LLM Agents as AI Scientists: A Survey

Peixuan Han *ph16@illinois.edu*
ph16

Zirui Cheng *ziruic4@illinois.edu*
ziruic4

Abstract

In recent years, the rapid development of Large Language Models (LLMs) has significantly reshaped the landscape of scientific research, providing powerful support throughout the research lifecycle. In this survey, we investigate current research concerning the transformative impact of agentic LLMs on the scientific process. Starting from an analysis of the limitations of human research, we examine the distinct contributions of LLMs across different key stages: hypothesis discovery, experiment implementation, paper writing, and peer reviewing. Our analysis highlights task-specific approaches and evaluation standards from a human-centered perspective, offering a detailed overview of the current state of the field. By outlining existing challenges and future directions, we hope this serves as a guide for researchers and practitioners seeking to harness these models to propel scientific discovery.

1 Introduction

For centuries, human curiosity has driven scientific research, leading to groundbreaking discoveries that have shaped our understanding of the world. From the earliest scientific inquiries to frontier innovations, research has been the cornerstone of progress across fields like physics, medicine, technology, computer science, etc. This enduring pursuit of knowledge reflects humanity's relentless drive to explore, understand, and improve the world around us.

Despite its successes, purely human-led research faces significant hurdles. The growing volume of scientific literature makes it difficult for researchers to stay current, while biases in study design and interpretation can skew results. Additionally, the reproducibility crisis has raised concerns about the reliability of many findings, slowing the pace of true scientific advancement. As research becomes more complex and interdisciplinary, these challenges threaten to limit the efficiency and impact of human efforts.

Recent advancements in artificial intelligence, particularly Large Language Models (LLMs), offer a promising way to address these challenges. Trained on vast amounts of text data, LLMs possess a wealth of knowledge and outstanding abilities in knowledge, reasoning Guo et al. (2025), and tool use Qin et al. (2023). These abilities make LLM-based agents powerful tools for automating tasks in scientific research. LLM agents can serve as augmentations to human research society, potentially accelerating the research process and enhancing the quality of scientific productions.

This survey explores how LLMs can be integrated into various stages of the research process, from idea generation and experimental design to writing and peer review. While highlighting the benefits and current progress, we also address the challenges in ethics and applications, aiming to provide a balanced view of how this technology can complement human efforts. Through this examination, we seek to illuminate the transformative role LLMs can play in shaping the future of research.

Survey Organization. The structure of this survey is as follows: §2 discusses the current limitations of human scientific discovery, demonstrating a background for the potential contributions LLM agents could make to human scientific discovery in realistic settings. §3 focuses on LLMs for scientific hypothesis discovery, including an overview of methodologies and key challenges. §4 discusses the potential capabilities of LLM agents in designing and implementing experiments. §5 delves into LLM-driven paper writing, covering key steps, including citation generation. Last but not least, §6 investigates current research in potential applications of LLM agents in peer reviewing. For each topic, we conclude with a summary of current challenges and a list of future directions.

2 Background — Limitations in Human Research

In this section, we examine the limitations of current human scientific research and highlights potential ways LLM agents could help address them.

We argue that current research practices suffer from fragmented workflows, rigid formats, and outdated reward systems that prioritize short-term outputs over innovation. Standardization can improve reproducibility and collaboration but may introduce rigidity and stifle creativity. Traditional publication formats like PDFs limit interactivity, while newer platforms (e.g., arXiv, GitHub) offer faster, more open sharing—but still face challenges in evaluation and engagement. Meanwhile, we also posit that current flawed incentive structures drive researchers toward safe, trend-driven topics and polished presentation, often at the expense of the depth and originality of their research output. Ethical oversight lags behind modern methodologies, failing to fully protect participants, researchers, or reviewers. Systemic biases in peer review and unequal access to funding further compound these issues, limiting diversity and fairness. Together, they call for a rethinking of the research pipeline toward more dynamic, transparent, and inclusive models where LLMs could play a transformative role.

2.1 Efficiency Issues

Time constrains. Human research efficiency is fundamentally constrained by the cognitive capacity of researchers. Unlike machines, humans have natural limitations in attention, memory, and decision-making. According to Miller's law Miller (1956), individuals can typically hold only 8 units of information in working memory at once. When researchers face an influx of data, literature, or concurrent tasks, this threshold is quickly surpassed, resulting in information overload. For example, a scientist attempting to synthesize dozens of papers or manage multiple project variables simultaneously may experience errors or delays, particularly in complex interdisciplinary studies. Moreover, the constant need to make decisions, ranging from experimental design to task prioritization, leads to decision fatigue Vohs et al. (2018). Research shows that this cumulative burden impairs executive functioning and reasoning, increasing reliance on biases and mental shortcuts Vohs et al. (2005). A striking illustration comes from studies of journal editors Stewart et al. (2012): as they reviewed more manuscripts per session (from 10–19 to over 20), rejection rates rose from 38% to 44%, with editors reviewing three or more manuscripts daily rejecting submissions without peer review 6% more often than those handling fewer. This demonstrates how decision fatigue amplifies biases and undermines decision quality, significantly impacting research efficiency.

Communication cost. Effective collaboration is vital for interdisciplinary and large-scale research, yet communication inefficiencies frequently undermine it. Disciplinary differences in terminology and frameworks create misunderstandings that stall progress. Without a shared vocabulary, integrating knowledge becomes arduous, especially in projects requiring precision. For example, a term like "model" might mean a biological system to a biologist but a computational algorithm to a computer scientist, leading to misaligned interpretations. In addition, challenges in coordination and information overload hinder efficient research in large research groups. Excessive emails and unstructured meetings overwhelm researchers, while hierarchical dynamics often silence junior scientists, whose novel ideas go unheard (Azoulay et al., 2018). Geographical and cultural barriers further complicate matters in international collaborations, with time zone disparities, differing institutional norms, and communication styles causing delays and conflicts.

Knowledge barrier. As technology progresses rapidly, the scope of knowledge required to become a mature scientist has grown substantially. Today's researchers must master not only foundational scientific principles but also cutting-edge fields like data analysis, advanced computing, and interdisciplinary applications. Fields like AI-assisted drug design, quantum computing, molecular robotics, and smart cities require integrating knowledge and methods across disciplines. This expanded knowledge base prolongs the education process and necessitates more sophisticated training programs, driving up the cost of preparing a fully equipped scientist. An extreme case in sci-fi novels can illustrate this limitation: in the future, human knowledge may grow to the point where the average person will not be able to specialize in a single field in their whole lifetime. If we stick to purely human research paradigms, human technology will cease to develop when the threshold is reached.

