Transforming Science with Large Language Models: A Survey on AI-assisted Scientific Discovery, Experimentation, Content Generation, and Evaluation

STEFFEN EGER, University of Technology Nuremberg, Germany
YONG CAO, University of Tübingen, Germany
JENNIFER D'SOUZA, TIB Leibniz Information Centre for Science and Technology University Library, Germany
ANDREAS GEIGER, University of Tübingen, Germany
CHRISTIAN GREISINGER, University of Technology Nuremberg, Germany
STEPHANIE GROSS, Austrian Research Institute for Artificial Intelligence, Austria
YUFANG HOU, Interdisciplinary Transformation University Austria, Austria
BRIGITTE KRENN, Austrian Research Institute for Artificial Intelligence, Austria
ANNE LAUSCHER, University of Hamburg, Germany
YIZHI LI, The University of Manchester, United Kingdom of Great Britain and Northern Ireland
CHENGHUA LIN, The University of Manchester, United Kingdom of Great Britain and Northern Ireland
NAFISE MOOSAVI, The University of Sheffield, United Kingdom of Great Britain and Northern Ireland
WEI ZHAO, University of Aberdeen, United Kingdom of Great Britain and Northern Ireland
TRISTAN MILLER, University of Manitoba, Canada

Yong Cao was supported by a VolkswagenStiftung Momentum grant. Jennifer D'Souza was supported by the SCINEXT project (BMBF, German Federal Ministry of Education and Research, Grant ID: 01lS22070). The NLLG Lab at UTN gratefully acknowledges support from the Federal Ministry of Education and Research (BMBF) via the research grant "Metrics4NLG" and the German Research Foundation (DFG) via the Heisenberg Grant EG 375/5-1. The work of Anne Lauscher is supported by the Excellence Strategy of the German Federal Government and the Federal States. Our AI use cases are documented in the supplemental material.
Authors' Contact Information: Steffen Eger (corresponding author), University of Technology Nuremberg, Nuremberg, Bavaria, Germany; e-mail: steffen.eger@utn.de; Yong Cao, University of Tübingen, Tübingen, BW, Germany; e-mail: yong.cao@uni-tuebingen.de; Jennifer D'Souza, TIB Leibniz Information Centre for Science and Technology University Library, Hanover, NDS, Germany; e-mail: jennifer.dsouza@tib.eu; Andreas Geiger, University of Tübingen, Tübingen, BW, Germany; e-mail: a.geiger@uni-tuebingen.de; Christian Greisinger, University of Technology Nuremberg, Nuremberg, BY, Germany; e-mail: christian.greisinger@utn.de; Stephanie Gross, Austrian Research Institute for Artificial Intelligence, Vienna, Vienna, Austria; e-mail: stephanie.gross@ofai.at; Yufang Hou, Interdisciplinary Transformation University Austria, Linz, Austria; e-mail: yufang.hou@it-u.at; Brigitte Krenn, Austrian Research Institute for Artificial Intelligence, Vienna, Vienna, Austria; e-mail: Brigitte.krenn@ofai.at; Anne Lauscher, University of Hamburg, Hamburg, HH, Germany; e-mail: anne.lauscher@uni-hamburg.de; Yizhi Li, The University of Manchester, Manchester, United Kingdom of Great Britain and Northern Ireland; e-mail: yizhi.li-2@manchester.ac.uk; Chenghua Lin, The University of Manchester, Manchester, United Kingdom of Great Britain and Northern Ireland; e-mail: chenghua.lin@manchester.ac.uk; Nafise Moosavi, The University of Sheffield, Sheffield, England, United Kingdom of Great Britain and Northern Ireland; e-mail: n.s.moosavi@Sheffield.ac.uk; Wei Zhao, University of Aberdeen, Aberdeen, Scotland, United Kingdom of Great Britain and Northern Ireland; e-mail: wei.zhao@abdn.ac.uk; Tristan Miller, University of Manitoba, Winnipeg, Manitoba, Canada; e-mail: Tristan.Miller@umanitoba.ca.



This work is licensed under a Creative Commons Attribution 4.0 International License.
© 2027 Copyright held by the owner/author(s).
ACM 0360-0300/2027/01-ART11
https://doi.org/10.1145/3845596

With the advent of large multimodal language models, science is now at a threshold of an AI-based technological transformation. An emerging ecosystem of models and tools aims to support researchers throughout the scientific lifecycle, including (1) searching for relevant literature, (2) generating research ideas and conducting experiments, (3) producing text-based content, (4) creating multimodal artifacts such as figures and diagrams, and (5) evaluating scientific work, as in peer review. In this survey, we provide a curated overview of literature representative of the core techniques, evaluation practices, and emerging trends in AI-assisted scientific discovery. Across the five tasks outlined above, we discuss datasets, methods, results, evaluation strategies, limitations, and ethical concerns, including risks to research integrity through the misuse of generative models. We aim for this survey to serve both as an accessible, structured orientation for newcomers to the field, as well as a catalyst for new AI-based initiatives and their integration into future "AI4Science" systems.

CCS Concepts: • **Social and professional topics** → *Assistive technologies*; • **Applied computing** → *Physical sciences and engineering*; *Life and medical sciences*; *Law, social and behavioral sciences*; • **Computing methodologies** → **Natural language processing**; **Artificial intelligence**; • **General and reference** → **Surveys and overviews**.

Additional Key Words and Phrases: Language language models, science, AI4Science, search, experimentation, idea generation, multimodal content generation, evaluation, peer review

ACM Reference Format:
Steffen Eger, Yong Cao, Jennifer D'Souza, Andreas Geiger, Christian Greisinger, Stephanie Gross, Yufang Hou, Brigitte Krenn, Anne Lauscher, Yizhi Li, Chenghua Lin, Nafise Moosavi, Wei Zhao, and Tristan Miller. 2027. Transforming Science with Large Language Models: A Survey on AI-assisted Scientific Discovery, Experimentation, Content Generation, and Evaluation. *ACM Comput. Surv.* 59, 1, Article 11 (January 2027), 36 pages. https://doi.org/10.1145/3845596

1 Introduction

Throughout history, science has undergone a number of paradigm shifts, culminating in today's era of data-intensive exploration [77]. Although new tools and frameworks have accelerated the pace of scientific discovery, its basic steps have remained unchanged for centuries. These include (1) conception of a research question or problem, typically arising from a gap in disseminated knowledge; (2) collection and study of existing literature or data relevant to the problem; (3) formulation of a falsifiable hypothesis; (4) design and execution of experiments to test this hypothesis; (5) analysis and interpretation of the resulting data; and (6) reporting on the findings, allowing for their exploitation in real-world applications or as a source of knowledge for a further iteration of the scientific cycle.

The advent of large multimodal foundation models, such as ChatGPT, Gemini, Qwen, and DeepSeek, is profoundly affecting many sectors of society, including scientific research. Empirical evidence suggests that this influence extends well beyond computer science: an analysis of approximately 148,000 papers from 22 non-CS disciplines has revealed a rapid increase in citations of **large language models** (**LLMs**) between 2018 and 2024 [161]. In parallel, a large global survey of researchers conducted by Wiley reported widespread expectations that use of AI will become mainstream in scientific practice in the next two years, despite its current use being often limited to writing assistance.[1]

While science has traditionally relied on human ingenuity and labor for generating research ideas, formulating hypotheses, searching for relevant literature, conducting experiments, and reporting results, recent AI systems have been promising support at every stage of this cycle. Examples include Elicit and ORKG ASK for literature search, The AI Scientist [141] for experimentation, and

[1] https://www.wiley.com/en-us/ai-study, https://www.nature.com/articles/d41586-025-00343-5

AutomaTikZ [16] and DeTikZify [17] for multimodal content generation. There is moreover growing interest in the use of AI for evaluating scientific outputs through automated peer review [239]. Collectively, these advancements suggest the emergence of an integrated AI-assisted research workflow with the potential to accelerate discovery and streamline the documentation and communication of results.[2] The consolidation of a research community around AI-assisted science is evidenced by the establishment in 2024 and 2025 of dedicated venues such as the workshops on Natural Scientific Language Processing and Research Knowledge Graphs (NSLP) [174], Foundation Models for Science (FM4Science), AI & Scientific Discovery (AISD), Towards a Knowledge-grounded Scientific Research Lifecycle (AI4Research), AI Agents for Science (Agents4Science), and Human–LLM Collaboration for Ethical and Responsible Science Production (SciProdLLM) [249].

Despite the rapid progress in this area, existing surveys typically focus on specific domains, such as applications in the social sciences or physics [e.g., 229, 246], or on a relatively narrow set of research tasks and ethical concerns [e.g., 73, 143, 247]. To address this gap, the present survey adopts a workflow-centric perspective, providing a broad, cross-cutting overview of five central aspects of AI support for the research cycle: (1) literature search and summarization (Section 3.1); (2) scientific experimentation and research idea generation (Section 3.2); (3) unimodal generation and refinement of textual content, including titles, abstracts, and citations (Section 3.3); (4) multimodal content generation and interpretation, including figures, tables, slides, and posters (Section 3.4); and (5) AI-assisted peer review (Section 3.5). Rather than aiming for comprehensive coverage within each area, we focus on representative approaches that capture core methodological ideas and allow meaningful comparison across tasks.

Ethical considerations are paramount in any discussion of AI in science. Current tools exhibit a number of problems and limitations, including "hallucination", bias, limited reasoning abilities, and substantial environmental costs, and mechanisms for evaluating their output remain underdeveloped. Broader concerns include the risks of "fake science", plagiarism, and erosion of research integrity through diminished human oversight. Recent policy guidance on the use of AI in science, such as that of the EU,[3] emphasizes both the transformative potential of these technologies and the risks they pose if deployed without appropriate safeguards. In this survey, these ethical considerations are addressed alongside our treatment of the appertaining research tasks, as well as in a dedicated discussion in Section 4.

As illustrated in Figure 1, the remainder of this article is structured as follows. In Section 2 we discuss the scope and methodological approach of our survey. The subsections of Section 3 review representative literature for individual tasks in the research lifecycle, presenting datasets, methods, evaluation practices, limitations and future directions, and connections across tasks. In Section 4 we address broader ethical and integrity concerns. In Section 5 we conclude with a synthesis of opportunities and challenges for AI-assisted scientific research. The article's appendix includes background material on the scientific discovery cycle, further elaboration on AI support for specific topics and tasks, and a list of abbreviations. At https://github.com/NL2G/TransformingScienceLLMs we maintain a periodically updated list of further resources relating to this survey.

2 Survey Scope and Methodology

This survey provides a broad, workflow-centric overview of AI methods and applications that support scientific research across the full research lifecycle. It is intended primarily for

[2]The benefits could be particularly significant for non-native speakers of English and those with lower technical skills, potentially increasing diversity and inclusivity in research.

