indicates that even advanced models (e.g., the OpenAI o-series) show marked weaknesses in causal reasoning tasks and typically rely on additional context strategies to yield only modest improvements (Kadziolka and Salehkaleybar, 2025). Despite the recent success of LLMs in high-difficulty math competitions, their capabilities in complex mathematical reasoning remain deficient. A recent study highlights this by demonstrating that these models systematically fail to solve a specific problem of similar difficulty, thereby exposing the brittleness of their reasoning processes (Frieder and Hart, 2025).

Overcoming these deficits requires moving beyond correlation-based modeling. Promising directions include incorporating explicit causal reasoning frameworks, advancing neuro-symbolic methods, and employing human-in-the-loop procedures to guide experimental design, causal discovery, and the interpretation of results. Absent such advances, LLMs alone cannot provide the rigorous reasoning support that scientific inquiry demands.

7.2.3 Multilinguality in AI-driven Research

Language barriers remain a major obstacle to global scientific progress (Amano et al., 2016; 2021; Márquez and Porras, 2020). Within AI-driven science and innovation, the dominant reliance on English widens knowledge gaps: AI systems struggle to learn from non-English knowledge systems (e.g., Traditional Chinese Medicine) due to scarce high-fidelity translations, obscuring potential breakthroughs (Kleidermacher and Zou, 2025; Nasser et al., 2025). Multilingual capability has therefore become a core determinant of whether AI can genuinely serve the global scientific community. A primary source of deficiency is the extreme imbalance in foundation-model training data. Approximately 80% of journal content used for LLM training is in English, whereas 95% of the world's population are not native English speakers (Huang et al., 2024a; Ramírez-Castañeda, 2020). This imbalance contributes to the "curse of multilinguality", whereby performance in a given language tends to degrade as support for more languages is added (Gurgurov et al., 2024). The resulting barriers are not only technical but also equity-related: the dominance of English in science reinforces the imprint of particular cultural viewpoints while marginalizing perspectives from non-English communities (Márquez and Porras, 2020). Observable asymmetries in public information access amplify this effectfor example, the number of English search results for "science" is reported to substantially exceed those in other languages, including in countries with strong scientific traditions such as Germany and Russia. Such design and data biases risk hardening into systemic "colonial bias" effectively walling off non-English knowledge systems and

widening the global knowledge divide (Shahid et al., 2025). These pressures are especially pronounced in domain-specific AI applications for research.

The core multilingual challenge is to support high-level scientific work while maintaining broad language coverage under constrained computational budgets. Addressing this tension requires moving beyond general-purpose modeling toward fine-grained architectural choices and resource-allocation strategies tailored to specific domains and multilingual contexts. This includes domain-conditioned tokenization, language-aware Mixture-of-Experts (MoE) routing, and adaptive retrieval pipelines that integrate non-English corpora. Such designs aim to improve coverage and fidelity without sacrificing rigor or inclusivity.

7.2.4 Multimodality in AI-driven Research

As scientific data grows more diversespanning text, charts, tables, code snippets, and experimental signals, the limitations of current foundation models in multimodal scientific applications have become a key bottleneck for AI-driven science and innovation (Cao et al., 2024; Wang et al., 2024c;e). A holistic evaluation across 30 tasks in HEMM reports systematic shortfalls on science-specific workloads, with even state-of-the-art multimodal models showing a 34.7% performance drop relative to general tasks (Liang et al., 2024). Evidence further links this gap to data scarcity: scientific terminology, experimental symbols, and figures are underrepresented in existing corpora, constraining generalization in scientific domains (Li et al., 2024a). Heterogeneity across modalities also exposes weaknesses in cross-modal fusion. Cai et al. (2025) found that multimodal large language models struggle to distinguish relevant from irrelevant cross-modal signals, making them sensitive to misleading inputs and potentially distorting downstream analyses. Preference for particular modalities exacerbates robustness issues (Ni et al., 2025), manifesting as "modality laziness," where certain modalities are systematically weakened during optimization and hinder comprehensive integration of heterogeneous scientific data (Ma et al., 2025).

These limitations arise from technical constraints and from issues in data quality and domain knowledge integration. Addressing these limitations requires a three-pronged approach. First, we must develop science-specific benchmarks that probe fine-grained competencies, such as symbol grounding and table-to-figure reasoning. Second, we need to conduct domain-focused pretraining to increase model coverage of scientific figures, diagrams, and specialized vocabularies. Finally, robust cross-modal fusion methods must be designed with explicit mechanisms for relevance filtering, modality balancing, and uncertainty calibration, which will enable the faithful integration and analysis of diverse scientific evidence.

7.3 Emergence

Beyond foundational capabilities, AI systems used in scientific contexts can display unpredictable higher-order behaviors when tackling complex tasks. Such emergent behaviors span the generation of novel hypotheses, optimization of research strategies, multi-agent collaboration, and forms of self-directed evolution. While these capacities may accelerate discovery and extend the frontiers of knowledge, they also introduce risks including reasoning errors, significant resource consumption, and limited collaboration efficiency. The following sec-

tions examine potential and limitations across three dimensions: creativity, evolution, and collaboration.

7.3.1 Creativity

Artificial intelligence systems can exhibit creative behaviors that exceed their initial design scope when proposing new hypotheses and scientific insights. This shift marks a transition from passive data processing toward more active participation in scientific innovation. For example, approaches that integrate large language models with dynamic planning and tree search can explore novel directions within vast research spaces(Yamada et al., 2025). These benefits come with notable hazards. Apparent novelty may be illusory when it arises from spurious pattern recognition rather than rigorous scientific reasoning (Reddy and Shojaee, 2025). Simultaneously, the combinatorial explosion of candidate hypotheses and experimental designs elevates the computational burden, limiting applicability in complex, real-world settings (Yamada et al., 2025; Zhou and Arel, 2025). Moreover, reliance on structured planning algorithms, while improving controllability, can restrict flexibility and adaptability, curtailing exploration in interdisciplinary or uncertain contexts (Yuan et al., 2025).

The scientific value of AI-generated ideas therefore hinges less on superficial novelty or volume and more on whether proposals are robust, verifiable, and integrative across domains. In practice, this implies coupling generative exploration with: (i) principled hypothesis priors and rejection criteria; (ii) resource-aware pruning and budgeting for search; and (iii) iterative validation pipelines that combine cross-modal evidence and uncertainty estimation.

7.3.2 Evolution

For AI research agents to move beyond auxiliary roles toward greater autonomy, a central requirement is the capacity for self-directed learning and evolution. A prevalent approach is self-evolution, in which the model acts as its own feedback source and iteratively improves research ability through cyclic evaluation and output optimization (Weng et al., 2024; Romera-Paredes et al., 2024). AlphaEvolve exemplifies this paradigm by coupling large language models with automated evaluation and an evolutionary algorithm loop to autonomously generate and refine algorithms and code, reporting advances across infrastructure, chip design, and mathematical problem solving (Novikov et al., 2025).

However, current self-evolving AI systems expose unforeseen vulnerabilities. Self-reflectiondriven optimization can amplify biases across iterations, reinforcing errors rather than correcting them. Continuous self-optimization and multi-agent simulations also impose substantial computational burdens that limit practical scalability. Progress therefore hinges on evolutionary strategies that are both efficient and verifiable: (i) resource-aware scheduling and early-stopping criteria to curb runaway search; (ii) diversification and stochastic regularization to counter iterative bias reinforcement; and (iii) error-correction mechanisms with externalized evaluatione.g., held-out tests, provenance tracking, and reproducible checkpointsto maintain stability and enable trustworthy improvement in complex research settings.

7.3.3 Collaboration

In multi-agent and cross-disciplinary settings, AI research agents can synchronize information, arbitrate conflicts, and coordinate complex tasks. Yet substantial risks persist. A core obstacle is the absence of standardized communication protocols across teams, which impedes efficient knowledge exchange. The inherent complexity of task allocation, authority transfer, and heterogeneous interaction channels further raises the likelihood of coordination failures, constraining collective intelligence and slowing the evolutionary trajectory of such systems (Hammond et al., 2025; Gomez et al., 2025; Holter and El-Assady, 2024). A second pressure point is the tension between privacy and access in cross-institutional collaboration. Strict anonymization and regulatory constraints protect individuals but can degrade data utility and limit a model's ability to capture representative patterns (Myakala et al., 2024). Conversely, weak governance elevates ethical and compliance risks. Practical frictions compound these issues, as discrepancies in access rights, network bandwidth, and legal frameworks across institutions introduce communication delays and inconsistent model updates during distributed training, undermining the efficiency and stability of federated learning (Guendouzi et al., 2023).

Realizing the benefits of collaboration therefore requires both technical and organizational advances: (i) interoperable communication schemas and standardized agent-to-agent protocols for intent, state, and evidence exchange; (ii) adaptive collaboration strategies, such as role assignment, authority delegation, and conflict-resolution policies that respond to task dynamics; and (iii) privacy-preserving sharing mechanisms that balance utility with compliance.

Taken together, the challenges reviewed in this chapter, from ethical blind spots and safety vulnerabilities to limitations of foundation models and uncertainties of emergent behaviors, underscore that fulfilling the promise of AI-driven science and innovation depends not only on technical advances but also on adaptive governance frameworks that safeguard integrity, inclusivity, and trustworthiness in future research.

8 Future Directions

Against the backdrop of these unresolved risks and challenges, we discuss future directions through three perspectives. First, the role of AI in science and innovation will not be static but dynamic, requiring flexible shifts across contexts, from acting as a computational assistant to serving as a creative partner or, in bounded cases, an autonomous investigator. Understanding when and how such role transitions are appropriate will be central to harnessing AI effectively. Second, large-scale models must be better aligned with the aims and norms of scientific inquiry, moving beyond statistical pattern recognition toward supporting hypothesis generation, causal reasoning, and the construction of verifiable knowledge. Achieving this alignment will be essential for ensuring that AI contributes not only to efficiency gains but also to genuine advances in discovery and innovation. Third, the long-term trajectory of AI-enabled science will depend on building robust communities, shared standards, and open platforms that promote transparency, reproducibility, and equitable participation.

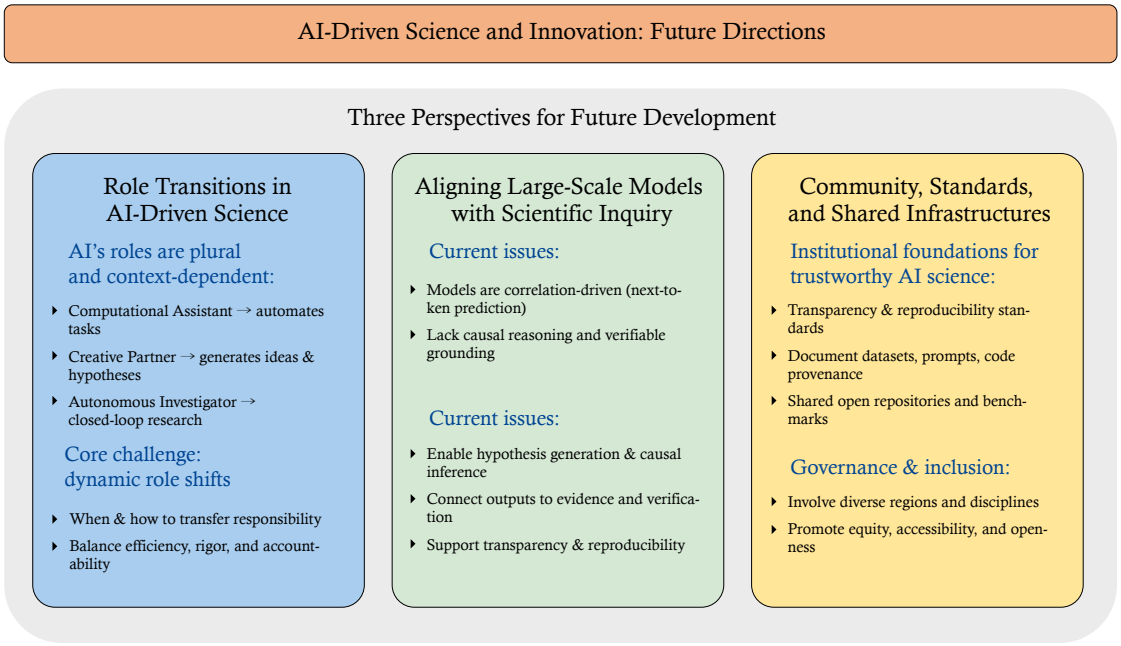

Figure 12: Three perspectives shaping the future of AI-driven science and innovation.

Such institutional and infrastructural foundations will be decisive in determining whether AI develops into a trustworthy and widely accepted partner in the scientific enterprise.

These three perspectives, AI's dynamic role transitions, the alignment of large-scale models with scientific inquiry, and the establishment of shared standards and infrastructures are conceptually summarized in Figure 12, which illustrates the interrelated directions shaping the future of AI-driven science and innovation.

