From Hypothesis to Publication: A Comprehensive Survey of AI-Driven Research Support Systems

Zekun Zhou[1], **Xiaocheng Feng**[1,2*], **Lei Huang**[1], **Xiachong Feng**[3*],
Ziyun Song[1], **Ruihan Chen**[1], **Liang Zhao**[1], **Weitao Ma**[1], **Yuxuan Gu**[1],
Baoxin Wang[4], **Dayong Wu**[4], **Guoping Hu**[4], **Ting Liu**[1], **Bing Qin**[1,2]
[1] Harbin Institute of Technology [2] Peng Cheng Laboratory
[3] The University of Hong Kong [4] iFLYTEK Research
{zkzhou,xcfeng,lhuang,yxgu,qinb}@ir.hit.edu.cn
fengxc@hku.hk {bxwang2,dywu2,gphu}@iflytek.com

Abstract

Research is a fundamental process driving the advancement of human civilization, yet it demands substantial time and effort from researchers. In recent years, the rapid development of artificial intelligence (AI) technologies has inspired researchers to explore how AI can accelerate and enhance research. To monitor relevant advancements, this paper presents a systematic review of the progress in this domain. Specifically, we organize the relevant studies into three main categories: hypothesis formulation, hypothesis validation, and manuscript publication. Hypothesis formulation involves knowledge synthesis and hypothesis generation. Hypothesis validation includes the verification of scientific claims, theorem proving, and experiment validation. Manuscript publication encompasses manuscript writing and the peer review process. Furthermore, we identify and discuss the current challenges faced in these areas, as well as potential future directions for research. Finally, we also offer a comprehensive overview of existing benchmarks and tools across various domains that support the integration of AI into the research process. We hope this paper serves as an introduction for beginners and fosters future research. Resources have been made publicly available[1].

1 Introduction

Research is creative and systematic work aimed at expanding knowledge and driving civilization's development (Eurostat, 2018). Researchers typically identify a topic, review relevant literature, synthesize existing knowledge, and formulate hypothesis, which are validated through theoretical and experimental methods. Findings are then documented in manuscripts that undergo peer review before publication (Benos et al., 2007; Boyko et al., 2023). However, this process is resource-intensive, requiring specialized expertise and posing entry barriers for researchers (Blaxter et al., 2010).

In recent years, artificial intelligence (AI) technologies, represented by large language models (LLMs), have experienced rapid development (Brown et al., 2020; OpenAI, 2023; Dubey et al., 2024; Yang et al., 2024a; DeepSeek-AI et al., 2024; Guo et al., 2025). These models exhibit exceptional capabilities in text understanding, reasoning, and generation (Schaeffer et al., 2023). In this context, AI is increasingly involving the entire research pipeline (Messeri and Crockett, 2024), sparking extensive discussion about its implications for research (Hutson, 2022; Williams et al., 2023; Morris, 2023; Fecher et al., 2023). Moreover, following the release of ChatGPT, approximately 20% of academic papers and peer-reviewed texts in certain fields have been modified by LLMs (Liang et al., 2024a,b). A study also reveals that 81% of researchers integrate LLMs into their workflows (Liao et al., 2024).

As the application of AI in research attracts increasing attention, a significant body of related studies has begun to emerge. To systematically synthesize existing research, we present comprehensive survey that emulates human researchers by using the research process as an organizing framework. Specifically, as depicted in Figure 1, the research process is divided into three key stages: (1) Hypothesis Formulation, involving knowledge synthesis and hypothesis generation; (2) Hypothesis Validation, encompassing scientific claim verification, theorem proving, and experiment validation; (3) Manuscript Publication, which focuses on academic publications and is further divided into manuscript writing and peer review.

Comparing with Existing Surveys Although Luo et al. (2025) reviews the application of AI in research, it predominantly focuses on LLMs, while neglecting the knowledge synthesis

*Co-Corresponding Author

[1] https://github.com/zkzhou126/AI-for-Research

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 11773–11803

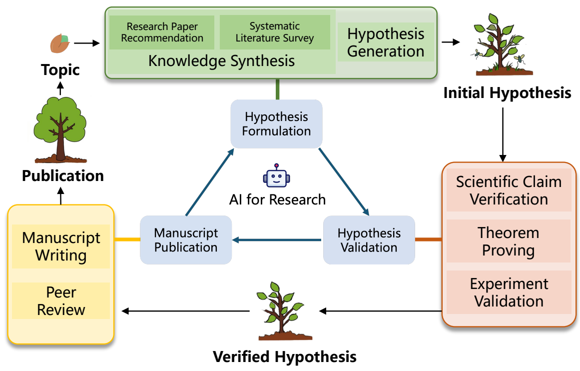

Figure 1: Overview of AI for research. The framework consists of three stages: hypothesis formulation, hypothesis validation, and manuscript publication. In the hypothesis formulation stage, knowledge integration leads to the proposal of an initial hypothesis after a topic is identified. The hypothesis validation stage involves verifying the hypothesis from three perspectives to ensure its correctness and validity. Finally, the manuscript publication stage focuses on drafting and publishing the validated hypothesis.

that precedes hypothesis generation and the theoretical validation of hypothesis. Other surveys concentrate on more specific areas, such as paper recommendation (Beel et al., 2016; Bai et al., 2019; Kreutz and Schenkel, 2022), scientific literature review (Altmami and Menai, 2022), hypothesis generation (Kulkarni et al., 2025), scientific claim verification (Vladika and Matthes, 2023; Dmonte et al., 2024), theorem proving (Li et al., 2024e), manuscript writing (Li and Ouyang, 2024), and peer review (Lin et al., 2023a; Kousha and Thelwall, 2024). Additionally, certain surveys emphasize the application of AI in scientific domains (Zheng et al., 2023b; Zhang et al., 2024d; Gridach et al., 2025).

Contributions Our contributions can be summarized as follows: (1) We align the relevant fields with the research process of human researchers, systematically integrating and extending these aspects while primarily focusing on the research process itself. (2) We introduce a meticulous taxonomy (shown in Figure 2). (3) We provide a summary of tools that can assist in the research process. (4) We discuss new frontiers, outline their challenges, and shed light on future research.

Survey Organization We first elaborate hypothesis formulation (§2), followed by hypothesis validation (§3) and manuscript publication (§4). Additionally, we present benchmarks (§5), and tools (§6) that utilized in research. Finally, we outline challenges as well as future directions (§7), and discusssion about relevant ethical considerations (§8). In the Appendix, we provide further discussion on open questions (§A), challenges faced in different domains (§B), and a comparison of capabilities among different methods (§C).

2 Hypothesis Formulation

This stage centers on the process of hypothesis formulation. As illustrated in Figure 3, it commences with developing a comprehensive understanding of the domain, followed by identifying a specific aspect and generating pertinent hypothesis. This section is further structured into two key components: Knowledge Synthesis and Hypothesis Generation.

2.1 Knowledge Synthesis

Knowledge synthesis constitutes the foundational step in the research process. During this phase, researchers are required to identify and critically evaluate existing literature to establish a thorough understanding of the field. This step is pivotal for uncovering new research directions, refining methodologies, and supporting evidence-based decision-making (Asai et al., 2024). In this section, the process of knowledge synthesis is divided into two modules: Research Paper Recommendation and Systematic Literature Review.

2.1.1 Research Paper Recommendation

Research paper recommendation (RPR) identifies and recommends novel and seminal articles aligned with researchers' interests. Prior studies have shown that recommendation systems outperform keyword-based search engines in terms of efficiency and reliability when extracting valuable insights from large-scale datasets (Bai et al., 2019). Existing methodologies are broadly categorized into four paradigms: content-based filtering, collaborative filtering, graph-based approaches, and hybrid systems (Beel et al., 2016; Li and Zou, 2019; Bai et al., 2019; Shahid et al., 2020). Recent advancements propose multi-dimensional classification frameworks based on data source utilization (Kreutz and Schenkel, 2022).

Recent trends in research suggest a decline in publication volumes related to RPR (Sharma et al., 2023), alongside an increasing focus on user-centric optimizations. Existing studies emphasize the limitations of traditional paper-centric interaction models and advocate for more effective utilization of author relationship graphs (Kang et al.,

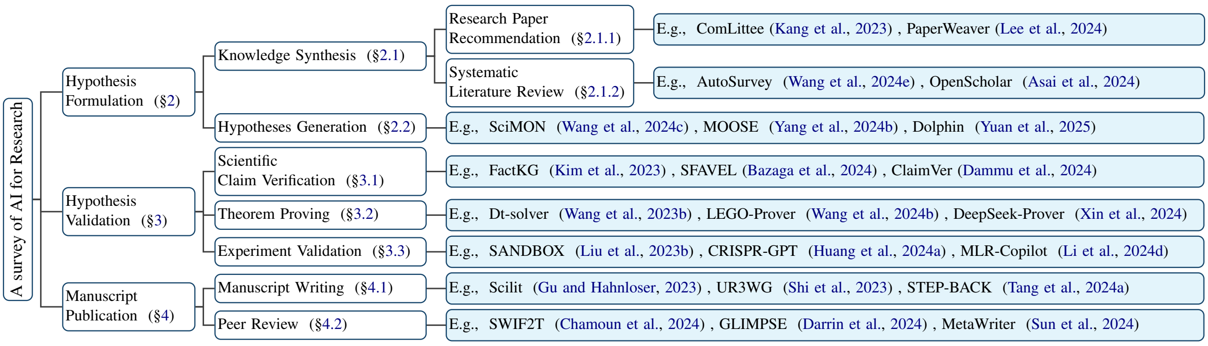

Figure 2: Taxonomy of Hypothesis Formulation, Hypothesis Validation and Manuscript Publication (This is a simplified version, full version in Figure 6).

2023). Multi-stage recommendation architectures that integrate diverse methodologies have been shown to achieve superior performance (Pinedo et al., 2024; Stergiopoulos et al., 2024). Visualization techniques that link recommended papers to users' publication histories via knowledge graphs (Kang et al., 2022) and LLMs-powered comparative analysis frameworks (Lee et al., 2024) represent emerging directions for enhancing interpretability and contextual relevance.

2.1.2 Systematic Literature Review

Systematic literature review (SLR) constitutes a rigorous and structured methodology for evaluating and integrating prior research on a specific topic (Webster and Watson, 2002; Zhu et al., 2023; Bolaños et al., 2024). In contrast to single-document summaries (Elhadad et al., 2005), SLR entails synthesizing information across multiple related scientific documents (Altmami and Menai, 2022). SLR can further be divided into two stages: outline generation and full-text generation (Shao et al., 2024; Agarwal et al., 2024b; Block and Kuckertz, 2024).

Outline generation, especially structured outline generation, is highlighted by recent studies as a pivotal factor in enhancing the quality of surveys. Zhu et al. (2023) demonstrated that hierarchical frameworks substantially enhance survey coherence. AutoSurvey (Wang et al., 2024e) extended conventional outline generation by recommending both sub-chapter titles and detailed content descriptions, ensuring comprehensive topic coverage. Additionally, multi-level topic generation via clustering methodologies has been proposed as an effective strategy for organizing survey structures (Katz et al., 2024). Advanced systems such as STORM (Shao et al., 2024) employed LLM-driven outline drafting combined with multi-agent discussion cycles to iteratively refine the generated outlines. Tree-based hierarchical architectures have gained increasing adoption in this domain. For instance, CHIME (Hsu et al., 2024) optimized LLM-generated hierarchies through human-AI collaboration, while HiReview (Hu et al., 2024b) demonstrated the efficacy of multi-layer tree representations for systematic knowledge organization.

Full-text generation follows the outline generation stage. AutoSurvey and Lai et al. (2024) utilized LLMs with carefully designed prompts to construct comprehensive literature reviews step-by-step. PaperQA2 (Skarlinski et al., 2024) introduced an iterative agent-based approach for generating reviews, while STORM employed multi-agent conversation data for this purpose. LitLLM (Agarwal et al., 2024a) and Agarwal et al. (2024b) adopted a plan-based search enhancement strategy. KGSum (Wang et al., 2022a) integrated knowledge graph information into paper encoding and used a two-stage decoder for summary generation. Bio-SIEVE (Robinson et al., 2023) and Susnjak et al. (2024) fine-tuned LLMs for automatic review generation. OpenScholar (Asai et al., 2024) developed a pipeline that trained a new model without relying on a dedicated survey-generation model.

2.2 Hypothesis Generation

Hypothesis generation, known as idea generation, refers to the process of coming up with new concepts, solutions, or approaches. It is the most important step in driving the progress of the entire research (Qi et al., 2023).

Early work focused more on predicting relationships between concepts, because researchers believed that new concepts come from links with old concepts (Henry and McInnes, 2017; Krenn et al., 2022). As language models became more powerful (Zhao et al., 2023a), researchers are beginning to focus on open-ended idea generation

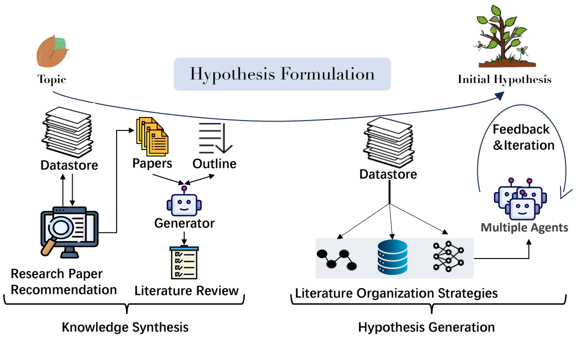

Figure 3: This figure illustrates the hypothesis formulation process, consisting of two stages: knowledge synthesis and hypothesis generation, which together produce an initial hypothesis related to a specific topic.

(Girotra et al., 2023; Si et al., 2024; Kumar et al., 2024). Recent advancements in AI-driven hypothesis generation highlight diverse approaches to research conceptualization. For instance, MOOSE-Chem (Yang et al., 2024c) and IdeaSynth (Pu et al., 2024) used LLMs to bridge inspiration-to-hypothesis transformation via interactive frameworks. The remaining research primarily falls into two areas: enhancing input data quality and improving the quality of generated hypothesis.

Input data quality improvement is demonstrated by Majumder et al. (2024a); Liu et al. (2024a), who showed that LLMs can generate comprehensive hypothesis from existing academic data. Literature organization strategies have evolved through various methodologies, including triplet representations (Wang et al., 2024c), chain-based architectures (Li et al., 2024a), and complex database systems (Wang et al., 2024d). Knowledge graphs emerge as critical infrastructure (Hogan et al., 2021), enabling semantic relationship mapping via subgraph identification (Buehler, 2024; Ghafarollahi and Buehler, 2024). Notably, SciMuse (Gu and Krenn, 2024) pioneered researcher-specific hypothesis generation by constructing personalized knowledge graphs.

Hypothesis quality improvement has been addressed through feedback and iteration (Rabby et al., 2025), as proposed by HypoGeniC (Zhou et al., 2024) and MOOSE (Yang et al., 2024b). Feedback mechanisms include direct responses to hypothesis (Baek et al., 2024), experimental outcome evaluations (Ma et al., 2024; Yuan et al., 2025), comparison with the existing literature (Schmidgall and Moor, 2025), and automated peer review comments (Lu et al., 2024). FunSearch (Romera-Paredes et al., 2024) generates solutions by iteratively combining the innovative capabilities of LLM with the verification capabilities of an evaluator. Beyond iterative feedback, collaborative efforts among researchers have also been recognized, leading to multi-agent hypothesis generation approaches (Nigam et al., 2024; Ghafarollahi and Buehler, 2024). VIRSCI (Su et al., 2024) further optimized this process by customizing knowledge for each agent. Additionally, the Nova framework (Hu et al., 2024a) was introduced to refine hypothesis by leveraging outputs from other research as input.