[3]https://research-and-innovation.ec.europa.eu/document/download/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en?filename=ec_rtd_ai-guidelines.pdf

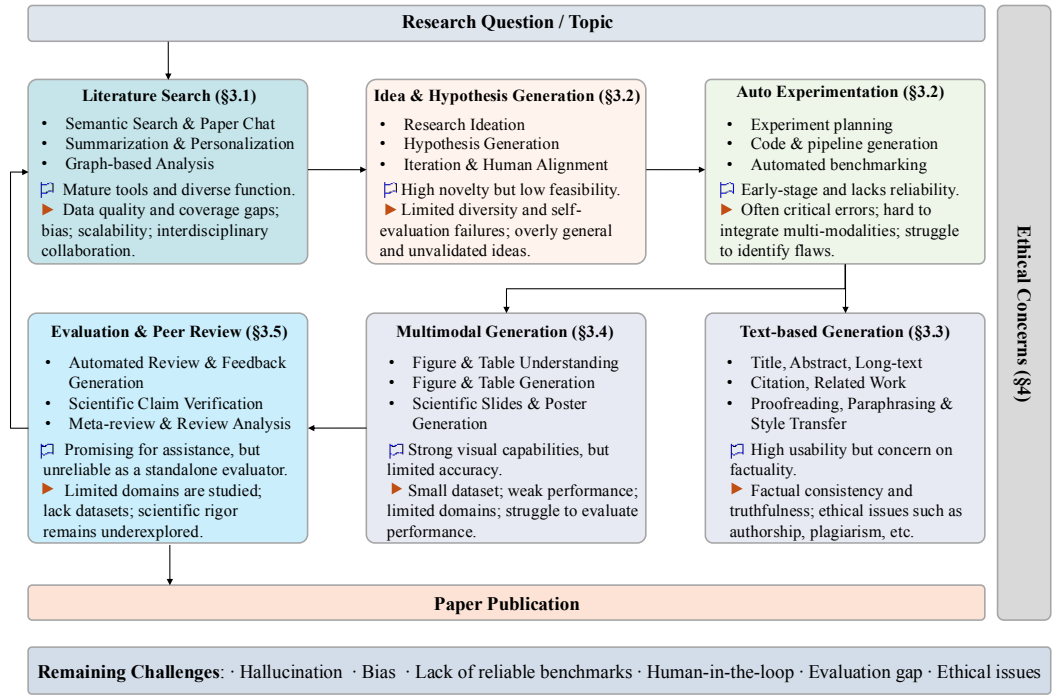


Fig. 1. Overview of the AI-assisted scientific research workflow and remaining challenges, illustrating how AI support different stages of the research process. Each block summarizes the current status of AI capabilities (⚐) and their limitations (▸) at that stage.

researchers in AI-related fields (e.g., natural language processing, computer vision, and machine learning) seeking a structured orientation to this rapidly evolving area, with clear entry points for deeper exploration. Some of the material will also be useful to policymakers, practitioners, and research collaborators in adjacent fields, including human–computer interaction, library and information science, communication studies, metascience, science journalism, and research ethics.

Given our topic's wide scope, rapid progress, and dependence on knowledge and methods from different domains, we adopt a narrative rather than a systematic survey methodology. This methodology allows for greater freedom in the selection and framing of relevant papers [100], which promotes "breadth of literature coverage and flexibility to deal with evolving knowledge and concepts" [22] as well as the ability to "bridge related areas of work, provoke thoughts, inspire new theoretical models, and direct future efforts in a research domain" [158]. Rather than imposing rigid inclusion and exclusion criteria, the narrative approach allows the survey to emphasize conceptual coherence, methodological representativeness, and comparability across tasks. At the same time, it requires transparency about scope and limitations: the survey does not aim to be exhaustive, nor does it claim to capture every recent publication.

The papers and tools discussed in each subsection were selected by diverse co-author teams with domain expertise, using a common set of guiding principles. Initial pools of candidate works were typically assembled by combining (i) seed papers known to the co-authors for their technical depth or influence, expanded through forward and backward citation analysis, and (ii) targeted

keyword searches in major scholarly databases such as Google Scholar. From these pools, works were selected on the basis of their relevance to the task under discussion, the maturity and clarity of their methodology, the reputation of the publication venue, and indicators of impact such as citation patterns relative to publication date. Preference was given to approaches that exemplify core techniques, employ well-defined evaluation protocols, or have served as reference points for subsequent work. This selection strategy was intended to foreground approaches that are representative of broader research trends, amenable to comparison, and likely to remain relevant at least in the near term, thereby supporting the survey's integrative goals. Throughout Section 3, individual subsections use a common structure to highlight common evaluation dimensions, methodological tradeoffs, and connections between tasks; this helps to situate individual contributions within a larger picture of AI-assisted scientific research and reduce any single-viewpoint bias inherent in our selection process.

3 AI Support for Individual Topics and Tasks

3.1 Literature Search, Summarization, and Comparison

The rapid growth of scientific literature presents a significant challenge for researchers to efficiently search, analyze, and summarize information. AI-powered tools are transforming these tasks by leveraging NLP, machine learning (ML), LLMs, citation and knowledge graphs (KGs) to automate the retrieval, extraction, and summarization of scientific information. This section surveys state-of-the-art AI-enhanced literature discovery tools, categorized according to their core functionality: (1) **AI-enhanced search**, which retrieves relevant literature from vast repositories; (2) **graph-based systems**, which map relationships between research concepts and publications; (3) **paper chat and QA**, which enable interactive exploration of scientific content; and (4) **recommender systems**, which suggest relevant papers based on user preferences. (We further discuss traditional search engines and benchmarks with leaderboards in Appendix B.1.)

3.1.1 Data. Scientific search engines rely on vast publisher databases to provide access to scientific literature. Understanding the classification of these repositories is essential for assessing search engines' coverage, reliability, and effectiveness in evidence-based research. Repositories vary by **access model**, **subject focus**, and **content type**, each serving a distinct role in academic discovery and knowledge dissemination. By **access model**, repositories fall into *open access repositories*, which provide unrestricted access to research articles (e.g., PubMed Central, arXiv); *subscription-based repositories*, requiring institutional or individual subscriptions (e.g., ScienceDirect, SpringerLink); and *hybrid repositories*, offering both free and paywalled content (e.g., Taylor & Francis Online, Oxford Academic). By **subject focus**, repositories are either *multidisciplinary*, covering broad disciplines (e.g., Web of Science, Scopus), or *subject-specific*, specializing in fields such as medicine (PubMed), physics (INSPIRE-HEP), and social sciences (SSRN). By **content type**, *institutional repositories* archive research outputs from specific organizations (e.g., MIT DSpace, Harvard DASH); *preprint repositories* enable early dissemination of research before peer review (e.g., bioRxiv, chemRxiv); and *government and public sector repositories* provide access to publicly funded research (e.g., NASA ADS, OpenAIRE). *Data repositories* (e.g., Dryad, Zenodo) store research datasets, supporting transparency and reproducibility, while *aggregator repositories* (e.g., BASE, CORE) index content from multiple sources for broader searches. Lastly, *grey literature repositories* (e.g., OpenGrey, EThOS) provide access to non-traditional research outputs such as theses, reports, and white papers, which may not be available through conventional publisher platforms.

The structure of scientific repositories shapes AI-enhanced search. While broad AI-based search engines like Elicit and ORKG ASK primarily query scholarly databases across multiple publisher repositories, similar to Google Scholar, tools like NotebookLM focus on user-selected documents,

and recommender systems such as Scholar Inbox rank new literature by relevance. AI-driven search enables customizable knowledge bases while optimizing discovery, retrieval, and personalization in research.

3.1.2 Methods and Results. Here, we discuss four types of state-of-the-art AI-enhanced search tools. A detailed overview of these tools is provided in Table 1.

AI-enhanced Search. Platforms such as Elicit, Consensus, OpenScholar [5], and SciSpace leverage AI, including LLMs, to extend beyond traditional search by enabling semantic search, paper summarization, evidence synthesis, and trend analysis. Unlike conventional search engines that rely on keyword matching, these tools use NLP and machine learning to extract key insights, synthesize information to answer research queries [66], and generate structured summaries [75, 223, 248]. Their ability to quickly summarize and categorize findings—such as study outcomes, methodologies, and limitations—helps researchers efficiently compare and interpret literature.

Graph-based Systems. Graph-based systems such as ORKG ASK [155] are designed to facilitate structured access to scientific knowledge. Unlike conventional paper search engines, they leverage a KG that organizes research contributions as structured data rather than unstructured text. Such contributions are typically extracted from the abstract, introduction, and result sections [48, 160]. Those systems enable users to ask complex, domain-specific questions and receive answers synthesized from semantically structured scientific data. They typically use techniques such as KG-based reasoning and **retrieval-augmented generation** (**RAG**) to extract relevant information from the KG, providing more interpretable and verifiable answers compared to traditional LLM-based QA systems. CiteSpace and Sci2 are specialized bibliometric analysis and network analysis tools to study the structure and evolution of scientific research. CiteSpace focuses on identifying research trends, keyword co-occurrence networks, and citation bursts, using visual analytics to highlight emerging topics and influential papers using graphs. Sci2 is a more general-purpose tool designed for analyzing scholarly datasets, enabling users to perform network analysis, geospatial mapping, and temporal modeling of scientific literature and collaboration patterns. Connected Papers is a visual literature exploration tool that maps papers related to a seed paper to provide an overview of a research field and support tasks such as bibliography construction and identification of prior and derivative work. Instead of building a direct citation tree, it organizes papers using similarity scores based primarily on *bibliographic coupling* and *co-citation*, typically via normalized overlap-based measures [98, 193]. The resulting weighted similarity graph visually clusters closely related papers and separates weaker ones, enabling interactive exploration of research clusters, and is powered by large-scale scholarly metadata (e.g., Semantic Scholar).

Paper Chat and QA. Paper chat and **question-answering** (**QA**) systems such as ChatGPT, DeepSeek Chat, NotebookLM, ExplainPaper, ChatPDF, and DocAnalyzer.AI allow users to interact with scientific papers by asking questions and receiving responses based on the document's content. They typically process a limited number of user-provided PDFs or text from specific websites. The core technology behind them is RAG [6, 95, 114], a technique that combines information retrieval with LLMs to improve accuracy and grounding. A typical RAG system first partitions the document into smaller sections and converts them into vector representations using embedding models. Upon a user query, the system retrieves the most relevant sections based on semantic similarity and passes them as context to an LLM, which then generates a response. This mechanism ensures that answers are directly grounded in the provided documents rather than relying solely on the model's pre-trained knowledge, enhancing factual reliability and interpretability. Some systems incorporate LLM agents [23, 125, 204] that can reason over retrieved information, summarize findings, or

Table 1. Overview of Popular Literature Search, Summarization, and Comparison Tools and Their Key Features