2.2 Research Reproduction

Reproducibility is a cornerstone of modern scientific research (Moonesinghe et al., 2007; Simons, 2014). By reproducing previous work independently, we confirm the soundness of conclusions and exclude random or human factors that might affect the results. Successful reproductions not only bolsters the confidence of the academic community in the original work but also lays a solid foundation for further studies and innovations. Moreover, reproducibility promotes transparency and accountability, enabling scientists to share their methodologies openly and learn from each other's successes and failures. Responsible scientists remind themselves to double-check their results before publishing since reliable and reproducible findings lead to a greater and more positive impact than unverifiable ones. On the contrary, lack of reproducing may cause untenable results to be accepted by the academia without scrutiny, as exemplified by the "Reproducibility Project: Psychology"[1] conducted to reproduce classic psychological experiments. In this project, only 39% of the experiments are successfully reproduced, raising people's concerns about the psychological theories based upon these experiments.

With the scope and scale of scientific research continuously expanding in recent years, reproducing becomes more and more challenging. Specifically, two drastic changes in scientific research led to the reproduction crisis:

The number of research papers. The ease of modern communication and the growth of the academic community have led to an explosion of research findings. We can observe such an explosion by the statistics of research papers, the prevalent form of research outcomes. ArXiv, one of the largest and most influential preprint repositories in academia, has more than doubled the number of monthly paper submissions since 2015, reaching over 20k research papers per month. The number of submissions and accepted papers in journals and conferences has also exploded in recent years. Namely, the Association for Computational Linguistics (ACL) received 3378 submissions in 2022, while only 692 papers were submitted in 2015. Manually verifying all these papers is extremely time-consuming, if even possible.

The complexity of research papers. As human continues to explore the world and science continues to develop, the complexity for verifying scientific findings has sharply increased. On one hand, frontier discoveries often involve prohibitively costly types of equipment. For instance, state-of-the-art language models utilize thousands of GPUs and trillions of text tokens to train (Guo et al., 2025; Achiam et al., 2023), which is far beyond the ability of all but a few Well-resourced institutes or corporations. Similar challenges arise in other fields of research: In physics, high-energy particle colliders are built to explore new particles, which cost billions of dollars to build Albajar et al. (1987); in materials science, advanced microscopes like TEM (Kübel et al., 2005) and ESEM (Zhang et al., 2020) are essential for analysing molecules.

2.3 Research Pipeline

Modern human research across disciplines typically proceeds through a series of stages, from initial conception to validation. While the specifics vary by field, a general pipeline includes steps like hypothesis generation, literature review, experiment implementation, analysis and interpretation, communication, and dissemination. Ideally, each step of this pipeline is executed with rigor and honesty to advance knowledge. However, current academic incentive structures often misalign with these stages, introducing distortions at multiple points. The well-known "**publish or perish**" culture means researchers are rewarded primarily for outputs – notably, published papers in high-impact venues – rather than for the quality of processes that lead to those outputs. Career advancement in academia is heavily dependent on metrics like the number of publications, journal prestige (impact factor), citation counts, and grant dollars obtained (Edwards &

[1] `https://osf.io/ezcuj/`

Roy, 2017). Likewise, the race to publish in prestigious journals that value novel, striking results encourages scientists to sensationalize findings and may even prompt them to hide null or inconclusive results that are deemed unlikely to be accepted Trueblood et al. (2025). As detailed below, such research format rewards (tied to how and where work is published) and outcome-based rewards (tied to the nature of the results) skew decisions throughout the research process – from which hypotheses get pursued, to how studies are conducted and reported, all the way to which findings see the light of day.

First, researchers are frequently evaluated based on formal research outputs—especially peer-reviewed publications in high-impact venues Edwards & Roy (2017). This emphasis on publication count, journal prestige, and grant acquisition creates pressure to prioritize quantity over quality, encourage salami-slicing, and focus on trendy or publishable topics rather than fundamental or high-risk questions. Key research stages such as thorough literature reviews, transparent methodological design, and open data sharing are undervalued as they offer little immediate reward in metric-driven evaluation systems. The focus on format often disincentivizes careful, time-intensive work that enhances reproducibility or long-term scientific value Trueblood et al. (2025).

Second, reward structures that prioritize positive, novel results over negative or null findings further distort the research pipeline Nosek et al. (2012); Bik (2024). This leads to selective reporting, outcome embellishment, and a systemic bias against replication. As studies with "publishable" outcomes are more likely to be accepted and cited, researchers may avoid high-risk or exploratory projects in favor of those that are more likely to yield confirmatory, marketable results. Replication studies—essential for scientific validation—are rarely rewarded and thus remain scarce. The result is a literature biased toward optimistic claims, contributing to poor reproducibility and undermining public trust in science.

In summary, the current human research pipeline still requires realigning incentives with each stage of that pipeline. Currently, research format rewards (e.g., counting publications and valuing prestigious venues) and outcome-based rewards (e.g., prioritizing positive, novel results) often conflict with methodological rigor and transparency. Recognizing these distortions is the first step toward reforms that encourage comprehensive literature reporting, honesty about negative outcomes, and routine replication, thereby realigning academic rewards with the core stages of the research process and the production of reliable knowledge.

2.4 Regulation Constraints

While ethical regulations in human research have been instrumental in promoting safety, fairness, and accountability, their practical limitations increasingly hinder their effectiveness across modern research contexts. Institutional Review Boards (IRBs), for example, are foundational to participant protection, yet their processes are often procedural and static, lacking the capacity to adapt to evolving risks in long-term or technologically complex studies (Stahl & Stahl, 2021; Grady, 2015). This rigidity can delay critical research and inadequately address novel ethical dilemmas arising in fields such as artificial intelligence and behavioral tracking.

Informed consent, though a central ethical safeguard, frequently falls short in practice. Documents are often laden with technical jargon, impeding participant comprehension, especially among vulnerable groups (Clark-Kazak, 2019). Furthermore, consent is usually treated as a one-time event, failing to account for the dynamic nature of data use. Participants rarely retain control over how their information is repurposed or shared in secondary analyses, raising concerns about autonomy and transparency (Yadav et al., 2023).

Efforts to ensure fairness in scientific evaluation, such as double-blind peer review and transparency platforms, are undermined by persistent structural inequities. Reviewer anonymity can often be breached through writing style or institutional references, and systemic biases related to race, gender, and geographic affiliation continue to shape publication and funding outcomes (Heidt, 2023; Freeman & Robbins, 2005). Ethical norms promoting equity remain aspirational without standardized enforcement or reform of entrenched power structures.