8.1 Role Transitions in AI-Driven Science

The roles that AI can assume within science and innovation are inherently plural rather than fixed. Depending on context, AI may function as a computational assistant that accelerates routine tasks, as a creative partner that contributes novel hypotheses or design strategies, or, in constrained domains, as an autonomous investigator capable of executing closed research loops. The central challenge is not to choose one role over the others, but to understand how responsibilities should shift across scenarios and how authority and accountability can be distributed between human and machine actors. Clarifying these role transitions will be crucial for building systems that are not only efficient, but also scientifically rigorous, and socially responsible.

8.2 Aligning Large-Scale Models with Scientific Inquiry

While large-scale language models have demonstrated remarkable ability across diverse tasks, their core training paradigm remains correlation-driven, oriented toward predicting the next

token rather than engaging in grounded reasoning. This creates a fundamental tension with the aims of science, which require more than surface-level plausibility arising from correlation-driven prediction: scientific inquiry demands explanatory adequacy, causal inference, and the generation of knowledge that can be independently verified and reproduced. The resulting gap highlights why current models, despite their breadth of capability, cannot yet be regarded as reliable scientific instruments. Addressing this gap requires a closer alignment between large-scale models and the methodological norms of science. Such alignment involves reorienting models from merely producing coherent text toward actively supporting hypothesis generation, facilitating causal reasoning, and linking outputs to explicit forms of evidence. It also entails embedding models within workflows that enforce transparency and reproducibility, so that claims generated with AI assistance can be systematically tested, falsified, or corroborated by the broader research community. In this sense, alignment is not only a technical adjustment, but also an epistemic one: it asks how AI systems can operate within the evidentiary standards that define scientific knowledge. Looking forward, the extent to which this alignment can be achieved will shape the trajectory of AI-enabled discovery. If large-scale models remain confined to correlation-driven pattern recognition, their contributions will be limited to accelerating routine tasks and synthesizing existing knowledge. But if they can be brought into closer conformity with scientific reasoning, helping to articulate hypotheses, identify causal structures, and construct verifiable claims. They may evolve from powerful language generators into genuine engines of discovery and innovation. Achieving this transformation will be decisive for determining whether AI reshapes science at the level of productivity alone, or at the deeper level of paradigm formation.

8.3 Community, Standards, and Shared Infrastructures

Beyond roles and methodological alignment, the long-term trajectory of AI-enabled science will depend critically on the institutional and communal foundations that support its practice. Scientific knowledge is not only produced by individuals or systems, but is validated, disseminated, and preserved within communities governed by shared norms. As AI becomes more deeply embedded in research workflows, there is a pressing need to establish standards, governance structures, and open infrastructures that ensure its contributions are trustworthy, reproducible, and equitable.

A first priority is the development of common standards for transparency and reproducibility. This includes establishing guidelines for documenting AI-assisted research processes, reporting provenance information such as datasets, prompts, and code, and providing sufficient detail for independent replication. Without such standards, the integration of AI into science risks amplifying opacity and undermining the credibility of research outputs.

Equally important is the construction of shared platforms and infrastructures that enable collective progress. Open repositories for AI models, benchmarks, and experimental protocols can foster comparability across studies, reduce duplication of effort, and accelerate the diffusion of best practices. Such platforms also provide an avenue for incorporating negative results and failure cases, which are critical for cumulative knowledge building but often neglected in traditional publication systems.

Finally, sustaining AI-enabled science as a trustworthy enterprise will require inclusive community governance. This involves not only technical experts but also policymakers, funding bodies, and diverse research communities across linguistic and cultural contexts. Attention to equity and accessibility will be crucial to prevent the consolidation of AI-driven science within a narrow set of institutions or regions. By embedding principles of openness, accountability, and inclusivity, the scientific community can ensure that AI serves as a catalyst for collective discovery rather than as a source of fragmentation or inequity.

9 Conclusion

Science and innovation have always advanced through their mutual reinforcement, where conceptual breakthroughs enable technological progress, and technological tools open new avenues for discovery. In today's landscape, the rise of AI reconfigures this dynamic by simultaneously acting as an accelerator of efficiency and as a partner in creativity. The collaborative paradigm underscores the indispensable value of human expertise: by integrating dialogue, iterative feedback, and co-design, it ensures that AI-driven processes remain interpretable, scientifically grounded, and enriched by the creativity and insight characteristic of human reasoning. In contrast, the autonomous paradigm has demonstrated remarkable power in scaling knowledge acquisition, generating hypotheses, and streamlining experimental design, thereby addressing the growing challenges of information overload and resource-intensive inquiry.

Despite these advances, realizing the full potential of AI-driven science and innovation requires overcoming persistent limitations. Opaqueness in reasoning, lack of standardized evaluation, and the difficulty of integrating fragmented knowledge into coherent theories remain pressing challenges. Addressing these issues will be crucial for moving beyond acceleration toward genuine transformation, where AI not only augments productivity but also contributes to the creation of new scientific paradigms.

This survey has mapped the field across its major stages, from knowledge acquisition and hypothesis generation to experimentation and dissemination, while contrasting collaborative and autonomous paradigms. By reviewing existing methods, platforms, and domain-specific innovations, we provide a structured foundation for future research. Looking forward, the convergence of human-centered collaboration with scalable automation offers a promising pathway toward an AI-enabled scientific ecosystem that is both rigorous and imaginative, capable of addressing the complexity and openness inherent in modern science.

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 King. 2025. Scientific hypothesis generation by large language models: laboratory validation in breast cancer treatment. *Journal of The Royal Society Interface* 22 (2025), 20240674. https://doi.org/10.1098/rsif.2024.0674

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael ONeill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvil emgulyt, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin ídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 630, 8016 (2024), 493–500. https://doi.org/10.1038/s41586-024-07487-w

Felix Adams, Austin McDannald, Ichiro Takeuchi, and A. Gilad Kusne. 2023. Human-In-the-Loop for Bayesian Autonomous Materials Phase Mapping. arXiv:2306.10406 [cond-mat.mtrl-sci] https://arxiv.org/abs/2306.10406

Dhruv Agarwal, Bodhisattwa Prasad Majumder, Reece Adamson, Megha Chakravorty, Satvika Reddy Gavireddy, Aditya Parashar, Harshit Surana, Bhavana Dalvi Mishra, Andrew McCallum, Ashish Sabharwal, et al. 2025. Open-ended Scientific Discovery via Bayesian Surprise. *arXiv preprint arXiv:2507.00310* (2025).

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

Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H Laradji, Krishnamurthy Dj Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. 2024b. LitLLMs, LLMs for Literature Review: Are we there yet? *arXiv preprint arXiv:2412.15249* (2024).

Iftekhar Ahmed, Aldeida Aleti, Haipeng Cai, Alexander Chatzigeorgiou, Pinjia He, Xing Hu, Mauro Pezzè, Denys Poshyvanyk, and Xin Xia. 2025. Artificial Intelligence for Software Engineering: The Journey So Far and the Road Ahead. 34, 5 (2025). https://doi.org/10.1145/3719006

Nurshat Fateh Ali, Md Mahdi Mohtasim, Shakil Mosharrof, and T Gopi Krishna. 2024. Automated literature review using nlp techniques and llm-based retrieval-augmented generation. In *2024 International Conference on Innovations in Science, Engineering and Technology (ICISET)*. IEEE, 1–6.

Atilla Kaan Alkan, Shashwat Sourav, Maja Jablonska, Simone Astarita, Rishabh Chakrabarty, Nikhil Garuda, Pranav Khetarpal, Maciej Pióro, Dimitrios Tanoglidis, Kartheik G Iyer, et al. 2025. A Survey on Hypothesis Generation for Scientific Discovery in the Era of Large Language Models. *arXiv preprint arXiv:2504.05496* (2025).

Tatsuya Amano, Juan P González-Varo, and William J Sutherland. 2016. Languages are still a major barrier to global science. *PLoS biology* 14, 12 (2016), e2000933.

Tatsuya Amano, Clarissa Rios Rojas, Yap Boum II, Margarita Calvo, and Biswapriya B Misra. 2021. Ten tips for overcoming language barriers in science. *Nature Human Behaviour* 5, 9 (2021), 1119–1122.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, et al. 2018. Construction of the literature graph in semantic scholar. *arXiv preprint arXiv:1805.02262* (2018).

O Ananyin. 2024. Economic science: The challenge of fragmentation. *Journal of the New Economic Association* 63, 2 (2024), 193–210.

Johan O. L. Andreasson, Michael R. Gotrik, Michelle J. Wu, Hannah K. Wayment-Steele, Wipapat Kladwang, Fernando Portela, Roger Wellington-Oguri, Eterna Participants, Rhiju Das, and William J. Greenleaf. 2022. Crowdsourced RNA design discovers diverse, reversible, efficient, self-contained molecular switches. *Proceedings of the National Academy of Sciences* 119, 18 (2022), e2112979119. https://doi.org/10.1073/pnas.2112979119 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2112979119

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

Sören Auer, Dante AC Barone, Cassiano Bartz, Eduardo G Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, et al. 2023. The sciqa scientific question answering benchmark for scholarly knowledge. *Scientific Reports* 13, 1 (2023), 7240.

Sören Auer, Allard Oelen, Muhammad Haris, Markus Stocker, Jennifer DSouza, Kheir Eddine Farfar, Lars Vogt, Manuel Prinz, Vitalis Wiens, and Mohamad Yaser Jaradeh. 2020. Improving access to scientific literature with knowledge graphs. *Bibliothek Forschung und Praxis* 44, 3 (2020), 516–529.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2024. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738* (2024).

Lei Bai, Zhongrui Cai, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, Yongkang Chen, et al. 2025. Intern-S1: A Scientific Multimodal Foundation Model. *arXiv preprint arXiv:2508.15763* (2025).

Mitra Baratchi, Can Wang, Steffen Limmer, Jan N. van Rijn, Holger Hoos, Thomas Bäck, and Markus Olhofer. 2024. Automated machine learning: past, present and future. *Artificial Intelligence Review* 57, 5 (2024), 122.

Jonas Belouadi, Eddy Ilg, Margret Keuper, Hideki Tanaka, Masao Utiyama, Raj Dabre, Steffen Eger, and Simone Paolo Ponzetto. 2025. TikZero: Zero-Shot Text-Guided Graphics Program Synthesis. *arXiv preprint arXiv:2503.11509* (2025).

Jonas Belouadi, Anne Lauscher, and Steffen Eger. 2024. AutomaTikZ: Text-Guided Synthesis of Scientific Vector Graphics with TikZ. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=v3K5TVP8kZ

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3615–3620. https://doi.org/10.18653/v1/D19-1371

Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Danielle Goldfarb, Hoda Heidari, Leila Khalatbari, et al. 2024. International scientific report on the safety of advanced ai (interim report). *arXiv preprint arXiv:2412.05282* (2024).

Prabhat Kumar Bharti, Meith Navlakha, Mayank Agarwal, and Asif Ekbal. 2024. PolitePEER: does peer review hurt? A dataset to gauge politeness intensity in the peer reviews. *Language Resources and Evaluation* 58, 4 (2024), 1291–1313.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. 2023. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* 619, 7970 (2023), 533–538. https://doi.org/10.1038/s41586-023-06185-3

Ahsan Bilal, David Ebert, and Beiyu Lin. 2025. LLMs for Explainable AI: A Comprehensive Survey. arXiv:2504.00125 [cs.AI] https://arxiv.org/abs/2504.00125

Agustín Borrego, Danilo Dessì, Daniel Ayala, Inma Hernández, Francesco Osborne, Diego Reforgiato Recupero, Davide Buscaldi, David Ruiz, and Enrico Motta. 2025. Research hypothesis generation over scientific knowledge graphs. *Knowledge-Based Systems* 315 (2025), 113280.

Peter Bretscher. 2022. Information overload and resilience in facing foundational issues. *Proceedings of the National Academy of Sciences* 119, 9 (2022), e2120180119. https://doi.org/10.1073/pnas.2120180119 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2120180119

Marie Davidsen Buhl, Ben Bucknall, and Tammy Masterson. 2025. Emerging Practices in Frontier AI Safety Frameworks. *arXiv preprint arXiv:2503.04746* (2025).

Benjamin Burger, Phillip M Maffettone, Vladimir V Gusev, Catherine M Aitchison, Yang Bai, Xiaoyan Wang, Xiaobo Li, Ben M Alston, Buyi Li, Rob Clowes, et al. 2020. A mobile robotic chemist. *Nature* 583, 7815 (2020), 237–241.

James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, et al. 2025. Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 19552–19564.

David Buterez, Jon Paul Janet, Steven J. Kiddle, Dino Oglic, and Pietro Lió. 2024. Transfer learning with graph neural networks for improved molecular property prediction in the multi-fidelity setting. *Nature Communications* 15, 1 (2024), 1517.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S Weld. 2020. TLDR: Extreme summarization of scientific documents. *arXiv preprint arXiv:2004.15011* (2020).