	Platform	Search	Recommendations	Collections	Citation Analysis	Trending Analysis	Author Profiles	Visualization Tools	Paper Chat	Idea Generation	Paper Writing	Summarization	Paper Review	Datasets	Code Repositories	LLM Integration	Web API	Personalization	Cost	Data Source
AI-Enhanced Search	Elicit	✓						✓	✓		✓	✓			✓				Freemium	125 million
	OpenScholar	✓	✓					✓			✓				✓				Free	45 million
	Undermind	✓	✓					✓			✓				✓		✓		Premium	over 200 million
	Perplexity	✓						✓	✓		✓	✓			✓				Freemium	
	Consensus	✓	✓					✓			✓				✓	✓			Freemium	over 200 million
	SciSpace	✓	✓					✓	✓		✓	✓			✓				Freemium	
	scienceQA	✓	✓	✓				✓	✓		✓	✓			✓				Freemium	220 million
	PaperQA2							✓						✓	✓				Free	
	Paperguide	✓	✓					✓	✓		✓	✓			✓				Freemium	
	HyperWrite	✓						✓	✓	✓	✓	✓			✓				Premium	
	ChatAcademia	✓						✓	✓	✓	✓	✓			✓		✓		Premium	
	Bohrium	✓	✓			✓		✓							✓				Freemium	170 million
	Paperpal	✓	✓					✓		✓	✓				✓				Freemium	over 3 million
	Scholar Labs	✓													✓				Freemium	
Graph-Based	Connected Papers	✓	✓				✓												Freemium	214 million
	ScholarGPS	✓		✓	✓	✓	✓												Free	over 200 million
	CiteSpace				✓		✓												Freemium	
	Sci2						✓												Free	
	NLP-KG	✓	✓	✓		✓	✓												Free	
	ORKG ASK	✓	✓								✓				✓				Free	76 million
	Litmaps	✓	✓				✓								✓				Freemium	
Paper Chat	ChatGPT	✓						✓	✓	✓	✓	✓			✓	✓			Freemium	10 PDF files
	Claude	✓						✓	✓	✓	✓	✓			✓	✓			Freemium	5 PDF files
	DeepSeek	✓						✓	✓	✓	✓	✓			✓	✓			Free	
	Research		✓					✓	✓		✓	✓			✓				Freemium	1 PDF file
	NotebookLM							✓	✓		✓	✓			✓		✓		Freemium	50 PDF files
	Enago Read	✓	✓					✓	✓		✓	✓			✓		✓		Freemium	1 PDF file
	DocAnalyzer.AI		✓					✓	✓		✓	✓			✓	✓	✓		Premium	few PDF files
	Coral		✓					✓	✓		✓	✓			✓				Freemium	1 PDF file
	ExplainPaper							✓	✓		✓	✓			✓				Freemium	1 PDF file
	ChatPDF	✓	✓					✓	✓		✓	✓			✓				Premium	1 PDF file
	AnswerThis	✓	✓					✓		✓	✓	✓			✓				Freemium	over 300 million
Recommender	arxiv-sanity-lite	✓	✓	✓													✓		Free	
	Scholar Inbox	✓	✓	✓		✓		✓							✓		✓		Free	
	ResearchTrend.ai	✓				✓													Freemium	
	TrendingPapers	✓	✓			✓					✓				✓		✓		Free	
	Bytez	✓				✓		✓	✓		✓	✓			✓	✓			Freemium	
	Notesum.ai	✓	✓	✓							✓				✓		✓		Freemium	
	Research Rabbit	✓		✓			✓												Free	

✓indicates feature availability; empty cells indicate lack of features or publicly documented support therefor.

extract key insights. These agents can follow multi-step reasoning strategies to provide more nuanced responses, such as synthesizing information from multiple sections or explaining technical terms in simpler language. By anchoring responses to document content, RAG-based systems mitigate hallucinations and make it easier for users to verify claims by checking the referenced passages. The effectiveness of these systems depends on the quality of document chunking, the efficiency of retrieval, and the model's ability to integrate information into coherent, context-aware answers.

Recommender Systems. Scientific paper recommender systems such as arxiv-sanity-lite, Scholar Inbox, ResearchTrend.ai, and Research Rabbit leverage machine learning and information retrieval

techniques to help researchers discover relevant literature. These systems generally fall into three main categories: content-based filtering, collaborative filtering, and hybrid approaches. Content-based methods [4, 19] analyze the text of papers to build representations that capture their meaning. Traditional approaches rely on sparse abstract or document representations such as TF–IDF [194], which assigns importance to words based on their frequency and distinctiveness in a corpus. More advanced models, such as SPECTER [38] and GTE [126], use dense abstract or document embeddings derived from neural networks; they map papers into a high-dimensional vector space where similar documents are close to each other. The **Massive Text Embedding Benchmark** (**MTEB**) [152] ranks many state-of-the-art embedding models on a comprehensive benchmark comprising various different datasets and tasks. These embeddings enable fast similarity searches and improve over simple keyword matching. In contrast, collaborative filtering [13, 213] relies on user interactions, such as downloads, bookmarks, and citations, to recommend papers based on the behavior of similar users. One challenge of pure collaborative filtering is the cold start problem, where new papers or users lack sufficient data for recommendations. To mitigate this, many modern systems employ hybrid approaches, such as two-tower architectures [40, 236, 238]. These models learn separate representations for papers and users, combining textual embeddings with user interaction data to generate more personalized recommendations. State-of-the-art systems often use a mix of these techniques to balance relevance, novelty, and diversity. The effectiveness of these systems depends on the quality of embeddings, the availability of interaction data, and the efficiency of ranking algorithms that surface the most useful papers.

3.1.3 Domains of Application. The tools discussed in this section are largely domain-agnostic and can be applied across scientific disciplines by adapting the underlying corpus and domain resources. For example, in medical and neuroscience research, Ben Ezzdine et al. [18] survey exercise and cognitive-training interventions for neurodegenerative disorders and discuss how AI methods are being used to support evidence synthesis and analysis. In ecology, LLMs have been evaluated for extracting structured ecological variables from the scientific literature to accelerate evidence synthesis [68]. In chemistry and materials science, NLP/LLM pipelines have been used to mine synthesis conditions and material-property records from papers to construct structured datasets [188, 250]. However, many benchmarks used to evaluate such systems remain concentrated in computer science and AI, reflecting current dataset availability.

3.1.4 Limitations and Future Directions. A primary challenge for scholarly search systems is *data quality and coverage gaps*: systems often struggle with incomplete, non-standard, or outdated data sources, which can lead to inaccuracies and inconsistencies in retrieved information. There is also the issue of *model bias*, where search and ranking algorithms adopt biases of their training data, potentially influencing the visibility of certain research areas and limiting the diversity of perspectives presented to users. Another major limitation lies in *scalability and real-time processing*—i.e., efficiently handling large-scale datasets while maintaining low latency and high retrieval accuracy. Finally, many AI-assisted research tools rely on proprietary data, closed APIs, or evolving LLM backends, which complicates strict reproducibility and long-term comparability. These limitations suggest several promising future directions. One potential avenue is *enhanced personalization*, which can be achieved by adapting search engines to user preferences, providing more tailored recommendations based on research interests and behavioral patterns. Fostering *interdisciplinary collaboration* through the integration of AI-powered search systems with other digital tools, such as data visualization platforms and research management software, could likewise facilitate more comprehensive and insightful research outcomes.

Table 2. Overview of Datasets for Idea and Hypothesis Generation and Experimentation

Dataset	Source	Data Size	Domain	Time Span	Task
SciMON [25]	ACL Anthology	135,814 papers	NLP	1952–2022	Idea Generation
IDEA Challenge [50]	University of Bristol	240 prototypes	Engineering	2022	Idea Generation
SPACE-IDEAS+ [60]	COLING	1020 ideas	Physics	2024	Idea Generation
TOMATO-Chem [233]	Nature and Science	51 papers	Chemistry	2024	Hypothesis Generation
LLM4BioHypoGen [168]	PubMed	2,900 papers	Medicine	2000–2024	Hypothesis Generation
CSKG-600 [46]	CSKG	600 hypotheses	AI	2010–2017	Hypothesis Generation
ScienceAgentBench [32]	OSU NLP	44 papers	Diverse	2024	Automated Experimentation
SWE-bench [89]	ICLR	2,294 issues	SWE	2024	Automated Experimentation
MLGym-Bench [153]	Meta	13 tasks	Diverse	2025	Automated Experimentation

3.2 AI-Driven Scientific Discovery: Ideation, Hypothesis Generation, and Experimentation

Ideation focuses on proposing new tools and/or analyzing existing ones, while hypothesis generation involves formulating specific, testable questions that guide empirical or theoretical justifications. In today's age of rapidly growing scientific literature, the effort of moving from literature review to idea or hypothesis formation has become increasingly time-consuming. Recently, LLMs have been employed to address this issue by making idea and hypothesis formation efficient: they are being leveraged both as generators (to autonomously produce ideas and hypotheses) and as evaluators (to assess their quality and select those that are meaningful, relevant, and novel). Experimentation adds further complexity, requiring careful methodological design, large-scale simulations, and in-depth results analysis. In this section, we first review ideation, hypothesis generation, and their (intrinsic and downstream) evaluation, then discuss how experimentation, framed as a form of downstream evaluation, can be automated through LLMs.

3.2.1 Data. Here we survey diverse datasets for evaluating LLMs in hypothesis generation, idea formation, and experimentation. These datasets, summarized in Table 2, were constructed from various scholarly sources representing a variety of scientific domains.

SciMON [25], a dataset for the idea generation task, is a subset of the **Semantic Scholar Open Research Corpus** (**S2ORC**) [140] focusing on abstracts of **Association for Computational Linguistics** (**ACL**) publications from 1952 to 2022. It contains 135,814 abstracts, divided into training (before 2021), validation (2021), and test (2022) sets. Each abstract was annotated using PL-Marker [235] and a structure classifier [37] that extract keywords and categorize sentences as providing background, research ideas, and so on. The tools were evaluated on a human-curated subset, and only high-confidence annotations were retained; despite this, some annotations may be incorrect, and errors can propagate and affect ideation quality.

SPACE-IDEAS+ [60] releases two versions of datasets. The smaller one contains 176 ideas sampled from the **Open Space Innovation Platform** (**OSIP**), an online repository of publicly available ideas related to space innovation. All the ideas were manually annotated by human experts, where each sentence was labeled with one of five roles (Challenge, Proposal, Elaboration, Benefits, Context) by two human annotators. Meetings were conducted to resolve disagreements between annotators. The larger version contains 1,020 ideas that were annotated by having GPT-3.5-turbo adopt the same annotation guidelines. A subset of the generated annotations was evaluated by comparing with human annotations; the agreement of 50% indicates the mediocre quality of GPT annotations.

TOMATO-Chem [233] is a hypothesis generation dataset containing 51 chemistry and material science papers published in *Nature* or *Science* in 2024. To these papers experts applied annotations concerning the background, research questions, works that potentially inspired the paper, hypotheses, and experiments for hypothesis justification. A limitation of the dataset is that

details concerning the annotation task (e.g., the number of annotators and their agreement) are not reported.

LLM4BioHypoGen [168], another hypothesis generation dataset, consists of 2,900 medical publications sourced from PubMed, where 2,500 papers were used for the training set and 200 papers for validation set (both published pre-January 2023), with 200 papers in the test set (post-August 2023). Each paper was annotated by using GPT instead of human experts to construct background and hypothesis pairs; this raises concerns about data quality, and consequently, the reliability of findings derived from this dataset.

ScienceAgentBench [32] is an automated experimentation dataset comprised of 102 tasks derived from 44 peer-reviewed publications across four disciplines: bioinformatics, computational chemistry, geographical information science, and psychology & cognitive neuroscience. Each task requires an agent to generate a self-contained Python program based on a natural language instruction, a dataset, and optional expert-provided knowledge. The benchmark employs multiple evaluation metrics, including the valid execution rate, success rate, CodeBERTScore, and API cost, to assess the generated programs' correctness, execution, and efficiency.

SWE-bench [89] is a similar dataset containing 2,294 tasks derived from 12 popular open-source Python repositories in real-world software engineering. Each task requires the model to edit a full codebase based on an issue description, producing a patch that must apply cleanly and pass fail-to-pass tests. The benchmark features long, complex inputs, robust evaluation via real-world testing, and the ability to be continually updated with new issues. However the execution-based testing can be misleading as it does not assess criteria such as comprehensiveness, efficiency, or readability. While there is an additional rubric-based human evaluation where final versions are revised by experts, these human annotators are mainly familiar with Python and tend to dismiss other languages.