The push for reproducibility and open science has also revealed significant barriers. Although open data mandates and pre-registration frameworks aim to improve transparency, many researchers face disincentives to comply, including fear of being scooped, lack of infrastructure, and the high cost of open-access publishing (Gundersen et al., 2018; Kwon, 2022). Data-sharing requirements are often inconsistently applied, and commercial or proprietary constraints further limit reproducibility in industry-influenced research (Bostrom, 2018).

In summary, while ethical regulations provide necessary frameworks, their current implementation is hampered by inflexibility, inconsistent enforcement, and limited capacity to respond to the complexities of contemporary human research. More adaptive, participatory, and enforceable approaches are needed to align ethical governance with the realities of modern scientific practice.

3 Hypothesis Discovery

Proposing a hypothesis is the first step in scientific research, which lays the foundation for further experiments and analysis. Although hypothesis discovery in frontier scientific research is primarily conducted by humans, we do see potential for LLMs in proposing valuable scientific hypotheses.

Before the age of LLMs, automated hypothesis discovery was rooted in literature-based discovery (LBD) and inductive reasoning. LBD, pioneered by Swanson (1986), seeks to uncover novel insights by connecting previously unlinked pieces of information within scientific literature. Early LBD methods relied on techniques like word vectors Tshitoyan et al. (2019) and link prediction models Sybrandt et al. (2020); Wang et al. (2019), which were effective for identifying pairwise relationships but struggled to capture the deeper contextual understanding that human researchers naturally employ. Modern advancements, such as SciMON Wang et al. (2024a), have overcome these limitations by incorporating natural language contexts, enabling the generation of more sophisticated and nuanced hypotheses. Alongside LBD, inductive reasoning has played a critical role in deriving general hypotheses from specific observations Norton (2003)—a process central to scientific breakthroughs. This approach requires hypotheses to be consistent with observations, reflective of reality, and generalizable Yang et al. (2022); Qiu et al. (2023).

Building on these foundations, the development of methods for scientific hypothesis discovery using large language models (LLMs) has progressed along a structured trajectory, integrating several key innovations. Inspiration retrieval, the process of identifying and gathering relevant knowledge or information from existing sources, has advanced from pulling semantically similar content or graph-based neighbors Wang et al. (2024a) to relying on LLMs to select relevant inspirations based on their parametric knowledge, like MOOSE Yang et al. (2023) and MOOSE-Chem Yang et al. (2024). In addition, several feedback mechanisms are proposed to address the uncertainty in LLMs and to ensure hypothesis quality. For instance, novelty checkers compare outputs to existing literature for originality, validity checkers use heuristics or experimental data to confirm accuracy, and clarity checkers refine hypotheses for precision and detail. Human researchers often reiterate and refine hypotheses multiple times; similarly, evolutionary algorithms inspired by biological principles are proposed for LLMs to optimize hypotheses through iterative mutation and selection Ma et al. (2024).

Besides this core trajectory, a range of alternative methods have emerged to tackle distinct challenges in scientific discovery. For instance, Pu et al. (2024) proposes IdeaSynth, a system designed for developing research ideas by representing concepts as interconnected nodes on a visual interface; Weng et al. (2024) proposes a dual framework that creates ideas and evaluates them in turns; Li et al. (2024a) optimizes LLMs for idea generation with post-training techniques, using a framework that blends Supervised Fine-Tuning (SFT) with Controllable Reinforcement Learning (RL). These diverse approaches demonstrated LLMs' ability to adapt to specialized domains and unify different discovery paradigms.

Challenges. One major challenge in automated hypothesis discovery is verification. As the nature of scientific discovery is to find novel knowledge that has not been verified by wet lab experiments, automatically evaluating the hypothesis is very challenging. Building accurate and well-structured benchmarks highly relies on experts, but the size of an expert-composed benchmark is usually very limited. In some disciplines, such as chemistry, even an expert's evaluation of the generated novel hypothesis is unreliable. This causes a need for automated experiments to verify the large-scale machine-generated hypotheses. Another challenge is the creativity of language models. Although proven to propose valid hypotheses, LLMs are known to have limited creativity Chakrabarty et al. (2024), given their next-token-prediction nature. Therefore, AI-generated hypotheses are often combinations of existing work or marginal improvements, which lack fundamental novelty.

4 Experiment Implementation

In addition to generating hypotheses, large language models (LLMs) are playing an increasingly vital role in scientific research by automating experimental design and enhancing workflow efficiency. With extensive built-in world knowledge, LLMs can make informed decisions in real-world contexts without requiring training on domain-specific data. To fully leverage their capabilities, LLMs are often structured as agents with two essential features (Kambhampati et al., 2024): modularity and tool integration. Modularity allows for smooth interaction with external systems such as databases, experimental platforms, and computational tools. At the same time, integration with specialized tools enables LLMs to function as central coordinators within research workflows, managing tasks like data access, computation, and experimental operations. This section focuses on how LLMs contribute to the strategic planning and execution of scientific research.

4.1 Experiment Design

Experimental design—the structured process of planning, organizing, and executing scientific investigations—is fundamental to producing reliable and meaningful results. Traditionally, human researchers have led this process, relying on domain expertise and intuition to define variables, control conditions, and select appropriate methodologies. However, human-led design is often constrained by cognitive biases, limited capacity to process vast datasets, and difficulty in managing the complexity of multifactorial experiments.

LLMs are reshaping this landscape by addressing many of these limitations. First, advanced prompting-based techniques such as Chain-of-Thought (Wei et al., 2023) and ReAct (Yao et al., 2022) are increasingly used in studies related to experiment design to enhance the accuracy in experiment workflows. Moreover, the growing capabilities of reflection and refinement (Madaan et al., 2023; Shinn et al., 2023) also allow LLMs to iteratively evaluate and refine the experimental plans. For example, previous studies tried to use LLMs to simulate expert discussions by letting LLMs engage in collaborative dialogue to challenge assumptions and analyze outputs (Li et al., 2024b).

With their abilities to analyze extensive datasets and generate insights at scale, LLMs support researchers in breaking down complex experimental questions into manageable sub-tasks, identifying optimal design strategies, and refining experimental structures (Boiko et al., 2023; M. Bran et al., 2024; Huang et al., 2024; Rasal & Hauer, 2024; Shen et al., 2023; Wu et al., 2023). For example, HuggingGPT Shen et al. (2023) uses LLMs to parse user queries into structured task lists while determining execution sequences and resource dependencies. Leveraging such capabilities of LLMs in scientific discovery has great potential in experiment design. In a recent work, ChemCrow (M. Bran et al., 2024) uses iterative reasoning and dynamic planning, using a structured framework to refine experimental approaches based on real-time feedback. ChemCrow (M. Bran et al., 2024) not only aids expert chemists and lowers barriers for non-experts but also fosters scientific advancement by bridging the gap between experimental and computational chemistry.