Rui Cai, Bangzheng Li, Xiaofei Wen, Muhao Chen, and Zhe Zhao. 2025. Diagnosing and Mitigating Modality Interference in Multimodal Large Language Models. *arXiv preprint arXiv:2505.19616* (2025).

Max W. Callaghan and Finn Müller-Hansen. 2020. Statistical stopping criteria for automated screening in systematic reviews. *Systematic Reviews* 9, 1 (2020), 273. https://doi.org/10.1186/s13643-020-01521-4

Ruisheng Cao, Fangyu Lei, Haoyuan Wu, Jixuan Chen, Yeqiao Fu, Hongcheng Gao, Xinzhuang Xiong, Hanchong Zhang, Wenjing Hu, Yuchen Mao, et al. 2024. Spider2-v: How far are multimodal agents from automating data science and engineering workflows? *Advances in Neural Information Processing Systems* 37 (2024), 107703–107744.

Neeloy Chakraborty, Melkior Ornik, and Katherine Driggs-Campbell. 2025. Hallucination detection in foundation models for decision-making: A flexible definition and review of the state of the art. *Comput. Surveys* 57, 7 (2025), 1–35.

Aofei Chang, Le Huang, Parminder Bhatia, Taha Kass-Hout, Fenglong Ma, and Cao Xiao. 2025. MedHEval: Benchmarking Hallucinations and Mitigation Strategies in Medical Large Vision-Language Models. *arXiv preprint arXiv:2503.02157* (2025).

Diego Chapinal-Heras and Carlos Díaz-Sánchez. 2023. A review of AI applications in Human Sciences research. *Digital Applications in Archaeology and Cultural Heritage* 30 (2023), e00288.

Kinsuk Chauhan, Girish N Nadkarni, Fergus Fleming, James McCullough, Cijiang J He, John Quackenbush, Barbara Murphy, Michael J Donovan, Steven G Coca, and Joseph V Bonventre. 2020. Initial Validation of a Machine Learning-Derived Prognostic Test (KidneyIntelX) Integrating Biomarkers and Electronic Health Record Data To Predict Longitudinal Kidney Outcomes. *Kidney360* 1, 8 (2020), 731–739.

Hui Chen, Miao Xiong, Yujie Lu, Wei Han, Ailin Deng, Yufei He, Jiaying Wu, Yibo Li, Yue Liu, and Bryan Hooi. 2025c. MLR-Bench: Evaluating AI Agents on Open-Ended Machine Learning Research. *arXiv preprint arXiv:2505.19955* (2025).

Nuo Chen, Andre Lin HuiKai, Jiaying Wu, Junyi Hou, Zining Zhang, Qian Wang, Xidong Wang, and Bingsheng He. 2025a. XtraGPT: LLMs for Human-AI Collaboration on Controllable Academic Paper Revision. *arXiv preprint arXiv:2505.11336* (2025).

Qiguang Chen, Mingda Yang, Libo Qin, Jinhao Liu, Zheng Yan, Jiannan Guan, Dengyun Peng, Yiyan Ji, Hanjing Li, Mengkang Hu, et al. 2025d. AI4Research: A Survey of Artificial Intelligence for Scientific Research. *arXiv preprint arXiv:2507.01903* (2025).

Renqi Chen, Haoyang Su, Shixiang Tang, Zhenfei Yin, Qi Wu, Hui Li, Ye Sun, Nanqing Dong, Wanli Ouyang, and Philip Torr. 2025b. AI-Driven Automation Can Become the Foundation of Next-Era Science of Science Research. *arXiv preprint arXiv:2505.12039* (2025).

Mouyang Cheng, Chu-Liang Fu, Ryotaro Okabe, Abhijatmedhi Chotrattanapituk, Artittaya Boonkird, Nguyen Tuan Hung, and Mingda Li. 2025. AI-driven materials design: a mini-review. arXiv:2502.02905 [cond-mat.mtrl-sci] https://arxiv.org/abs/2502.02905

Xusen Cheng and Lulu Zhang. 2025. AI-generated literature reviews threaten scientific progress. *Nature* 641, 8064 (2025), 852–852.

Sathya R. Chitturi, Akash Ramdas, Yue Wu, Brian Rohr, Stefano Ermon, Jennifer Dionne, Felipe H. da Jornada, Mike Dunne, Christopher Tassone, Willie Neiswanger, and Daniel Ratner. 2024. Targeted materials discovery using Bayesian algorithm execution. *npj Computational Materials* 10, 1 (2024), 156.

Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging Large Language Model for Thematic Analysis. arXiv:2310.15100 [cs.CL] https://arxiv.org/abs/2310.15100

Tianwei Dai, Sriram Vijayakrishnan, Filip T. Szczypiski, Jean-François Ayme, Ehsan Simaei, Thomas Fellowes, Rob Clowes, Lyubomir Kotopanov, Caitlin E. Shields, Zhengxue Zhou, John W. Ward, and Andrew I. Cooper. 2024. Autonomous mobile robots for exploratory synthetic chemistry. *Nature* 635, 8040 (2024), 890–897. https://doi.org/10.1038/s41586-024-08173-7

Jonathan de Bruin, Peter Lombaers, Casper Kaandorp, Jelle Teijema, Timo van der Kuil, Berke Yazan, Angie Dong, and Rens van de Schoot. 2025. ASReview LAB v.2: Open-source text screening with multiple agents and a crowd of experts. *Patterns* 6, 7 (2025), 101318. https://doi.org/10.1016/j.patter.2025.101318

Saaketh Desai, Sadhvikas Addamane, Jeffrey Y Tsao, Igal Brener, Laura P Swiler, Remi Dingreville, and Prasad P Iyer. 2025. AutoSciLab: A Self-Driving Laboratory For Interpretable Scientific Discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 146–154.

Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. 2017. Nextflow enables reproducible computational workflows. *Nature biotechnology* 35, 4 (2017), 316–319.

Zijian Ding, Michelle Brachman, Joel Chan, and Werner Geyer. 2025. "The Diagram is like Guardrails": Structuring GenAI-assisted Hypotheses Exploration with an Interactive Shared Representation. In *Proceedings of the 2025 Conference on Creativity and Cognition (C&C '25)*. Association for Computing Machinery, New York, NY, USA, 606625. https://doi.org/10.1145/3698061.3726935

Zhihua Du, Jiale Yi, Jianqiang Li, Hai-Ru You, Zhu-Hong You, Zhi-An Huang, and Yu-An Huang. 2025. scExGraph: Explainable graph neural network for predicting tumor environment components with single-cell sequencing data. *Knowledge-Based Systems* 329 (2025), 114416.

Alexander Dunn, John Dagdelen, Nicholas Walker, Sanghoon Lee, Andrew S Rosen, Gerbrand Ceder, Kristin Persson, and Anubhav Jain. 2022. Structured information extraction from complex scientific text with fine-tuned large language models. *arXiv preprint arXiv:2212.05238* (2022).

Nicholas Edwards, Yukyung Lee, Yujun Audrey Mao, Yulu Qin, Sebastian Schuster, and Najoung Kim. 2025. RExBench: Can coding agents autonomously implement AI research extensions? *arXiv preprint arXiv:2506.22598* (2025).

Steffen Eger, Yong Cao, Jennifer D'Souza, Andreas Geiger, Christian Greisinger, Stephanie Gross, Yufang Hou, Brigitte Krenn, Anne Lauscher, Yizhi Li, et al. 2025. Transforming science with large language models: A survey on ai-assisted scientific discovery, experimentation, content generation, and evaluation. *arXiv preprint arXiv:2502.05151* (2025).

Xingli Fang, Jianwei Li, Varun Mulchandani, and Jung-Eun Kim. 2025b. Trustworthy AI: Safety, Bias, and Privacy – A Survey. arXiv:2502.10450 [cs.CR] https://arxiv.org/abs/2502.10450

You-Le Fang, Dong-Shan Jian, Xiang Li, and Yan-Qing Ma. 2025a. AI-Newton: A Concept-Driven Physical Law Discovery System without Prior Physical Knowledge. arXiv:2504.01538 [cs.AI] https://arxiv.org/abs/2504.01538

Michael Felderer and Rudolf Ramler. 2021. *Quality Assurance for AI-Based Systems: Overview and Challenges (Introduction to Interactive Session).* Springer International Publishing, 3342. https://doi.org/10.1007/978-3-030-65854-0_3

KJ Feng, Kevin Pu, Matt Latzke, Tal August, Pao Siangliulue, Jonathan Bragg, Daniel S Weld, Amy X Zhang, and Joseph Chee Chang. 2024. Cocoa: Co-planning and co-execution with ai agents. *arXiv preprint arXiv:2412.10999* (2024).

Kaiyue Feng, Yilun Zhao, Yixin Liu, Tianyu Yang, Chen Zhao, John Sous, and Arman Cohan. 2025b. PHYSICS: Benchmarking Foundation Models on University-Level Physics Problem Solving. *arXiv preprint arXiv:2503.21821* (2025).

Tao Feng, Yihang Sun, and Jiaxuan You. 2025a. GraphEval: A Lightweight Graph-Based LLM Framework for Idea Evaluation. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=5RUM1aIdok

Emilio Ferrara. 2024. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci* 6, 1 (2024), 3.

Raymond Fok, Joseph Chee Chang, Marissa Radensky, Pao Siangliulue, Jonathan Bragg, Amy X. Zhang, and Daniel S. Weld. 2025. Facets, Taxonomies, and Syntheses: Navigating

Structured Representations in LLM-Assisted Literature Review. arXiv:2504.18496 [cs.HC] https://arxiv.org/abs/2504.18496

Simon Frieder and William Hart. 2025. No LLM Solved Yu Tsumura's 554th Problem. *arXiv preprint arXiv:2508.03685* (2025).

Raoul Frijters, Marianne Van Vugt, Ruben Smeets, René Van Schaik, Jacob De Vlieg, and Wynand Alkema. 2010. Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS computational biology* 6, 9 (2010), e1000943.

Leonardo Gambacorta, Yiping Huang, Han Qiu, and Jingyi Wang. 2024. How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm. *Journal of Financial Stability* 73 (2024), 101284.

Kanishk Gandhi, Michael Y Li, Lyle Goodyear, Louise Li, Aditi Bhaskar, Mohammed Zaman, and Noah D Goodman. 2025. BoxingGym: Benchmarking Progress in Automated Experimental Design and Model Discovery. *arXiv preprint arXiv:2501.01540* (2025).

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: a survey and perspectives. *Humanities and Social Sciences Communications* 11, 1 (2024), 1259.

Xian Gao, Zongyun Zhang, Ting Liu, and Yuzhuo Fu. 2025. GoAI: Enhancing AI Students' Learning Paths and Idea Generation via Graph of AI Ideas. arXiv:2503.08549 [cs.AI] https://arxiv.org/abs/2503.08549

Aniketh Garikaparthi, Manasi Patwardhan, Lovekesh Vig, and Arman Cohan. 2025. IRIS: Interactive Research Ideation System for Accelerating Scientific Discovery. arXiv:2504.16728 [cs.AI] https://arxiv.org/abs/2504.16728

Moncef Garouani. 2025. An experimental survey and Perspective View on Meta-Learning for Automated Algorithms Selection and Parametrization. arXiv:2504.06207 [cs.LG] https://arxiv.org/abs/2504.06207

Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for science writing using language models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*. 1002–1019.

Alireza Ghafarollahi and Markus J Buehler. 2025. SciAgents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials* 37, 22 (2025), 2413523.

Vahid Ghafouri, Faisal Alatawi, Mansooreh Karami, Jose Such, and Guillermo Suarez-Tangil. 2024. Transformer-Based Quantification of the Echo Chamber Effect in Online Communities. 8, CSCW2, Article 467 (Nov. 2024), 27 pages.

Akash Ghosh, Aparna Garimella, Pritika Ramu, Sambaran Bandyopadhyay, and Sriparna Saha. 2025. Infogen: Generating Complex Statistical Infographics from Documents. *arXiv preprint arXiv:2507.20046* (2025).

Yolanda Gil. 2017. Thoughtful artificial intelligence: Forgingăaănewăpartnershipăforădata scienceăandăscientificădiscovery. *Data Science* 1, 1-2 (2017), 119–129. https://doi.org/10.3233/DS-170011 arXiv:https://doi.org/10.3233/DS-170011

Mark Glickman and Yi Zhang. 2024. AI and Generative AI for Research Discovery and Summarization. *Harvard Data Science Review* 6, 2 (apr 30 2024). https://hdsr.mitpress.mit.edu/pub/xedo5giw.

Catalina Gomez, Sue Min Cho, Shichang Ke, Chien-Ming Huang, and Mathias Unberath. 2025. Human-AI collaboration is not very collaborative yet: a taxonomy of interaction patterns in AI-assisted decision making from a systematic review. *Frontiers in Computer Science* 6 (2025), 1521066.