3.2.2 Methods and Results. Here, we discuss state-of-the-art methods and results in hypotheses generation, idea formation, and automated experimentation. Figure 2 provides some examples for each approach.

Idea Generation. Several methods have been proposed to generate research ideas, variously using iterative refinement, human alignment, end-to-end autonomous systems, or multi-agent systems [7, 80, 88, 105, 120, 141, 170, 183, 200, 223, 231]. For **iterative refinement**, Hu et al. [80] introduce an iterative planning and search framework aimed at enhancing the novelty and diversity of ideas generated by LLMs. By systematically retrieving external knowledge, the approach addresses the limitations of existing models in producing simplistic or repetitive suggestions. Similarly, IdeaSynth [163] focus on iterative refinement by providing literature-grounded feedback. It represents research ideas as nodes on a canvas to facilitate the idea iteration and composition of different idea facets, enabling researchers to develop more detailed and diverse ideas. Work involving **human alignment** seeks to organize information in ways that mirror human research processes. **Chain of Ideas** (**CoI**) [120] proposes structuring literature into a chain to emulate the progressive development of research domains. This facilitates the identification of meaningful directions and has been shown to outperform existing methods in generating ideas comparable in quality to those produced by human researchers. Scideator [170], in contrast, focuses on recombining facets (e.g., purposes, mechanisms, and evaluations) from existing research papers to synthesize novel ideas. By incorporating automated novelty assessments, Scideator enables users to identify overlaps and refine their ideas. The AI Scientist [141] is a prominent example of an **end-to-end autonomous system**. It is a framework for automating the entire research pipeline, including idea generation, experiment execution, and paper writing. The AI Scientist-v2 [231] employs an agentic framework using tree search and iterative refinement. Among **multi-agent systems**,

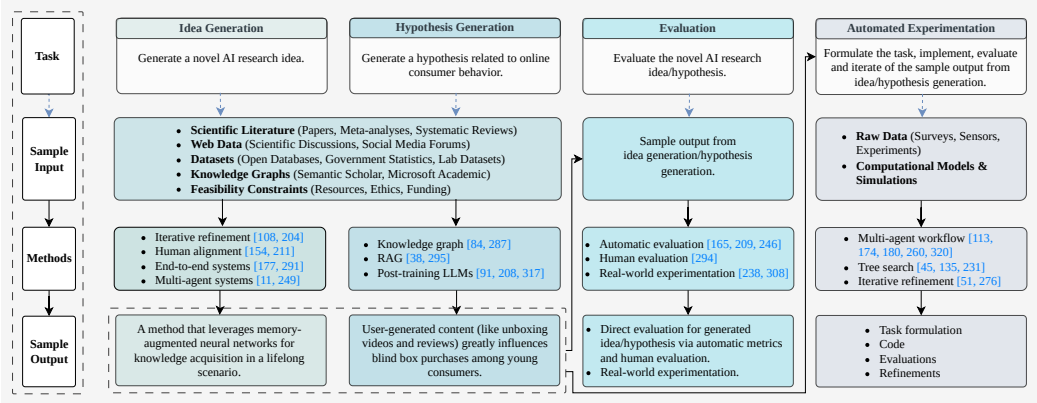


Fig. 2. Examples of idea generation, hypothesis generation, evaluation, and automated experimentation follow a four-component structure: *task*, *sample input*, *methods*, and *sample output*. The *task* defines the goal of each process. For idea/hypothesis generation, *sample input* consists of benchmark datasets for each task, whereas evaluation uses its *sample output*. Automated experimentation uses *sample output* from idea/hypothesis generation for the *task* and raw data/computational models and simulation as *sample input*. *Methods* encompass a taxonomy of approaches. *Sample output* differs by process: idea/hypothesis generation and automated experimentation yield textual outputs (descriptions, explanations, or code), whereas evaluation produces quantitative scores.

VirSci [200] employs an ensemble of virtual agents to collectively generate, evaluate, and refine ideas. It outperforms individual LLMs, underscoring the potential of teamwork in enhancing scientific innovation. ResearchAgent [7] is another a multi-agent LLM framework that enables ideation by generating new research problems, methods, and experiments from existing literature and iteratively refining them through LLM-assisted peer review.

Hypothesis Generation. A considerable number of studies [1, 8, 12, 28, 29, 62, 67, 90, 135, 136, 151, 157, 167, 205, 206, 212, 227, 232, 233, 251, 251] have employed LLMs to generate hypotheses, all of which differ in their approaches. These can be broadly categorized as employing (a) knowledge graphs, (b) RAG, and/or (c) post-training of LLMs. Among **knowledge graph** approaches is KG-CoI [227], which uses an external KG that grounds the model's reasoning in structured knowledge, enhancing the quality of generated hypotheses and reducing hallucination during hypothesis generation. SciAgents [62] is a framework that combines KGs with multi-agent systems and LLMs; it organizes scientific concepts into a graph and enables agents to reason over this structure to generate and iteratively refine hypotheses. For **RAG**, MOOSE-Chem [233] is a retrieval-based system for the rediscovery of high-impact chemistry hypotheses given only a research background. It first retrieves relevant papers from a large corpus, then uses these papers together with the background to generate candidate hypotheses, and finally ranks the hypotheses by quality. Chemist-X [29] is a retrieval-augmented agent for reaction condition discovery in chemical synthesis. It uses LLMs to retrieve information (e.g., reaction conditions and molecules) from chemical literature and molecular databases, helping to narrow the search space for plausible reaction conditions, which are then treated as candidate hypotheses about optimal experimental settings. Major strategies for **post-training LLMs**, include few-shot learning, fine-tuning, and iterative refinement. Qi et al. [167] find that hypotheses generated by few-shot learning are judged by humans as more testable than those generated in a zero-shot setup. They report that while fine-tuning improves the overall quality of hypotheses, the improvement is limited to the domain of training data; in unseen domains,

fine-tuning harms hypothesis quality, particularly the novelty aspect. Zhou et al. [251] iteratively refine hypotheses through **reinforcement learning** (**RL**), aiming to increase the similarity between a research problem and a generated hypothesis. The AI co-scientist [67] employs a multi-agent system relying on a generate–debate–evolve framework to iteratively enhance generated hypotheses.

Evaluation. The evaluation of generated ideas and hypotheses ensures that they are scientifically meaningful and feasible. Evaluation approaches [e.g., 25, 109, 168, 184, 190, 197, 215, 232, 233, 245] can be generally categorized as intrinsic (automatic or human evaluation) or downstream (simulation-based evaluation or real-world experimentation).

For **automatic evaluation**, metrics such as ROUGE [130] have been used to assess the quality of generated hypotheses (or ideas) by measuring their similarity to human-annotated ground truths [197]. Qi et al. [168] leverage LLMs-as-judge to evaluate hypotheses based on four scientific aspects: novelty, relevance to the given background, significance within the research community, and verifiability (i.e., testability). In **human evaluation**, domain experts are involved to assess the quality of generated ideas and hypotheses, especially when ground truth is unavailable. For instance, Yang et al. [232] invited social scientists to provide feedback on the LLM-generated hypothesis, "In collectivist cultures, individuals engage in more conspicuous consumption behaviors compared to individualistic cultures." The social scientists found the hypothesis potentially novel and counterintuitive, as prior research suggests that collectivist cultures typically discourage conspicuous displays of personal wealth. Such expert feedback helps identify hypotheses that are meaningful and worth further investigation. **Real-world experimentation** is employed by Zhang et al. [245], who evaluate candidate hypotheses (protein variants) within an automated biofoundry. Each variant is constructed and evaluated using instruments such as liquid handlers and thermocyclers, together with peripheral devices like plate sealers, shakers, and incubators, to measure enzymatic activity, protein yield, and success rates. Si et al. [190] invited 43 expert researchers to experiment with LLM-generated ideas. Each researcher spent over 100 hours implementing the assigned ideas and wrote a short paper to report the results. Ideas proposed by human researchers were also implemented as a control group. All the papers were then reviewed blindly by expert reviewers. Experimental results from LLM-generated ideas were judged by expert reviewers to be less novel, exciting, or effective compared to those from human ideas.

Automated Experimentation. Experimentation, which can serve as downstream evaluation for empirically validating AI-generated ideas and hypotheses, encompasses task formulation, implementation, evaluation, and iteration. Automated experimentation aims to streamline this workflow, with approaches like neural architecture search [52] and AutoML [74]. LLMs further enhance this by enabling automation through natural language prompts. AutoML-GPT [209] and MLcopilot [244] use LLMs for hyperparameter tuning, while MLAgentBench [85] benchmarks fundamental automation tasks. Recent work explores advanced frameworks incorporating multi-agent collaboration, tree search, and iterative refinement for scientific experimentation.

For **multi-agent workflow**, GVIM [144] enhances chemical research with domain-specific functions, while DrugAgent [138] employs LLMs for task planning in drug discovery. AutoML-Agent [208] integrates retrieval-augmented planning for AutoML tasks, and MLAgentBench [85] benchmarks LLM-driven agents in machine learning experimentation. The Agent-as-a-Judge framework [253] introduces structured agent evaluation. For **tree search**, AIDE [185] applies solution space tree search to refine solutions in Kaggle challenges. The Tree Search for Language Model Agents framework [102] enables LLM agents to plan multi-step interactions using best-first tree search, pruning less promising options. SELA [35] combines LLM-generated insights with Monte Carlo tree search, iteratively refining machine learning experiments by selecting promising configurations and executing them. For **iterative refinement**, APEx [39] automates LLM-based

experimentation with an orchestrator, execution engine, benchmark generator, and model library. OpenHands [219] enables AI agents to interact with software, execute actions in a sandboxed runtime, and collaborate across tasks using predefined benchmarks.

3.2.3 Domains of Application. Studies have addressed ideation and hypothesis generation in NLP, engineering, physics, chemistry, the social sciences, and medicine. Work on automated experimentation has similarly relied on domain-specific datasets to guide the process of designing and testing experiments. However, the designs underlying these systems are typically domain-agnostic. For instance, iterative refinement, human alignment, multi-agent systems, and tree search are low-level methodologies that are applicable across multiple domains. Regarding evaluation, except for automatic evaluation, most evaluation approaches are limited to the domains for which they are designed: human evaluation relies on domain-specific evaluation guidelines, simulation-based evaluation requires domain-specific laboratory configurations, and infrastructure in which real-world experimentation is conducted may differ across domains.

3.2.4 Limitations and Future Directions. A large-scale study [191] comparing human researchers and LLMs finds that LLMs generate ideas judged to be more novel but slightly less feasible, highlighting challenges like limited diversity and self-evaluation failures. Additionally, given that ideas and hypotheses are theoretical and costly to validate, it is unclear whether they could lead to scientific discovery. Furthermore, previous methods lack due diligence through data, and therefore generated ideas and hypotheses are often too general [233]. Moreover, LLMs may end up re-generating recently discovered ideas and hypotheses, since they lack access to recent scientific papers [136]. Their outputs are moreover very sensitive to the framing of input prompts [157]. Future work could focus on improving feasibility and diversity of ideas and hypotheses, consulting scientific papers in real time, and refining ideation and hypothesis generation through data inspection. LLM-automated experimentation has several additional challenges. First, LLMs' propensity to hallucinate results or references disrupt the precise steps required for experimental workflows. They also struggle to integrate and align different modalities, such as video, audio, or sensory data, which are increasingly essential in modern scientific experimentation. Moreover, LLMs lack the critical analysis capabilities necessary to identify flaws or refine hypotheses during experimentation. Particularly in biology and chemistry, they may also struggle with precise reasoning and tool usage, which are vital for ensuring experimental success [172].