4.2 Experiment Execution

Experiment execution—the process of carrying out planned procedures to gather data and evaluate hypotheses—is a fundamental yet demanding phase of scientific research. It involves a range of meticulous tasks, from preparing data and managing protocols to conducting trials and recording results. Human-led execution often struggles with inefficiencies, errors, and the cognitive load of coordinating complex or large-scale experiments. Data preparation, which encompasses cleaning, labeling, and feature engineering, is notoriously time-consuming and labor-intensive. These limitations can hinder progress, reduce reproducibility, and constrain the scope of inquiry. Addressing these challenges requires methods that not only streamline execution but also enhance consistency, scalability, and interpretability across experimental workflows.

Previous studies have demonstrated that LLMs have great potential in dealing with such tasks that usually consume amounts of human effort. First, in terms of data preparation, LLMs can automate the process of data cleaning, data labeling, and even feature engineering (Chen et al., 2024; Zhang et al., 2024; Tan et al., 2024; Ziems et al., 2024). In addition, human researchers often face great challenges when data is difficult to obtain. LLMs can also be used to synthesize experimental data directly (Li et al., 2023; Liu et al., 2023a). Second, previous studies have also shown that LLMs can play a diverse role in automating experimental workflows to execute the experiments across different disciplines (M. Bran et al., 2024; Boiko et al., 2023; Wang et al., 2025; Ramos et al., 2023; Liu et al., 2023b; Ye et al., 2023; Rives et al., 2021; Lin et al., 2023) since they can acquire task-specific capabilities through pretraining, finetuning, and tool-augmented learning. For example, Coscientist (Boiko et al., 2023) uses LLMs to autonomously design, plan, and perform complex experiments by incorporating LLMs empowered by tools in experimental workflows, showcasing its potential for accelerating research across diverse tasks like the successful reaction optimization of palladium-catalyzed cross-couplings. Meanwhile, Wang et al. (Wang et al., 2025) demonstrate that the joint usage of LLMs with evolutionary algorithms yields superior performances by improving both the quality and the final solution and convergence speed, thereby reducing the number of required objective evaluations for many chemical experiments.

4.3 Challenges

LLMs show promise for experiment design and execution but face key challenges, especially in autonomous planning. However, they often lack the structured reasoning and domain-specific insight needed for scientific tasks, frequently hallucinate facts, and struggle with prompt sensitivity and ethical nuance (Kambhampati et al., 2024; Zhuo et al., 2023). Addressing these issues will require smarter modular architectures, real-time fact-checking, adaptive prompting, and the integration of expert reasoning. With these improvements, LLMs can evolve from text generators into reliable collaborators in complex, high-stakes scientific research.

5 Paper Writing

Large language models (LLMs) have seen widespread adoption in academic writing contexts, where they serve as valuable tools to support and enhance the scientific writing process. Researchers have actively investigated the use of LLMs across several core components of scholarly paper composition, including the automatic generation of citation contexts, the drafting of related work sections, and the overall structuring and writing of full manuscripts. These applications demonstrate the potential of LLMs to alleviate the cognitive and time-intensive aspects of writing, offering assistance in both content creation and stylistic refinement. This section delves into the specific roles that LLMs play in facilitating various academic writing tasks, highlighting the methodologies employed, the benefits they offer, and the limitations or challenges that remain in their integration into scientific authorship workflows.

5.1 Citation Generation

Citation text generation refers to the task of automatically crafting concise and contextually relevant summaries of referenced works within a citing paper. This process is crucial for synthesizing prior research in a coherent and informative manner. Traditionally, scholars must manually read, interpret, and integrate numerous sources—a time-consuming and cognitively demanding task that can lead to oversight, inconsistency, or superficial coverage of related work.

With recent advances in LLMs, previous studies have shown that they can enhance the efficiency, consistency, and depth of citation writing through their ability to process large volumes of information and maintain contextual accuracy (Xing et al., 2020; Li & Ouyang, 2024a; Wang et al., 2021; Ge et al., 2021; Gu & Hahnloser, 2024). For example, Li and Ouyang Li & Ouyang (2024a) prompt an LLM to generate a natural language description that emphasizes the relationships between pairs of papers in the citation network. Meanwhile, previous researchers have also developed models to produce rich citation texts. For example, AutoCite (Wang et al., 2021) is an automatic writing assistant model that not only infers potentially related work but also automatically generates the citation context at the same time. To create an efficient experience to support researchers, Gu et al. Gu & Hahnloser (2024) further integrate the manuscript context, the context of the referenced paper, and the desired control attributes into a structured template and use it to finetune the LLMs, providing humans with more control in citation generation with LLMs.

5.2 Literature Review Generation

Literature review generation in academic writing refers to the automated creation of coherent, contextually relevant summaries that synthesize findings from multiple scholarly sources. This task plays a central role in framing the background, significance, and intellectual lineage of a research work. However, conducting a comprehensive literature review is one of the most demanding aspects of scholarly writing—it requires significant time, effort, and cognitive load to identify, interpret, and integrate a wide array of research papers. Human researchers often struggle with information overload, limited memory, and unintentional biases, which can lead to incomplete or surface-level reviews.

Automated literature review generation, particularly with the aid of LLMs, can help address these limitations by processing large volumes of text, identifying thematic connections, and generating structured, fluent summaries. Studying literature review generation began before the development of LLMs (Hoang & Kan, 2010). Recently, researchers have developed case studies to explore the use of ChatGPT for literature review tasks and related work generation, showcasing its capabilities to assist researchers in completing these tasks (Zimmermann et al., 2024). Among current research in this field, retrieval-augmented generation (RAG), which enhances LLM-based literature review generation by grounding in factual content retrieved from external sources, is widely used to help address the challenges such as hallucinations (Agarwal et al., 2025; Hu et al., 2025; Shi et al., 2023; Susnjak et al., 2025; Yu et al., 2024). For example, LitLLM introduced a toolkit that operates on RAG principles, specialized prompting and instructing techniques with the help of LLMs, thus reducing the time and effort needed for comprehensive literature reviews while minimizing hallucinations (Agarwal et al., 2025).