Knowledge synthesis and hypothesis generation comprise the hypothesis formulation phase. Research paper recommendation supports knowledge acquisition, while systematic literature review aids organization within knowledge synthesis. Recent advances integrate LLMs (de la Torre-López et al., 2023) to enhance knowledge integration (Huang and Tan, 2023; Gupta et al., 2023; Kacena et al., 2024; Tang et al., 2024b). By developing a deep understanding of a domain through knowledge synthesis, researchers can identify research directions and use hypothesis generation techniques to formulate hypothesis. Additionally, the distinction between scientific discovery and hypothesis generation is discussed in §A.

3 Hypothesis Validation

In scientific research, any proposed hypothesis must undergo rigorous validation to establish its validity. In some studies, this process is also referred to as 'falsification' (Liu et al., 2024d; Huang et al., 2025). As illustrated in Figure 4, this section explores the application of AI in verifying scientific hypothesis through three methodological components: Scientific Claim Verification, Theorem Proving, and Experiment Validation.

3.1 Scientific Claim Verification

Scientific claim verification, also referred to as scientific fact-checking or scientific contradiction detection, aims to assess the veracity of claims related to scientific knowledge. This process assists scientists in verifying research hypothesis, discovering evidence, and advancing scientific work (Wadden et al., 2020; Vladika and Matthes, 2023; Skarlinski et al., 2024). Research on scientific claim verification primarily focuses on three key elements: the claim, the evidence, and the validity of the claim

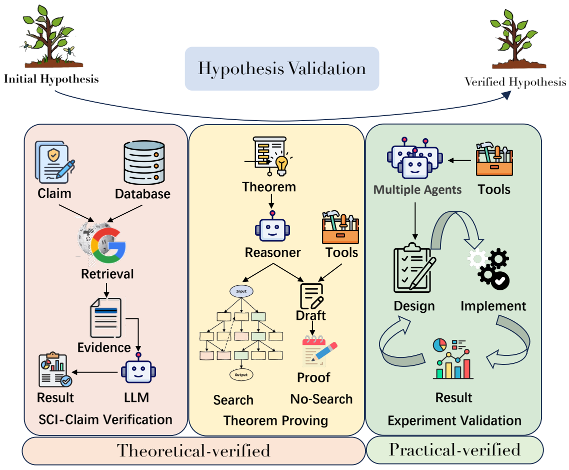

Figure 4: This figure illustrates the various perspectives for hypothesis validation during the hypothesis validation stage. A hypothesis is typically divided into scientific claims and theorems, with SCI-claim verification (scientific claim verification) and theorem proving ensuring theoretical correctness, while experiment validation assesses practical feasibility.

(Dmonte et al., 2024).

Claim Studies have highlighted that certain claims lack supporting evidence (Wührl et al., 2024a), while others have demonstrated the ability to perform claim-evidence alignment without annotated data (Bazaga et al., 2024). Additionally, methods such as HiSS (Zhang and Gao, 2023) and ProToCo (Zeng and Gao, 2023) proposed generating multiple claim variants to enhance verification.

Evidence Existing research has explored various aspects related to evidence, including evidentiary sources (Vladika and Matthes, 2024a), retrieval configurations (Vladika and Matthes, 2024b), strategies for identifying and mitigating flawed evidence (Glockner et al., 2022; Wührl et al., 2024b; Glockner et al., 2024a), and approaches for processing sentence-level (Pan et al., 2023b) versus document-level indicators (Wadden et al., 2022b).

Verification In the verification results generation phase, MAGIC (Kao and Yen, 2024) and SERIf (Cao et al., 2024b) proposed utilizing LLMs to synthesize evidence into more comprehensive information. FactKG (Kim et al., 2023) and Muharram and Purwarianti (2024) structured evidence into knowledge graphs, enabling claim attribution (Dammu et al., 2024; Wu et al., 2023). Furthermore, Atanasova et al. (2020); Krishna et al. (2022); Pan et al. (2023a); Eldifrawi et al. (2024); Zhang et al. (2024b) advocated for generating explanatory annotations alongside experimental outcomes during the verification process. Meanwhile, Das et al. (2023); Altuncu et al. (2023) emphasized the critical role of domain expertise in ensuring accurate verification.

3.2 Theorem Proving

Theorem proving constitutes a subtask of logical reasoning, aimed at reinforcing the validity of the underlying theory within a hypothesis (Pease et al., 2019; Yang et al., 2023c; Li et al., 2024e).

Following the proposal of GPT-f (Polu and Sutskever, 2020) to utilize generative language models for theorem proving, researchers initially combined search algorithms with language models (Lample et al., 2022; Wang et al., 2023b). However, a limitation in search-based approaches was later identified by Wang et al. (2024a), who highlighted their tendency to explore insignificant intermediate conjectures. This led some teams to abandon search algorithms entirely. Subsequently, alternative methods emerged, such as the two-stage framework proposed by Jiang et al. (2023) and Lin et al. (2024), which prioritized informal conceptual generation before formal proof construction. Thor (Jiang et al., 2022a) introduced theorem libraries to accelerate proof generation, an approach enhanced by Logo-power (Wang et al., 2024b) through dynamic libraries. Studies like Baldur (First et al., 2023), Mustard (Huang et al., 2024c), and DeepSeek-Prover (Xin et al., 2024) demonstrated improvements via targeted data synthesis and fine-tuning, though COPRA (Thakur et al., 2024) questioned their generalizability and proposed an environment-agnostic alternative. Complementary strategies included theoretical decomposition into sub-goals (Zhao et al., 2023b) and leveraging LLMs as collaborative assistants in interactive environments (Song et al., 2024).

3.3 Experiment Validation

Experiment validation involves designing and conducting experiments based on the hypothesis. The empirical validity of the hypothesis is then determined through analysis of the experimental results (Huang et al., 2024b).

Experiment validation represents a time-consuming component of scientific research. Recent advancements in LLMs have enhanced their ability to plan and reason about experimental tasks (Kambhampati et al., 2024), prompting researchers to use these models for designing and implementing experiments (Ruan et al., 2024b). To ensure accuracy, studies such as Zhang et al. (2023) and Arlt et al. (2024) imposed input/output constraints, though this reduced generalizability. To address this, Boiko et al. (2023); Bran et al. (2024); Huang et al. (2024a) integrated tools to expand model capabilities. Full automation was achieved by Ni and Buehler (2023); Li et al. (2024a); Lu et al. (2024) through prompt-guided multi-agent collaboration. Madaan et al. (2023); Yuan et al. (2025) further highlighted that the integration of feedback mechanisms demonstrated potential for enhancing design quality, while Zhang et al. (2024a); Liu et al. (2024c); Ni et al. (2024) employed experimental outcomes to refine hyperparameter configurations, and Szymanski et al. (2023); Li et al. (2024d); Baek et al. (2024) leveraged agent-generated analytical insights to facilitate iterative hypothesis refinement. In contrast, social science research often uses LLMs as experimental subjects to simulate human participants (Liu et al., 2023b; Manning et al., 2024; Mou et al., 2024).

A hypothesis can be conceptualized as consisting of two key components: claims and theorems. Scientific claim verification and theorem proving offer theoretical validation of hypothesis through formal reasoning and logical deduction, whereas experiment validation provides comprehensive practical validation via empirical testing.

4 Manuscript Publication

Upon validating a hypothesis as feasible, researchers generally progress to the publication stage. As depicted in Figure 5, this section categorizes Manuscript Publication into two primary components: Manuscript Writing and Peer Review.

4.1 Manuscript Writing

Manuscript writing, also referred to as scientific or research writing. At this stage, researchers articulate the hypothesis they have formulated and the results they have validated in the form of a scholarly paper. This process is crucial, as it not only disseminates findings but also deepens researchers' understanding of their work (Colyar, 2009).

Early shared tasks focused on assisting researchers in writing or analyzing linguistic features (Dale and Kilgarriff, 2010; Daudaravicius, 2015). Recent advances have led to three main directions: citation text generation, related work

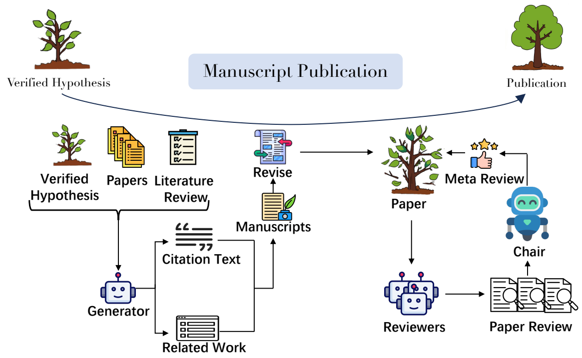

Figure 5: This figure shows the transformation of a validated hypothesis into a publication, leveraging outputs from the hypothesis formulation and validation stages.

generation, and complete manuscript generation.

Citation Text Generation (Sentence Level) A subset of research on AI in scientific writing has focused on citation text generation, which addresses the academic need for referencing prior work while mitigating model inaccuracies (Gao et al., 2023b; Gu and Hahnloser, 2023). For instance, Wang et al. (2022b) developed an automated citation generation system by integrating manuscript content with citation graphs. However, its reliance on rigid template-based architectures led to inflexible citation formats. This limitation motivated subsequent studies to propose incorporating citation intent as a control parameter during text generation, aiming to improve contextual relevance and rhetorical adaptability (Yu et al., 2022; Jung et al., 2022; Koo et al., 2023; Gu and Hahnloser, 2024).

Related Work Generation (Paragraph Level) In contrast to citation text generation, several studies have focused on related work generation in scholarly writing, emphasizing the production of multiple citation texts and the systematic analysis of inter-citation relationships (Li and Ouyang, 2022, 2024). The ScholaCite framework (Martin-Boyle et al., 2024) leveraged GPT-4 to cluster citation sources and generate draft literature review sections, although it required manually provided reference lists. By contrast, the UR3WG system (Shi et al., 2023) adopted a retrieval-augmented architecture integrated with large language models to autonomously acquire relevant references. To improve the quality of generated related work sections, Yu et al. (2024b) utilized GNNs to model complex relational dynamics between target manuscripts and cited literature, while Nishimura et al. (2024) initiative advocated for explicit novelty assertions regarding referenced publications.

Complete Manuscripts Generation (Full-text Level) The aforementioned investigations primarily focused on specific components of scientific writing, while a study by Lai et al. (2024) explored the progressive generation of complete manuscripts via structured workflows. The AI-Scientist system (Lu et al., 2024) further introduced section-wise self-reflection mechanisms to enhance compositional coherence. Several studies emphasized human-AI collaborative frameworks for improving writing efficiency (Lin, 2024; Feng et al., 2024; Ifargan et al., 2024), whereas Tang et al. (2024a) concentrated on enabling personalized content generation in multi-author collaborative environments. Following initial manuscript drafting, subsequent text revision processes were systematically examined (Du et al., 2022b; Jourdan et al., 2023; Dang et al., 2025). The OREO system (Li et al., 2022) utilized attribute classification for iterative in-situ editing, while Du et al. (2022a); Pividori and Greene (2024) incorporated researcher feedback loops for progressive text optimization. Notably, Kim et al. (2022); Chamoun et al. (2024); D'Arcy et al. (2024b) proposed replacing manual feedback with automated evaluation metrics.

4.2 Peer Review

Peer review serves as a critical mechanism for improving the quality of academic manuscripts through feedback and evaluation, forming the cornerstone of quality control in scientific research. However, the process is hindered by its slow pace, high time consumption, and increasing strain due to the growing academic workload (Lin et al., 2023a; Kousha and Thelwall, 2024; Thelwall and Yaghi, 2024). To address these challenges and enhance manuscript quality, researchers have investigated the application of AI in peer review (Yuan et al., 2022; Liu and Shah, 2023; Niu et al., 2023; Kuznetsov et al., 2024; Thakkar et al., 2025). Peer review can be categorized into two main types: paper review generation and meta-review generation.

Paper Review Generation In paper review generation, reviewers provide both scores and evaluations for manuscripts. For instance, Setio and Tsuchiya (2022) formulated score prediction as a regression task, Muangkammuen et al. (2022) utilized semi-supervised learning, and Couto et al. (2024) treated the task as a classification problem to evaluate the alignment between manuscripts and review criteria. While these approaches focused on label prediction for paper reviews, Yuan and Liu (2022) extended the scope by directly generating reviews through the construction of a concept graph integrated with a citation graph.

Subsequently, a pilot study conducted by Robertson (2023) demonstrated the capability of GPT-4 to generate paper reviews. Further investigations, such as those by AI-Scientist (Lu et al., 2024) and Liang et al. (2023), evaluated its performance as a review agent. Additionally, systems like MARG (D'Arcy et al., 2024a) and SWIF2T (Chamoun et al., 2024) employed multi-agent frameworks to generate reviews via internal discussions and task decomposition. In contrast, AgentReview (Jin et al., 2024) and Tan et al. (2024) modeled the review process as a dynamic, multi-turn dialogue. Furthermore, CycleResearcher (Weng et al., 2024) and OpenReviewer (Idahl and Ahmadi, 2024) fine-tuned models for comparative reviews and structured outputs aligned with conference guidelines.

Meta-Review Generation In meta-review generation, chairs are tasked with identifying a paper's core contributions, strengths, and weaknesses while synthesizing expert opinions on manuscript quality. Meta-reviews are conceptualized as abstractions of comments, discussions, and paper abstracts (Li et al., 2023). Santu et al. (2024) investigated the use of LLMs for automated meta-review generation, while Zeng et al. (2023) proposed a guided, iterative prompting approach. MetaWriter (Sun et al., 2024) utilized LLMs to extract key reviewer arguments, whereas GLIMPSE (Darrin et al., 2024) and Kumar et al. (2023) focused on reconciling conflicting statements to ensure fairness. Additionally, Li et al. (2024b) introduced a three-layer sentiment consolidation framework for meta-review generation, and PeerArg (Sukpanichnant et al., 2024) integrated LLMs with knowledge representation to address subjectivity and bias via a multiparty argumentation framework (MPAF). DeepReview (Zhu et al., 2025) generates a comprehensive meta-review by simulating expert evaluation across multiple dimensions.

During the Manuscript Publication phase, researchers can leverage AI to systematically complete manuscript writing by incorporating validated hypothesis, related papers, and literature reviews. The manuscript is subsequently subjected to peer review, involving iterative revisions before culminating in its final publication.

5 Benchmarks

Given that AI for research spans multiple disciplines, the tasks addressed within each domain vary significantly. To facilitate cross-domain exploration, we provide a summary of benchmarks associated with various areas, including research paper recommendation, systematic literature review, hypothesis generation, scientific claim verification, theorem proving, experiment verification, manuscript writing, and peer review. An overview of these benchmarks is presented in Table 9.

6 Tools

To accelerate the research workflow, we have curated a collection of tools designed to support various stages of the research process, with their applicability specified for each stage. To ensure practical relevance, our selection criteria emphasize tools that are publicly accessible or demonstrate significant influence on GitHub. A comprehensive overview of these tools is presented in Table 10.

7 Challenges

We identify several intriguing and promising avenues for future research.