3.3 Text-Based Content Generation

Generating scientific content includes generating texts of various types and lengths, each demanding different skills. In this section, we focus on selected subtasks to emphasize the varied challenges inherent in each. **Titles** of scientific papers need to reflect the content of a paper in a few catchy words, while **abstracts** are concise, stand-alone summaries. Approaches to generating **long texts** face challenges such as structuring arguments and maintaining factual consistency. **Related work** generation also requires text summarization skills, though in a more concise form. Generating **bibliographic references** depends on scientific discovery and literature research and has limited room for variation in phrasing, unlike in tasks such as **proof-reading and paraphrasing**. As we discuss, current generative models exhibit varying performance across these subtasks.

3.3.1 Data. Open access research articles are a valuable data source for text-based content generation. These include scientific publisher repositories offering at least some open access content (e.g., Nature portfolio, Taylor & Francis) as well as preprint repositories (e.g., arXiv, bioRxiv). These repositories can be leveraged to develop datasets with pairs of titles and abstracts or abstract and conclusion/future work pairs. Wang et al. [216], for example, extract from PumMed title to abstract

Table 3. Annotated or Task-Specific Datasets for Scientific Text Generation

Dataset	Size	Sources	Application
Abstract-title humor [31]	2,638 humor annotated titles	ML and NLP domain	Title generation
PaperRobot [216]	27K title–abstract, 27K abstract–conclusion/future work, 20K conclusion/future work–title pairs	PubMed	Title, abstract, conclusion, and future work generation
ScisummNet [234]	1,000 papers + 20 citation sentences each	ACL Anthology	Related work generation
CORWA [122]	927 related work sections	NLP domain	Related work generation
CiteBench [57]	358,765 documents + citations	arXiv et al.	Related work generation
LongWriter [9]	6,000 texts (literature, academic writing, popular science, news)	Open datasets	Long text generation
LongWriter-Zero [226]	8,610 instruction tuning requiring outputs exceeding 10,000 words	Open datasets	Long text generation
LongEval [225]	166 high-quality human-authored samples exceeding 2,000 words	arXiv, blogs, and Wikipedia	Long text generation
Casimir [92]	15,646 papers (consecutive versions)	OpenReview	Paraphrasing
ParaRev [91]	48,203 paper paragraphs (consec. versions)	OpenReview	Paraphrasing

pairs, abstract to conclusion and future work pairs, and conclusion and future work to title pairs. Annotated, task-specific datasets for scientific text generation are presented in Table 3.

3.3.2 Methods and Results. Here we survey approaches to generating textual content for scientific papers, such as titles, abstracts, related work sections, and bibliographies. An overview of these processes is given in Appendix B.3.

Title Generation. Generating appropriate titles for scientific papers is an important task because titles are the first access point of a paper and can attract substantial reader interest; titles can also influence the reception of a paper [113]. Consequently, several works have targeted generating titles automatically, often using paper abstracts as input. For example, Mishra et al. [148] use a pipeline of three modules, viz. generation by transformer based (GPT-2) models, selection (from multiple candidates) and refinement. Chen and Eger [31] also leverage transformers for title generation from abstracts, including humorous title generation. Their results show that none of the applied models (BART, GPT-2, T5, GPT-3.5) can adequately generate humorous titles. However, these findings may partially reflect the subjective nature of humor in scientific communication. Sebo et al. [186] find that GPT-4o generated titles are preferred by human raters over human titles, given the abstract. Wang et al. [216] address the problem differently by drafting titles based on future work sections of previous related papers.

Abstract Generation. There are several approaches trying to assess the capabilities of LLMs to generate abstracts based on context information such as paper titles, journal names, keywords, or the full text of the paper. Hwang et al. [87] assess the ability of GPT-3.5 and -4 to generate abstracts based on a full text. The results are manually evaluated using the Consolidated Standards of Reporting Trials for abstracts, a standard that aims to enhance the overall quality of scientific abstracts [79]. While the readability of GPT-generated abstracts is rated higher, their overall quality is inferior to the originals. Wang et al. [216] generate abstracts from titles, leveraging transformers and knowledge bases. Gao et al. [58] collect 50 publications from five medical journals and use ChatGPT to generate abstracts based on their titles and journal names. Both original and generated abstracts are evaluated using AI output detectors and human reviewers. Human reviewers are able to identify 68% of the generated abstracts, but misclassified 14% of original abstracts as LLM-generated. Farhat et al. [53] evaluate the performance of ChatGPT generating abstracts based on three keywords, the name of a database (Scopus or Web of Science), and the task to analyze bibliographic data in the domain indicated by the keywords. Manually comparing the generated

abstracts to original abstracts on the same topic, the authors conclude that at the time of the study, ChatGPT was not a trustworthy tool.

Long Text Generation. The AI Scientist [141, 231] produces complete scientific manuscripts by leveraging outputs from earlier stages of the scientific lifecycle, such as experimental results, intermediate analyses, and candidate citations. By conditioning on these intermediate outputs, the system drafts papers largely conforming to domain-specific conventions, including citation and disciplinary writing norms. However, the system does not explicitly address the challenge of maintaining global narrative coherence, nor does it provide principled mechanisms for modeling long-range logical dependencies and argument consistency across extended texts. Beyond end-to-end research automation frameworks, other work focuses more specifically on long-form text generation itself. LongWriter [9] addresses long-form text generation by targeting long-range coherence and structural consistency. The model introduces hierarchical attention mechanisms to enhance thematic consistency across extended texts and employs tailored fine-tuning strategies to better align generated outputs with user prompts. LongEval [225] provides a systematic evaluation of long-text generation under both direct and plan-based generation paradigms across academic, encyclopedic, and blog-style domains. The findings suggest that larger, general-purpose, instruction-tuned models often perform comparably to specialized, smaller long-text models (e.g., LongWriter), raising questions about the marginal benefits of domain-specific fine-tuning for long-form generation. Motivated by the need for stronger structural control, recent work has increasingly moved beyond standard **supervised fine-tuning** (**SFT**) toward approaches based on RL. LongWriter-Zero [226] demonstrates that RL without SFT can enable ultra-long text generation (i.e., 10,000+ words). By employing composite reward models that jointly evaluate length, quality, and formatting constraints, LongWriter-Zero achieves competitive or even superior performance relative to proprietary models (i.e., Claude-Sonnet-4, Gemini-2.5-Pro) in long-form generation tasks. Similarly, LongReward [242] leverages RL with custom-designed reward signals that emphasize coherence, factual accuracy, and linguistic quality. These reward mechanisms are particularly relevant for scientific text generation, where accuracy and adherence to domain-specific conventions are crucial.

Related Work Generation. There has been a substantial body of work on related work generation through text summarization, with variances in the approach (extractive or abstractive) and the citation text length (sentence- or paragraph-level). Extractive approaches focus on selecting sentences from cited papers and reordering them to form a paragraph of related work. For instance, Hoang and Kan [78] propose an extractive summarization approach that selects sentences describing the cited papers. Subsequent extractive approaches differ from this approach in how they order the extracted sentences: while Wang et al. [214] assume that the sentence order is given, Hu and Wan [82] and Deng et al. [44] take advantage of an automatic approach to reorder sentences based on topic coherence. Most abstractive approaches are based on language models and focus on generating either a single sentence from a single reference, or a single paragraph from multiple references. Typically, the abstractive process is repeated until a related work section is complete. AbuRa'ed et al. [2] introduce an abstractive summarization approach to generate citation sentences in a single-reference setup. Their approach has been trained on the **ScisummNet** corpus [234] with paper abstracts as inputs and citation sentences as outputs. Li et al. [122] further extend this idea to a multiple-reference setup, namely generating a paragraph of citation sentences from various cited papers. Their approach has been trained on the **CORWA** corpus [122] to generate both citation and transition sentences. Recently, several works have explored LLMs for related work generation. For instance, Şahinuç et al. [177] use instruction promoting with LLMs, an alternative to extractive and abstractive approaches, to generate citation sentences. Martin-Boyle et al. [146] employ a

citation graph coupled with LLMs to produce a related work section. Li and Ouyang [124] use LLM prompting to extract features that capture relationships between citing and cited papers; these features are then composed into a new prompt that enables the LLM to generate a related work section. Şahinuç et al. [176] introduce a multi-turn evaluation framework for assessing the quality of AI-generated related work sections. The framework uses expert preferences to align with human judgment, and iteratively evaluates section drafts and generates natural-language feedback for revision. Overall, extractive approaches, while factual, often lack fluency and coherence. In contrast, abstractive approaches and instruction prompting, which are based on (large) language models, do not struggle with these issues, however, they suffer from factual errors, known as hallucination.

Citation Generation. Bibliographic references are important for ensuring the scientific integrity of a paper. However, in many cases, cited references generated by LLMs such as ChatGPT are reported not to exist—that is, they are hallucinated or incorrect [53, 84, 123]. Most studies on this phenomenon are case studies presenting one or more examples. In a study by Walters and Wilder [211], GPT-3.5 and -4 are used to generate 84 literature reviews containing 636 bibliographic citations. 55% of the GPT-3.5 citations were fabricated, compared to 18% of GPT-4's. Additionally, 43% of non-fabricated GPT-3.5 citations contain substantive citation errors, compared to 24% for GPT-4. Despite this notable improvement, issues with citation fabrication and errors persist. In ScholarPilot [220], retrieval tokens are generated to query citation databases and the retrieved references are directly fed into the model to augment the text generation process. The model is based on Qwen-2.5-7B and fine-tuned on a corpus of 500K arXiv papers. It outperforms other Qwen-2.5 variants according to model-based evaluation using GPT-4o.

Proof-reading and Paraphrasing. LLMs have been reported to provide useful assistance for scientific writing tasks, such as proof-reading, or providing suggestions for improving the writing style [179]. Additionally, some authors emphasize that LLMs can be helpful especially for non-native English speakers with regards to grammar, sentence structure, vocabulary and even translation, effectively serving as an English editing service [24, 99]. Most papers on this topic are case studies, with results qualitatively evaluated by a human expert. Jourdan et al. [92] introduce Casimir, a dataset of 15,646 OpenReview articles with sentence-level paired revisions. A later extension, ParaRev [91], provides paragraph-level pairs, with a subset manually annotated with revision instructions. Experiments show that detailed instructions substantially improve automated revision quality under both statistical and model-based evaluations.

Evaluation. Generated scientific texts are evaluated both with task-specific methods and those general to LLMs. Traditional reference-based metrics, such as BLEU [156] or ROUGE [130], require associated reference output as a ground truth and have shown low correlation with human judgments [134]. With the rise of deep learning and LLMs, model-based approaches are gaining increasing importance [59, 115]. Although not exclusive to scientific texts, they focus on highly relevant dimensions such as coherence, consistency, and fluency [111, 139]. In a more domain-specific approach, Huang et al. [86] propose **PaperEval**, a multi-agent system powered by LLMs to assess scientific paper quality across various dimensions (question, method, result, and conclusion). Its feasibility and effectiveness in distinguishing high-quality from low-quality papers have been tested on three evaluation datasets across four scientific fields (mathematics, physics, chemistry, and medicine).