Rubén González-Sendino, Emilio Serrano, and Javier Bajo. 2024. Mitigating bias in artificial intelligence: Fair data generation via causal models for transparent and explainable decision-making. *Future Generation Computer Systems* 155 (2024), 384–401.

Royston Goodacre. 2003. Explanatory analysis of spectroscopic data using machine learning of simple, interpretable rules. *Vibrational Spectroscopy* 32, 1 (2003), 33–45.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. 2025. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864* (2025).

Mourad Gridach, Jay Nanavati, Khaldoun Zine El Abidine, Lenon Mendes, and Christina Mack. 2025. Agentic AI for Scientific Discovery: A Survey of Progress, Challenges, and Future Directions. arXiv:2503.08979 [cs.CL] https://arxiv.org/abs/2503.08979

Xuemei Gu and Mario Krenn. 2025. Interesting Scientific Idea Generation using Knowledge Graphs and LLMs: Evaluations with 100 Research Group Leaders. arXiv:2405.17044 [cs.AI] https://arxiv.org/abs/2405.17044

Kevin Immanuel Gubbi, Sayed Aresh Beheshti-Shirazi, Tyler Sheaves, Soheil Salehi, Sai Manoj PD, Setareh Rafatirad, Avesta Sasan, and Houman Homayoun. 2022. Survey of Machine Learning for Electronic Design Automation. In *Proceedings of the Great Lakes Symposium on VLSI 2022* (Irvine, CA, USA) *(GLSVLSI '22)*. Association for Computing Machinery, New York, NY, USA, 513518. https://doi.org/10.1145/3526241.3530834

Badra Souhila Guendouzi, Samir Ouchani, Hiba EL Assaad, and Madeleine EL Zaher. 2023. A systematic review of federated learning: Challenges, aggregation methods, and development tools. *Journal of Network and Computer Applications* 220 (2023), 103714.

Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Myles Kim, Corey M. Williams, Stefan Bekiranov, and Aidong Zhang. 2025. IdeaBench: Benchmarking Large Language Models for Research Idea Generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2* (Toronto ON, Canada) *(KDD '25)*. Association for Computing Machinery, New York, NY, USA, 58885899. https://doi.org/10.1145/3711896.3737419

Sunil Gupta, Alistair Shilton, Arun Kumar AV, Shannon Ryan, Majid Abdolshah, Hung Le, Santu Rana, Julian Berk, Mahad Rashid, and Svetha Venkatesh. 2023. BO-Muse: A human expert and AI teaming framework for accelerated experimental design. *arXiv preprint arXiv:2303.01684* (2023).

Onder Gurcan. 2024. LLM-Augmented Agent-Based Modelling for Social Simulations: Challenges and Opportunities. arXiv:2405.06700 [physics.soc-ph] https://arxiv.org/abs/2405.06700

Daniil Gurgurov, Tanja Bäumel, and Tatiana Anikina. 2024. Multilingual large language models and curse of multilinguality. *arXiv preprint arXiv:2406.10602* (2024).

Thilo Hagendorff. 2024. Mapping the ethics of generative AI: A comprehensive scoping review. *Minds and Machines* 34, 4 (2024), 39.

Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, et al. 2025. Multi-agent risks from advanced ai. *arXiv preprint arXiv:2502.14143* (2025).

Yuqiang Han, Xiaoyang Xu, Chang-Yu Hsieh, Keyan Ding, Hongxia Xu, Renjun Xu, Tingjun Hou, Qiang Zhang, and Huajun Chen. 2024. Retrosynthesis prediction with an iterative string editing model. *Nature Communications* 15, 1 (2024), 6404.

Mark A. Hanson, Pablo Gómez Barreiro, Paolo Crosetto, and Dan Brockington. 2024. The strain on scientific publishing. *Quantitative Science Studies* 5, 4 (2024), 823843. https://doi.org/10.1162/qss_a_00327

Kenneth D Harris. 2025. AIRUS: a simple workflow for AI-assisted exploration of scientific data. *bioRxiv* (2025), 2025–02.

Xin He, Kaiyong Zhao, and Xiaowen Chu. 2021. AutoML: a survey of the state-of-the-art. *Knowledge-Based Systems* 212 (2021), 106622.

Yifeng He, Jicheng Wang, Yuyang Rong, and Hao Chen. 2025. FuzzAug: Data Augmentation by Coverage-guided Fuzzing for Neural Test Generation. arXiv:2406.08665 [cs.SE] https://arxiv.org/abs/2406.08665

Zhen He, Shuofeng Hu, Yaowen Chen, Sijing An, Jiahao Zhou, Runyan Liu, Junfeng Shi, Jing Wang, Guohua Dong, Jinhui Shi, Jiaxin Zhao, Le Ou-Yang, Yuan Zhu, Xiaochen Bo, and Xiaomin Ying. 2024. Mosaic integration and knowledge transfer of single-cell multimodal data with MIDAS. *Nature Biotechnology* 42, 10 (2024), 1594–1605.

Sam Henry and Bridget T McInnes. 2017. Literature based discovery: models, methods, and trends. *Journal of biomedical informatics* 74 (2017), 20–32.

Dong-hyuk Heo, Inyoung Kim, Heejae Seo, Seong-Gwang Kim, Minji Kim, Jiin Park, Hongsil Park, Seungmo Kang, Juhee Kim, Soonmyung Paik, and Seong-Eui Hong. 2024. DEEPOMICS FFPE, a deep neural network model, identifies DNA sequencing artifacts from formalin fixed paraffin embedded tissue with high accuracy. *Scientific Reports* 14, 1 (2024), 2559.

Danielle Hitch. 2024. Artificial Intelligence Augmented Qualitative Analysis: The Way of the Future? *Qualitative Health Research* 34, 7 (2024), 595–606.

Steffen Holter and Mennatallah El-Assady. 2024. Deconstructing Human-AI Collaboration: Agency, Interaction, and Adaptation. In *Computer graphics forum*, Vol. 43. Wiley Online Library, e15107.

Dimitar Hristovski, Carol Friedman, Thomas C Rindflesch, and Borut Peterlin. 2006. Exploiting semantic relations for literature-based discovery. In *AMIA annual symposium proceedings*, Vol. 2006. 349.

Ting-Yao Hsu, C Lee Giles, and Ting-Hao'Kenneth' Huang. 2021. SciCap: Generating captions for scientific figures. *arXiv preprint arXiv:2110.11624* (2021).

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. *arXiv preprint arXiv:2410.14255* (2024).

Jiayu Huang, Ruoxin Ritter Wang, Jen-Hao Liu, Boming Xia, Yue Huang, Ruoxi Sun, Jason Minhui Xue, and Jinan Zou. 2025a. A Meta-Analysis of LLM Effects on Students across Qualification, Socialisation, and Subjectification. arXiv:2509.22725 [cs.CY] https://arxiv.org/abs/2509.22725

Jincai Huang, Yongjun Xu, Qi Wang, Qi Cheems Wang, Xingxing Liang, Fei Wang, Zhao Zhang, Wei Wei, Boxuan Zhang, Libo Huang, et al. 2025b. Foundation models and intelligent decision-making: Progress, challenges, and perspectives. *The Innovation* (2025).

Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, et al. 2024a. A survey on large language models with multilingualism: Recent advances and new frontiers. *arXiv preprint arXiv:2405.10936* (2024).

Kaixuan Huang, Yuanhao Qu, Henry Cousins, William A Johnson, Di Yin, Mihir Shah, Denny Zhou, Russ Altman, Mengdi Wang, and Le Cong. 2024b. Crispr-gpt: An llm agent for automated design of gene-editing experiments. *arXiv preprint arXiv:2404.18021* (2024).

Maximilian Idahl and Zahra Ahmadi. 2024. Openreviewer: A specialized large language model for generating critical scientific paper reviews. *arXiv preprint arXiv:2412.11948* (2024).

Timofey V Ivanisenko, Pavel S Demenkov, and Vladimir A Ivanisenko. 2024. An accurate and efficient approach to knowledge extraction from scientific publications using structured ontology models, graph neural networks, and large language models. *International Journal of Molecular Sciences* 25, 21 (2024), 11811.

Md Abrar Jahin, Md Sakib Hossain Shovon, M. F. Mridha, Md Rashedul Islam, and Yutaka Watanobe. 2024. A hybrid transformer and attention based recurrent neural network for robust and interpretable sentiment analysis of tweets. *Scientific Reports* 14, 1 (2024), 24882.

A Jain, SP Ong, G Hautier, W Chen, WD Richards, S Dacek, S Cholia, D Gunter, D Skinner, G Ceder, et al. [n. d.]. The Materials Project: a materials genome approach to accelerating materials innovation, APL Mater. 1 (2013) 011002. *Available from DOI* 10, 1.4812323 ([n. d.]).

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. *arXiv preprint arXiv:2503.22708* (2025).

Mohamad Yaser Jaradeh, Allard Oelen, Kheir Eddine Farfar, Manuel Prinz, Jennifer D'Souza, Gábor Kismihók, Markus Stocker, and Sören Auer. 2019. Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge. In *Proceedings of the 10th International Conference on Knowledge Capture* (Marina Del Rey, CA, USA) *(K-CAP '19)*. Association for Computing Machinery, New York, NY, USA, 243246. https://doi.org/10.1145/3360901.3364435

Haoran Jiang, Shaohan Shi, Yunjie Yao, Chang Jiang, and Quan Li. 2025b. HypoChainer: A Collaborative System Combining LLMs and Knowledge Graphs for Hypothesis-Driven Scientific Discovery. arXiv:2507.17209 [cs.HC] https://arxiv.org/abs/2507.17209

Xuhui Jiang, Chengjin Xu, Yinghan Shen, Xun Sun, Lumingyuan Tang, Saizhuo Wang, Zhongwu Chen, Yuanzhuo Wang, and Jian Guo. 2023. On the evolution of knowledge graphs: A survey and perspective. *arXiv preprint arXiv:2310.04835* (2023).

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

Zexun Jiang, Yafang Shi, Maoxu Li, Hongjiang Xiao, Yunxiao Qin, Qinglan Wei, Ye Wang, and Yuan Zhang. 2024. Casevo: A Cognitive Agents and Social Evolution Simulator. arXiv:2412.19498 [cs.SI] https://arxiv.org/abs/2412.19498

Cui-Na Jiao, Ying-Lian Gao, Dao-Hui Ge, Junliang Shang, and Jin-Xing Liu. 2024. Multi-modal imaging genetics data fusion by deep auto-encoder and self-representation network for Alzheimer's disease diagnosis and biomarkers extraction. *Engineering Applications of Artificial Intelligence* 130 (2024), 107782.

Ming Jin and Hyunin Lee. 2025. Position: AI Safety Must Embrace an Antifragile Perspective. arXiv:2509.13339 [cs.AI] https://arxiv.org/abs/2509.13339

Matthew Jin, Syed Shahriar, Michele Tufano, Xin Shi, Shuai Lu, Neel Sundaresan, and Alexey Svyatkovskiy. 2023. InferFix: End-to-End Program Repair with LLMs. arXiv:2303.07263 [cs.SE] https://arxiv.org/abs/2303.07263

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. Agentreview: Exploring peer review dynamics with llm agents. *arXiv preprint arXiv:2406.12708* (2024).

Sebastian Antony Joseph, Syed Murtaza Husain, Stella SR Offner, Stéphanie Juneau, Paul Torrey, Adam S Bolton, Juan P Farias, Niall Gaffney, Greg Durrett, and Junyi Jessy Li. 2025. AstroVisBench: A Code Benchmark for Scientific Computing and Visualization in Astronomy. *arXiv preprint arXiv:2505.20538* (2025).

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *nature* 596, 7873 (2021), 583–589.

Kacper Kadziolka and Saber Salehkaleybar. 2025. Causal Reasoning in Pieces: Modular In-Context Learning for Causal Discovery. *arXiv preprint arXiv:2507.23488* (2025).

Maha Kalai, Hamdi Becha, and Kamel Helali. 2024. Effect of artificial intelligence on economic growth in European countries: a symmetric and asymmetric cointegration based on linear and non-linear ARDL approach. *Journal of Economic Structures* 13, 1 (2024), 22. https://doi.org/10.1186/s40008-024-00345-y

Takuro Kawada, Shunsuke Kitada, Sota Nemoto, and Hitoshi Iyatomi. 2025. SciGA: A Comprehensive Dataset for Designing Graphical Abstracts in Academic Papers. *arXiv preprint arXiv:2507.02212* (2025).

Yujing Ke, Kevin George, Kathan Pandya, David Blumenthal, Maximilian Sprang, Gerrit GroSSmann, Sebastian Vollmer, and David Antony Selby. 2025. BioDisco: Multi-agent hypothesis generation with dual-mode evidence, iterative feedback and temporal evaluation. arXiv:2508.01285 [cs.AI] https://arxiv.org/abs/2508.01285

Rajat Keshri, Arun George Zachariah, and Michael Boone. 2025. Enhancing Code Consistency in AI Research with Large Language Models and Retrieval-Augmented Generation. arXiv:2502.00611 [cs.SE] https://arxiv.org/abs/2502.00611

Wasif Khan, Seowung Leem, Kyle B See, Joshua K Wong, Shaoting Zhang, and Ruogu Fang. 2025. A comprehensive survey of foundation models in medicine. *IEEE Reviews in Biomedical Engineering* (2025).