7.1 Integration of Diverse Research Tasks

The research process is an integrated pipeline of interdependent stages. Paper recommendation and literature review provide an AI tool with a field overview and relevant works, ensuring that hypothesis generation is informed and of higher quality. Hypothesis validation assesses feasibility both logically and practically, with results feeding back to refine the hypothesis (Penadés et al., 2025). In manuscript writing, validated hypotheses and prior outputs serve as key inputs. Peer review evaluates the manuscript and offers feedback across modules, enabling the hypothesis generator to adjust content accordingly (Lu et al., 2024). In addition, combinations can also be made between some small fields, for instance, meta-review generation could be integrated with scientific claim verification, experiment verification could be linked with hypothesis formulation (Jansen et al., 2025; Yuan et al., 2025; Liu et al., 2024d), and research paper recommendation systems could be connected with manuscript writing processes (Gu and Hahnloser, 2023). Furthermore, some studies have begun to emphasize the development of systems capable of covering multiple stages of the research process (Jansen et al., 2024; Weng et al., 2024; Yu et al., 2024a).

7.2 Integration with Reasoning-Oriented Language Models

Research is a process that places a significant emphasis on logic and reasoning. Theorem proving serves as a subtask within logical reasoning (Li et al., 2024e), while hypothesis generation is widely recognized as the primary form of reasoning employed by scientists when observing the world and proposing hypothesis to explain these observations (Yang et al., 2024b). Experiment verification, in turn, demands a high degree of planning capability from models (Kambhampati et al., 2024). Recent advances in reasoning-oriented language models, such as OpenAI-o1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025), have substantially enhanced the reasoning abilities of these models. Consequently, we posit that integrating reasoning language models with reasoning tasks is a promising future direction. This prediction was validated by experiments conducted by Schmidgall et al. (2025) using o1-Preview.

Furthermore, in Appendix §B, we provide a summary of the challenges in hypothesis formulation, validation, and manuscript publication.

8 Ethical Considerations

AI has demonstrated significant potential in enhancing productivity by mitigating human limitations, thereby motivating increased investigation into its capacity to accelerate the research process (Messeri and Crockett, 2024). Nevertheless, its integration into scientific workflows introduces a range of ethical concerns (Fecher et al., 2023; Morris, 2023), including algorithmic biases, data privacy issues, risks of plagiarism, and the broader implications of AI-generated content for research communities. In this work, we examine these ethical challenges across the key stages of the research lifecycle: hypothesis formulation, validation, and publication.

During the hypothesis formulation stage, research paper recommendation systems and literature reviews are commonly employed; however, they often suffer from limitations that can lead to the formation of information bubbles and restrict exposure to diverse viewpoints. Furthermore, these systems tend to reinforce recognition disparities between prominent and lesser-known researchers and may inadvertently contribute to the dissemination of misinformation (Polonioli, 2021; Bolaños et al., 2024). To address these biases, recommendation algorithms can be enhanced by emphasizing content-based rather than author-based recommendations and by incorporating robust evaluation mechanisms to strengthen the credibility of suggested materials.

In contrast, AI-driven hypothesis generation presents more pronounced ethical challenges. First, the attribution of intellectual property rights and authorship for AI-generated hypotheses remains ambiguous (Majumder et al., 2024a). Additionally, the widespread generation of low-quality content poses a risk of diluting the integrity of the academic landscape (Hu et al., 2024a), while the potential misuse of such technologies for illicit purposes cannot be overlooked (Si et al., 2024). Addressing these concerns necessitates the development of robust accountability frameworks, the assignment of clear responsibility for AI-generated outputs to researchers, and the establishment of appropriate legal and regulatory mechanisms.

During the hypothesis validation phase, automated systems for scientific fact-checking remain underdeveloped. This limitation may be exploited by malicious actors to create advanced misinformation generators capable of circumventing existing fact-checking tools (Wadden et al., 2022b). Likewise, in the context of experimental validation, there is a risk of unethical or legally questionable experiments being designed (Eger et al., 2025). These concerns underscore the need for continued research into model safety.

During the manuscript publication stage, several challenges remain. Text generated by AI models may carry a risk of plagiarism (Salvagno et al., 2023; Gupta and Pruthi, 2025), while AI-assisted peer reviews often offer vague feedback and exhibit inherent biases (Schintler et al., 2023; Drori and Te'eni, 2024; Pataranutaporn et al., 2025). To address these issues, the development of robust detection methods is essential. However, current detection tools are still in the early stages of maturity (Gupta and Pruthi, 2025).

9 Conclusion

This paper provides a systematic survey of existing research on AI for research, offering a comprehensive review of the advancements in the field. Within each category, we offer detailed descriptions of the associated subfields. In addition, we examine current challenges, ethical considerations, and potential directions for future research. To support researchers in exploring AI-driven research applications and enhancing workflow efficiency, we also summarize existing benchmarks and tools, accompanied by a comparative analysis of representative methods and their capabilities.

Furthermore, in the course of investigating various subfields within AI for research, we observed that this domain remains in its infancy. Research in numerous directions remains at an experimental stage, and substantial progress is necessary before these approaches can be effectively applied in practical scenarios. We hope that this survey serves as an introduction to the field for researchers and contributes to its continued advancement.

Limitation

This study presents a comprehensive survey of AI for research, based on the framework of the research process conducted by human researchers.

We have made our best effort, but there may still be some limitations. Due to space constraints, we provide only concise summaries of each method without detailed technical elaboration. Given the rapid progress in AI and the expanding research landscape, we primarily focus on works published after 2022, with earlier studies receiving less attention. To emphasize areas that closely mimic the human research process, some topics are excluded from the main text but briefly discussed in Appendix §A. Moreover, as AI for Research is still an emerging field, the lack of standardized benchmarks and evaluation metrics hinders direct comparison. Nonetheless, we offer a comparative analysis of representative methods across domains using attribute graphs in Appendix §C.

Acknowledgments

Xiachong Feng and Xiaocheng Feng are the co-corresponding author of this work. We thank the anonymous reviewers for their insightful comments. This work was supported by the National Natural Science Foundation of China (NSFC) (grant 62276078, U22B2059), the Key R&D Program of Heilongjiang via grant 2022ZX01A32, and the Fundamental Research Funds for the Central Universities (XNJKKGYDJ2024013).

References

Shubham Agarwal, Issam H. Laradji, Laurent Charlin, and Christopher Pal. 2024a. Litllm: A toolkit for scientific literature review. *CoRR*, abs/2402.01788.

Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H Laradji, Krishnamurthy DJ Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. 2024b. Llms for literature review: Are we there yet? *arXiv preprint arXiv:2412.15249*.

Microsoft Research AI4Science and Microsoft Azure Quantum. 2023. The impact of large language models on scientific discovery: a preliminary study using GPT-4. *CoRR*, abs/2311.07361.

Fadi Aljamaan, Mohamad-Hani Temsah, Ibraheem Altamimi, Ayman Al-Eyadhy, Amr Jamal, Khalid Alhasan, Tamer A Mesallam, Mohamed Farahat, Khalid H Malki, and 1 others. 2024. Reference hallucination score for medical artificial intelligence chatbots: development and usability study. *JMIR Medical Informatics*, 12(1):e54345.

Nouf Ibrahim Altmami and Mohamed El Bachir Menai. 2022. Automatic summarization of scientific articles: A survey. *J. King Saud Univ. Comput. Inf. Sci.*, 34(4):1011–1028.

Enes Altuncu, Jason R. C. Nurse, Meryem Bagriacik, Sophie Kaleba, Haiyue Yuan, Lisa Bonheme, and Shujun Li. 2023. aedfact: Scientific fact-checking made easier via semi-automatic discovery of relevant expert opinions. *CoRR*, abs/2305.07796.

Sören Arlt, Haonan Duan, Felix Li, Sang Michael Xie, Yuhuai Wu, and Mario Krenn. 2024. Meta-designing quantum experiments with language models. *CoRR*, abs/2406.02470.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'Arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, and 6 others. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *CoRR*, abs/2411.14199.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7352–7364. Association for Computational Linguistics.

Sai Anirudh Athaluri, Sandeep Varma Manthena, VSR Krishna Manoj Kesapragada, Vineel Yarlagadda, Tirth Dave, and Rama Tulasi Siri Duddumpudi. 2023. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through chatgpt references. *Cureus*, 15(4).

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

Xiaomei Bai, Mengyang Wang, Ivan Lee, Zhuo Yang, Xiangjie Kong, and Feng Xia. 2019. Scientific paper recommendation: A survey. *IEEE Access*, 7:9324–9339.

André Bauer, Simon Trapp, Michael Stenger, Robert Leppich, Samuel Kounev, Mark Leznik, Kyle Chard, and Ian Foster. 2024. Comprehensive exploration of synthetic data generation: A survey. *arXiv preprint arXiv:2401.02524*.

Adrián Bazaga, Pietro Lio, and Gos Micklem. 2024. Unsupervised pretraining for fact verification by language model distillation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. 2016. Paper recommender systems: a literature survey. *International Journal on Digital Libraries*, 17:305–338.

Dale J Benos, Edlira Bashari, Jose M Chaves, Amit Gaggar, Niren Kapoor, Martin LaFrance, Robert Mans, David Mayhew, Sara McGowan, Abigail Polter, and 1 others. 2007. The ups and downs of peer review. *Advances in physiology education*, 31(2):145–152.

Loraine Blaxter, Christina Hughes, and Malcolm Tight. 2010. *How to research*. McGraw-Hill Education (UK).

Joern Block and Andreas Kuckertz. 2024. What is the future of human-generated systematic literature reviews in an age of artificial intelligence? *Management Review Quarterly*, pages 1–6.

Ben Bogin, Kejuan Yang, Shashank Gupta, Kyle Richardson, Erin Bransom, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. SUPER: evaluating agents on setting up and executing tasks from research repositories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 12622–12645. Association for Computational Linguistics.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nat.*, 624(7992):570–578.

Francisco Bolaños, Angelo A. Salatino, Francesco Osborne, and Enrico Motta. 2024. Artificial intelligence for literature reviews: opportunities and challenges. *Artif. Intell. Rev.*, 57(9):259.

James Boyko, Joseph Cohen, Nathan Fox, Maria Han Veiga, Jennifer I-Hsiu Li, Jing Liu, Bernardo Modenesi, Andreas H. Rauch, Kenneth N. Reid, Soumi

Tribedi, Anastasia Visheratina, and Xin Xie. 2023. An interdisciplinary outlook on large language models for scientific research. *CoRR*, abs/2311.04929.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nat. Mac. Intell.*, 6(5):525–535.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Markus J. Buehler. 2024. Accelerating scientific discovery with generative knowledge extraction, graph-based representation, and multimodal intelligent graph reasoning. *Mach. Learn. Sci. Technol.*, 5(3):35083.

Ruisheng Cao, Fangyu Lei, Haoyuan Wu, Jixuan Chen, Yeqiao Fu, Hongcheng Gao, Xinzhuang Xiong, Hanchong Zhang, Wenjing Hu, Yuchen Mao, Tianbao Xie, Hongshen Xu, Danyang Zhang, Sida I. Wang, Ruoxi Sun, Pengcheng Yin, Caiming Xiong, Ansong Ni, Qian Liu, and 4 others. 2024a. Spider2-v: How far are multimodal agents from automating data science and engineering workflows? In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Yupeng Cao, Aishwarya Muralidharan Nair, Elyon Eyimife, Nastaran Jamalipour Soofi, K. P. Subbalakshmi, John R. Wullert II, Chumki Basu, and David Shallcross. 2024b. Can large language models detect misinformation in scientific news reporting? *CoRR*, abs/2402.14268.

Eric Chamoun, Michael Sejr Schlichtkrull, and Andreas Vlachos. 2024. Automated focused feedback generation for scientific writing assistance. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 9742–9763. Association for Computational Linguistics.

Tzeng-Ji Chen. 2023. Chatgpt and other artificial intelligence applications speed up scientific writing. *Journal of the Chinese Medical Association*, 86(4):351–353.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. 2025. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Gautam Choudhary, Natwar Modani, and Nitish Maurya. 2021. React: A review comment dataset for actionability (and more). In *Web Information Systems Engineering - WISE 2021 - 22nd International Conference on Web Information Systems Engineering, WISE 2021, Melbourne, VIC, Australia, October 26-29, 2021, Proceedings, Part II*, volume 13081 of *Lecture Notes in Computer Science*, pages 336–343. Springer.

Julia Colyar. 2009. Becoming writing, becoming writers. *Qualitative Inquiry*, 15(2):421–436.

Paulo Henrique Couto, Quang Phuoc Ho, Nageeta Kumari, Benedictus Kent Rachmat, Thanh Gia Hieu Khuong, Ihsan Ullah, and Lisheng Sun-Hosoya. 2024. Relevai-reviewer: A benchmark on AI reviewers for survey paper relevance. *CoRR*, abs/2406.10294.

Robert Dale and Adam Kilgarriff. 2010. Helping our own: Text massaging for computational linguistics as a new shared task. In *Proceedings of the 6th International Natural Language Generation Conference*.

Preetam Prabhu Srikar Dammu, Himanshu Naidu, Mouly Dewan, YoungMin Kim, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. Claimver: Explainable claim-level verification and evidence attribution of text through knowledge graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 13613–13627. Association for Computational Linguistics.

Hai Dang, Chelse Swoopes, Daniel Buschek, and Elena L. Glassman. 2025. Corpusstudio: Surfacing emergent patterns in A corpus of prior work while writing. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI 2025, YokohamaJapan, 26 April 2025- 1 May 2025*, pages 1211:1–1211:19. ACM.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024a. MARG: multi-agent review generation for scientific papers. *CoRR*, abs/2401.04259.

Mike D'Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. 2024b. ARIES: A corpus of scientific paper edits made in response to peer reviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 6985–7001. Association for Computational Linguistics.

Maxime Darrin, Ines Arous, Pablo Piantanida, and Jackie Chi Kit Cheung. 2024. GLIMPSE: pragmatically informative multi-document summarization for scholarly reviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 12737–12752. Association for Computational Linguistics.

Anubrata Das, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. 2023. The state of human-centered NLP technology for fact-checking. *Inf. Process. Manag.*, 60(2):103219.

Vidas Daudaravicius. 2015. Automated evaluation of scientific writing: Aesw shared task proposal. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–63.

José de la Torre-López, Aurora Ramírez, and José Raúl Romero. 2023. Artificial intelligence to automate the systematic review of scientific literature. *Computing*, 105(10):2171–2194.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 81 others. 2024. Deepseek-v3 technical report. *CoRR*, abs/2412.19437.

Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. 2024. Claim verification in the age of large language models: A survey. *CoRR*, abs/2408.14317.

Iddo Drori and Dov Te'eni. 2024. Human-in-the-loop AI reviewing: Feasibility, opportunities, and risks. *J. Assoc. Inf. Syst.*, 25(1):7.

Wanyu Du, Zae Myung Kim, Vipul Raheja, Dhruv Kumar, and Dongyeop Kang. 2022a. Read, revise, repeat: A system demonstration for human-in-the-loop iterative text revision. *CoRR*, abs/2204.03685.

Wanyu Du, Vipul Raheja, Dhruv Kumar, Zae Myung Kim, Melissa Lopez, and Dongyeop Kang. 2022b. Understanding iterative revision from human-written text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3573–3590. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. Nlpeer: A unified resource for the computational study of peer review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5049–5073. Association for Computational Linguistics.