3.3.3 Domains of Application. Text-based content generation is relevant for all scientific domains. Liang et al. [128] conduct a large-scale analysis across 1M papers published between January 2020 and September 2024 to measure the prevalence of LLM-modified content over time. The papers they investigated were from a variety of disciplines and published on arXiv, bioRxiv, or Nature portfolio.

Their results show the largest and fastest growth in computer science, with up to 22% of the papers containing LLM-modified content; mathematics had the lowest prevalence (up to 9%). According to arXiv's September 2024 Natural Language Learning & Generation report [112], top-cited papers show notably fewer markers of AI-generated content as compared to random samples.

3.3.4 Limitations and Future Directions. LLMs and LLM applications demonstrate strong capabilities in tasks such as proofreading and paraphrasing, but still exhibit notable limitations. In particular, factual consistency, truthfulness, and bibliographic citations are affected by hallucination and require human oversight, even though the hallucination rate decreases in newer models. Rapid advances in LLMs further complicate evaluation, quickly rendering existing methods outdated and hindering reproducibility. Evaluating AI-generated text remains challenging: statistical metrics like BLEU and ROUGE lack semantic depth, model-based evaluations can be unreliable, and many tasks depend on small-scale human evaluations due the scarcity of appropriate benchmarks. Consequently, future research must prioritize robust benchmarks and datasets for scientific text generation. Beyond technical challenges, ethical concerns, including authorship, plagiarism, bias, and truthfulness, underscore the need for trustworthy and responsible AI systems.

3.4 Multimodal Content Generation and Understanding

Multimodal content generation in the scientific domain refers to generating multimodal scientific content such as figures and tables in scientific papers, or slides and posters in a post-publication process. Automating such tasks via AI is important for multiple reasons: (i) generating high-quality figures, tables, slides and posters is costly in terms of effort and time; (ii) high-quality multimodal content in a paper can positively influence citation or acceptance decisions [110]; (iii) tables, figures, posters, and slides make scientific content more accessible and compact. **Multimodal content understanding** refers to interpreting scientific images and tables, a prerequisite for answering questions about them or providing captions or summaries. These tasks are likewise effortful and time-consuming for human authors, pointing to a role for AI assistance.

3.4.1 Data. In this section we detail datasets and benchmarks for selected multimodal tasks. Further tasks, including table understanding and generation, can be found in Appendix B.4, along with an overview table.

Scientific Figure Understanding. Scientific figure understanding benchmarks typically contain summaries or QA pairs for scientific figures. Kembhavi et al. [96] provide a dataset with over 5K richly annotated diagrams and over 15K questions and answers. **FigureQA** [93] is a synthetic dataset of over 100K scientific-style (dot-)line plots, vertical and horizontal bar graphs, and pie charts, along with 1M associated questions generated using 15 different templates. Later research focuses on more challenging and realistic QA pairs. **ChartQA** [147], for instance, provides complex reasoning questions concerning charts from various science-related sources. **CharXiv** [221] is a manually curated dataset of descriptive and reasoning questions about 2.3K "natural, challenging, and diverse" charts from aXiv papers. **ArXivQA** [119], a dataset of 35K scientific figures sourced from arXiv, contains 100K GPT-4o-generated, manually filtered QA pairs. **SPIQA** [162] is a dataset of 270K manually and automatically created QA pairs that interpret complex scientific figures and tables. Xu et al. [228] treat the problem of chart summarization with a 190K-instance dataset that builds on **ChartSumm** [171], itself containing more than 84K charts along with their metadata and descriptions.

Scientific Figure Generation. Several datasets for scientific figure generation have been proposed. **DaTikZ** [15–17] provides captions of scientific figures as instructions along with corresponding TikZ code, sourced from the TEX Stack Exchange and arXiv submissions. A later and larger

version [70] improves the data quality through VLM-generated descriptions instead of noisy captions. For the task of converting sketches into scientific figures, the **SketchFig** [17] benchmark provides 549 figure–sketch pairs sourced from the TEX Stack Exchange. **DiagramGenBenchmark** [222] contains 6,713 training and 270 testing samples for diagram generation, and 1,400 training and 200 testing samples for diagram editing, sourced from freely licensed DOT and TikZ diagrams in VGQA and DaTikZ. **Plot2XML** [41] includes 247 complex diagrams sourced from conference papers spanning multiple domains. **ChartMimic** [189] is a manually curated benchmark of 1,000 triplets of (instruction, code, figure) instances for chart generation across various domains, including physics and economics. The data comes from human annotators writing Python code for prototype charts. **ScImage** [243] contains targeted template instructions focusing on different comprehension dimensions (spatial, numeric, attribute); for a subset of the data, the authors also provide reference figures that were manually evaluated as being of high quality. In contrast to the other benchmarks, ScImage also contains instructions in languages other than English. **SciDoc2DiagramBench** [149] is a benchmark comprised of 1,000 diagrams extracted from the presentation slides of 89 ACL papers, along with human-written "intents". The intents describe how the content from each paper can be translated into the diagrams for presentation purposes. **nvBench** [142], a benchmark of 25K tuples of natural language queries and corresponding visualizations, is drawn from 153 databases and contains more than 7K visualizations across seven chart types. nvBench is synthesized from natural language to SQL benchmarks.

Scientific Slide and Poster Generation. Early efforts at automating the generation of presentation slides from scientific papers relied on relatively small datasets for development and evaluation. For example, Sravanthi et al. [195] generate presentations from a modest collection of eight papers. Similarly, Hu and Wan [81] and Wang et al. [218] use 1,200 and 175 paper–slide pairs, respectively. For scientific poster generation, Qiang et al. [169] collect 25 pairs of scientific papers and their corresponding posters. Such early datasets are often inaccessible due to various restrictions on distribution.

Two free-content datasets, **DOC2PPT** [56] and **SciDuet** [201], have since emerged as widely used resources for scientific slide generation. The former consists of 5,873 scientific documents and their associated presentation slide decks, totalling around 100K slides, drawn from three research communities: computer vision, natural language processing, and machine learning. SciDuet has 1,088 papers and 10,034 slides from conferences such as ICML, NeurIPS, and the ACL Anthology. More recently, **SlidesBench** [61] has provided 7K training and 585 test examples, each containing 20 slides on average, sourced from the web and covering ten broad domains (art, marketing, environment, technology, etc.).

3.4.2 Methods and Results. Here we survey approaches to multimodal content generation and understanding; a summary table, along with additional related works, is provided in Appendix B.4.

Scientific Figure Understanding. Scientific figure understanding is typically framed in terms of (visual) QA—e.g., whether models are able to adequately answer questions on a given input figure [96]. Several recent studies train baseline models such as transformers or relation networks [93, 147], such as relation networks or apply recent LLMs to benchmarks [119, 221]. They generally show large gaps between proprietary models like GPT-4o and the strongest open models, and between all models and human performance. For chart summarization, Rahman et al. [171] find that older pre-trained language models such as BART and T5 suffer from hallucination and disregard important data points. Xu et al. [228] propose ChartAdapter, a lightweight transformer module that can be combined with LLMs for improved modeling of chart summarization. **Evaluation** of scientific figure understanding benchmarks tends to employ automatic metrics, many of which are now

considered outdated or unreliable. For example, Xu et al. [228] use BLEU and ROUGE. By contrast, Pramanick et al. [162] report both human and automatic evaluation, the latter including not just traditional QA metrics such as METEOR, ROUGE, and BERTScore, but also novel LLM-based metrics.

Scientific Figure Generation. Although work on automating visualization for science dates back to the 1980s at least, most recent work, including that covered here, involves multimodal LLMs. **AutomaTikZ** [16] and **TikZero** [15] fine-tune custom LLMs on DaTikZ for the text-to-TikZ task. **TikZilla** [70] extends this line of work by applying RL using a custom image encoder for more semantically meaningful rewards. **VisCoder** [154] uses Python libraries for visualization code generation. **DeTikZify** [17] reconstructs scientific figures in TikZ from sketches or images. **Draw with Thought** [41] uses multimodal LLMs for a preliminary coarse-to-fine planning approach, followed by structure-aware code generation with a self-refine mechanism for reconstruction. **DiagramAgent** [222] combines generation, reconstruction, and editing by introducing AI agents (plan, code, diagram-to-code, check). Shi et al. [189] aim to generate Python code from instructions and/or images, specifically focusing on charts. They evaluate three proprietary and 11 open-weight LLMs on their ChartMimic benchmark, finding that even the best models (GPT-4 and Claude-3-opus) have substantial room for improvement. Zhang et al. [243] provide a template-based approach to evaluate various multimodal LLMs in generating scientific images. They explore those that generate TikZ and Python code for images, as well as those that generate images directly,[4] and also consider different input languages (English, German, Chinese, Farsi). They find that, except for GPT-4o, most models struggle substantially in generating adequate scientific images. Zala et al. [240] explore the diagram generation task where LLMs first generate diagram plans and then the diagrams themselves. Mondal et al. [149] explore the same task with an additional refinement—namely, feedback from multiple critic models—to enhance factual correctness. **Evaluation** uses automatic metrics including DreamSim [55] for image similarity, Crystal BLEU [51] for code similarity, and CLIPScore [76] for text–image similarity, as well as manual evaluation by domain experts. The former are typically reported to have low or medium correlation with the latter, establishing the need for domain-specific evaluation in future work.

Scientific Slide and Poster Generation. Early work on scientific slide generation [81, 116, 195, 218] used extractive approaches, copying text from papers to serve as slide content. With advances in neural text generation, which enabled models to paraphrase and condense source text rather than simply copy it, later systems explored abstractive approaches to generate the textual content of slides, such as with sequence-to-sequence models [56, 201]. Researchers are now exploring (multimodal) LLMs for generating slides using natural language prompts [11, 145, 150]. AutoPresent [61], for example, fine-tunes an LLM using SlidesBench to generate slides from detailed natural language instructions with images, detailed instructions only, or high-level instructions. However, modern systems still take an extractive approach to multimodal content, copying images or tables directly from the original papers instead of generating new ones [11, 56, 150, 195, 201]. The task of poster generation has received less attention, with studies mainly exploring ML-based methods for generating key content and panel layouts [169, 230]. **Evaluation** of slide generation has involved both human judgments and automatic metrics (most commonly ROUGE). Fu et al.

[4]Most scientific figure generation approaches are based on code generation, while direct image generation approaches are rare (but see also the recent Google publication PaperBanana [252], which opts for direct image generation, albeit without comprehensive comparison). The likely reason is that code generation methods usually work better (e.g., allowing higher precision when plotting numeric inputs) and their output is easier for users to post-edit. Within code generation, those employing general-purpose languages like Python seem to perform better than those targeting specialized drawing languages like TikZ, though current comparisons are restricted to individual use cases [243].

[56] introduce some novel metrics: Longest Common Figure Subsequence, which measures the quality of figures in the generated slides; Text–Figure Relevance, which assesses the similarity between the text of the ground truth slide and the generated slide containing the same figure; and Mean Intersection over Union, which evaluates layout quality. Recent studies have also used LLMs to assess the quality of generated slides [11, 145]. For scientific poster generation, in addition to conducting user studies, Qiang et al. [169] measure the **mean squared error** (**MSE**) of the panel parameters such as panel size and aspect ratio.