Jaeyoung Kim, Jongho Lee, Hong-Jun Choi, Ting-Yao Hsu, Chieh-Yang Huang, Sungchul Kim, Ryan Rossi, Tong Yu, Clyde Lee Giles, Ting-HaoKenneth Huang, et al. 2025. Multi-LLM Collaborative Caption Generation in Scientific Documents. In *International Workshop on AI for Transportation*. Springer, 142–160.

Daniel King, Doug Downey, and Daniel S Weld. 2020. High-precision extraction of emerging concepts from scientific literature. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1549–1552.

Ross D. King, Jem Rowland, Wayne Aubrey, Maria Liakata, Magdalena Markham, Larisa N. Soldatova, Ken E. Whelan, Amanda Clare, Mike Young, Andrew Sparkes, Stephen G. Oliver, and Pinar Pir. 2009. The Robot Scientist Adam. *Computer* 42, 8 (2009), 46–54. https://doi.org/10.1109/MC.2009.270

Lawrence C Kingsland III, Gordon C Sharp, Donald R Kay, Sholom M Weiss, Gerald C Roeseler, and Donald AB Lindberg. 1982. An expert consultant system in rheumatology: AI/RHEUM. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*. 748.

Hannah Calzi Kleidermacher and James Zou. 2025. Science across languages: Assessing llm multilingual translation of scientific papers. *arXiv preprint arXiv:2502.17882* (2025).

Petr Knoth, Drahomira Herrmannova, Matteo Cancellieri, Lucas Anastasiou, Nancy Pontika, Samuel Pearce, Bikash Gyawali, and David Pride. 2023. CORE: A global aggregation service for open access papers. *Scientific Data* 10, 1 (2023), 366.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 1613, 15 pages.

Johannes Köster and Sven Rahmann. 2012. Snakemakea scalable bioinformatics workflow engine. *Bioinformatics* 28, 19 (2012), 2520–2522.

Atharva Kulkarni, Yuan Zhang, Joel Ruben Antony Moniz, Xiou Ge, Bo-Hsiang Tseng, Dhivya Piraviperumal, Swabha Swayamdipta, and Hong Yu. 2025. Evaluating Evaluation Metrics–The Mirage of Hallucination Detection. *arXiv preprint arXiv:2504.18114* (2025).

Sandeep Kumar, Tirthankar Ghosal, Vinayak Goyal, and Asif Ekbal. 2024. Can Large Language Models Unlock Novel Scientific Research Ideas? *arXiv preprint arXiv:2409.06185* (2024).

Shrinidhi Kumbhar, Venkatesh Mishra, Kevin Coutinho, Divij Handa, Ashif Iquebal, and Chitta Baral. 2025. Hypothesis generation for materials discovery and design using goal-driven and constraint-guided llm agents. *arXiv preprint arXiv:2501.13299* (2025).

A. Gilad Kusne, Heshan Yu, Changming Wu, Huairuo Zhang, Jason Hattrick-Simpers, Brian DeCost, Suchismita Sarker, Corey Oses, Cormac Toher, Stefano Curtarolo, Albert V. Davydov, Ritesh Agarwal, Leonid A. Bendersky, Mo Li, Apurva Mehta, and Ichiro Takeuchi. 2020. On-the-fly closed-loop materials discovery via Bayesian active learning. *Nature Communications* 11, 1 (2020), 5966. https://doi.org/10.1038/s41467-020-19597-w

Zheyuan Lai and Yingming Pu. 2025. PriM: Principle-Inspired Material Discovery through Multi-Agent Collaboration. arXiv:2504.08810 [cs.LG] https://arxiv.org/abs/2504.08810

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. *arXiv preprint arXiv:2312.07559* (2023).

Nathan Lambert. 2025. Reinforcement Learning from Human Feedback. arXiv:2504.12501 [cs.LG] https://arxiv.org/abs/2504.12501

Jürgen Landauer and Sarah Klassen. 2025. Visual Foundation Models for Archaeological Remote Sensing: A Zero-Shot Approach. *Geomatics* 5, 4 (2025). https://www.mdpi.com/2673-7418/5/4/52

Minhyeong Lee, Suyoung Hwang, Seunghyun Moon, Geonho Nah, Donghyun Koh, Youngjun Cho, Johyun Park, Hojin Yoo, Jiho Park, Haneul Choi, et al. 2025. Spacer: Towards Engineered Scientific Inspiration. *arXiv preprint arXiv:2508.17661* (2025).

Yoonjoo Lee, Hyeonsu B Kang, Matt Latzke, Juho Kim, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2024. Paperweaver: Enriching topical paper alerts by contextualizing recommended papers with user-collected papers. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.

William Leeney and Ryan McConville. 2023. Uncertainty in GNN Learning Evaluations: The Importance of a Consistent Benchmark for Community Detection. arXiv:2305.06026 [cs.LG] https://arxiv.org/abs/2305.06026

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. 2024a. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision* 16, 1-2 (2024), 1–214.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=3IyL2XWDkG

Jiatong Li, Weida Wang, Qinggang Zhang, Junxian Li, Di Zhang, Changmeng Zheng, Shufei Zhang, Xiaoyong Wei, and Qing Li. 2025b. Mol-R1: Towards Explicit Long-CoT Reasoning in Molecule Discovery. *arXiv preprint arXiv:2508.08401* (2025).

Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, et al. 2024b. Chain of ideas: Revolutionizing research via novel idea development with llm agents. *arXiv preprint arXiv:2410.13185* (2024).

Y. Li, T. Gu, C. Yang, M. Li, C. Wang, L. Yao, W. Gu, and D. Sun. 2025a. AI-Assisted Hypothesis Generation to Address Challenges in Cardiotoxicity Research: Simulation Study Using ChatGPT With GPT-4o. *Journal of Medical Internet Research* 27 (2025), e66161. https://doi.org/10.2196/66161

Paul Pu Liang, Akshay Goindani, Talha Chafekar, Leena Mathur, Haofei Yu, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2024. Hemm: Holistic evaluation of multimodal foundation models. *Advances in Neural Information Processing Systems* 37 (2024), 42899–42940.

Xun Liang, Jiawei Yang, Yezhaohui Wang, Chen Tang, Zifan Zheng, Shichao Song, Zehao Lin, Yebin Yang, Simin Niu, Hanyu Wang, et al. 2025. Surveyx: Academic survey automation via large language models. *arXiv preprint arXiv:2502.14776* (2025).

Zhicheng Lin. 2025a. Beyond principlism: practical strategies for ethical AI use in research practices. *AI and Ethics* 5, 3 (2025), 2719–2731.

Zhicheng Lin. 2025b. Hidden Prompts in Manuscripts Exploit AI-Assisted Peer Review. *arXiv preprint arXiv:2507.06185* (2025).

Robert K Lindsay, Bruce G Buchanan, Edward A Feigenbaum, and Joshua Lederberg. 1993. DENDRAL: a case study of the first expert system for scientific hypothesis formation. *Artificial intelligence* 61, 2 (1993), 209–261.

Adam Dahlgren Lindström, Leila Methnani, Lea Krause, Petter Ericson, Íñigo Martínez de Rituerto de Troya, Dimitri Coelho Mollo, and Roel Dobbe. 2024. AI Alignment through Reinforcement Learning from Human Feedback? Contradictions and Limitations. arXiv:2406.18346 [cs.AI] https://arxiv.org/abs/2406.18346

Chunjiang Liu, Yikun Han, Haiyun Xu, Shihan Yang, Kaidi Wang, and Yongye Su. 2024b. A Community Detection and Graph Neural Network Based Link Prediction Approach for Scientific Literature. arXiv:2401.02542 [cs.SI] https://arxiv.org/abs/2401.02542

Chengzhi Liu, Yuzhe Yang, Kaiwen Zhou, Zhen Zhang, Yue Fan, Yannan Xie, Peng Qi, and Xin Eric Wang. 2025d. Presenting a Paper is an Art: Self-Improvement Aesthetic Agents for Academic Presentations. *arXiv preprint arXiv:2510.05571* (2025).

Jingwei Liu, Ling Yang, Hao Luo, Fan Wang Hongyan Li, and Mengdi Wang. 2025b. Preacher: Paper-to-Video Agentic System. *arXiv preprint arXiv:2508.09632* (2025).

Yiren Liu, Si Chen, Haocong Cheng, Mengxia Yu, Xiao Ran, Andrew Mo, Yiliu Tang, and Yun Huang. 2024a. How ai processing delays foster creativity: Exploring research question

co-creation with an llm-based agent. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–25.

Yiren Liu, Pranav Sharma, Mehul Oswal, Haijun Xia, and Yun Huang. 2025a. PersonaFlow: Designing LLM-Simulated Expert Perspectives for Enhanced Research Ideation. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS 25)*. ACM, 506534. https://doi.org/10.1145/3715336.3735789

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. *arXiv preprint arXiv:2503.21248* (2025).

Stanley Lo, Sterling G Baird, Joshua Schrier, Ben Blaiszik, Nessa Carson, Ian Foster, Andrés Aguilar-Granda, Sergei V Kalinin, Benji Maruyama, Maria Politi, et al. 2024. Review of low-cost self-driving laboratories in chemistry and materials science: the frugal twin concept. *Digital Discovery* 3, 5 (2024), 842–868.

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 preprint arXiv:2408.06292* (2024).

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems* 35 (2022), 2507–2521.

Ming Luo, Wenyu Yang, Long Bai, Lin Zhang, Jia-Wei Huang, Yinhong Cao, Yuhua Xie, Liping Tong, Haibo Zhang, Lei Yu, Li-Wei Zhou, Yi Shi, Panke Yu, Zuoyun Wang, Zuoqiang Yuan, Peijun Zhang, Youjun Zhang, Feng Ju, Hongbin Zhang, Fang Wang, Yuanzheng Cui, Jin Zhang, Gongxue Jia, Dan Wan, Changshun Ruan, Yue Zeng, Pengpeng Wu, Zhaobing Gao, Wenrui Zhao, Yongjun Xu, Guangchuang Yu, Caihuan Tian, Ling N. Jin, Ji Dai, Bingqing Xia, Baojun Sun, Fei Chen, Yi-Zhou Gao, Haijun Wang, Bing Wang, Dake Zhang, Xin Cao, Huaiyu Wang, and Tao Huang. 2024. Artificial intelligence for life sciences: A comprehensive guide and future trends. *The Innovation Life* 2, 4 (2024), 100105.

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

Xiaoyu Ma, Hao Chen, and Yongjian Deng. 2025. Improving Multimodal Learning Balance and Sufficiency through Data Remixing. *arXiv preprint arXiv:2506.11550* (2025).

Benjamin P MacLeod, Fraser GL Parlane, Thomas D Morrissey, Florian Häse, Loïc M Roch, Kevan E Dettelbach, Raphaell Moreira, Lars PE Yunker, Michael B Rooney, Joseph R Deeth, et al. 2020. Self-driving laboratory for accelerated discovery of thin-film materials. *Science Advances* 6, 20 (2020), eaaz8867.

Simon Makin. 2024. AI is vulnerable to attack. Can it ever be used safely? *Nature* (Jul 2024). https://www.nature.com/articles/d41586-024-02419-0 News & Views / commentary.

Nour Makke and Sanjay Chawla. 2024. Interpretable scientific discovery with symbolic regression: a review. *Artificial Intelligence Review* 57, 1 (2024), 2.

Daniel J Mankowitz, Andrea Michi, Anton Zhernov, Marco Gelmi, Marco Selvi, Cosmin Paduraru, Edouard Leurent, Shariq Iqbal, Jean-Baptiste Lespiau, Alex Ahern, et al. 2023. Faster sorting algorithms discovered using deep reinforcement learning. *Nature* 618, 7964 (2023), 257–263.

Melissa C Márquez and Ana Maria Porras. 2020. Science communication in multiple languages is critical to its effectiveness. *Frontiers in Communication* 5 (2020), 31.

Shray Mathur, Noah van der Vleuten, Kevin G Yager, and Esther HR Tsai. 2025. VISION: a modular AI assistant for natural human-instrument interaction at scientific user facilities. *Machine Learning: Science and Technology* 6, 2 (2025), 025051.

James Clerk Maxwell. 1865. VIII. A dynamical theory of the electromagnetic field. *Philosophical transactions of the Royal Society of London* 155 (1865), 459–512.