Steffen Eger, Yong Cao, Jennifer D'Souza, Andreas Geiger, Christian Greisinger, Stephanie Gross, Yufang Hou, Brigitte Krenn, Anne Lauscher, Yizhi Li, Chenghua Lin, Nafise Sadat Moosavi, Wei Zhao, and Tristan Miller. 2025. Transforming science with large language models: A survey on ai-assisted scientific discovery, experimentation, content generation, and evaluation. *CoRR*, abs/2502.05151.

Islam Eldifrawi, Shengrui Wang, and Amine Trabelsi. 2024. Automated justification production for claim veracity in fact checking: A survey on architectures and approaches. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 6679–6692. Association for Computational Linguistics.

Noemie Elhadad, M-Y Kan, Judith L Klavans, and Kathleen R McKeown. 2005. Customization in a unified framework for summarizing medical literature. *Artificial intelligence in medicine*, 33(2):179–198.

Eurostat. 2018. *The measurement of scientific, technological and innovation activities Oslo manual 2018 guidelines for collecting, reporting and using data on innovation*. OECD publishing.

Benedikt Fecher, Marcel Hebing, Melissa Laufer, Jörg Pohle, and Fabian Sofsky. 2023. Friend or foe? exploring the implications of large language models on the science system. *CoRR*, abs/2306.09928.

K. J. Kevin 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. *CoRR*, abs/2412.10999.

Emily First, Markus N. Rabe, Talia Ringer, and Yuriy Brun. 2023. Baldur: Whole-proof generation and repair with large language models. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2023, San Francisco, CA, USA, December 3-9, 2023*, pages 1229–1241. ACM.

Martin Funkquist, Ilia Kuznetsov, Yufang Hou, and Iryna Gurevych. 2023. Citebench: A benchmark for scientific citation text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7337–7353. Association for Computational Linguistics.

Conner Ganjavi, Michael B Eppler, Asli Pekcan, Brett Biedermann, Andre Abreu, Gary S Collins, Inderbir S Gill, and Giovanni E Cacciamani. 2024. Publishers' and journals' instructions to authors on use of generative artificial intelligence in academic and scientific publishing: bibliometric analysis. *bmj*, 384.

Catherine A. Gao, Frederick M. Howard, Nikolay S. Markov, Emma C. Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T. Pearson. 2023a. Comparing scientific abstracts generated by chatgpt to real abstracts with detectors and blinded human reviewers. *npj Digit. Medicine*, 6.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6465–6488. Association for Computational Linguistics.

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

Karan Girotra, Lennart Meincke, Christian Terwiesch, and Karl T Ulrich. 2023. Ideas are dimes a dozen: Large language models for idea generation in innovation. *Available at SSRN 4526071*.

Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. Missing counter-evidence renders NLP fact-checking unrealistic for misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5916–5936. Association for Computational Linguistics.

Max Glockner, Yufang Hou, Preslav Nakov, and Iryna Gurevych. 2024a. Grounding fallacies misrepresenting scientific publications in evidence. *CoRR*, abs/2408.12812.

Max Glockner, Yufang Hou, Preslav Nakov, and Iryna Gurevych. 2024b. Missci: Reconstructing fallacies in misrepresented science. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 4372–4405. Association for Computational Linguistics.

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 preprint arXiv:2503.08979*.

Dritjon Gruda. 2024. Three ways chatgpt helps me in my academic writing. *Nature*, 10.

Nianlong Gu and Richard Hahnloser. 2024. Controllable citation sentence generation with language models. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 22–37. Association for Computational Linguistics.

Nianlong Gu and Richard H. R. Hahnloser. 2023. Scilit: A platform for joint scientific literature discovery, summarization and citation generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2023, Toronto, Canada, July 10-12, 2023*, pages 235–246. Association for Computational Linguistics.

Xuemei Gu and Mario Krenn. 2024. Generation and human-expert evaluation of interesting research ideas using knowledge graphs and large language models. *CoRR*, abs/2405.17044.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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

Rohun Gupta, Isabel Herzog, Joseph Weisberger, John Chao, Kongkrit Chaiyasate, and Edward S Lee. 2023. Utilization of chatgpt for plastic surgery research: friend or foe? *Journal of Plastic, Reconstructive & Aesthetic Surgery*, 80:145–147.

Tarun Gupta and Danish Pruthi. 2025. All that glitters is not novel: Plagiarism in AI generated research. *CoRR*, abs/2502.16487.

Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S. Weld, and Marti A. Hearst. 2021. Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, pages 413:1–413:18. ACM.

Sam Henry and Bridget T. McInnes. 2017. Literature based discovery: Models, methods, and trends. *J. Biomed. Informatics*, 74:20–32.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge graphs. *ACM Comput. Surv.*, 54(4).

Chao-Chun Hsu, Erin Bransom, Jenna Sparks, Bailey Kuehl, Chenhao Tan, David Wadden, Lucy Lu Wang, and Aakanksha Naik. 2024. CHIME: llm-assisted hierarchical organization of scientific studies for literature review support. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 118–132. Association for Computational Linguistics.

Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili

Pan, and Zhenzhong Lan. 2024a. Nova: An iterative planning and search approach to enhance novelty and diversity of LLM generated ideas. *CoRR*, abs/2410.14255.

Yuntong Hu, Zhuofeng Li, Zheng Zhang, Chen Ling, Raasikh Kanjiani, Boxin Zhao, and Liang Zhao. 2024b. Hireview: Hierarchical taxonomy-driven automatic literature review generation. *arXiv preprint arXiv:2410.03761*.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument mining for understanding peer reviews. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2131–2137. Association for Computational Linguistics.

Jingshan Huang and Ming Tan. 2023. The role of chatgpt in scientific communication: writing better scientific review articles. *American journal of cancer research*, 13(4):1148.

Kaixuan Huang, Yuanhao Qu, Henry Cousins, William A. Johnson, Di Yin, Mihir Shah, Denny Zhou, Russ B. Altman, Mengdi Wang, and Le Cong. 2024a. CRISPR-GPT: an LLM agent for automated design of gene-editing experiments. *CoRR*, abs/2404.18021.

Kexin Huang, Ying Jin, Ryan Li, Michael Y. Li, Emmanuel J. Candès, and Jure Leskovec. 2025. Automated hypothesis validation with agentic sequential falsifications. *CoRR*, abs/2502.09858.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *CoRR*, abs/2311.05232.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2024b. Mlagentbench: Evaluating language agents on machine learning experimentation. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Yinya Huang, Xiaohan Lin, Zhengying Liu, Qingxing Cao, Huajian Xin, Haiming Wang, Zhenguo Li, Linqi Song, and Xiaodan Liang. 2024c. MUSTARD: mastering uniform synthesis of theorem and proof data. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Matthew Hutson. 2022. Could ai help you to write your next paper? *Nature*, 611(7934):192–193.

Maximilian Idahl and Zahra Ahmadi. 2024. Openreviewer: A specialized large language model for generating critical scientific paper reviews. *CoRR*, abs/2412.11948.

Tal Ifargan, Lukas Hafner, Maor Kern, Ori Alcalay, and Roy Kishony. 2024. Autonomous llm-driven research from data to human-verifiable research papers. *CoRR*, abs/2404.17605.

Jessica Irons, Patrick Cooper, Melanie McGrath, and Andreas Duenser. 2024. Ai for scientists: Using cognitive task analysis to guide scientific ai.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, and 80 others. 2024. Openai o1 system card. *CoRR*, abs/2412.16720.

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*.

Peter A. Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. 2024. Discoveryworld: A virtual environment for developing and evaluating automated scientific discovery agents. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Albert Qiaochu Jiang, Wenda Li, Szymon Tworkowski, Konrad Czechowski, Tomasz Odrzygózdz, Piotr Milos, Yuhuai Wu, and Mateja Jamnik. 2022a. Thor: Wielding hammers to integrate language models and automated theorem provers. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Albert Qiaochu Jiang, Sean Welleck, Jin Peng Zhou, Timothée Lacroix, Jiacheng Liu, Wenda Li, Mateja Jamnik, Guillaume Lample, and Yuhuai Wu. 2023. Draft, sketch, and prove: Guiding formal theorem provers with informal proofs. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Chao Jiang, Wei Xu, and Samuel Stevens. 2022b. arxivedits: Understanding the human revision process in scientific writing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9420–9435. Association for Computational Linguistics.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. Agentreview: Exploring peer review dynamics with LLM agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 1208–1226. Association for Computational Linguistics.

Léane Jourdan, Florian Boudin, Richard Dufour, and Nicolas Hernandez. 2023. Text revision in scientific writing assistance: An overview. *CoRR*, abs/2303.16726.

Léane Jourdan, Nicolas Hernandez, Richard Dufour, Florian Boudin, and Akiko Aizawa. 2025. Pararev: Building a dataset for scientific paragraph revision annotated with revision instruction. *arXiv preprint arXiv:2501.05222*.

Léane Isabelle Jourdan, Florian Boudin, Nicolas Hernandez, and Richard Dufour. 2024. CASIMIR: A corpus of scientific articles enhanced with multiple author-integrated revisions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 2883–2892. ELRA and ICCL.

Shing-Yun Jung, Ting-Han Lin, Chia-Hung Liao, Shyan-Ming Yuan, and Chuen-Tsai Sun. 2022. Intent-controllable citation text generation. *Mathematics*, 10(10):1763.

Melissa A Kacena, Lilian I Plotkin, and Jill C Fehrenbacher. 2024. The use of artificial intelligence in writing scientific review articles. *Current Osteoporosis Reports*, 22(1):115–121.

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Saldyt, and Anil Murthy. 2024. Position: Llms can't plan, but can help planning in llm-modulo frameworks. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard H. Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (peerread): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1647–1661. Association for Computational Linguistics.

Dongyeop Kang, Andrew Head, Risham Sidhu, Kyle Lo, Daniel S. Weld, and Marti A. Hearst. 2020. Document-level definition detection in scholarly documents: Existing models, error analyses, and future directions. In *Proceedings of the First Workshop on Scholarly Document Processing, SDP@EMNLP 2020, Online, November 19, 2020*, pages 196–206. Association for Computational Linguistics.

Hyeonsu B. Kang, Rafal Kocielnik, Andrew Head, Jiangjiang Yang, Matt Latzke, Aniket Kittur, Daniel S. Weld, Doug Downey, and Jonathan Bragg. 2022. From who you know to what you read: Augmenting scientific recommendations with implicit social networks. In *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022*, pages 302:1–302:23. ACM.

Hyeonsu B. Kang, Nouran Soliman, Matt Latzke, Joseph Chee Chang, and Jonathan Bragg. 2023. Comlittee: Literature discovery with personal elected author committees. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, pages 738:1–738:20. ACM.

Ying Kang, Aiqin Hou, Zimin Zhao, and Daguang Gan. 2021. A hybrid approach for paper recommendation. *IEICE TRANSACTIONS on Information and Systems*, 104(8):1222–1231.

Wei-Yu Kao and An-Zi Yen. 2024. MAGIC: multi-argument generation with self-refinement for domain generalization in automatic fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 10891–10902. ELRA and ICCL.

Tetsu Kasanishi, Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. 2023. Scireviewgen: A large-scale dataset for automatic literature review generation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6695–6715. Association for Computational Linguistics.

Uri Katz, Mosh Levy, and Yoav Goldberg. 2024. Knowledge navigator: Llm-guided browsing framework for exploratory search in scientific literature. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 8838–8855. Association for Computational Linguistics.

Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and Edward Choi. 2023. Factkg: Fact verification via reasoning on knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16190–16206. Association for Computational Linguistics.

Zae Myung Kim, Wanyu Du, Vipul Raheja, Dhruv Kumar, and Dongyeop Kang. 2022. Improving iterative text revision by learning where to edit from other revision tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9986–9999. Association for Computational Linguistics.

Ryan Koo, Anna Martin, Linghe Wang, and Dongyeop Kang. 2023. Decoding the end-to-end writing trajectory in scholarly manuscripts. *CoRR*, abs/2304.00121.

Kayvan Kousha and Mike Thelwall. 2024. Artificial intelligence to support publishing and peer review: A summary and review. *Learn. Publ.*, 37(1):4–12.

Mario Krenn, Lorenzo Buffoni, Bruno C. Coutinho, Sagi Eppel, Jacob Gates Foster, Andrew Gritsevskiy, Harlin Lee, Yichao Lu, João P. Moutinho, Nima Sanjabi, Rishi Sonthalia, Ngoc Mai Tran, Francisco Valente, Yangxinyu Xie, Rose Yu, and Michael Kopp. 2022. Predicting the future of AI with AI: high-quality link prediction in an exponentially growing knowledge network. *CoRR*, abs/2210.00881.

Christin Katharina Kreutz and Ralf Schenkel. 2022. Scientific paper recommendation systems: a literature review of recent publications. *Int. J. Digit. Libr.*, 23(4):335–369.

Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2022. Proofver: Natural logic theorem proving for fact verification. *Trans. Assoc. Comput. Linguistics*, 10:1013–1030.

Adithya Kulkarni, Fatimah Alotaibi, Xinyue Zeng, Longfeng Wu, Tong Zeng, Barry Menglong Yao, Minqian Liu, Shuaicheng Zhang, Lifu Huang, and Dawei Zhou. 2025. Scientific hypothesis generation and validation: Methods, datasets, and future directions. *arXiv preprint arXiv:2505.04651*.

Sandeep Kumar, Tirthankar Ghosal, and Asif Ekbal. 2023. When reviewers lock horns: Finding disagreements in scientific peer reviews. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 16693–16704. Association for Computational Linguistics.

Sandeep Kumar, Tirthankar Ghosal, Vinayak Goyal, and Asif Ekbal. 2024. Can large language models unlock novel scientific research ideas? *CoRR*, abs/2409.06185.

Ilia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K. Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, Sheng Lu, Mausam, Margot Mieskes, Aurélie Névéol, Danish Pruthi, Lizhen Qu, Roy Schwartz, Noah A. Smith, Thamar Solorio, and 5 others. 2024. What can natural language processing do for peer review? *CoRR*, abs/2405.06563.

Yuxuan Lai, Yupeng Wu, Yidan Wang, Wenpeng Hu, and Chen Zheng. 2024. Instruct large language models to generate scientific literature survey step by step. In *Natural Language Processing and Chinese Computing - 13th National CCF Conference, NLPCC 2024, Hangzhou, China, November 1-3, 2024, Proceedings, Part V*, volume 15363 of *Lecture Notes in Computer Science*, pages 484–496. Springer.

Guillaume Lample, Timothée Lacroix, Marie-Anne Lachaux, Aurélien Rodriguez, Amaury Hayat, Thibaut Lavril, Gabriel Ebner, and Xavier Martinet. 2022. Hypertree proof search for neural theorem proving. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Jon M. Laurent, Joseph D. Janizek, Michael Ruzo, Michaela M. Hinks, Michael J. Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D. White, and Samuel G. Rodriques. 2024. Lab-bench: Measuring capabilities of language models for biology research. *CoRR*, abs/2407.10362.

Ju Yoen Lee. 2023. Can an artificial intelligence chatbot be the author of a scholarly article? *Journal of educational evaluation for health professions*, 20.

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 CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 19:1–19:19. ACM.

Jingjing Li, Zichao Li, Tao Ge, Irwin King, and Michael R. Lyu. 2022. Text revision by on-the-fly representation optimization. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10956–10964. AAAI Press.

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

Miao Li, Eduard H. Hovy, and Jey Han Lau. 2023. Summarizing multiple documents with conversational structure for meta-review generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 7089–7112. Association for Computational Linguistics.