5.3 Draft and Writing

Draft or manuscript writing involves the structured composition of scientific content—including definitions, explanations, and visual descriptions—into a coherent, publication-ready form. This process can be highly labor-intensive for researchers, requiring clarity, precision, and audience awareness. To ease this burden, researchers have also explored potential opportunities to use LLMs to assist humans in different draft writing settings. August et al. August et al. (2022) introduced a new task and dataset for defining scientific terms and controlling the complexity of generated definitions as a way of adapting academic writing to a specific reader's background knowledge. To augment authors' writing process, Hsu et al. (Hsu et al., 2021) automates the generation of captions for scientific figures, enabling qtime-consuminguick and accurate descriptions of visual data. Despite such specific techniques, researchers also developed holistic systems such as PaperRobot (Wang et al., 2019) to use LLMs to help organize and draft sections of papers based on user inputs. Similarly, CoAuthor (Lee et al., 2022) takes a human-AI collaborative approach to let LLMs help authors by generating suggestions and expanding text. In addition, AI Scientist (Lu et al., 2024) developed a broader system that integrates different workflows into the paper writing.

5.4 Challenges

While LLMs offer promise in academic writing, they face key challenges in maintaining factual accuracy, contextual coherence, and analytical depth. Issues like hallucinated citations, limited context windows, and shallow reasoning undermine the rigor required for scholarly work (Wang et al., 2024b). Ethical concerns—such as plagiarism and misrepresentation—further complicate their use (Li & Ouyang, 2024b). Addressing these problems calls for better retrieval systems, longer context handling, improved citation validation, and domain-specific fine-tuning. Integrating human oversight and enforcing clear ethical standards will be essential to ensure responsible, high-quality AI-assisted academic writing.

6 Peer Reviewing

As mentioned in 2, verifying and reviewing scientific findings is a highly specialized and time-consuming process. In order to free researchers from the timely peer review, researchers have utilized LLMs to assist them in evaluating papers, generating meta-reviews, detecting errors, and addressing ethical concerns.

There are two styles of building AI reviewers: utilizing a single LLM and building a system with multiple modules. In the single-LLM line of work, researchers find LLMs are capable of generating review comments by analyzing paper content against human-defined criteria, such as significance, methodological rigidity, and novelty. Based on that, more advanced reviewing methods are proposed: MetaGen Bhatia et al. (2020) first generates extractive summarization and then provide careful feedback; Kumar et al. (2021) trains a neural architecture for joint decision prediction and review generation; MReD Shen et al. (2021) introduced structure-controlled generation using sentence-level functional labels; ReviewRobot Wang et al. (2020) utilizes knowledge graphs to systematically identify and structure knowledge elements.

To provide high-quality feedback for longer and more complex research papers, researchers developed more sophisticated systems by leveraging multiple specialized models to handle different aspects of the review process. These architectures integrate diverse LLMs, each fine-tuned or prompted for specific subtasks, such as assessing novelty, summarizing reviewer comments, or detecting methodological errors. For instance, Zeng et al. (2024) proposes a framework where one LLM extracts key arguments from reviewer narratives while another synthesizes them into a structured meta-review, ensuring a balanced summary that captures critical feedback. Reviewer2 Gao et al. (2024) is another multi-LLM framework where one LLM generates a context-aware scoring rubric, and the other LLM provides targeted responses. Similarly, Kuznetsov et al. (2024) highlights the use of multi-model systems to cross-validate findings, where one model evaluates technical accuracy and another checks for ethical concerns, reducing biases inherent in single-model outputs. By distributing tasks across models, these architectures improve robustness and mitigate limitations like overgeneralization Yuan et al. (2022).

Additionally, LLMs can also assist in meta-review generation, where they synthesize multiple peer reviews into a cohesive summary. For instance, (Santu et al., 2024; Zeng et al., 2024) highlight LLMs' ability to produce structured meta-reviews by summarizing reviewer narratives, ensuring key points are captured accurately. LLMs' ability to aggregate multiple information sources is particularly useful for editors who need to consolidate diverse feedback.

Besides efficiency, LLM reviewers can also help reduce bias and unprofessional behaviors in human research by providing objective, standardized evaluations that minimize subjective influences often present in human reviews Cortes & Lawrence (2021); Goldberg et al. (2025). LLMs can also flag unprofessional conduct, such as derogatory comments or plagiarism, ensuring a more respectful and ethical review process.

Challenges. Although LLMs can provide high-quality feedback to research papers efficiently, some issues still exist, which limits the broader application of LLM reviewers. Although LLMs possess vast commonsense knowledge, reviewing papers often requires deep expertise and nuanced understanding, which exceeds the capability of current models. For example, LLMs are not good at theorem proving and mathematical calculations, which makes them fail to recognize subtle but critical assumptions in papers related to theoretical physics Zhou et al. (2024).

The complexity of academic writing also presents unique challenges, especially with longer documents. Although context windows in language models are growing, LLMs still have difficulty sustaining coherent analysis throughout lengthy texts, often losing the thread of intricate arguments that span several sections. This can lead to evaluations that are inconsistent or even contradictory Chamoun et al. (2024).

7 Limitations

In this survey, our objective is to provide a human-centered investigation of the potential capabilities of LLM agents for accelerating scientific discovery. Therefore, our analysis is based on the current limitations of human research, and our discussion is toward addressing such challenges across different stages of the research pipeline. However, we acknowledge that the general concept of "LLM for Science Discovery" is a huge topic. Our survey does not aim to provide a comprehensive view of LLM agents for scientific discovery in different fields, including mathematics, physics, and chemistry. Meanwhile, since our analysis is rooted in the potential capabilities of LLM agents in replicating the human research process, our goal is not to provide an overview of novel pathways in which LLM could reshape the research process.

8 Conclusion

Our survey provides a comprehensive examination of how Large Language Models (LLMs) are reshaping the entire scientific research workflow—from the early stages of hypothesis discovery and experiment implementation to paper writing and peer evaluation. We investigate the emerging opportunities and ongoing challenges associated with deploying LLMs in these contexts, shedding light on their current strengths, constraints, and broader impact on research efficiency. Starting from a human-centered point of view, we discuss current limitations in the human research process and then analyze potential ways LLM agents could address these challenges. Although LLMs offer novel tools to support and streamline diverse research activities, their application is still limited by technical, contextual, and ethical concerns. However, rapid progress in LLM development suggests a future where these models become integral to scientific inquiry, accelerating knowledge generation and enabling new forms of interdisciplinary collaboration.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H. Laradji, Krishnamurthy DJ Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. LitLLM: A Toolkit for Scientific Literature Review, March 2025. URL `http://arxiv.org/abs/2402.01788`. arXiv:2402.01788 [cs].

C Albajar, MG Albrow, OC Allkofer, A Astbury, B Aubert, T Axon, C Bacci, T Bacon, N Bains, JR Batley, et al. Events with large missing transverse energy at the cern collider: Iii. mass limits on supersymmetric particles. *Physics Letters B*, 198(2):261–270, 1987.