Alex Mei, Sharon Levy, and William Yang Wang. 2023. ASSERT: Automated Safety Scenario Red Teaming for Evaluating the Robustness of Large Language Models. arXiv:2310.09624 [cs.CL] https://arxiv.org/abs/2310.09624

Melkamu Mersha, Khang Lam, Joseph Wood, Ali K. AlShami, and Jugal Kalita. 2024. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing* 599 (2024), 128111.

Rohan Mishra and Bin Li. 2020. The Application of Artificial Intelligence in the Genetic Study of Alzheimer's Disease. *Aging and Disease* 11, 6 (2020), 1567–1584.

Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review* 56, 4 (2023), 3005–3054. https://doi.org/10.1007/s10462-022-10246-w

Christopher E. Mower and Haitham Bou-Ammar. 2025. Al-Khwarizmi: Discovering Physical Laws with Foundation Models. arXiv:2502.01702 [cs.LG] https://arxiv.org/abs/2502.01702

KK Mueen Ahmed and Bandar E Al Dhubaib. 2011. Zotero: A bibliographic assistant to researcher. *Journal of Pharmacology and Pharmacotherapeutics* 2, 4 (2011), 304–305.

H Müller, S Pachnanda, F Pahl, and C Rosenqvist. 2022. The application of artificial intelligence on different types of literature reviews–a comparative study. In *2022 International Conference on Applied Artificial Intelligence (ICAPAI)*.

Praveen Kumar Myakala, Anil Kumar Jonnalagadda, and Chiranjeevi Bura. 2024. Federated learning and data privacy: A review of challenges and opportunities. *International Journal of Research Publication and Reviews* 5, 12 (2024), 10–55248.

Yasmine Nahal, Janosch Menke, Julien Martinelli, Markus Heinonen, Mikhail Kabeshov, Jon Paul Janet, Eva Nittinger, Ola Engkvist, and Samuel Kaski. 2024. Human-in-the-loop active learning for goal-oriented molecule generation. *Journal of Cheminformatics* 16, 1 (2024), 138. https://doi.org/10.1186/s13321-024-00924-y

Arvind Narayanan and Sayash Kapoor. 2024. AI snake oil: What artificial intelligence can do, what it cant, and how to tell the difference. In *AI Snake Oil*. Princeton University Press.

Mahdi Nasser, Laura Sayyah, and Fadi A Zaraket. 2025. Towards Terminology Management Automation for Arabic. *arXiv preprint arXiv:2503.19211* (2025).

Vladimir Naumov, Diana Zagirova, Sha Lin, Yupeng Xie, Wenhao Gou, Anatoly Urban, Nina Tikhonova, Khadija Alawi, Mike Durymanov, Fedor Galkin, et al. 2025. Dora ai scientist: Multi-agent virtual research team for scientific exploration discovery and automated report generation. *bioRxiv* (2025).

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. *arXiv preprint arXiv:2410.22360* (2024).

Isaac Newton. 1833. *Philosophiae naturalis principia mathematica*. Vol. 1. G. Brookman.

Haotian Ni, Yake Wei, Hang Liu, Gong Chen, Chong Peng, Hao Lin, and Di Hu. 2025. RollingQ: Reviving the Cooperation Dynamics in Multimodal Transformer. *arXiv preprint arXiv:2506.11465* (2025).

Ziqi Ni, Yahao Li, Kaijia Hu, Kunyuan Han, Ming Xu, Xingyu Chen, Fengqi Liu, Yicong Ye, and Shuxin Bai. 2024. MatPilot: an LLM-enabled AI Materials Scientist under the Framework of Human-Machine Collaboration. *ArXiv* abs/2411.08063 (2024). https://api.semanticscholar.org/CorpusID:273993701

Josh M Nicholson, Milo Mordaunt, Patrice Lopez, Ashish Uppala, Domenic Rosati, Neves P Rodrigues, Peter Grabitz, and Sean C Rife. 2021. scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. *Quantitative science studies* 2, 3 (2021), 882–898.

Erlend Nilsen, Diana Bowler, and John Linnell. 2020. Exploratory and confirmatory research in the open science era. *Journal of Applied Ecology* 57 (02 2020). https://doi.org/10.1111/1365-2664.13571

Juran Noh, Hieu A Doan, Heather Job, Lily A Robertson, Lu Zhang, Rajeev S Assary, Karl Mueller, Vijayakumar Murugesan, and Yangang Liang. 2024. An integrated high-throughput robotic platform and active learning approach for accelerated discovery of optimal electrolyte formulations. *Nature Communications* 15, 1 (2024), 2757.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. 2025. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131* (2025).

Frank Noé, Alexandre Tkatchenko, Klaus-Robert Müller, and Cecilia Clementi. 2020. Machine learning for molecular simulation. *Annual Review of Physical Chemistry* 71 (2020), 361–390. https://doi.org/10.1146/annurev-physchem-042018-052331

Soroush Omranpour, Guillaume Rabusseau, and Reihaneh Rabbany. 2024. Higher Order Transformers: Enhancing Stock Movement Prediction On Multimodal Time-Series Data. arXiv:2412.10540 [cs.LG] https://arxiv.org/abs/2412.10540

Charles O'Neill, Tirthankar Ghosal, Roberta Rileanu, Mike Walmsley, Thang Bui, Kevin Schawinski, and Ioana Ciuc. 2025. Sparks of Science: Hypothesis Generation Using Structured Paper Data. arXiv:2504.12976 [cs.CL] https://arxiv.org/abs/2504.12976

Hassan Oukhouya, Aziz Lmakri, Mohamed El Yahyaoui, Raby Guerbaz, Said El Melhaoui, Moustapha Faizi, and Khalid El Himdi. 2025. Predictive modeling for the Moroccan financial market: a nonlinear time series and deep learning approach. *Future Business Journal* 11, 1 (2025), 218.

Shrey Pandit, Jiawei Xu, Junyuan Hong, Zhangyang Wang, Tianlong Chen, Kaidi Xu, and Ying Ding. 2025. Medhallu: A comprehensive benchmark for detecting medical hallucinations in large language models. *arXiv preprint arXiv:2502.14302* (2025).

Wei Pang, Kevin Qinghong Lin, Xiangru Jian, Xi He, and Philip Torr. 2025. Paper2Poster: Towards Multimodal Poster Automation from Scientific Papers. *arXiv preprint arXiv:2505.21497* (2025).

Andrea Passerini, Aryo Gema, Pasquale Minervini, Burcu Sayin, and Katya Tentori. 2025. Fostering effective hybrid human-LLM reasoning and decision making. *Frontiers in Artificial Intelligence* Volume 7 - 2024 (2025). https://doi.org/10.3389/frai.2024.1464690

Yunhua Pei, John Cartlidge, Anandadeep Mandal, Daniel Gold, Enrique Marcilio, and Riccardo Mazzon. 2025. Cross-Modal Temporal Fusion for Financial Market Forecasting. arXiv:2504.13522 [cs.LG] https://arxiv.org/abs/2504.13522

Iris Cristina Peláez-Sánchez, Davis Velarde-Camaqui, and Leonardo David Glasserman-Morales. 2024. The impact of large language models on higher education: exploring the connection between AI and Education 4.0. In *Frontiers in Education*, Vol. 9. Frontiers Media SA, 1392091.

Ian M Pendleton, Gary Cattabriga, Zhi Li, Mansoor Ani Najeeb, Sorelle A Friedler, Alexander J Norquist, Emory M Chan, and Joshua Schrier. 2019. Experiment Specification, Capture and Laboratory Automation Technology (ESCALATE): a software pipeline for automated chemical experimentation and data management. *MRS Communications* 9, 3 (2019), 846–859.

Huimin Peng. 2020. A Comprehensive Overview and Survey of Recent Advances in Meta-Learning. arXiv:2004.11149 [cs.LG] https://arxiv.org/abs/2004.11149

Piktochart. [n. d.]. Piktochart AI. https://piktochart.com/ai. Accessed: 2025-08-21.

Dillan Prasad, Aditya Khandeshi, Spencer Sartin, Rishi Jain, Nader Dahdaleh, Maciej Lesniak, Yuan Luo, and Christopher Ahuja. 2025. Will AI become our Co-PI? *npj Digital Medicine* 8, 1 (2025), 440.

Utkarsh Pratiush, Kevin M Roccapriore, Yongtao Liu, Gerd Duscher, Maxim Ziatdinov, and Sergei V Kalinin. 2025. Building workflows for an interactive human-in-the-loop automated experiment (hAE) in STEM-EELS. *Digital Discovery* 4, 5 (2025), 1323–1338.

Kevin Pu, K. J. Kevin Feng, Tovi Grossman, Tom Hope, Bhavana Dalvi Mishra, Matt Latzke, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2025a. IdeaSynth: Iterative Research Idea Development Through Evolving and Composing Idea Facets with Literature-Grounded Feedback. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI 25)*. ACM, 131. https://doi.org/10.1145/3706598.3714057

Yingming Pu, Tao Lin, and Hongyu Chen. 2025b. PiFlow: Principle-aware Scientific Discovery with Multi-Agent Collaboration. *arXiv preprint arXiv:2505.15047* (2025).

Sukannya Purkayastha, Zhuang Li, Anne Lauscher, Lizhen Qu, and Iryna Gurevych. 2025. LazyReview A Dataset for Uncovering Lazy Thinking in NLP Peer Reviews. *arXiv preprint arXiv:2504.11042* (2025).

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 preprint arXiv:2311.05965* (2023).

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. *arXiv preprint arXiv:2407.08940* (2024).

Wei Qiu, Ayse B. Dincer, Joseph D. Janizek, Safiye Celik, Mikael J. Pittet, Kamila Naxerova, and Su-In Lee. 2025a. Deep profiling of gene expression across 18 human cancers. *Nature Biomedical Engineering* 9, 3 (2025), 333–355.

Yansheng Qiu, Haoquan Zhang, Zhaopan Xu, Ming Li, Diping Song, Zheng Wang, and Kaipeng Zhang. 2025b. AI Idea Bench 2025: AI Research Idea Generation Benchmark. arXiv:2504.14191 [cs.AI] https://arxiv.org/abs/2504.14191

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. arXiv:2409.14634 [cs.HC] https://arxiv.org/abs/2409.14634

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* 378 (2019), 686–707.

Valeria Ramírez-Castañeda. 2020. Disadvantages in preparing and publishing scientific papers caused by the dominance of the English language in science: The case of Colombian researchers in biological sciences. *PloS one* 15, 9 (2020), e0238372.

Jacob T. Rapp, Bennett J. Bremer, and Philip A. Romero. 2024. Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nature Chemical Engineering* 1, 1 (2024), 97–107. https://doi.org/10.1038/s44286-023-00002-4

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

Shuo Ren, Pu Jian, Zhenjiang Ren, Chunlin Leng, Can Xie, and Jiajun Zhang. 2025. Towards scientific intelligence: A survey of llm-based scientific agents. *arXiv preprint arXiv:2503.24047* (2025).

Janosh Riebesell, Rhys EA Goodall, Philipp Benner, Yuan Chiang, Bowen Deng, Alpha A Lee, Anubhav Jain, and Kristin A Persson. 2023. Matbench Discovery–A framework to evaluate machine learning crystal stability predictions. *arXiv preprint arXiv:2308.14920* (2023).

Loïc M Roch, Florian Häse, Christoph Kreisbeck, Teresa Tamayo-Mendoza, Lars PE Yunker, Jason E Hein, and Alán Aspuru-Guzik. 2018. ChemOS: orchestrating autonomous experimentation. *Science Robotics* 3, 19 (2018), eaat5559.

Juan A Rodriguez, Abhay Puri, Shubham Agarwal, Issam H Laradji, Pau Rodriguez, Sai Rajeswar, David Vazquez, Christopher Pal, and Marco Pedersoli. 2025. Starvector: Generating scalable vector graphics code from images and text. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 16175–16186.

Juan A Rodriguez, David Vazquez, Issam Laradji, Marco Pedersoli, and Pau Rodriguez. 2023. FigGen: Text to scientific figure generation. *arXiv preprint arXiv:2306.00800* (2023).

Carol A Rohl, Charlie EM Strauss, Kira MS Misura, and David Baker. 2004. Protein structure prediction using Rosetta. In *Methods in enzymology*. Vol. 383. Elsevier, 66–93.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. 2024. Mathematical discoveries from program search with large language models. *Nature* 625, 7995 (2024), 468–475.

Paolo Rosati et al. 2025. Enki: AI for Archaeology Datasets for Automatic Site Recognition. Zenodo dataset. https://doi.org/10.5281/zenodo.14950565

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. arXiv:2412.17596 [cs.CL] https://arxiv.org/abs/2412.17596

Yixiang Ruan, Chenyin Lu, Ning Xu, Yuchen He, Yixin Chen, Jian Zhang, Jun Xuan, Jianzhang Pan, Qun Fang, Hanyu Gao, et al. 2024. An automatic end-to-end chemical synthesis development platform powered by large language models. *Nature communications* 15, 1 (2024), 10160.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. Unveiling Hallucination in Text, Image, Video, and Audio Foundation Models: A Comprehensive Review. (2024).