Miao Li, Jey Han Lau, and Eduard H. Hovy. 2024b. A sentiment consolidation framework for meta-review generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 10158–10177. Association for Computational Linguistics.

Ruochen Li, Liqiang Jing, Chi Han, Jiawei Zhou, and Xinya Du. 2024c. Learning to generate research idea with dynamic control. *CoRR*, abs/2412.14626.

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

Weisheng Li, Chao Chang, Chaobo He, Zhengyang Wu, Jiongsheng Guo, and Bo Peng. 2020. Academic paper recommendation method combining heterogeneous network and temporal attributes. In *Computer Supported Cooperative Work and Social Computing - 15th CCF Conference, ChineseCSCW 2020, Shenzhen, China, November 7-9, 2020, Revised Selected Papers*, volume 1330 of *Communications in Computer and Information Science*, pages 456–468. Springer.

Xiangci Li and Jessica Ouyang. 2022. Automatic related work generation: A meta study. *CoRR*, abs/2201.01880.

Xiangci Li and Jessica Ouyang. 2024. Related work and citation text generation: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 13846–13864. Association for Computational Linguistics.

Zhaoyu Li, Jialiang Sun, Logan Murphy, Qidong Su, Zenan Li, Xian Zhang, Kaiyu Yang, and Xujie Si. 2024e. A survey on deep learning for theorem proving. *CoRR*, abs/2404.09939.

Zhi Li and Xiaozhu Zou. 2019. A review on personalized academic paper recommendation. *Comput. Inf. Sci.*, 12(1):33–43.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. 2024a. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on AI conference peer reviews. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D. Manning, and James Y. Zou. 2024b. Mapping the increasing use of llms in scientific papers. *CoRR*, abs/2404.01268.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, Daniel A. McFarland, and James Zou. 2023. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *CoRR*, abs/2310.01783.

Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X. Zhang. 2024. Llms as research tools: A large scale survey of researchers' usage and perceptions. *CoRR*, abs/2411.05025.

Haohan Lin, Zhiqing Sun, Yiming Yang, and Sean Welleck. 2024. Lean-star: Learning to interleave thinking and proving. *CoRR*, abs/2407.10040.

Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. 2023a. Automated scholarly paper review: Concepts, technologies, and challenges. *Inf. Fusion*, 98:101830.

Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. 2023b. MOPRD: A multidisciplinary open peer review dataset. *Neural Comput. Appl.*, 35(34):24191–24206.

Zhicheng Lin. 2024. Techniques for supercharging academic writing with generative ai. *Nature Biomedical Engineering*, pages 1–6.

Chengwu Liu, Jianhao Shen, Huajian Xin, Zhengying Liu, Ye Yuan, Haiming Wang, Wei Ju, Chuanyang Zheng, Yichun Yin, Lin Li, Ming Zhang, and Qun Liu. 2023a. FIMO: A challenge formal dataset for automated theorem proving. *CoRR*, abs/2309.04295.

Haokun Liu, Yangqiaoyu Zhou, Mingxuan Li, Chenfei Yuan, and Chenhao Tan. 2024a. Literature meets data: A synergistic approach to hypothesis generation. *CoRR*, abs/2410.17309.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M. Dai, Diyi Yang, and Soroush Vosoughi. 2023b. Training socially aligned language models in simulated human society. *CoRR*, abs/2305.16960.

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

Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. 2024b. Conversational drug editing using retrieval and domain feedback. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Shuaiqi Liu, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. 2022. Generating a structured summary of numerous academic papers: Dataset and method. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4259–4265. ijcai.org.

Siyi Liu, Chen Gao, and Yong Li. 2024c. Large language model agent for hyper-parameter optimization. *CoRR*, abs/2402.01881.

Zijun Liu, Kaiming Liu, Yiqi Zhu, Xuanyu Lei, Zonghan Yang, Zhenhe Zhang, Peng Li, and Yang Liu. 2024d. AIGS: generating science from ai-powered automated falsification. *CoRR*, abs/2411.11910.

Kyle Lo, Joseph Chee Chang, Andrew Head, Jonathan Bragg, Amy X. Zhang, Cassidy Trier, Chloe Anastasiades, Tal August, Russell Authur, Danielle Bragg, Erin Bransom, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Yen-Sung Chen, Evie Yu-Yen Cheng, Yvonne Chou, Doug Downey, Rob Evans, and 36 others. 2023. The semantic reader project: Augmenting scholarly documents through ai-powered interactive reading interfaces. *CoRR*, abs/2303.14334.

Renze Lou, Hanzi Xu, Sijia Wang, Jiangshu Du, Ryo Kamoi, Xiaoxin Lu, Jian Xie, Yuxuan Sun, Yusen Zhang, Jihyun Janice Ahn, Hongchao Fang, Zhuoyang Zou, Wenchao Ma, Xi Li, Kai Zhang, Congying Xia, Lifu Huang, and Wenpeng Yin. 2024. AAAR-1.0: assessing ai's potential to assist research. *CoRR*, abs/2410.22394.

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. *CoRR*, abs/2408.06292.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. 2023. SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7787–7813. Association for Computational Linguistics.

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*.

Pingchuan Ma, Tsun-Hsuan Wang, Minghao Guo, Zhiqing Sun, Joshua B. Tenenbaum, Daniela Rus, Chuang Gan, and Wojciech Matusik. 2024. LLM and simulation as bilevel optimizers: A new paradigm to advance physical scientific discovery. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

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. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Sanchaita Hazra, Ashish Sabharwal, and Peter Clark. 2024a. Position: Data-driven discovery with large generative models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

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

Benjamin S Manning, Kehang Zhu, and John J Horton. 2024. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research.

Anna Martin-Boyle, Aahan Tyagi, Marti A. Hearst, and Dongyeop Kang. 2024. Shallow synthesis of knowledge in gpt-generated texts: A case study in automatic related work composition. *CoRR*, abs/2402.12255.

Rui Meng, Khushboo Thaker, Lei Zhang, Yue Dong, Xingdi Yuan, Tong Wang, and Daqing He. 2021. Bringing structure into summaries: a faceted summarization dataset for long scientific documents. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 1080–1089. Association for Computational Linguistics.

Lisa Messeri and MJ Crockett. 2024. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58.

Meredith Ringel Morris. 2023. Scientists' perspectives on the potential for generative AI in their fields. *CoRR*, abs/2304.01420.

Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, and Zhongyu Wei. 2024. From individual to society: A survey on social simulation driven by large language model-based agents. *CoRR*, abs/2412.03563.

Panitan Muangkammuen, Fumiyo Fukumoto, Jiyi Li, and Yoshimi Suzuki. 2022. Exploiting labeled and unlabeled data via transformer fine-tuning for peer-review score prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2233–2240. Association for Computational Linguistics.

Arief Purnama Muharram and Ayu Purwarianti. 2024. Enhancing natural language inference performance with knowledge graph for COVID-19 automated fact-checking in indonesian language. *CoRR*, abs/2409.00061.

Bo Ni and Markus J. Buehler. 2023. Mechagents: Large language model multi-agent collaborations can solve mechanics problems, generate new data, and integrate knowledge. *CoRR*, abs/2311.08166.

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. *CoRR*, abs/2411.08063.

Harshit Nigam, Manasi Patwardhan, Lovekesh Vig, and Gautam Shroff. 2024. Acceleron: A tool to accelerate research ideation. *CoRR*, abs/2403.04382.

Kazuya Nishimura, Kuniaki Saito, Tosho Hirasawa, and Yoshitaka Ushiku. 2024. Toward related work generation with structure and novelty statement. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 38–57.

Liang Niu, Nian Xue, and Christina Pöpper. 2023. Unveiling the sentinels: Assessing AI performance in cybersecurity peer review. *CoRR*, abs/2309.05457.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Srishti Palani, Aakanksha Naik, Doug Downey, Amy X. Zhang, Jonathan Bragg, and Joseph Chee Chang. 2023. Relatedly: Scaffolding literature reviews with existing related work sections. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, pages 742:1–742:20. ACM.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023a. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6981–7004. Association for Computational Linguistics.

Liangming Pan, Yunxiang Zhang, and Min-Yen Kan. 2023b. Investigating zero- and few-shot generalization in fact verification. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP 2023 -Volume 1: Long Papers, Nusa Dua, Bali, November 1 - 4, 2023*, pages 511–524. Association for Computational Linguistics.

Pat Pataranutaporn, Nattavudh Powdthavee, and Pattie Maes. 2025. Can AI solve the peer review crisis? A large scale experiment on llm's performance and biases in evaluating economics papers. *CoRR*, abs/2502.00070.

Alison Pease, Simon Colton, Chris Warburton, Athanasios Nathanail, Irina Preda, Daniel Arnold, Daniel Winterstein, and Mike Cook. 2019. The importance of applying computational creativity to scientific and mathematical domains. In *10th International Conference on Computational Creativity, ICCC 2019*, pages 250–257. Association for Computational Creativity.

José R Penadés, Juraj Gottweis, Lingchen He, Jonasz B Patkowski, Alexander Daryin, Wei-Hung Weng, Tao Tu, Anil Palepu, Artiom Myaskovsky, Annalisa Pawlosky, and 1 others. 2025. Ai mirrors experimental science to uncover a mechanism of gene transfer crucial to bacterial evolution. *Cell*.

Iratxe Pinedo, Mikel Larrañaga, and Ana Arruarte. 2024. Arzigo: A recommendation system for scientific articles. *Inf. Syst.*, 122:102367.

Milton Pividori and Casey S. Greene. 2024. A publishing infrastructure for artificial intelligence (ai)-assisted academic authoring. *J. Am. Medical Informatics Assoc.*, 31(9):2103–2113.

Andrea Polonioli. 2021. The ethics of scientific recommender systems. *Scientometrics*, 126(2):1841–1848.

Stanislas Polu and Ilya Sutskever. 2020. Generative language modeling for automated theorem proving. *CoRR*, abs/2009.03393.

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

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. *CoRR*, abs/2311.05965.

Hanhao Qu, Yu Cao, Jun Gao, Liang Ding, and Ruifeng Xu. 2022. Interpretable proof generation via iterative backward reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2968–2981. Association for Computational Linguistics.

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

Dragomir R. Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The ACL anthology network corpus. *Lang. Resour. Evaluation*, 47(4):919–944.

Zachary Robertson. 2023. GPT4 is slightly helpful for peer-review assistance: A pilot study. *CoRR*, abs/2307.05492.

Ambrose Robinson, William Thorne, Ben P. Wu, Abdullah Pandor, Munira Essat, Mark Stevenson, and Xingyi Song. 2023. Bio-sieve: Exploring instruction tuning large language models for systematic review automation. *CoRR*, abs/2308.06610.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR

Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, and 1 others. 2024. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475.

Kai Ruan, Xuan Wang, Jixiang Hong, and Hao Sun. 2024a. Liveideabench: Evaluating llms' scientific creativity and idea generation with minimal context. *CoRR*, abs/2412.17596.

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

Michele Salvagno, Fabio Silvio Taccone, and Alberto Giovanni Gerli. 2023. Can artificial intelligence help for scientific writing? *Critical care*, 27(1):75.

Shubhra Kanti Karmaker Santu, Sanjeev Kumar Sinha, Naman Bansal, Alex Knipper, Souvika Sarkar, John Salvador, Yash Mahajan, Sri Guttikonda, Mousumi Akter, Matthew Freestone, and Matthew C. Williams Jr. 2024. Prompting llms to compose meta-review drafts from peer-review narratives of scholarly manuscripts. *CoRR*, abs/2402.15589.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3499–3512. Association for Computational Linguistics.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Laurie A. Schintler, Connie L. McNeely, and James Witte. 2023. A critical examination of the ethics of ai-mediated peer review. *CoRR*, abs/2309.12356.

Samuel Schmidgall and Michael Moor. 2025. Agentrxiv: Towards collaborative autonomous research. *arXiv preprint arXiv:2503.18102*.

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

Yakub Sebastian, Eu-Gene Siew, and Sylvester O. Orimaye. 2017. Emerging approaches in literature-based discovery: techniques and performance review. *The Knowledge Engineering Review*, 32:e12.

Basuki Setio and Masatoshi Tsuchiya. 2022. The quality assist: A technology-assisted peer review based on citation functions to predict the paper quality. *IEEE Access*, 10:126815–126831.

Abdul Shahid, Muhammad Tanvir Afzal, Moloud Abdar, Mohammad Ehsan Basiri, Xujuan Zhou, Neil Y Yen, and Jia-Wei Chang. 2020. Insights into relevant knowledge extraction techniques: a comprehensive review. *The Journal of Supercomputing*, 76:1695–1733.

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. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6252–6278. Association for Computational Linguistics.

Ritu Sharma, Dinesh Gopalani, and Yogesh Kumar Meena. 2023. An anatomization of research paper recommender system: Overview, approaches and challenges. *Eng. Appl. Artif. Intell.*, 118:105641.

Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2022. Mred: A meta-review dataset for structure-controllable text generation. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2521–2535. Association for Computational Linguistics.

Yongliang Shen, Kaitao Song, Xu Tan, Wenqi Zhang, Kan Ren, Siyu Yuan, Weiming Lu, Dongsheng Li, and Yueting Zhuang. 2024. Taskbench: Benchmarking large language models for task automation. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Zhengliang Shi, Shen Gao, Zhen Zhang, Xiuying Chen, Zhumin Chen, Pengjie Ren, and Zhaochun Ren. 2023. Towards a unified framework for reference retrieval and related work generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 5785–5799. Association for Computational Linguistics.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? A large-scale human study with 100+ NLP researchers. *CoRR*, abs/2409.04109.

Zachary S. Siegel, Sayash Kapoor, Nitya Nagdir, Benedikt Stroebl, and Arvind Narayanan. 2024. Core-bench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *CoRR*, abs/2409.11363.

Michael D. Skarlinski, Sam Cox, Jon M. Laurent, James D. Braza, Michaela M. Hinks, Michael J.

Hammerling, Manvitha Ponnapati, Samuel G. Rodriques, and Andrew D. White. 2024. Language agents achieve superhuman synthesis of scientific knowledge. *CoRR*, abs/2409.13740.

Peiyang Song, Kaiyu Yang, and Anima Anandkumar. 2024. Towards large language models as copilots for theorem proving in lean. *CoRR*, abs/2404.12534.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, and 1 others. 2025. Paperbench: Evaluating ai's ability to replicate ai research. *arXiv preprint arXiv:2504.01848*.

Vaios Stergiopoulos, Michael Vassilakopoulos, Eleni Tousidou, and Antonio Corral. 2024. An academic recommender system on large citation data based on clustering, graph modeling and deep learning. *Knowl. Inf. Syst.*, 66(8):4463–4496.

Haoyang Su, Renqi Chen, Shixiang Tang, Xinzhe Zheng, Jingzhe Li, Zhenfei Yin, Wanli Ouyang, and Nanqing Dong. 2024. Two heads are better than one: A multi-agent system has the potential to improve scientific idea generation. *CoRR*, abs/2410.09403.

Purin Sukpanichnant, Anna Rapberger, and Francesca Toni. 2024. Peerarg: Argumentative peer review with llms. *CoRR*, abs/2409.16813.

Lu Sun, Stone Tao, Junjie Hu, and Steven P. Dow. 2024. Metawriter: Exploring the potential and perils of AI writing support in scientific peer review. *Proc. ACM Hum. Comput. Interact.*, 8(CSCW1):1–32.