Tal August, Katharina Reinecke, and Noah A. Smith. Generating Scientific Definitions with Controllable Complexity. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8298–8317, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.569. URL `https://aclanthology.org/2022.acl-long.569/`.

Pierre Azoulay, Joshua Graff-Zivin, Brian Uzzi, Dashun Wang, Heidi Williams, James A. Evans, Ginger Zhe Jin, Susan Feng Lu, Benjamin F. Jones, Katy Börner, Karim R. Lakhani, Kevin J. Boudreau, and Eva C. Guinan. Toward a more scientific science. *Science*, 361(6408):1194–1197, September 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aav2484. URL `https://www.science.org/doi/10.1126/science.aav2484`.

Chaitanya Bhatia, Tribikram Pradhan, and Sukomal Pal. Metagen: An academic meta-review generation system. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1653–1656, 2020.

Elisabeth M. Bik. Publishing negative results is good for science. *Access Microbiology*, 6(4):000792, April 2024. ISSN 2516-8290. doi: 10.1099/acmi.0.000792. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11083460/`.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, December 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06792-0. URL `https://www.nature.com/articles/s41586-023-06792-0`. Publisher: Nature Publishing Group.

Nick Bostrom. Strategic implications of openness in ai development. In *Artificial intelligence safety and security*, pp. 145–164. Chapman and Hall/CRC, 2018.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–34, 2024.

Eric Chamoun, Michael Schlichktrull, and Andreas Vlachos. Automated focused feedback generation for scientific writing assistance. *arXiv preprint arXiv:2405.20477*, 2024.

Daoyuan Chen, Yilun Huang, Zhijian Ma, Hesen Chen, Xuchen Pan, Ce Ge, Dawei Gao, Yuexiang Xie, Zhaoyang Liu, Jinyang Gao, Yaliang Li, Bolin Ding, and Jingren Zhou. Data-Juicer: A One-Stop Data Processing System for Large Language Models. In *Companion of the 2024 International Conference on Management of Data*, SIGMOD/PODS '24, pp. 120–134, New York, NY, USA, June 2024. Association for Computing Machinery. ISBN 9798400704222. doi: 10.1145/3626246.3653385. URL `https://doi.org/10.1145/3626246.3653385`.

Christina Clark-Kazak. Developing ethical guidelines for research. *Forced Migration Review*, (61), 2019.

Corinna Cortes and Neil D Lawrence. Inconsistency in conference peer review: Revisiting the 2014 neurips experiment. *arXiv preprint arXiv:2109.09774*, 2021.

Marc A. Edwards and Siddhartha Roy. Academic Research in the 21st Century: Maintaining Scientific Integrity in a Climate of Perverse Incentives and Hypercompetition. *Environmental Engineering Science*, 34(1):51–61, January 2017. ISSN 1092-8758. doi: 10.1089/ees.2016.0223. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5206685/`.

Phyllis Freeman and Anthony Robbins. Closing the 'publishing gap' between rich and poor. *SciDev. net-Communication*, 2005.

Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. Reviewer2: Optimizing review generation through prompt generation. *arXiv preprint arXiv:2402.10886*, 2024.

Yubin Ge, Ly Dinh, Xiaofeng Liu, Jinsong Su, Ziyao Lu, Ante Wang, and Jana Diesner. BACO: A Background Knowledge- and Content-Based Framework for Citing Sentence Generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1466–1478, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.116. URL `https://aclanthology.org/2021.acl-long.116/`.

Alexander Goldberg, Ivan Stelmakh, Kyunghyun Cho, Alice Oh, Alekh Agarwal, Danielle Belgrave, and Nihar B Shah. Peer reviews of peer reviews: A randomized controlled trial and other experiments. *PloS one*, 20(4):e0320444, 2025.

Christine Grady. Institutional Review Boards. *Chest*, 148(5):1148–1155, November 2015. ISSN 0012-3692. doi: 10.1378/chest.15-0706. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4631034/`.

Nianlong Gu and Richard Hahnloser. Controllable Citation Sentence Generation with Language Models. In Tirthankar Ghosal, Amanpreet Singh, Anita Waard, Philipp Mayr, Aakanksha Naik, Orion Weller, Yoonjoo Lee, Shannon Shen, and Yanxia Qin (eds.), *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pp. 22–37, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL `https://aclanthology.org/2024.sdp-1.4/`.

Odd Erik Gundersen, Yolanda Gil, and David W Aha. On reproducible ai: Towards reproducible research, open science, and digital scholarship in ai publications. *AI magazine*, 39(3):56–68, 2018.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Amanda Heidt. Racial inequalities in journals highlighted in giant study. *Nature*, 2023.

Cong Duy Vu Hoang and Min-Yen Kan. Towards Automated Related Work Summarization. In Chu-Ren Huang and Dan Jurafsky (eds.), *Coling 2010: Posters*, pp. 427–435, Beijing, China, August 2010. Coling 2010 Organizing Committee. URL https://aclanthology.org/C10-2049/.

Ting-Yao Hsu, C Lee Giles, and Ting-Hao Huang. SciCap: Generating Captions for Scientific Figures. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3258–3264, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.277. URL https://aclanthology.org/2021.findings-emnlp.277/.

Yuntong Hu, Zhuofeng Li, Zheng Zhang, Chen Ling, Raasikh Kanjiani, Boxin Zhao, and Liang Zhao. Taxonomy Tree Generation from Citation Graph, February 2025. URL http://arxiv.org/abs/2410.03761. arXiv:2410.03761 [cs].

Kaixuan Huang, Yuanhao Qu, Henry Cousins, William A. Johnson, Di Yin, Mihir Shah, Denny Zhou, Russ Altman, Mengdi Wang, and Le Cong. CRISPR-GPT: An LLM Agent for Automated Design of Gene-Editing Experiments, April 2024. URL http://arxiv.org/abs/2404.18021. arXiv:2404.18021 [cs].

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Saldyt, and Anil Murthy. LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks, June 2024. URL http://arxiv.org/abs/2402.01817. arXiv:2402.01817 [cs].

Christian Kübel, Andreas Voigt, Remco Schoenmakers, Max Otten, David Su, Tan-Chen Lee, Anna Carlsson, and John Bradley. Recent advances in electron tomography: Tem and haadf-stem tomography for materials science and semiconductor applications. *Microscopy and Microanalysis*, 11(5):378–400, 2005.

Asheesh Kumar, Tirthankar Ghosal, and Asif Ekbal. A deep neural architecture for decision-aware meta-review generation. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 222–225. IEEE, 2021.

Ilia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, et al. What can natural language processing do for peer review? *arXiv preprint arXiv:2405.06563*, 2024.