Ahmed M Salih, Zahra Raisi-Estabragh, Ilaria Boscolo Galazzo, Petia Radeva, Steffen E Petersen, Karim Lekadir, and Gloria Menegaz. 2025. A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems* 7, 1 (2025), 2400304.

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

Paul G Schmidt and Amnon J Meir. 2024. Using generative AI for literature searches and scholarly writing: is the integrity of the scientific discourse in Jeopardy. *Not Am Math Soc* 71, 1 (2024), 93–104.

Farhana Shahid, Mona Elswah, and Aditya Vashistha. 2025. Think outside the data: Colonial biases and systemic issues in automated moderation pipelines for low-resource languages. *arXiv preprint arXiv:2501.13836* (2025).

Erzhuo Shao, Yifang Wang, Yifan Qian, Zhenyu Pan, Han Liu, and Dashun Wang. 2025. SciSciGPT: Advancing Human-AI Collaboration in the Science of Science. arXiv:2504.05559 [cs.AI] https://arxiv.org/abs/2504.05559

Yijia Shao, Yucheng Jiang, Theodore A Kanell, Peter Xu, Omar Khattab, and Monica S Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. *arXiv preprint arXiv:2402.14207* (2024).

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. *arXiv preprint arXiv:1906.03741* (2019).

Hua Shen, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao Kenneth Huang. 2023a. ConvXAI: Delivering heterogeneous AI explanations via conversations to support human-AI scientific writing. In *Companion publication of the 2023 conference on computer supported cooperative work and social computing*. 384–387.

Yiqing Shen, Zan Chen, Michail Mamalakis, Luhan He, Haiyang Xia, Tianbin Li, Yanzhou Su, Junjun He, and Yu Guang Wang. 2024. A fine-tuning dataset and benchmark for large language models for protein understanding. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2390–2395.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023b. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems* 36 (2023), 38154–38180.

Hongyuan Sheng, Jingwen Sun, Oliver Rodríguez, Benjamin B. Hoar, Weitong Zhang, Danlei Xiang, Tianhua Tang, Avijit Hazra, Daniel S. Min, Abigail G. Doyle, Matthew S. Sigman, Cyrille Costentin, Quanquan Gu, Joaquín Rodríguez-López, and Chong Liu. 2024. Autonomous closed-loop mechanistic investigation of molecular electrochemistry via automation. *Nature Communications* 15, 1 (2024), 2781. https://doi.org/10.1038/s41467-024-47210-x

Hyungyu Shin, Jingyu Tang, Yoonjoo Lee, Nayoung Kim, Hyunseung Lim, Ji Yong Cho, Hwajung Hong, Moontae Lee, and Juho Kim. 2025. Mind the Blind Spots: A Focus-Level Evaluation Framework for LLM Reviews. *arXiv preprint arXiv:2502.17086* (2025).

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

Aarush Sinha, Viraj Virk, Dipshikha Chakraborty, and PS Sreeja. 2025. ArxEval: Evaluating Retrieval and Generation in Language Models for Scientific Literature. *arXiv preprint arXiv:2501.10483* (2025).

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. arXiv:2409.13740 [cs.CL] https://arxiv.org/abs/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 (1998), 149–153.

Sebastian Steiner, Jakob Wolf, Stefan Glatzel, Anna Andreou, Jarosław M Granda, Graham Keenan, Trevor Hinkley, Gerardo Aragon-Camarasa, Philip J Kitson, Davide Angelone, et al. 2019. Organic synthesis in a modular robotic system driven by a chemical programming language. *Science* 363, 6423 (2019), eaav2211.

Jonathan M. Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M. Donghia, Craig R. MacNair, Shawn French, Lindsey A. Carfrae, Zohar Bloom-Ackermann, Victoria M. Tran, Anush Chiappino-Pepe, Ahmed H. Badran, Ian W. Andrews, Emma J. Chory, George M. Church, Eric D. Brown, Tommi S. Jaakkola, Regina Barzilay, and James J. Collins. 2020. A Deep Learning Approach to Antibiotic Discovery.

Cell 180, 4 (2020), 688–702.e13. https://doi.org/10.1016/j.cell.2020.01.021

Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, et al. 2024. Many Heads Are Better Than One: Improved Scientific Idea Generation by A LLM-Based Multi-Agent System. *arXiv preprint arXiv:2410.09403* (2024).

Jérémie Sublime. 2024. The Return of Pseudosciences in Artificial Intelligence: Have Machine Learning and Deep Learning Forgotten Lessons from Statistics and History? *arXiv preprint arXiv:2411.18656* (2024).

Teo Susnjak, Peter Hwang, Napoleon Reyes, Andre L. C. Barczak, Timothy McIntosh, and Surangika Ranathunga. 2025. Automating Research Synthesis with Domain-Specific Large Language Model Fine-Tuning. *ACM Transactions on Knowledge Discovery from Data* 19, 3 (March 2025), 139. https://doi.org/10.1145/3715964

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

Don R Swanson and Neil R Smalheiser. 1997. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial intelligence* 91, 2 (1997), 183–203.

Justin Sybrandt, Ilya Tyagin, Michael Shtutman, and Ilya Safro. 2020. AGATHA: automatic graph mining and transformer based hypothesis generation approach. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 2757–2764.

Nathan J Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E Kumar, Tanjin He, David Milsted, Matthew J McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, et al. 2023a. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* 624, 7990 (2023), 86–91.

Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, Tanjin He, David Milsted, Matthew J. McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, Haegyeom Kim, Anubhav Jain, Christopher J. Bartel, Kristin Persson, Yan Zeng, and Gerbrand Ceder. 2023b. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* 624, 7990 (2023), 86–91. https://doi.org/10.1038/s41586-023-06734-w

Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. 2025a. AI-Researcher: Autonomous Scientific Innovation. *arXiv preprint arXiv:2505.18705* (2025).

Xiangru Tang, Zhuoyun Yu, Jiapeng Chen, Yan Cui, Daniel Shao, Fang Wu, Kexu Li, Wangchunshu Zhou, Weixu Wang, Zhi Huang, Arman Cohan, Smita Krishnaswamy, and Mark Gerstein. 2025b. scAgents: A Multi-Agent Framework for Fully Autonomous End-to-End Single-Cell Perturbation Analysis. In *ICML 2025 Generative AI and Biology (GenBio) Workshop*. https://openreview.net/forum?id=HGJQvwGtfJ

InternAgent Team, Bo Zhang, Shiyang Feng, Xiangchao Yan, Jiakang Yuan, Runmin Ma, Yusong Hu, Zhiyin Yu, Xiaohan He, Songtao Huang, Shaowei Hou, Zheng Nie, Zhilong Wang, Jinyao Liu, Tianshuo Peng, Peng Ye, Dongzhan Zhou, Shufei Zhang, Xiaosong Wang, Yilan Zhang, Meng Li, Zhongying Tu, Xiangyu Yue, Wangli Ouyang, Bowen Zhou, and Lei Bai. 2025a. InternAgent: When Agent Becomes the Scientist – Building Closed-Loop System from Hypothesis to Verification. arXiv:2505.16938 [cs.AI] https://arxiv.org/abs/2505.16938

NovelSeek Team, Bo Zhang, Shiyang Feng, Xiangchao Yan, Jiakang Yuan, Zhiyin Yu, Xiaohan He, Songtao Huang, Shaowei Hou, Zheng Nie, et al. 2025b. NovelSeek: When Agent Becomes the Scientist–Building Closed-Loop System from Hypothesis to Verification. *arXiv preprint arXiv:2505.16938* (2025).

Yunsheng Tian, Mina Konakovi Lukovi, Timothy Erps, Michael Foshey, and Wojciech Matusik. 2021. AutoOED: Automated Optimal Experiment Design Platform. arXiv:2104.05959 [cs.AI] https://arxiv.org/abs/2104.05959

Gary Tom, Stefan P Schmid, Sterling G Baird, Yang Cao, Kourosh Darvish, Han Hao, Stanley Lo, Sergio Pablo-García, Ella M Rajaonson, Marta Skreta, et al. 2024. Self-driving laboratories for chemistry and materials science. *Chemical Reviews* 124, 16 (2024), 9633–9732.

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 (2024), 1–14.

Junior Cedric Tonga, KV Aditya Srivatsa, Kaushal Kumar Maurya, Fajri Koto, and Ekaterina Kochmar. 2025. Simulating LLM-to-LLM Tutoring for Multilingual Math Feedback. arXiv:2506.04920 [cs.CL] https://arxiv.org/abs/2506.04920

Juan Diego Toscano, Vivek Oommen, Alan John Varghese, Zongren Zou, Nazanin Ahmadi Daryakenari, Chenxi Wu, and George Em Karniadakis. 2024. From PINNs to PIKANs: Recent Advances in Physics-Informed Machine Learning. arXiv:2410.13228 [cs.LG] https://arxiv.org/abs/2410.13228

Trieu H. Trinh, Yuhuai Wu, Quoc V. Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature* 625, 7995 (2024), 476–482.

Alan Mathison Turing et al. 1936. On computable numbers, with an application to the Entscheidungsproblem. *J. of Math* 58, 345-363 (1936), 5.

Silviu-Marian Udrescu and Max Tegmark. 2020. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances* 6, 16 (2020), eaay2631.

Rens van de Schoot, Jonathan de Bruin, Raoul Schram, Parisa Zahedi, Jan de Boer, Felix Weijdema, Bianca Kramer, Martijn Huijts, Maarten Hoogerwerf, Gerbrich Ferdinands, Albert Harkema, Joukje Willemsen, Yongchao Ma, Qixiang Fang, Sybren Hindriks, Lars Tummers, and Daniel L. Oberski. 2021. An open source machine learning framework for

efficient and transparent systematic reviews. *Nature Machine Intelligence* 3, 2 (2021), 125–133. https://doi.org/10.1038/s42256-020-00287-7

Yuwei Wan, Yixuan Liu, Aswathy Ajith, Clara Grazian, Bram Hoex, Wenjie Zhang, Chunyu Kit, Tong Xie, and Ian Foster. 2024. SciQAG: A Framework for Auto-Generated Science Question Answering Dataset with Fine-grained Evaluation. arXiv:2405.09939 [cs.CL] https://arxiv.org/abs/2405.09939

Fengxiang Wang, Hongzhen Wang, Zonghao Guo, Di Wang, Yulin Wang, Mingshuo Chen, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, et al. 2025e. Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery?. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 14325–14336.

Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2024c. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 19162–19170.

Peng Wang, Yongheng Zhang, Hao Fei, Qiguang Chen, Yukai Wang, Jiasheng Si, Wenpeng Lu, Min Li, and Libo Qin. 2024e. S3 agent: unlocking the power of VLLM for zero-shot multi-modal sarcasm detection. *ACM Transactions on Multimedia Computing, Communications and Applications* (2024).

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)*. Association for Computational Linguistics, 279299. https://doi.org/10.18653/v1/2024.acl-long.18

Wenxiao Wang, Lihui Gu, Liye Zhang, Yunxiang Luo, Yi Dai, Chen Shen, Liang Xie, Binbin Lin, Xiaofei He, and Jieping Ye. 2024b. Scipip: An llm-based scientific paper idea proposer. *arXiv preprint arXiv:2410.23166* (2024).

Weishi Wang, Yue Wang, Shafiq Joty, and Steven C. H. Hoi. 2023b. RAP-Gen: Retrieval-Augmented Patch Generation with CodeT5 for Automatic Program Repair. arXiv:2309.06057 [cs.SE] https://arxiv.org/abs/2309.06057

Xingbo Wang, Samantha L Huey, Rui Sheng, Saurabh Mehta, and Fei Wang. 2024d. SciDaSynth: Interactive structured knowledge extraction and synthesis from scientific literature with large language model. *arXiv preprint arXiv:2404.13765* (2024).

Xin Wang, Jiyao Liu, Yulong Xiao, Junzhi Ning, Lihao Liu, Junjun He, Botian Shi, and Kaicheng Yu. 2025c. THE-Tree: Can Tracing Historical Evolution Enhance Scientific Verification and Reasoning? *arXiv preprint arXiv:2506.21763* (2025).

Xin Wang and Wenwu Zhu. 2024. Advances in neural architecture search. *National Science Review* 11, 8 (08 2024), nwae282. https://doi.org/10.1093/nsr/nwae282

Yike Wang, Shangbin Feng, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2025b. ScienceMeter: Tracking Scientific Knowledge Updates in Language Models. *arXiv preprint arXiv:2505.24302* (2025).

Yubo Wang, Xueguang Ma, Ping Nie, Huaye Zeng, Zhiheng Lyu, Yuxuan Zhang, Benjamin Schneider, Yi Lu, Xiang Yue, and Wenhu Chen. 2025d. Scholarcopilot: Training large language models for academic writing with accurate citations. *arXiv preprint arXiv:2504.00824* (2025).