Teo Susnjak, Peter Hwang, Napoleon H. Reyes, Andre L. C. Barczak, Timothy R. McIntosh, and Surangika Ranathunga. 2024. Automating research synthesis with domain-specific large language model fine-tuning. *CoRR*, abs/2404.08680.

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

Nathan J Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E Kumar, Tanjin He, David Milsted, Matthew J McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, and 1 others. 2023. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624(7990):86–91.

Cheng Tan, Dongxin Lyu, Siyuan Li, Zhangyang Gao, Jingxuan Wei, Siqi Ma, Zicheng Liu, and Stan Z. Li. 2024. Peer review as A multi-turn and long-context dialogue with role-based interactions. *CoRR*, abs/2406.05688.

Xiangru Tang, Xingyao Zhang, Yanjun Shao, Jie Wu, Yilun Zhao, Arman Cohan, Ming Gong, Dongmei Zhang, and Mark Gerstein. 2024a. Step-back profiling: Distilling user history for personalized scientific writing. *CoRR*, abs/2406.14275.

Xuemei Tang, Xufeng Duan, and Zhenguang G Cai. 2024b. Are llms good literature review writers? evaluating the literature review writing ability of large language models. *arXiv preprint arXiv:2412.13612*.

Min Tao, Xinmin Yang, Gao Gu, and Bohan Li. 2020. Paper recommend based on lda and pagerank. In *Artificial Intelligence and Security: 6th International Conference, ICAIS 2020, Hohhot, China, July 17–20, 2020, Proceedings, Part III 6*, pages 571–584. Springer.

Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. 2025. Can llm feedback enhance review quality? a randomized study of 20k reviews at iclr 2025. *arXiv preprint arXiv:2504.09737*.

Amitayush Thakur, George Tsoukalas, Yeming Wen, Jimmy Xin, and Swarat Chaudhuri. 2024. An in-context learning agent for formal theorem-proving. In *First Conference on Language Modeling*.

Mike Thelwall and Abdullah Yaghi. 2024. Evaluating the predictive capacity of chatgpt for academic peer review outcomes across multiple platforms. *CoRR*, abs/2411.09763.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 809–819. Association for Computational Linguistics.

Yangjie Tian, Xungang Gu, Aijia Li, He Zhang, Ruohua Xu, Yunfeng Li, and Ming Liu. 2024. Overview of the NLPCC2024 shared task 6: Scientific literature survey generation. In *Natural Language Processing and Chinese Computing - 13th National CCF Conference, NLPCC 2024, Hangzhou, China, November 1-3, 2024, Proceedings, Part V*, volume 15363 of *Lecture Notes in Computer Science*, pages 400–408. Springer.

Venktesh V, Abhijit Anand, Avishek Anand, and Vinay Setty. 2024. Quantemp: A real-world open-domain benchmark for fact-checking numerical claims. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 650–660. ACM.

Juraj Vladika and Florian Matthes. 2023. Scientific fact-checking: A survey of resources and approaches. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6215–6230. Association for Computational Linguistics.

Juraj Vladika and Florian Matthes. 2024a. Comparing knowledge sources for open-domain scientific claim verification. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 2103–2114. Association for Computational Linguistics.

Juraj Vladika and Florian Matthes. 2024b. Improving health question answering with reliable and time-aware evidence retrieval. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 4752–4763. Association for Computational Linguistics.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. Healthfc: Verifying health claims with evidence-based medical fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 8095–8107. ELRA and ICCL.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7534–7550. Association for Computational Linguistics.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022a. Scifact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4719–4734. Association for Computational Linguistics.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022b. Multivers: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 61–76. Association for Computational Linguistics.

David Wadden, Kejian Shi, Jacob Morrison, Aakanksha Naik, Shruti Singh, Nitzan Barzilay, Kyle Lo, Tom Hope, Luca Soldaini, Shannon Zejiang Shen, Doug Downey, Hannaneh Hajishirzi, and Arman Cohan. 2024. Sciriff: A resource to enhance language model instruction-following over scientific literature. *CoRR*, abs/2406.07835.

Gengyu Wang, Kate Harwood, Lawrence Chillrud, Amith Ananthram, Melanie Subbiah, and Kathleen R. McKeown. 2023a. Check-covid: Fact-checking COVID-19 news claims with scientific evidence. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14114–14127. Association for Computational Linguistics.

Haiming Wang, Huajian Xin, Zhengying Liu, Wenda Li, Yinya Huang, Jianqiao Lu, Zhicheng Yang, Jing Tang, Jian Yin, Zhenguo Li, and Xiaodan Liang. 2024a. Proving theorems recursively. *CoRR*, abs/2405.14414.

Haiming Wang, Huajian Xin, Chuanyang Zheng, Zhengying Liu, Qingxing Cao, Yinya Huang, Jing Xiong, Han Shi, Enze Xie, Jian Yin, Zhenguo Li, and Xiaodan Liang. 2024b. Lego-prover: Neural theorem proving with growing libraries. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Haiming Wang, Ye Yuan, Zhengying Liu, Jianhao Shen, Yichun Yin, Jing Xiong, Enze Xie, Han Shi, Yujun Li, Lin Li, Jian Yin, Zhenguo Li, and Xiaodan Liang. 2023b. Dt-solver: Automated theorem proving with dynamic-tree sampling guided by proof-level value function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12632–12646. Association for Computational Linguistics.

Linghe Wang, Minhwa Lee, Ross Volkov, Luan Tuyen Chau, and Dongyeop Kang. 2025. Scholawrite: A dataset of end-to-end scholarly writing process. *CoRR*, abs/2502.02904.

Pancheng Wang, Shasha Li, Kunyuan Pang, Liangliang He, Dong Li, Jintao Tang, and Ting Wang. 2022a. Multi-document scientific summarization from a knowledge graph-centric view. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 6222–6233. International Committee on Computational Linguistics.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024c. Scimon: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 279–299. Association for Computational Linguistics.

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

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. 2024e. Autosurvey: Large language models can automatically write surveys. *CoRR*, abs/2406.10252.

Yifan Wang, Yiping Song, Shuai Li, Chaoran Cheng, Wei Ju, Ming Zhang, and Sheng Wang. 2022b. Disencite: Graph-based disentangled representation learning for context-specific citation generation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11449–11458. AAAI Press.

Jane Webster and Richard T. Watson. 2002. Analyzing the past to prepare for the future: Writing a literature review. *MIS Q.*, 26(2).

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

Nigel L. Williams, Stanislav Ivanov, and Dimitrios Buhalis. 2023. Algorithmic ghost in the research shell: Large language models and academic knowledge creation in management research. *CoRR*, abs/2303.07304.

Jinxuan Wu, Wenhan Chao, Xian Zhou, and Zhunchen Luo. 2023. Characterizing and verifying scientific claims: Qualitative causal structure is all you need. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 13428–13439. Association for Computational Linguistics.

Zijian Wu, Jiayu Wang, Dahua Lin, and Kai Chen. 2024. Lean-github: Compiling github LEAN repositories for a versatile LEAN prover. *CoRR*, abs/2407.17227.

Amelie Wührl, Yarik Menchaca Resendiz, Lara Grimminger, and Roman Klinger. 2024a. What makes medical claims (un)verifiable? analyzing entity and relation properties for fact verification. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 2046–2058. Association for Computational Linguistics.

Amelie Wührl, Dustin Wright, Roman Klinger, and Isabelle Augenstein. 2024b. Understanding fine-grained distortions in reports of scientific findings. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 6175–6191. Association for Computational Linguistics.

Huajian Xin, Daya Guo, Zhihong Shao, Zhizhou Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda Li, and Xiaodan Liang. 2024. Deepseek-prover: Advancing theorem proving in llms through large-scale synthetic data. *CoRR*, abs/2405.14333.

Jing Xiong, Jianhao Shen, Ye Yuan, Haiming Wang, Yichun Yin, Zhengying Liu, Lin Li, Zhijiang Guo, Qingxing Cao, Yinya Huang, Chuanyang Zheng, Xiaodan Liang, Ming Zhang, and Qun Liu. 2023. TRIGO: benchmarking formal mathematical proof reduction for generative language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 11594–11632. Association for Computational Linguistics.

Ziyang Xu. 2025. Patterns and purposes: A cross-journal analysis of ai tool usage in academic writing. *Preprint*, arXiv:2502.00632.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024a. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Kaiyu Yang and Jia Deng. 2019. Learning to prove theorems via interacting with proof assistants. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6984–6994. PMLR.

Kaiyu Yang, Aidan M. Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan J. Prenger, and Animashree Anandkumar. 2023a. Leandojo: Theorem proving with retrieval-augmented language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zhishen Yang, Raj Dabre, Hideki Tanaka, and Naoaki Okazaki. 2023b. Scicap+: A knowledge augmented dataset to study the challenges of scientific figure captioning. In *Proceedings of the Workshop on Scientific Document Understanding co-located with 37th AAAI Conference on Artificial Inteligence (AAAI 2023), Remote, February 14, 2023*, volume 3656 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024b. Large language models for automated open-domain scientific hypotheses discovery. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 13545–13565. Association for Computational Linguistics.

Zonglin Yang, Xinya Du, Rui Mao, Jinjie Ni, and Erik Cambria. 2023c. Logical reasoning over natural language as knowledge representation: A survey. *CoRR*, abs/2303.12023.

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. 2024c. Moosechem: Large language models for rediscovering unseen chemistry scientific hypotheses. *CoRR*, abs/2410.07076.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R. Fabbri, Irene Li, Dan Friedman, and Dragomir R. Radev. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7386–7393. AAAI Press.

Geyan Ye, Xibao Cai, Houtim Lai, Xing Wang, Junhong Huang, Longyue Wang, Wei Liu, and Xiangxiang Zeng. 2024. Drugassist: A large language model for molecule optimization. *CoRR*, abs/2401.10334.

Haofei Yu, Zhaochen Hong, Zirui Cheng, Kunlun Zhu, Keyang Xuan, Jinwei Yao, Tao Feng, and Jiaxuan You. 2024a. Researchtown: Simulator of human research community. *CoRR*, abs/2412.17767.

Luyao Yu, Qi Zhang, Chongyang Shi, An Lao, and Liang Xiao. 2024b. 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*, volume 14884 of *Lecture Notes in Computer Science*, pages 201–213. Springer.

Mengxia Yu, Wenhao Yu, Lingbo Tong, and Meng Jiang. 2022. Scientific comparative argument generation.

Jiakang Yuan, Xiangchao Yan, Botian Shi, Tao Chen, Wanli Ouyang, Bo Zhang, Lei Bai, Yu Qiao, and Bowen Zhou. 2025. Dolphin: Closed-loop open-ended auto-research through thinking, practice, and feedback. *arXiv preprint arXiv:2501.03916*.

Weizhe Yuan and Pengfei Liu. 2022. Kid-review: Knowledge-guided scientific review generation with oracle pre-training. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11639–11647. AAAI Press.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can we automate scientific reviewing? *J. Artif. Intell. Res.*, 75:171–212.

Fengzhu Zeng and Wei Gao. 2023. Prompt to be consistent is better than self-consistent? few-shot and zero-shot fact verification with pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4555–4569. Association for Computational Linguistics.

Qi Zeng, Mankeerat Sidhu, Ansel Blume, Hou Pong Chan, Lu Wang, and Heng Ji. 2024. 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*, page 20. Springer Nature.

Qi Zeng, Mankeerat Sidhu, Hou Pong Chan, Lu Wang, and Heng Ji. 2023. Meta-review generation with checklist-guided iterative introspection. *CoRR*, abs/2305.14647.

Lei Zhang, Yuge Zhang, Kan Ren, Dongsheng Li, and Yuqing Yang. 2024a. Mlcopilot: Unleashing the power of large language models in solving machine learning tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 2931–2959. Association for Computational Linguistics.

Shujian Zhang, Chengyue Gong, Lemeng Wu, Xingchao Liu, and Mingyuan Zhou. 2023. Automl-gpt: Automatic machine learning with GPT. *CoRR*, abs/2305.02499.

Xiaocheng Zhang, Xi Wang, Yifei Lu, Zhuangzhuang Ye, Jianing Wang, Mengjiao Bao, Peng Yan, and Xiaohong Su. 2024b. Augmenting the veracity and explanations of complex fact checking via iterative self-revision with llms. *CoRR*, abs/2410.15135.

Xingjian Zhang, Yutong Xie, Jin Huang, Jinge Ma, Zhaoying Pan, Qijia Liu, Ziyang Xiong, Tolga Ergen, Dongsub Shim, Honglak Lee, and Qiaozhu Mei. 2024c. MASSW: A new dataset and benchmark tasks for ai-assisted scientific workflows. *CoRR*, abs/2406.06357.

Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP 2023 -Volume 1: Long Papers, Nusa Dua, Bali, November 1 - 4, 2023*, pages 996–1011. Association for Computational Linguistics.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024d. A comprehensive survey of scientific large language models and their applications in scientific discovery. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 8783–8817. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023a. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Xueliang Zhao, Wenda Li, and Lingpeng Kong. 2023b. Decomposing the enigma: Subgoal-based demonstration learning for formal theorem proving. *CoRR*, abs/2305.16366.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. 2022. minif2f: a cross-system benchmark for formal olympiad-level mathematics. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Lei Shen, Zihan Wang, Andi Wang, Yang Li, and 1 others. 2023a. Codegeex: A pre-trained model for code generation with multilingual benchmarking on humaneval-x. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5673–5684.

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

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. Hypothesis generation with large language models. *CoRR*, abs/2404.04326.

Kun Zhu, Xiaocheng Feng, Xiachong Feng, Yingsheng Wu, and Bing Qin. 2023. Hierarchical catalogue generation for literature review: A benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 6790–6804. Association for Computational Linguistics.

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

A Further Discussion

Open Question: What is the difference between AI for science and AI for research? We posit that AI for research constitutes a subset of AI for science. While AI for research primarily focuses on supporting or automating the research process, it is not domain-specific and places greater emphasis on methodological advancements. In contrast, AI for science extends beyond the research process to include result-oriented discovery processes within specific domains, such as materials design, drug discovery, biology, and the solution of partial differential equations (Zheng et al., 2023b; AI4Science and Quantum, 2023; Zhang et al., 2024d).

Open Question: What is the difference between hypothesis generation and scientific discovery? Hypothesis generation, which is primarily based on literature-based review (LBD) (Swanson, 1986; Sebastian et al., 2017), emphasizing the process by which researchers generate new concepts, solutions, or approaches through existing research and their own reasoning. Scientific discovery encompasses not only hypothesis generation, but also innovation in fields like molecular optimization and drug development (Ye et al., 2024; Liu et al., 2024b), driven by outcome-oriented results.

Open Question: What is the difference between systematic literature review and related work generation? Existing research frequently addresses the systematic literature survey, which constitutes a component of the knowledge synthesis process during hypothesis formulation, alongside the related work generation phase in manuscript writing (Luo et al., 2025). However, we argue that these two tasks are distinct in nature. The systematic literature survey primarily focuses on summarizing knowledge extracted from diverse scientific documents, thereby assisting researchers in acquiring an initial understanding of a specific field (Altmami and Menai, 2022). In contrast, related work generation focuses on the writing process, emphasizing selection of pertinent literature and effective content structuring (Nishimura et al., 2024).