Diana Kwon. Open-access publishing fees deter researchers in the global south. *Nature*, February 2022. doi: 10.1038/d41586-022-00342-w. URL https://www.nature.com/articles/d41586-022-00342-w. Bandiera_abtest: a Cg_type: News Publisher: Nature Publishing Group Subject_term: Developing world, Publishing.

Mina Lee, Percy Liang, and Qian Yang. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pp. 1–19, New York, NY, USA, April 2022. Association for Computing Machinery. ISBN 978-1-4503-9157-3. doi: 10.1145/3491102.3502030. URL https://dl.acm.org/doi/10.1145/3491102.3502030.

Ruochen Li, Liqiang Jing, Chi Han, Jiawei Zhou, and Xinya Du. Learning to generate research idea with dynamic control. *arXiv preprint arXiv:2412.14626*, 2024a.

Ruosen Li, Ziming Luo, and Xinya Du. FG-PRM: Fine-grained Hallucination Detection and Mitigation in Language Model Mathematical Reasoning, November 2024b. URL http://arxiv.org/abs/2410.06304. arXiv:2410.06304 [cs].

Siyu Li, Jin Yang, and Kui Zhao. Are you in a Masquerade? Exploring the Behavior and Impact of Large Language Model Driven Social Bots in Online Social Networks, July 2023. URL http://arxiv.org/abs/2307.10337. arXiv:2307.10337 [cs].

Xiangci Li and Jessica Ouyang. Explaining Relationships Among Research Papers, February 2024a. URL http://arxiv.org/abs/2402.13426. arXiv:2402.13426 [cs].

Xiangci Li and Jessica Ouyang. Related Work and Citation Text Generation: A Survey, April 2024b. URL https://arxiv.org/abs/2404.11588v1.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/10.1126/science.ade2574. Publisher: American Association for the Advancement of Science.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M. Dai, Diyi Yang, and Soroush Vosoughi. Training Socially Aligned Language Models on Simulated Social Interactions, October 2023a. URL http://arxiv.org/abs/2305.16960. arXiv:2305.16960 [cs].

Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. Conversational Drug Editing Using Retrieval and Domain Feedback. October 2023b. URL https://openreview.net/forum?id=yRrPfKyJQ2.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery, September 2024. URL http://arxiv.org/abs/2408.06292. arXiv:2408.06292 [cs].

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, May 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00832-8. URL https://www.nature.com/articles/s42256-024-00832-8. Publisher: Nature Publishing Group.

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

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative Refinement with Self-Feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, December 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

Ramal Moonesinghe, Muin J Khoury, and A Cecile J W Janssens. Most published research findings are false—but a little replication goes a long way. *PLoS medicine*, 4(2):e28, 2007.

John D Norton. A little survey of induction. 2003.

Brian A. Nosek, Jeffrey R. Spies, and Matt Motyl. Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on psychological science : a journal of the Association for Psychological Science*, 7(6):615–631, November 2012. ISSN 1745-6916. doi: 10.1177/1745691612459058. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10540222/.

Kevin Pu, KJ Feng, Tovi Grossman, Tom Hope, Bhavana Dalvi Mishra, Matt Latzke, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. Ideasynth: Iterative research idea development through evolving and composing idea facets with literature-grounded feedback. *arXiv preprint arXiv:2410.04025*, 2024.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*, 2023.

Linlu Qiu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, et al. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. *arXiv preprint arXiv:2310.08559*, 2023.

Mayk Caldas Ramos, Shane S. Michtavy, Marc D. Porosoff, and Andrew D. White. Bayesian Optimization of Catalysts With In-context Learning, April 2023. URL `http://arxiv.org/abs/2304.05341`. arXiv:2304.05341 [physics].

Sumedh Rasal and E. J. Hauer. Navigating Complexity: Orchestrated Problem Solving with Multi-Agent LLMs, July 2024. URL `http://arxiv.org/abs/2402.16713`. arXiv:2402.16713 [cs].

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, April 2021. doi: 10.1073/pnas.2016239118. URL `https://www.pnas.org/doi/full/10.1073/pnas.2016239118`. Publisher: Proceedings of the National Academy of Sciences.

Shubhra Kanti Karmaker Santu, Sanjeev Kumar Sinha, Naman Bansal, Alex Knipper, Souvika Sarkar, John Salvador, Yash Mahajan, Sri Guttikonda, Mousumi Akter, Matthew Freestone, et al. Prompting llms to compose meta-review drafts from peer-review narratives of scholarly manuscripts. *arXiv preprint arXiv:2402.15589*, 2024.

Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. Mred: A meta-review dataset for structure-controllable text generation. *arXiv preprint arXiv:2110.07474*, 2021.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. *Advances in Neural Information Processing Systems*, 36:38154–38180, December 2023. URL `https://papers.nips.cc/paper_files/paper/2023/hash/77c33e6a367922d003ff102ffb92b658-Abstract-Conference.html`.

Zhengliang Shi, Shen Gao, Zhen Zhang, Xiuying Chen, Zhumin Chen, Pengjie Ren, and Zhaochun Ren. Towards a Unified Framework for Reference Retrieval and Related Work Generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5785–5799, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.385. URL `https://aclanthology.org/2023.findings-emnlp.385/`.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, December 2023. URL `https://papers.nips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html`.

Daniel J Simons. The value of direct replication. *Perspectives on psychological science*, 9(1):76–80, 2014.

Bernd Carsten Stahl and Bernd Carsten Stahl. Ethical issues of ai. *Artificial Intelligence for a better future: An ecosystem perspective on the ethics of AI and emerging digital technologies*, pp. 35–53, 2021.

Adam F Stewart, Donna M Ferriero, S Andrew Josephson, Daniel H Lowenstein, Robert O Messing, Jorge R Oksenberg, S Claiborne Johnston, and Stephen L Hauser. Fighting decision fatigue, 2012.

Teo Susnjak, Peter Hwang, Napoleon H. Reyes, Andre L. C. Barczak, Timothy R. McIntosh, and Surangika Ranathunga. Automating Research Synthesis with Domain-Specific Large Language Model Fine-Tuning. *ACM Transactions on Knowledge Discovery from Data*, 19(3):1–39, April 2025. ISSN 1556-4681, 1556-472X. doi: 10.1145/3715964. URL `http://arxiv.org/abs/2404.08680`. arXiv:2404.08680 [cs].

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

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

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large Language Models for Data Annotation and Synthesis: A Survey, December 2024. URL `http://arxiv.org/abs/2402.13446`. arXiv:2402.13446 [cs].