Yu Wang, Chao Pang, Yuzhe Wang, Junru Jin, Jingjie Zhang, Xiangxiang Zeng, Ran Su, Quan Zou, and Leyi Wei. 2023a. Retrosynthesis prediction with an interpretable deep-learning framework based on molecular assembly tasks. *Nature Communications* 14, 1 (2023), 6155.

Zora Zhiruo Wang, Akari Asai, Xinyan Velocity Yu, Frank F. Xu, Yiqing Xie, Graham Neubig, and Daniel Fried. 2025a. CodeRAG-Bench: Can Retrieval Augment Code Generation?. In *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics, Albuquerque, New Mexico, 3199–3214.

James D Watson and Francis HC Crick. 1953. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* 171, 4356 (1953), 737–738.

Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. 2019. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic acids research* 47, W1 (2019), W587–W593.

Jiaqi Wei, Yuejin Yang, Xiang Zhang, Yuhan Chen, Xiang Zhuang, Zhangyang Gao, Dongzhan Zhou, Guangshuai Wang, Zhiqiang Gao, Juntai Cao, et al. 2025. From AI for Science to Agentic Science: A Survey on Autonomous Scientific Discovery. *arXiv preprint arXiv:2508.14111* (2025).

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2024. Cycleresearcher: Improving automated research via automated review. *arXiv preprint arXiv:2411.00816* (2024).

Sebastian Johann Wetzel, Seungwoong Ha, Raban Iten, Miriam Klopotek, and Ziming Liu. 2025. Interpretable machine learning in physics: A review. *arXiv preprint arXiv:2503.23616* (2025).

Kevin Williams, Elizabeth Bilsland, Andrew Sparkes, Wayne Aubrey, Michael Young, Larisa N Soldatova, Kurt De Grave, Jan Ramon, Michaela De Clare, Worachart Sirawaraporn, et al. 2015. Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. *Journal of the Royal society Interface* 12, 104 (2015), 20141289.

Chun-Ka Wong, Ali Choo, Eugene C. C. Cheng, Wing-Chun San, Kelvin Chak-Kong Cheng, Yee-Man Lau, Minqing Lin, Fei Li, Wei-Hao Liang, Song-Yan Liao, Kwong-Man Ng, Ivan Fan-Ngai Hung, Hung-Fat Tse, and Jason Wing-Hon Wong. 2024. Lomics: Generation

of Pathways and Gene Sets using Large Language Models for Transcriptomic Analysis. arXiv:2407.09089 [q-bio.MN] https://arxiv.org/abs/2407.09089

Jian Cheng Wong, Abhishek Gupta, Chin Chun Ooi, Pao-Hsiung Chiu, Jiao Liu, and Yew-Soon Ong. 2025. Evolutionary Optimization of Physics-Informed Neural Networks: Evo-PINN Frontiers and Opportunities. arXiv:2501.06572 [cs.NE] https://arxiv.org/abs/2501.06572

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. arXiv:2308.08155 [cs.AI] https://arxiv.org/abs/2308.08155

Shican Wu, Xiao Ma, Dehui Luo, Lulu Li, Xiangcheng Shi, Xin Chang, Xiaoyun Lin, Ran Luo, Chunlei Pei, Changying Du, et al. 2025a. Automated literature research and review-generation method based on large language models. *National Science Review* 12, 6 (2025), nwaf169.

Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024. How easily do irrelevant inputs skew the responses of large language models? *arXiv preprint arXiv:2404.03302* (2024).

Xingyu Wu, Kui Yu, Jibin Wu, and Kay Chen Tan. 2025b. LLM Cannot Discover Causality, and Should Be Restricted to Non-Decisional Support in Causal Discovery. *arXiv preprint arXiv:2506.00844* (2025).

Chunqiu Steven Xia, Matteo Paltenghi, Jia Le Tian, Michael Pradel, and Lingming Zhang. 2024. Fuzz4All: Universal Fuzzing with Large Language Models. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering* (Lisbon, Portugal) *(ICSE '24)*. Association for Computing Machinery, New York, NY, USA, Article 126, 13 pages.

Yufei Xia, Zhiyin Han, Yawen Li, and Lingyun He. 2025. Credit scoring model for fintech lending: An integration of large language models and FocalPoly loss. *International Journal of Forecasting* 41, 3 (2025), 894–919.

Qiujie Xie, Yixuan Weng, Minjun Zhu, Fuchen Shen, Shulin Huang, Zhen Lin, Jiahui Zhou, Zilan Mao, Zijie Yang, Linyi Yang, et al. 2025. How Far Are AI Scientists from Changing the World? *arXiv preprint arXiv:2507.23276* (2025).

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. arXiv:2411.02382 [cs.CL] https://arxiv.org/abs/2411.02382

Ruoxi Xu, Yingfei Sun, Mengjie Ren, Shiguang Guo, Ruotong Pan, Hongyu Lin, Le Sun, and Xianpei Han. 2024. AI for social science and social science of AI: A survey. *Information Processing & Management* 61, 3 (2024), 103665.

Wanghan Xu, Xiangyu Zhao, Yuhao Zhou, Xiaoyu Yue, Ben Fei, Fenghua Ling, Wenlong Zhang, and Lei Bai. 2025. EarthSE: A Benchmark for Evaluating Earth Scientific Exploration Capability of LLMs. *CoRR* (2025).

Ziyang Xu. 2025. Patterns and Purposes: A Cross-Journal Analysis of AI Tool Usage in Academic Writing. *arXiv preprint arXiv:2502.00632* (2025).

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. *arXiv preprint arXiv:2504.08066* (2025).

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto MartinezMaldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gaevi. 2023. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology* 55, 1 (2023), 90112. https://doi.org/10.1111/bjet.13370

Shuo Yan, Ruochen Li, Ziming Luo, Zimu Wang, Daoyang Li, Liqiang Jing, Kaiyu He, Peilin Wu, George Michalopoulos, Yue Zhang, et al. 2025. LMR-BENCH: Evaluating LLM Agent's Ability on Reproducing Language Modeling Research. *arXiv preprint arXiv:2506.17335* (2025).

Jianke Yang, Manu Bhat, Bryan Hu, Yadi Cao, Nima Dehmamy, Robin Walters, and Rose Yu. 2025a. Discovering Symbolic Differential Equations with Symmetry Invariants. arXiv:2505.12083 [cs.LG] https://arxiv.org/abs/2505.12083

Yifei Yang, Runhan Shi, Zuchao Li, Shu Jiang, Bao-Liang Lu, Yang Yang, and Hai Zhao. 2024c. BatGPT-Chem: A Foundation Large Model For Retrosynthesis Prediction. arXiv:2408.10285 [cs.LG] https://arxiv.org/abs/2408.10285

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024a. Large Language Models for Automated Open-domain Scientific Hypotheses Discovery. In *Findings of the Association for Computational Linguistics ACL 2024*. 13545–13565.

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. 2024b. Moose-chem: Large language models for rediscovering unseen chemistry scientific hypotheses. *arXiv preprint arXiv:2410.07076* (2024).

Zerui Yang, Yuwei Wan, Yinqiao Li, Yudai Matsuda, Tong Xie, and Linqi Song. 2025b. DrugMCTS: a drug repurposing framework combining multi-agent, RAG and Monte Carlo Tree Search. *arXiv preprint arXiv:2507.07426* (2025).

Lyumanshan Ye, Xiaojie Cai, Xinkai Wang, Junfei Wang, Xiangkun Hu, Jiadi Su, Yang Nan, Sihan Wang, Bohan Zhang, Xiaoze Fan, et al. 2025. Interaction as Intelligence: Deep Research With Human-AI Partnership. *arXiv preprint arXiv:2507.15759* (2025).

Rui Ye, Xianghe Pang, Jingyi Chai, Jiaao Chen, Zhenfei Yin, Zhen Xiang, Xiaowen Dong, Jing Shao, and Siheng Chen. 2024. Are we there yet? revealing the risks of utilizing large language models in scholarly peer review. *arXiv preprint arXiv:2412.01708* (2024).

Qi Ying, Xin Xing, Liangliang Liu, Ai-Ling Lin, Nathan Jacobs, and Gongbo Liang. 2021. Multi-Modal Data Analysis for Alzheimers Disease Diagnosis: An Ensemble Model Using Imagery and Genetic Features. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 3586–3591.

Jason Youn, Navneet Rai, and Ilias Tagkopoulos. 2022. Knowledge integration and decision support for accelerated discovery of antibiotic resistance genes. *Nature communications* 13, 1 (2022), 2360.

Botao Yu, Frazier N Baker, Ziqi Chen, Xia Ning, and Huan Sun. 2024. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. *arXiv preprint arXiv:2402.09391* (2024).

Jiakang Yuan, Xiangchao Yan, Shiyang Feng, Bo Zhang, Tao Chen, Botian Shi, Wanli Ouyang, Yu Qiao, Lei Bai, and Bowen Zhou. 2025. Dolphin: Moving Towards Closed-loop Auto-research through Thinking, Practice, and Feedback. *arXiv preprint arXiv:2501.03916* (2025).

Jooyeol Yun, Heng Wang, Yotaro Shimose, Jaegul Choo, and Shingo Takamatsu. 2025. DesignLab: Designing Slides Through Iterative Detection and Correction. *arXiv preprint arXiv:2507.17202* (2025).

Daifeng Zhang, Guoqiang Bian, Yuanbin Zhang, Jiadong Xie, and Chenjun Hu. 2025a. MOLUNGN: a multi-omics graph neural network for biomarker discovery and accurate lung cancer classification. *Frontiers in Genetics* 16 (2025), 1610284.

Haoxuan Zhang, Ruochi Li, Yang Zhang, Ting Xiao, Jiangping Chen, Junhua Ding, and Haihua Chen. 2025d. The Evolving Role of Large Language Models in Scientific Innovation: Evaluator, Collaborator, and Scientist. *arXiv preprint arXiv:2507.11810* (2025).

Lingyu Zhang, Zhengran Ji, and Boyuan Chen. 2025c. CREW: Facilitating Human-AI Teaming Research. arXiv:2408.00170 [cs.HC] https://arxiv.org/abs/2408.00170

Wentao Zhang, Ce Cui, Yilei Zhao, Rui Hu, Yang Liu, Yahui Zhou, and Bo An. 2025b. Agentorchestra: A hierarchical multi-agent framework for general-purpose task solving. *arXiv preprint arXiv:2506.12508* (2025).

Yuqi Zhang, Adam Perer, and Will Epperson. 2024. Guided Statistical Workflows with Interactive Explanations and Assumption Checking. arXiv:2410.00365 [cs.HC] https://arxiv.org/abs/2410.00365

Zhilin Zhang, Xiang Zhang, Jiaqi Wei, Yiwei Xu, and Chenyu You. 2025e. PosterGen: Aesthetic-Aware Paper-to-Poster Generation via Multi-Agent LLMs. *arXiv preprint arXiv:2508.17188* (2025).

Xiangyu Zhao, Wanghan Xu, Bo Liu, Yuhao Zhou, Fenghua Ling, Ben Fei, Xiaoyu Yue, Lei Bai, Wenlong Zhang, and Xiao-Ming Wu. 2025a. MSEarth: A Benchmark for Multimodal Scientific Comprehension of Earth Science. *arXiv preprint arXiv:2505.20740* (2025).

Yilun Zhao, Kaiyan Zhang, Tiansheng Hu, Sihong Wu, Ronan Le Bras, Taira Anderson, Jonathan Bragg, Joseph Chee Chang, Jesse Dodge, Matt Latzke, et al. 2025b. SciArena: An Open Evaluation Platform for Foundation Models in Scientific Literature Tasks. *arXiv preprint arXiv:2507.01001* (2025).

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. 2025. From automation to autonomy: A survey on large language models in scientific discovery. *arXiv preprint arXiv:2505.13259* (2025).

Zhiling Zheng, Zichao Rong, Nakul Rampal, Christian Borgs, Jennifer T. Chayes, and Omar M. Yaghi. 2023. A GPT4 Reticular Chemist for Guiding MOF Discovery**. *Angewandte Chemie International Edition* 62, 46 (Oct. 2023). https://doi.org/10.1002/anie.202311983

Andy Zhou and Ron Arel. 2025. Tempest: Autonomous Multi-Turn Jailbreaking of Large Language Models with Tree Search. *arXiv preprint arXiv:2503.10619* (2025).

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025c. Deepreview: Improving llm-based paper review with human-like deep thinking process. *arXiv preprint arXiv:2503.08569* (2025).

Shuchen Zhu, Heyang Hua, and Shengquan Chen. 2025a. Rigorous integration of single-cell ATAC-seq data using regularized barycentric mapping. *Nature Machine Intelligence* 7, 9 (2025), 1461–1477.

Zeyu Zhu, Kevin Qinghong Lin, and Mike Zheng Shou. 2025b. Paper2Video: Automatic Video Generation from Scientific Papers. *arXiv preprint arXiv:2510.05096* (2025).