Discussion: Potential links between artificial intelligence systems and human research practice

- In research paper recommendation, PaperWeaver (Lee et al., 2024) offers an interactive page that allows users to modify the topics they are interested in.
- In systematic literature review, Block and Kuckertz (2024) highlights the significant role of humans, including setting correct questions and individualized problem-solving and theorizing. Meanwhile, Hsu et al. (2024) emphasizes manual correction during the outline generation process.
- In hypothesis generation, AI engages more closely with human researchers, ranging from scenarios where humans provide the core ideas and AI contributes by iteratively refining them (Pu et al., 2024), to more collaborative settings where humans and AI engage in dialogue to facilitate new scientific discoveries

(Ni et al., 2024; Liu et al., 2024b; Ye et al., 2024).

- In scientific claim verification, Altuncu et al. (2023); Das et al. (2023) highlight the critical role of experts in countering fake scientific news and advocate for the incorporation of expert opinions as a form of evidence.

- In theorem proving, Song et al. (2024) proposes leveraging LLMs as assistants to human researchers by generating suggested proof steps throughout the proving process.

- In experiment Validation, Ni et al. (2024) enhances the experimental setup through human-AI dialogue, whereas Li et al. (2024d) incorporates human input and real-time adjustments during the execution phase to optimize experimental design.

- In manuscript writing, Ifargan et al. (2024); Feng et al. (2024); Du et al. (2022a) require human intervention to suggest improvements to AI-generated paragraphs and enhance their quality through interactive methods.

- In peer review, Kumar et al. (2023); Darrin et al. (2024) advocate for assigning the responsibility of generating meta-reviews to human researchers. The role of AI is to assist by identifying conflicts among reviewers' opinions and supporting the chair in the scoring process, rather than independently assigning scores.

At present, AI for research remains in its early stages, and most systems still rely heavily on human priors and conventional research workflows. This is due to both the limitations of current models and the lack of trust from the research community. However, as model capabilities significantly improve, we may witness a paradigm shift. Systems like AlphaFold have already demonstrated that impactful scientific contributions can be made without fully replicating human research processes. In the future, AI may become an autonomous scientific agent, pursuing its own pathways of discovery, potentially letting humans learn from the model.

Discussion: The involvement of AI in manuscript writing The application of AI in manuscript writing has been accompanied by significant controversy. As LLMs demonstrated advanced capabilities, an increasing number of researchers began adopting these systems for scholarly composition (Liang et al., 2024b; Gao et al., 2023a). This trend raised concerns within the academic community (Salvagno et al., 2023), with scholars explicitly opposing the attribution of authorship to AI systems (Lee, 2023). Despite these reservations, the substantial time efficiencies offered by this technology led researchers to gradually accept AI-assisted writing practices (Gruda, 2024; Huang and Tan, 2023; Chen, 2023).This shift ultimately led to formal guidelines issued by leading academic journals (Ganjavi et al., 2024; Xu, 2025).

Discussion: Some areas that have not been discussed In addition to the eight areas discussed above, there are other lines of work that also aim to support scientific research, such as reading assistance (Kang et al., 2020; Head et al., 2021; Lo et al., 2023), which helps researchers read academic papers; literature processing[2], which handles documents in various formats to provide effective data for subsequent tasks; as well as code and data generation (Bauer et al., 2024; Zheng et al., 2023a), which serve as a foundation for experimental validation. However, as our focus is on the core process of scientific research, we have chosen to omit these aspects from the main text.

Discussion: Unified system and domain-specific system to automate research There is a clear distinction between unified and domain-specific AI research systems. Some efforts aim to develop general-purpose frameworks capable of supporting scientific discovery across domains (e.g., AI Scientist (Lu et al., 2024)), while others target domain-specific challenges (e.g., AI in biology (Irons et al., 2024)). Given the current limitations of AI capabilities, general-purpose systems have not yet fully replaced domain-specific approaches. Both directions remain valuable, but a long-term vision may favor general systems, as they hold the potential to integrate cross-disciplinary knowledge and push the boundaries of scientific understanding.

B Challenges

B.1 Hypothesis Formulation

Knowledge Synthesize Existing paper recommendation tools predominantly rely on the metadata of existing publications to suggest related arti-

[2] https://sdproc.org

cles, which often results in a lack of user-specific targeting and insufficiently detailed presentation that hampers comprehension. Leveraging LLMs can facilitate the construction of dynamic user profiles, enabling personalized literature recommendations and enhancing the richness of the contextual information provided for each recommended article, ultimately improving the user experience. In the process of generating systematic literature reviews, our practical experience reveals that the outline generation tools often produces redundant results with insufficient hierarchical structure. Moreover, the full-text generation process is prone to hallucinations—for instance, statements may not correspond to the cited articles—a pervasive issue in large language models (Huang et al., 2023; Bolaños et al., 2024; Susnjak et al., 2024). This problem can be ameliorated by enhancing the foundational model capabilities or by incorporating citation tracing.

Hypothesis Generation Most existing tools generate hypotheses by designing prompts or constructing systematic frameworks, which heavily rely on the capabilities of pre-trained models. However, these methods struggle to balance the novelty, feasibility, and validity of the hypotheses (Li et al., 2024c).Furthermore, our investigation reveals that many current approaches adopt novelty and feasibility as evaluation metrics; these metrics are either difficult to quantify or require manual scoring, which can introduce bias. To date, there is no unified benchmark to compare the various methods, and we believe that future research should prioritize establishing a unified metric that objectively reflects the strengths and weaknesses of different approaches.

B.2 Hypothesis Validation

Most existing scientific claim verification tools are largely confined to specific domains, exhibiting poor generalizability, which limits their practical applicability (Vladika and Matthes, 2023). Theorem proving, the scarcity of relevant data adversely affects performance improvements through training , results across different proof assistants are not directly comparable, and the lack of standardized evaluation benchmarks presents numerous challenges. Moreover, current approaches remain predominantly in the research stage and lack practical tools that facilitate interaction with researchers (Li et al., 2024e). Experiment Validation, as automatically generated experiments often suffer from a lack of methodological rigor, practical feasibility, and alignment with the original research objectives (Lou et al., 2024). All these fields require rigorous logical reasoning, and I believe that the recent surge in advanced reasoning technologies could potentially address these issues.

B.3 Manuscript Publication

Similar to systematic literature surveys, manuscript writing is also adversely affected by hallucination issues (Athaluri et al., 2023; Huang et al., 2023). Even when forced citation generation is employed, incorrect references may still be introduced (Aljamaan et al., 2024). Furthermore, the text generated by models requires meticulous examination by researchers to avoid ethical concerns, such as plagiarism risks (Salvagno et al., 2023). AI-generated manuscript reviews frequently provide vague suggestions and are susceptible to biases (Chamoun et al., 2024; Drori and Te'eni, 2024). Additionally, during meta-review generation, models can be misled by erroneous information arising from the manuscript review process (Kumar et al., 2023). To address these issues, it may be necessary for the industry to establish appropriate regulations or to employ AI-based methods for detecting AI-generated papers and reviews (Lin et al., 2023a).

C Ability Comparison

An effective survey should not only summarize existing methods within a field but also provide comparative analyses of different approaches. However, the domain of AI for Research remains in its early stages, with many areas lacking standardized benchmarks and even established evaluation metrics. To facilitate a clearer understanding of the distinctions among various methods, we draw on existing literature (Kang et al., 2023; Bolaños et al., 2024; Luo et al., 2025; Vladika and Matthes, 2023; Yang et al., 2023c; Li and Ouyang, 2022, 2024; Lin et al., 2023a) and adopt attribute graphs to compare representative approaches within each subfield, as illustrated in table §1 to table §8.

Method	Human-Computer Interaction	LLM	Required Information	Return Information	Relevance Source
ComLittee (Kang et al., 2023)	✓	-	Authorship Graphs	Meta data with relevant authors	R, Co, Ci
ArZiGo (Pinedo et al., 2024)	✓	-	User Interest	Meta data	R
PaperWeaver (Lee et al., 2024)	✓	✓	Collected Papers	Meta data with description	R
Kang et al. (2022)	-	-	Author's social network relationships +Reference relationship	Meta data with relevant authors	R

Table 1: Research Paper Recommendation, we referred to Kang et al. (2023) for comparing different methods, where R represents Paper recommender score, Co represents Co-author relationship, and Ci represents Cited author relationship.

Method	Research Field	Across Stages	Human Interaction	Task	Input	Output	Evaluation Method
AutoSurvey (Wang et al., 2024e)	Any	✓	-	Outline Generation, +Full-text Generation	Title & Full Content	Literature Survey	LLM & Human
CHIME (Hsu et al., 2024)	Biomedicine	-	✓	Outline Generation	Title & Full Content	Hierarchical Outline	Automatic Metrics
Knowledge Navigator (Katz et al., 2024)	Any	-	-	Outline Generation	Title & Full Content	Hierarchical Outline	LLM & Human
Relatedly (Palani et al., 2023)	Any	-	-	Full-text Generation	Title & Related Work	Literature Survey	Human
STORM (Shao et al., 2024)	Any	-	-	Outline Generation, +Full-text Generation	Title & Full Content	Literature Survey	LLM & Automatic Metrics

Table 2: Scientific Literature Review, we referred to Bolaños et al. (2024) and made modifications, thereby comparing different methods.

Method	Research Field	Across Stages	Human Interaction	Multi-agent	Trained Model	Online RAG	Novelty	Feasibility	Validity
COI (Li et al., 2024a)	Any	✓	-	-	-	✓	✓	✓	✓
Learn2Gen (Li et al., 2024c)	Artification Intelligence	✓	-	-	✓	-	✓	✓	✓
MatPilot (Ni et al., 2024)	Materials Science	✓	✓	✓	-	-	✓	✓	-
SciAgents (Ghafarollahi and Buehler, 2024)	Any	-	✓	✓	-	-	✓	✓	-
SciMON (Wang et al., 2024c)	Any	-	-	-	✓	-	✓	-	-

Table 3: Hypothesis Generation, we referred to Luo et al. (2025) and made modifications, thereby comparing different methods.

Method	Input	Document Retrieval	Human Interaction	Rationale Selection	Evidence Format	Output
MULTIVERS (Wadden et al., 2022b)	Claim & scientific abstract	Provided	-	Longformer	Document	Label & sentence-level rationales
SFAVEL (Bazaga et al., 2024)	Claim	Pre-trained Language Model	-	-	knowledge graph	Top-K Facts & Corresponding Relevance Scores
ProToCo (Zeng and Gao, 2023)	Claim-Evidence Pair	Provided	-	-	Sentence	Label
MAGIC (Kao and Yen, 2024)	Claim	Provided	-	Dense Passage Retriever	Sentence	Label
aedFaCT (Altuncu et al., 2023)	News Article	Google Search	✓	Human	Document	Evidence

Table 4: Scientific Claim Verification, we referred to Vladika and Matthes (2023) and made modifications, thereby comparing different methods.

Method	Generation Based	Stepwise	Heuristic Search	Informal or Formal	Human-authored Realistic Proof
IBR (Qu et al., 2022)	-	✓	✓	Informal	-
GPT-f (Polu and Sutskever, 2020)	✓	✓	-	Formal	✓
DT-Solver (Wang et al., 2023b)	✓	✓	✓	Formal	✓
POETRY (Wang et al., 2024a)	✓	-	-	Formal	✓

Table 5: Theorem proving, we referred to Yang et al. (2023c) and made modifications, thereby comparing different methods.

Method	Research Field	Across Stages	Human Interaction	Multi-agent	Task	Input	External tools
AutoML-GPT (Zhang et al., 2023)	Artification Intelligence	-	-	-	Automated Machine Learning	Task-oriented Prompts	-
Chemcrow (Bran et al., 2024)	Chemistry	-	✓	-	Chemical Task	Task Description	✓
DOLPHIN (Yuan et al., 2025)	Any	✓	-	✓	Automated Scientific Research	Idea	✓
MechAgents (Ni and Buehler, 2023)	Physics	-	-	✓	Mechanical Problem	-	-
Manning et al. (2024)	Social Science	✓	-	✓	Simulating Human	-	-

Table 6: Experiment Validation: we use attribute diagrams to compare different schemes, and the table design refers to Hypothesis Generation.

Method	Across Stages	Human Interaction	Task	Input	Evaluation Method
AI Scientist (Lu et al., 2024)	✓	-	Full-text Generation	Manuscript Template & Experimental Results & Hypothesis	LLM
data-to-paper (Ifargan et al., 2024)	✓	✓	Full-text Generation	Experimental Results & Hypothesis	-
ScholaCite (Martin-Boyle et al., 2024)	-	-	Related Work Generation	Title & Abstract & Citation	Citation Graph Metrics
SciLit (Gu and Hahnloser, 2023)	✓	-	Citation Generation	Keywords	Automatic Metrics
Gu and Hahnloser (2024)	-	-	Citation Generation	Citation Intent & Keywords	Human

Table 7: Manuscript Writing, we referred to Li and Ouyang (2022, 2024) and made modifications, thereby comparing different methods.

Method	Across Stages	Human Interaction	Paper Review	Meta Review	Multi-agent	Trained Model	Output
Gamma-Trans (Muangkammuen et al., 2022)	-	-	✓	-	-	✓	Peer-review Score
MARG (D'Arcy et al., 2024a)	-	-	✓	-	✓	-	Peer-review Comments
CycleResearcher (Weng et al., 2024)	✓	-	✓	-	-	✓	Peer-review Comments & Score
PeerArg (Sukpanichnant et al., 2024)	-	-	-	✓	-	-	Final Decision
GLIMPSE (Darrin et al., 2024)	-	✓	-	✓	-	-	Summary of Peer-review

Table 8: Peer Review, we referred to Lin et al. (2023a) and made modifications, thereby comparing different methods.