3.4.3 Domains of Application. Many recent datasets for multimodal content generation and understanding are drawn from arXiv and more generally the STEM domain. Models such as DeTikZify and AutomaTikZ have also been fine-tuned on such data. This indicates a limitation both in terms of application scenarios and model assessments, as these may perform worse when applied in cross-domain scenarios.

3.4.4 Limitations and Future Directions. Limitations common to the approaches we have discussed include (1) the comparatively small datasets for fine-tuning models; (2) sub–human-level performance on recent benchmarks, particularly for non-proprietary models; (3) over-representation of arXiv and STEM domains in training and evaluation; (4) models' lack of reasoning abilities; and (5) lack of reliable, task-specific evaluation methods, particularly for generative tasks. Some problems are task-specific: for example, for table generation, the input text may be very long, which constitutes a problem for many current LLMs; for slide generation, there are no approaches that can generate slides from multiple documents (e.g., for tutorials) or that generate content beyond that contained in a reference paper (which may be necessary for including relevant background material); for figure generation, models like AutomaTikZ are trained on captions, which are often not appropriate for generating the corresponding figure (e.g., a caption may simply be "Proof of Theorem X"), which is an instance not only of small but also deficient training and evaluation data. Reproducibility is hampered by some studies' use of datasets that are not freely distributable—for example, one study [70] reports that only a third of all arXiv-harvested papers are permissively licensed.

3.5 Peer Review

The highest standard in scientific quality control is **peer reviewing**. In this process, authors present their scientific arguments (e.g., the findings of a study, or a grant proposal) in form of a manuscript to their peers, who then assess its scientific validity and quality. Often this process has multiple stages, as shown in Figure 3. For instance, in the ACL Rolling Review (ARR) system, reviewers write detailed assessments whose arguments and questions the authors may then rebut and clarify to convince the reviewers to raise their scores. A meta-reviewer then evaluates this discussion and submits to the program chairs an acceptance/rejection recommendation, which may or may not be adopted. During this process, multiple (potentially multi-modal) artifacts are processed and created—mainly the manuscript under review, the written reviews, the author–reviewer discussion, and the meta-review. In general, peer review is considered a challenging, subjective process, where reviewers are prone to unfair biases like sexism and racism, and often rely on expedient heuristics [173, 199]. In some fields, these problems are compounded by an exploding number of submissions [106], pushing review systems to their limits. To counteract this situation, researchers have addressed various several problems under the umbrella of AI-supported peer review. Related overviews on the topic, or on some of its aspects, point to its importance and timeliness [27, 47, 103, 107, 131, 198]. Here, we focus on approaches to the most established tasks related to peer review, following the same structure as in previous sections.

3.5.1 Data. Peer reviewing data is scarce: few scientific communities publish reviewing artifacts at all, let alone under permissive licenses. Exceptions include PeerRead [94], which collects review data from various sources (e.g., ACL, ICRL), and CiteTracked [159], which also contains citation

Table 4. Annotated and/or Task-Specific Datasets for Analyzing Peer Reviewing

Dataset	Size	Sources	Application
HedgePeer [64]	2,966 documents	ICLR 2018	Uncertainty detection
PolitePeer [20]	2,500 sentences	ICLR et al.	Politeness analysis
COMPARE [192]	1,800 sentences	ICLR	Comparison analysis
ReAct [36]	1,250 comments	ICLR	Actionability analysis
MReD [187]	7,089 meta-reviews	ICLR	Meta-review analysis and generation
CiteTracked [159]	3,427 papers, 12K reviews	NeurIPS	Citation prediction
MOPRD [132]	6,578 papers	PeerJ	Review comment generation
Revise and Resubmit [108]	5.4K papers	F1000Research	Tagging, linking, version alignment
ORB [203]	92,879 reviews	OpenReview, SciPost	Acceptance prediction
ARIES [43]	3.9K comments	OpenReview	Feedback–edits alignment, revision generation
DISAPERE [97]	506 review–rebuttal pairs	ICLR	Review action analysis, polarity prediction, review aspect
PeerReviewAnalyze [63]	1,199 reviews	ICLR	Review paper section correspondence, paper aspect category detection, review statement role prediction, review statement significance detection, meta-review generation
JitsuPeer [166]	9,946 review and 11,103 rebuttal sentences	ICLR	argumentation analysis, canonical rebuttal scoring, review description generation, end2end canonical rebuttal generation

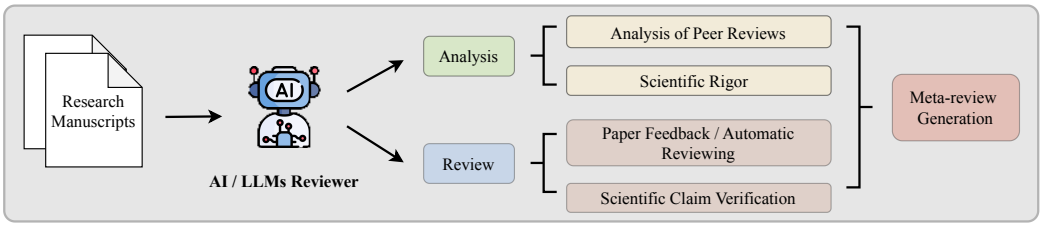


Fig. 3. Process of AI-enhanced peer review. In the analysis step, the LLM reviewer examines research manuscripts and evaluates peer reviews to assess scientific rigor. The review step involves providing feedback on the paper and verifying scientific claims. Finally, the gathered information is synthesized to generate a final meta-review.

information. NLPeer [49], a model for how larger-scale open publishing of raw peer reviewing data could work, uses a corpus of ARR reviews where the consent of the respective actors was explicitly obtained. Several annotated datasets support tasks in peer review analysis, an overview of which is provided in Table 4. Recent curated resources have focused on complex tasks such as understanding the effect of peer review feedback on manuscript revisions [43] or identifying the attitudes underlying specific criticisms in reviews [166].

3.5.2 Methods and Results. Pre-LLM approaches [e.g., 64] were mostly based on traditional ML methods, targeting simpler analyses involving sentence classification tasks. Later, deep learning approaches [e.g., 83] (including those based on pre-trained language models) and more complex analyses (e.g., of argumentation) defined the state of the art in computational peer review processing. Researchers have now started exploring LLMs in prompting-based frameworks, for complex tasks like peer review generation and meta-review generation [e.g., 137]. Below we provide an overview of the methods and results for the most common tasks under the umbrella of peer reviewing. Information on related tasks, such as scientific claim verification, can be found in Appendix B.5.

Analysis of Peer Reviews. Prior works have analyzed peer reviews for a multitude of aspects, like uncertainty [64], politeness [20], and sentiment [26]. However, given that science as a whole and especially peer review relies to a large extent on convincing peers, large efforts have been spent on understanding arguments or argument-related aspects (e.g., substantiation of arguments) in peer review artifacts [e.g., 54, 83]. Here, most approaches have used pre-trained language models. For instance, Hua et al. [83] work on mining the arguments in peer reviews using conditional random fields, LSTMs, and BERT. In contrast, Guo et al. [71] and Fromm et al. [54] fully rely on (domain-adjusted) pre-trained language models for argument mining: SciBERT, ArgBERT, and PeerBERT. Cheng et al. [34] leverage multi-task learning approaches based on LSTMs and BERT. In a similar vein, Purkayastha et al. [166] study the generation of rebuttals for author–reviewer discussions based on jiu-jitsu argumentation, a specific argumentation theory.

Paper Feedback and Automatic Reviewing. Several works have explored methods to provide general feedback on scientific publications to fully or partially automate peer reviews. For instance, Li et al. [118] propose a multi-task learning approach for peer review score prediction, where different aspect score prediction tasks (e.g., novelty) can inform each other. Ghosal et al. [65] leverage the concept of sentiment to predict scores based on review texts. In a similar vein, Bharti et al. [21] leverage paper–review interactions to predict final decisions of a review process. Wang et al. [217] focus on explainability during review score prediction for several review categories by constructing knowledge graphs (e.g., representing the background of a paper). More recent works have included generation of feedback texts into the problem setup. Bartoli and Medvet [14] frame the problem as exploring the potential of GPT-2 for conducting academic fraud by generating fake reviews. In contrast, Yuan et al. [239] ask whether it would be possible to automate reviewing leveraging targeted summarization models, a recently trending topic. Liu and Shah [137] explore prompting-based review generation with various LLMs, finding that GPT-4 performs best and that task granularity matters. Robertson [175] similarly finds GPT-4 to be "slightly" helpful for peer reviewing, and Liang et al. [127] demonstrate in a comparative study that users of a GPT-4–based peer review system found the feedback to be (very) helpful more than half the time. D'Arcy et al. [42] show a multi-agent approach with LLMs that engage in a discussion to produce better results than a single model.

Meta Review Generation. Kumar et al. [104] tackle meta-review generation using a multi-encoder transformer network, and Li et al. [121] use a multi-task learning approach for refining pre-trained language models for the task. Stappen et al. [196] explore the aggregation of reviews for providing additional computational decision support to editors based on uncertainty-aware methods like soft labeling. Both Zeng et al. [241] and Santu et al. [180] rely on LLMs, which they specifically prompt for the task. In contrast, Purkayastha et al. [165] propose a conversational agent that interactively supports meta-reviewers in their decision-making.

3.5.3 Domains of Application. Peer review setups vary in the aspects to review for, scoring schemes, expected review length, and the stages and dynamics of the reviewer–author and reviewer–reviewer discussions. Thus, while none of the studies presented above targets a problem unique to any scientific discipline, the particularities will likely be very different for each specific community and existing systems must be evaluated or adapted before deployment.

3.5.4 Limitations and Future Directions. The variety of scientific domains whose peer review has been studied is still limited. Most work relies on data from OpenReview, a platform used primarily by the representation learning and NLP communities; other disciplines may be wholly unrepresented in existing peer review datasets. Even within communities, there can be great variance in how peer reviews are conducted, which limits the comparability of approaches. Scientific rigor in particular

remains unexplored, despite being a critical aspect of peer review; most existing studies rely on predefined rigor checklists that are not easily scalable or transferable across domains. Given these gaps, future research could benefit from exploring new domains of peer review, developing domain adaptation approaches, and advancing models for assessing scientific rigor. Reproducibility studies and larger benchmarks could further advance the field. Additionally, ethical concerns demand the prioritization of research on *trustworthy* AI support for peer review, ensuring that human experts retain autonomy in the process.

4 Ethical Concerns

There is by now a growing body of work addressing major ethical concerns related to generative AI. Baldassarre et al. [10], for instance, present a systematic literature review regarding the social impact of generative AI, especially taking into account 71 papers on ChatGPT. They identify privacy, inequality, bias, discrimination, and stereotypes as areas of concern. Another literature review on ethics and generative AI [72] identifies jailbreaking, hallucination, alignment, harmful content, copyright, private data leakage, and impacts on human creativity as topics of increasing interest. This review furthermore identifies 19 distinct clusters of ethics topics, with fairness/bias being the most frequently mentioned, followed by safety, harmful content/toxicity, hallucination, privacy, interaction risks, security/robustness on ranks two to six, with writing/research on rank 18. Ali and Aysan [3] review 364 recent papers on generative AI and ethics published from 2022 to 2024 in different domains including the use of generative AI in scientific research. Topics identified as critical to academia are authenticity, intellectual property, and academic integrity. Sun et al. [202] argue that in application areas such as scientific research, ensuring the trustworthiness of LLMs is crucial. Dergaa et al. [45] find the use of AI chatbots in academic research heavily associated with stigma and suggest (inter alia) transparent disclosure as a mitigation strategy. However, there is evidence from more recent research that AI disclosure has negative effects on users' trust [181].