Jennifer S. Trueblood, David B. Allison, Sarahanne M. Field, Ayelet Fishbach, Stefan D. M. Gaillard, Gerd Gigerenzer, William R. Holmes, Stephan Lewandowsky, Dora Matzke, Mary C. Murphy, Sebastian Musslick, Vencislav Popov, Adina L. Roskies, Judith ter Schure, and Andrei R. Teodorescu. The misalignment of incentives in academic publishing and implications for journal reform. *Proceedings of the National Academy of Sciences*, 122(5):e2401231121, February 2025. doi: 10.1073/pnas.2401231121. URL `https://www.pnas.org/doi/10.1073/pnas.2401231121`. Publisher: Proceedings of the National Academy of Sciences.

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98, 2019.

Kathleen D Vohs, Roy F Baumeister, Jean M Twenge, Brandon J Schmeichel, Dianne M Tice, and Jennifer Crocker. Decision fatigue exhausts self-regulatory resources—but so does accommodating to unchosen alternatives. *Manuscript submitted for publication*, 1:1–55, 2005.

Kathleen D Vohs, Roy F Baumeister, Brandon J Schmeichel, Jean M Twenge, Noelle M Nelson, and Dianne M Tice. Making choices impairs subsequent self-control: A limited-resource account of decision making, self-regulation, and active initiative. In *Self-regulation and self-control*, pp. 45–77. Routledge, 2018.

Haorui Wang, Marta Skreta, Cher-Tian Ser, Wenhao Gao, Lingkai Kong, Felix Strieth-Kalthoff, Chenru Duan, Yuchen Zhuang, Yue Yu, Yanqiao Zhu, Yuanqi Du, Alán Aspuru-Guzik, Kirill Neklyudov, and Chao Zhang. Efficient Evolutionary Search Over Chemical Space with Large Language Models, March 2025. URL `http://arxiv.org/abs/2406.16976`. arXiv:2406.16976 [cs].

Qingqin Wang, Yun Xiong, Yao Zhang, Jiawei Zhang, and Yangyong Zhu. AutoCite: Multi-Modal Representation Fusion for Contextual Citation Generation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, pp. 788–796, New York, NY, USA, March 2021. Association for Computing Machinery. ISBN 978-1-4503-8297-7. doi: 10.1145/3437963.3441739. URL `https://dl.acm.org/doi/10.1145/3437963.3441739`.

Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. Paperrobot: Incremental draft generation of scientific ideas. *arXiv preprint arXiv:1905.07870*, 2019.

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. Reviewrobot: Explainable paper review generation based on knowledge synthesis. *arXiv preprint arXiv:2010.06119*, 2020.

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

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. AutoSurvey: Large Language Models Can Automatically Write Surveys, June 2024b. URL `http://arxiv.org/abs/2406.10252`. arXiv:2406.10252 [cs].

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, January 2023. URL `http://arxiv.org/abs/2201.11903`. arXiv:2201.11903 [cs].

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

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. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation, October 2023. URL `http://arxiv.org/abs/2308.08155`. arXiv:2308.08155 [cs].

Xinyu Xing, Xiaosheng Fan, and Xiaojun Wan. Automatic Generation of Citation Texts in Scholarly Papers: A Pilot Study. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6181–6190, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.550. URL `https://aclanthology.org/2020.acl-main.550/`.

Neel Yadav, Saumya Pandey, Amit Gupta, Pankhuri Dudani, Somesh Gupta, and Krithika Rangarajan. Data privacy in healthcare: In the era of artificial intelligence. *Indian Dermatology Online Journal*, 14(6): 788–792, 2023.

Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. Language models as inductive reasoners. *arXiv preprint arXiv:2212.10923*, 2022.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. Large language models for automated open-domain scientific hypotheses discovery. *arXiv preprint arXiv:2309.02726*, 2023.

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

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. September 2022. URL `https://openreview.net/forum?id=WE_vluYUL-X`.

Geyan Ye, Xibao Cai, Houtim Lai, Xing Wang, Junhong Huang, Longyue Wang, Wei Liu, and Xiangxiang Zeng. DrugAssist: A Large Language Model for Molecule Optimization, December 2023. URL `http://arxiv.org/abs/2401.10334`. arXiv:2401.10334 [q-bio].

Luyao Yu, Qi Zhang, Chongyang Shi, An Lao, and Liang Xiao. Reinforced Subject-Aware Graph Neural Network for Related Work Generation. In *Knowledge Science, Engineering and Management: 17th International Conference, KSEM 2024, Birmingham, UK, August 16–18, 2024, Proceedings, Part I*, pp. 201–213, Berlin, Heidelberg, August 2024. Springer-Verlag. ISBN 978-981-9754-91-5. doi: 10.1007/978-981-97-5492-2_16. URL `https://doi.org/10.1007/978-981-97-5492-2_16`.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212, 2022.

Qi Zeng, Mankeerat Sidhu, Ansel Blume, Hou Pong Chan, Lu Wang, and Heng Ji. Scientific opinion summarization: Paper meta-review generation dataset, methods, and evaluation. In *Artificial Intelligence for Research and Democracy: First International Workshop, AI4Research 2024, and 4th International Workshop, DemocrAI 2024, Held in Conjunction with IJCAI 2024, Jeju, South Korea, August 5, 2024, Proceedings*, pp. 20. Springer Nature, 2024.

Haochen Zhang, Yuyang Dong, Chuan Xiao, and Masafumi Oyamada. Jellyfish: A Large Language Model for Data Preprocessing, October 2024. URL `http://arxiv.org/abs/2312.01678`. arXiv:2312.01678 [cs].

Zhouyang Zhang, Yangbo Zhou, Xinli Zhu, Linfeng Fei, Haitao Huang, and Yu Wang. Applications of esem on materials science: Recent updates and a look forward. *Small Methods*, 4(2):1900588, 2020.

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

Terry Yue Zhuo, Zhuang Li, Yujin Huang, Fatemeh Shiri, Weiqing Wang, Gholamreza Haffari, and Yuan-Fang Li. On Robustness of Prompt-based Semantic Parsing with Large Pre-trained Language Model: An Empirical Study on Codex. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1090–1102, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.77. URL https://aclanthology.org/2023.eacl-main.77/.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, 50(1):237–291, March 2024. ISSN 0891-2017. doi: 10.1162/coli_a_00502. URL https://doi.org/10.1162/coli_a_00502.

Robert Zimmermann, Marina Staab, Mehran Nasseri, and Patrick Brandtner. Leveraging Large Language Models for Literature Review Tasks - A Case Study Using ChatGPT. In Teresa Guarda, Filipe Portela, and Jose Maria Diaz-Nafria (eds.), *Advanced Research in Technologies, Information, Innovation and Sustainability*, pp. 313–323, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-48858-0. doi: 10.1007/978-3-031-48858-0_25.