Task	Benchmark	Domain	Size	Input	Output	Metric
Hypothesis Formulation	SCHOLAT (Li et al., 2020)	Research Paper Recommendation	34,518	-	-	Recall, Precission, F1-score
	ACL selection network (Tao et al., 2020)	Research Paper Recommendation	18,718	Topics	Related Papers	Accuracy
	CiteSeer (Kang et al., 2021)	Research Paper Recommendation	1,100	Paper	Related Papers	Correlation Coefficient
	SciReviewGen (Kasanishi et al., 2023)	Systematic Literature Review	10,000+	Abstracts	literature review	ROUGE
	FacetSum (Meng et al., 2021)	Systematic Literature Review	60,024	Source Text+Facet	Summary of Facet	ROUGE
	BigSurvey (Liu et al., 2022)	Systematic Literature Review	7,000+	Abstracts	Survey Paragraph	ROUGE, F1-score
	SCHOLARQABENCH (Asai et al., 2024)	Systematic Literature Review	2,200	Question	Answer with Citations	Accuracy, Coverage, Citations + Relevance, Usefulness
	HiCaD (Zhu et al., 2023)	Systematic Literature Review	7,600	Reference Papers	Catalogues	Catalogue Edit Distance Similarity (CEDS) + Catalogue Quality Estimate (CQE)
	CLUSTREC-COVID (Katz et al., 2024)	Systematic Literature Review	2,284	Titles, Abstracts	Topic	Clusters per Topic
	CHIME (Hsu et al., 2024)	Systematic Literature Review	2,174	Topic	Hierarchies	F1-score
	Tian et al. (2024)	Systematic Literature Review	700	Subject, Reference	Title,Content	-
	MASSW (Zhang et al., 2024c)	Hypothesis Generation	152000	Context of Literature	Hypothesis	BLEU, ROUGE, BERTScore, + Cosine Similarity, BLEURT
	IdeaBench (Guo et al., 2024)	Hypothesis Generation	2,374	Instruction, Background Information	Hypothesis	Insight Score, BERTScore, Novelty, + LLM Similarity Rating, Feasibility
	SCIMON (Wang et al., 2024c)	Hypothesis Generation	-	Background Context	Idea	ROUGE, BERTScore +BARTScore, Novelty
	MOOSEYang et al. (2024b)	Hypothesis Generation	50	Background, Inspiration	Hypothesis	Validness, Novelty + Helpfulness
	DISCOVERYBENCH (Majumder et al., 2024b)	Hypothesis Generation	1,167	Data	Discovery	Hypothesis Match Score
	LiveIdeaBench (Ruan et al., 2024a)	Hypothesis Generation	-	Scientific Keywords	Idea	Originality, Feasibility + Fluency, Flexibilit
	Kumar et al. (2024)	Hypothesis Generation	100	Paper without Future Work	Idea	Idea Alignment Score, Idea Distinctness Index
Hypothesis Validation	SciRIFF (Wadden et al., 2024)	Scientific Claim Verification	137,000	Evidence, Task prompt	Structured Paragraph	F1, BLEU
	SCIFACT (Wadden et al., 2020)	Scientific Claim Verification	1,409	Claim, Evidence	Rationale Sentences, Label	Precision, Recall, Micro-F1
	SCIFACT-OPEN (Wadden et al., 2022a)	Scientific Claim Verification	279	Claim, Evidence	Rationale Sentences, Label	Precision, Recall,Micro-F1
	MISSCI (Glockner et al., 2024b)	Scientific Claim Verification	435	Claim, Premise, Context	Verification	Micro F1-score,P@1,Arg@1 + METEOR Score,BERTScore +NLI-A, NLI-S, Matches@1
	FEVER (Thorne et al., 2018)	Scientific Claim Verification	185,445	Claim, Evidence	Label, Necessary Evidence	F1-Score,Oracle Accuracy + Accuracy,Recall
	XClaimCheck (Kao and Yen, 2024)	Scientific Claim Verification	16,177	Claim, Evidence	Label, Argument	Macro-F1, Accuracy
	HEALTHVER (Sarrouti et al., 2021)	Scientific Claim Verification	14330	Claim, Evidence	Label	Macro Precision, Macro Recall + Macro F1-score, Accuracy
	QuanTemp (V et al., 2024)	Scientific Claim Verification	15,514	Claim, Evidence	Label	Weighted-F1 Score, Macro-F1, BLEU, + BERTScore, Cohen's Kappa Score + Human Evaluation
	SCITAB (Lu et al., 2023)	Scientific Claim Verification	1,225	Claim, Evidence	Label	Macro-F1
	Check-COVID (Wang et al., 2023a)	Scientific Claim Verification	1,504	Claim	Evidence	Accuracy, Precision, Recall, Macro-F1
	HealthFC (Vladika et al., 2024)	Scientific Claim Verification	750	Claim, Evidence	Label	Precision, Recall, F1-Macro
	FACTKG (Kim et al., 2023)	Scientific Claim Verification	108,000	Claim, Evidence	Label	Accuracy
	BEAR-FACT (Wührl et al., 2024a)	Scientific Claim Verification	1,448	Claim, Evidence +Entity/Relation Information	Label	F1-Score
	MINIF2F (Zheng et al., 2022)	Theorem Proving	488	Problem, Theorem	Proof	Pass Rate
	FIMO (Liu et al., 2023a)	Theorem Proving	149	Problem, Theorem, statements	Proof	Pass Rate
	LeanDojo (Yang et al., 2023a)	Theorem Proving	98,734	Problem, Theorem	Proof	R@k, MRR, Pass Rate
	Lean-github (Wu et al., 2024)	Theorem Proving	28,597	Problem, Theorem	Proof	Accuracy, Pass Rate
	TRIGO-real (Xiong et al., 2023)	Theorem Proving	427	Problem, Theorem	Proof	Pass Rate, Accuracy, EM@n
	TRIGO-web (Xiong et al., 2023)	Theorem Proving	453	Problem, Theorem	Proof	Pass Rate, Accuracy, EM@n
	TRIGO-gen (Xiong et al., 2023)	Theorem Proving	-	Problem, Theorem	Proof	Pass Rate, Accuracy, EM@n
	CoqGym (Yang and Deng, 2019)	Theorem Proving	71,000	Problem, Theorem	Proof	Success Rate
	MLAgentBench (Huang et al., 2024b)	Experiment Validation	13	-	-	Competence, Efficiency
	AAAR-1.0 (Lou et al., 2024)	Experiment Validation	-	Instance, Papers	Design, Explanation	S-F1, S-Precision, S-Recall + S-Match, ROUGE
	TASKBENCH (Shen et al., 2024)	Experiment Validation	17,331	-	-	ROUGE, t-F1, v-F1 +Normalized Edit Distance
	Spider2-V (Cao et al., 2024a)	Experiment Validation	494	Task	Experiment Execution	Success Rate
	CORE-Bench (Siegel et al., 2024)	Experiment Validation	270	Task Requirements	Experiment Result	Accuracy
	LAB-Bench (Laurent et al., 2024)	Experiment Validation	2400	Multiple-choice Question	Answer	Accuracy, Precision, Coverage
	PaperBench (Starace et al., 2025)	Experiment Validation	20	Paper, Additional Information	Code	Replication Score
	SUPER (Bogin et al., 2024)	Experiment Validation	801	Task Requirements	-	Accuracy, Landmark-Based Evaluation
	ScienceAgentBench (Chen et al., 2025)	Experiment Validation	102	Task Instruction, Dataset Information +Expert-Provided Knowledge	Program	Valid Execution Rate, Success Rate, CodeBERTScore, API Cost
Manuscript Publication	SciCap+ (Yang et al., 2023b)	Manuscript Writing	414,000	Figure, OCR tokens + Mention Paragraph	Caption	BLEU, ROUGE, METEOR + CIDEr, SPICE
	AAN Corpus (Radev et al., 2013)	Manuscript Writing	-	-	-	-
	SciSummNet (Yasunaga et al., 2019)	Manuscript Writing	1,000	Paper,Citation Sentence	Summary	ROUGE
	CiteBench (Funkquist et al., 2023)	Manuscript Writing	358,765	Cited Papers, Context	Citation Text	ROUGE, BERTScore
	ALCE (Gao et al., 2023b)	Manuscript Writing	3,000	Question	Answer with Citations	Recall, Precision
	GCite (Wang et al., 2022b)	Manuscript Writing	2,500	Citing/Cited Paper	Citation Text	BLEU, ROUGE
	ARXIVEDITS (Jiang et al., 2022b)	Manuscript Writing	1,000	Sentence Pairs	Sentence, Intent	Precision,Recall,F1-score
	CASIMIR (Jourdan et al., 2024)	Manuscript Writing	15,646	Original Sentence	Revised Sentence	Exact-match (EM),SARI, BLEU, + ROUGE-L,Bertscore
	ParaRev (Jourdan et al., 2025)	Manuscript Writing	48,203	Original Paragraph	Revised Paragraph	ROUGE-L,SARI + BertScore
	SCHOLAWRITE (Wang et al., 2025)	Manuscript Writing	62,000	Before-text	Writing Intention, After-text	F1-score, Lexical Diversity, Topic Consistency, Intention Coverage
	MReD (Shen et al., 2022)	Peer Review	7,089	Reviews	Meta-Review	ROUGE
	ORSUM(Zeng et al., 2024)	Peer Review	15,062	Reviews	Meta-Review	ROUGE-L, BERTScore, FACTCC + SummaC, DiscoScore
	PeerRead v1 (Kang et al., 2018)	Peer Review	107,000	Reviews	Accept/Reject	Accuracy
	NLPeer (Dycke et al., 2023)	Peer Review	5,000	Reviews,Paper	Review Score, Connection, + Review Category	MRSE, F1-macro + Precision, Recall
	AMPERE (Hua et al., 2019)	Peer Review	400	Review	Review with Type	Precision, Recall, F1-score
	MOPRD (Lin et al., 2023b)	Peer Review	6,578	Reviews,Paper	Editorial Decision, Review, + Meta-Review, Author Rebuttal	ROUGE, BARTScore
	ARIES (D'Arcy et al., 2024b)	Peer Review	1,720	Review Comment, Edits	Comment-Edit Pairs	Precision, Recall, F1-score
	ASAP-Review (Yuan et al., 2022)	Peer Review	-	Paper	Review	Aspect Coverage, Aspect Recall, +Semantic Equivalence +Human: Recommendation Accuracy(RAcc), +Informativeness(Info),Aspect-level, +Constructiveness(ACon) and Summary accuracy
	ReviewMT (Tan et al., 2024)	Peer Review	26,841	Paper	Review Dialogue	ROUGE,BLEU,METEOR
	ReAct (Choudhary et al., 2021)	Peer Review	6,250	Review	Classification of Review	Accuracy
	PEERSUM (Li et al., 2023)	Peer Review	-	Reviews	Meta-Review	ROUGE,BERTScore,UniEval,ACC

Table 9: An overview of benchmarks on AI for research. In the Input, Output, and Metric columns, the '+' symbol indicates that the row is a continuation of the previous row.

Tool	Research Paper Recommendation	Systematic Literature Review	Hypothesis Generation	Scientific Claim Verification	Theorem Proving	Experiment Verification	Manuscript Writing	Peer Review	Reading Assistance
Connected Paper	✓								
Inciteful	✓								
Litmaps	✓								
Pasa	✓								
Research Rabbit	✓								
Semantic Scholar	✓								✓
GenGO	✓								✓
Jenni AI	✓						✓		✓
Elicit	✓	✓							
Undermind	✓	✓							
OpenScholar	✓	✓							
ResearchBuddies	✓	✓							
Hyperwrite	✓	✓					✓		
Concensus	✓	✓		✓					
Iris.ai	✓	✓		✓					✓
MirrorThink	✓	✓		✓		✓			✓
SciSpace	✓	✓					✓	✓	✓
AskYourPDF	✓	✓		✓			✓	✓	✓
Iflytek	✓	✓		✓	✓	✓	✓		✓
FutureHouse	✓	✓	✓			✓			
Enago Read	✓	✓	✓	✓	✓				✓
Aminer	✓	✓	✓	✓	✓	✓	✓		✓
OpenRsearcher	✓	✓	✓		✓	✓	✓	✓	✓
ResearchFlow	✓	✓		✓	✓	✓	✓	✓	✓
You.com	✓	✓	✓	✓	✓	✓	✓	✓	✓
GPT Researcher		✓							
PICO Portal		✓							
SurveyX		✓							
Scinence42:Dora		✓					✓		
STORM		✓					✓		
ChatDOC		✓							✓
Scite		✓							✓
Silatus		✓							✓
Agent Laboratory		✓				✓	✓		
Sider		✓					✓		✓
Quillbot		✓					✓	✓	✓
Scholar AI		✓		✓		✓	✓	✓	✓
AI-Researcher		✓	✓			✓	✓	✓	
AI Scientist			✓			✓	✓	✓	
Isabelle					✓				
LeanCopilot					✓				
Llmstep					✓				
Proverbot9001					✓				
chatgpt_academic							✓		
gpt_academic							✓		
HeadlineAnalyzer							✓		
Langsmith Editor							✓		
Textero.ai							✓		
Wordvice.AI							✓		
Writesonic							✓		
Writefull							✓	✓	
Covidence								✓	
Penelope.ai								✓	
Byte-science									✓
Cool Papers									✓
Explainpaper									✓
Uni-finder									✓

Table 10: Tools for Research Paper Assistance

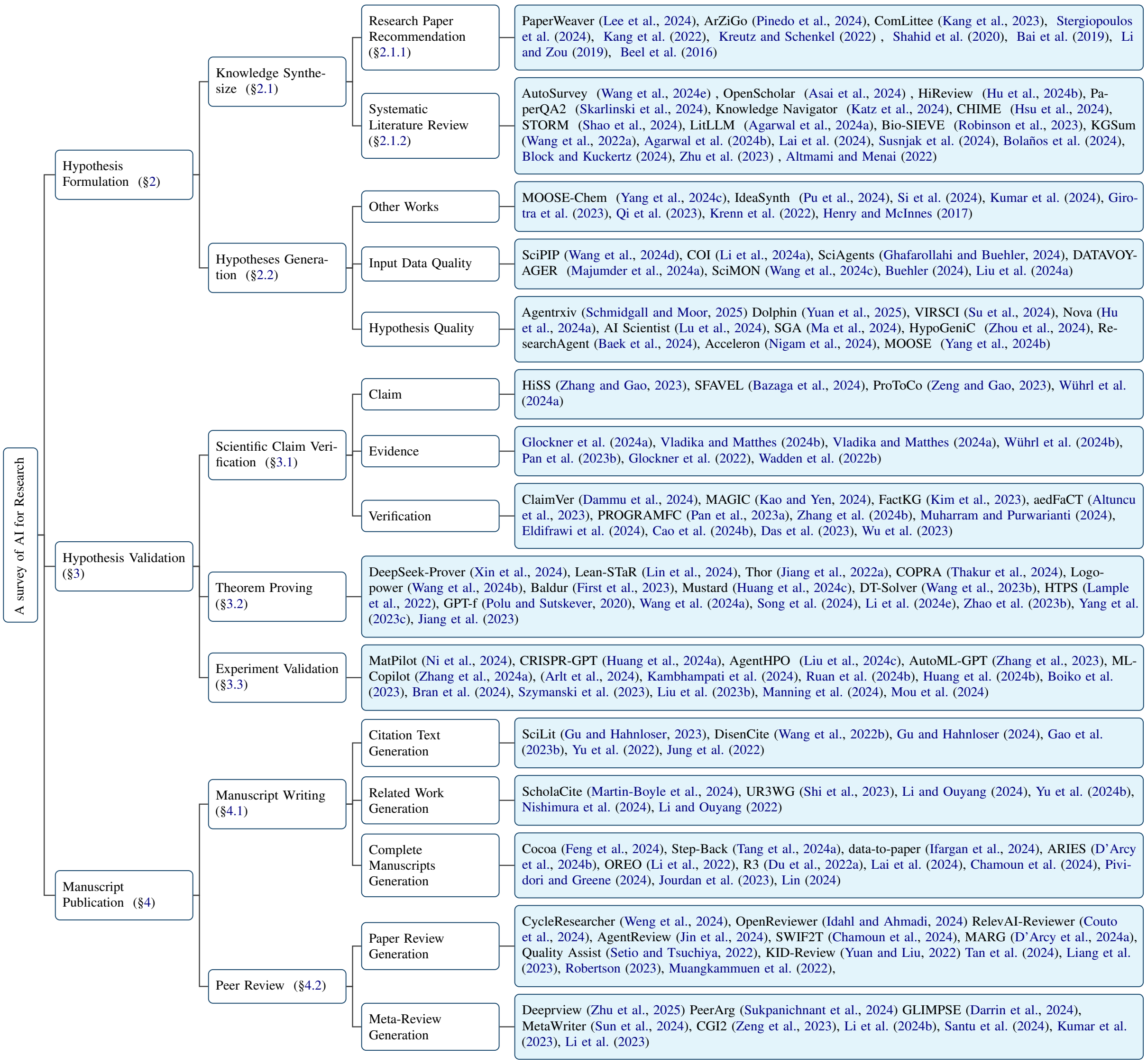

Figure 6: Taxonomy of Hypothesis Formulation, Hypothesis Validation and Manuscript Publication (Full Edition).