Truthfulness—i.e., the accurate representation of information, facts and results—is a particularly essential challenge for LLMs. Benchmarks and datasets developed to evaluate different aspects of it include TruthfulQA [133], HaluEval [117], and FELM [30], for identifying hallucinations; Self-Aware [237] for assessing awareness of knowledge limitations; FreshQA [210] and Pinocchio [237] for exploring adaptability to rapidly evolving information; and TrustLLM [202], which incorporates existing and new datasets not just on truthfulness but also safety, fairness, robustness, privacy, and machine ethics. Evaluations with TrustLLM show that proprietary LLMs generally outperform open-source LLMs in trustworthiness, Llama2 [207] being a notable exception. However, proprietary LLMs (including Llama2) often struggle to provide truthful responses when relying solely on internal knowledge. Their performance does improve significantly with additional external knowledge, and there exists a positive correlation between trustworthiness and the functional effectiveness of the model in downstream tasks.

Editors of scientific publications are particularly challenged by the increasing proportion of AI-generated text in manuscripts [33, 69, 101, 129] and by its potential use in peer reviewing (cf. Section 3.5), with AI hallucination paired with author misconduct being noted as a major problem. Sakai et al. [178], for instance, find that nearly 300 of the papers published at ACL, NAACL, and EMNLP in 2024 and 2025 contain at least one hallucinated reference. The editors of the *Journal of Information Technology* have elaborated on the limitations and risks of using generative AI in the production of scientific publications [182], referring to an eight-point "Artificial Imperfections" test to illustrate current limitations of generative AI: AI is (1) brittle, (2) opaque, (3) greedy, (4) shallow and tone-deaf, (5) manipulative and hackable, (6) biased, (7) invasive, (8) "faking it" [224, p. 107]. Guidelines proposed by the editors of *iMeta* similarly hold authors fully responsible for the integrity of their manuscripts, and for addressing ethical concerns and ensuring the accuracy and fairness

of AI-generated content, complying with data protection and privacy laws, and considering the relevant copyright and intellectual property issues [164]. They furthermore state that AI-assisted technologies cannot be recognized as authors, that the use of generative AI must be transparently disclosed (including the prompts and specific versions of the tools used), that AI-generated images and multimedia should be accepted only when specifically allowed, and that the use of AI in the reviewing process is expressly prohibited.

5 Conclusion

In this article, we surveyed approaches in the area of AI4Science, with a particular focus on recent LLM-based methods. We examined five key aspects of the research cycle: (1) search, (2) experimentation and research idea generation, (3) text-based content production, (4) multimodal content production, and (5) peer review. For each topic, we discussed relevant datasets, methods, and results, including evaluation strategies, while highlighting limitations and avenues for future research. Ethical concerns featured prominently in our survey, given the potential for misuse and challenges in maintaining scientific integrity in the face of AI-assisted content generation.

Overall, while recent advances suggest that AI systems can meaningfully support certain components of the scientific workflow, their current capabilities remain limited and uneven. Many methods rely on narrow benchmarks, struggle with generalization, or require substantial human oversight to avoid errors, bias, or misinterpretation. (An example is the heavy reliance of current computational approaches within peer reviewing on data coming from OpenReview.) Consequently, AI4Science should presently be viewed as a complementary set of tools rather than a transformative replacement for human expertise. We hope that this survey inspires new initiatives in AI4Science, driving faster, more efficient, and more inclusive scientific discovery, experimentation, reporting and content synthesis—while upholding the highest ethical standards. As the ultimate goal of science is to serve humanity, we hope these advancements will accelerate knowledge creation and enhance the accessibility and reliability of research, leading to improved healthcare, medical treatments, economic processes, among a myriad of other societal benefits.

References

[1] Abbi Abdel-Rehim, Hector Zenil, Oghenejokpeme Orhobor, Marie Fisher, Ross J. Collins, Elizabeth Bourne, Gareth W. Fearnley, Emma Tate, Holly X. Smith, Larisa N. Soldatova, et al. 2025. Scientific hypothesis generation by large language models: Laboratory validation in breast cancer treatment. *Journal of the Royal Society Interface* 22, 227 (2025), 20240674.

[2] Ahmed AbuRa'ed, Horacio Saggion, Alexander Shvets, and Àlex Bravo. 2020. Automatic related work section generation: Experiments in scientific document abstracting. *Scientometrics* 125 (2020), 3159–3185.

[3] Hassnian Ali and Ahmet Faruk Aysan. 2024. Ethical dimensions of generative AI: A cross-domain analysis using machine learning structural topic modeling. *International Journal of Ethics and Systems* 41, 1 (2024), 3–34.

[4] Maha Amami, Gabriella Pasi, Fabio Stella, and Rim Faiz. 2016. An LDA-based approach to scientific paper recommendation. In *Natural Language Processing and Information Systems: 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016.* Springer, 200–210.

[5] Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Tian, D'arcy Mike, et al. 2024. OpenScholar: Synthesizing scientific literature with retrieval-augmented language models. arXiv:2411.14199. Retrieved from https://arxiv.org/abs/2411.14199.

[6] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of the 12th International Conference on Learning Representations.*

[7] Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2025. ResearchAgent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics.* 6709–6738.

[8] Jiaxin Bai, Yicheng Wang, Tianshi Zheng, Yue Guo, Xin Liu, and Yangqiu Song. 2024. Advancing abductive reasoning in knowledge graphs through complex logical hypothesis generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* 1 (2024), 1312–1329.

[9] Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongWriter: Unleashing 10,000+ word generation from long context LLMs. In *Proceedings of the 13th International Conference on Learning Representations*.

[10] Maria Teresa Baldassarre, Danilo Caivano, Berenice Fernandez Nieto, Domenico Gigante, and Azzurra Ragone. 2023. The social impact of generative AI: An analysis on ChatGPT. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*. 363–373.

[11] Sambaran Bandyopadhyay, Himanshu Maheshwari, Anandhavelu Natarajan, and Apoorv Saxena. 2024. Enhancing presentation slide generation by LLMs with a multi-staged end-to-end approach. In *Proceedings of the 17th International Natural Language Generation Conference*. 222–229.

[12] Sachin Banker, Promothesh Chatterjee, Himanshu Mishra, and Arul Mishra. 2024. Machine-assisted social psychology hypothesis generation. *American Psychologist* 79, 6 (2024), 789.

[13] Trapit Bansal, David Belanger, and Andrew McCallum. 2016. Ask the GRU: Multi-task learning for deep text recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*. 107–114.

[14] Alberto Bartoli and Eric Medvet. 2020. Exploring the potential of GPT-2 for generating fake reviews of research papers. In *Proceedings of FSDM 2020*, Antonio J. Tallón-Ballesteros (Ed.). Springer, 390–396.

[15] Jonas Belouadi, Eddy Ilg, Margret Keuper, Hideki Tanaka, Masao Utiyama, Raj Dabre, Steffen Eger, and Simone Ponzetto. 2025. TikZero: Zero-shot text-guided graphics program synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 17793–17806.

[16] Jonas Belouadi, Anne Lauscher, and Steffen Eger. 2024. AutomaTikZ: Text-guided synthesis of scientific vector graphics with Tikz. In *Proceedings of the 12th International Conference on Learning Representations*.

[17] Jonas Belouadi, Simone Paolo Ponzetto, and Steffen Eger. 2024. DeTikZify: Synthesizing graphics programs for scientific figures and sketches with TikZ. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems*.

[18] Lamia Ben Ezzdine, Wissem Dhahbi, Ismail Dergaa, Halil İbrahim Ceylan, Noomen Guelmami, Helmi Ben Saad, Karim Chamari, Valentina Stefanica, and Abdelfatteh El Omri. 2025. Physical activity and neuroplasticity in neurodegenerative disorders: A comprehensive review of exercise interventions, cognitive training, and AI applications. *Frontiers in Neuroscience* 19 (2025), 1502417.

[19] Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. Content-based citation recommendation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, 238–251.

[20] Prabhat Bharti, Meith Navlakha, Mayank Agarwal, and Asif Ekbal. 2023. PolitePEER: Does peer review hurt? A dataset to gauge politeness intensity in the peer reviews. *Language Resources and Evaluation* 58 (2023), 1291–1313.

[21] Prabhat Kumar Bharti, Shashi Ranjan, Tirthankar Ghosal, Mayank Agrawal, and Asif Ekbal. 2021. PEERAssist: Leveraging on paper-review interactions to predict peer review decisions. In *Towards Open and Trustworthy Digital Societies: 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021*. 421–435.

[22] Jennifer A. Byrne. 2016. Improving the peer review of narrative literature reviews. *Research Integrity and Peer Review* 1, 12 (2016).

[23] Fengyu Cai, Xinran Zhao, Tong Chen, Sihao Chen, Hongming Zhang, Iryna Gurevych, and Heinz Koeppl. 2024. MixGR: Enhancing retriever generalization for scientific domain through complementary granularity. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 10369–10391.

[24] Andres Castellanos-Gomez. 2023. Good practices for scientific article writing with ChatGPT and other artificial intelligence language models. *Nanomanufacturing* 3, 2 (2023), 135–138.

[25] Miaosen Chai, Emily Herron, Erick Cervantes, and Tirthankar Ghosal. 2024. Exploring scientific hypothesis generation with Mamba. In *Proceedings of the 1st Workshop on NLP for Science*. 197–207.

[26] Souvic Chakraborty, Pawan Goyal, and Animesh Mukherjee. 2020. Aspect-based sentiment analysis of scientific reviews. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. 207–216.

[27] A. Checco, L. Bracciale, P. Loreti, S. Pinfield, and G. Bianchi. 2021. AI-assisted peer review. *Humanities and Social Sciences Communications* 8, 1 (January 2021), 1–11.

[28] Chen Chen, Ayman Maqsood, Zhuang Zhang, Xiaobing Wang, Linrui Duan, Huanhuan Wang, Tianyang Chen, Siyu Liu, Qiutong Li, Jingshan Luo, et al. 2024. The use of ChatGPT to generate experimentally testable hypotheses for improving the surface passivation of perovskite solar cells. *Cell Reports Physical Science* 5, 7 (2024).

[29] Kexin Chen, Jiamin Lu, Junyou Li, Xiaoran Yang, Yuyang Du, Kunyi Wang, Qiannuan Shi, Jiahui Yu, Lanqing Li, Jiezhong Qiu, et al. 2023. Chemist-X: Large language model-empowered agent for reaction condition recommendation in chemical synthesis. arXiv:2311.10776. Retrieved from https://arxiv.org/abs/2311.10776

[30] Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2023. FELM: Benchmarking factuality evaluation of large language models. In *Advances in Neural Information Processing Systems*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). Vol. 36, Curran Associates, Inc.

[31] Yanran Chen and Steffen Eger. 2023. Transformers go for the LOLs: Generating (Humourous) titles from scientific abstracts end-to-end. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, Daniel Deutsch, Rotem Dror, Steffen Eger, Yang Gao, Christoph Leiter, Juri Opitz, and Andreas Rücklé (Eds.). Association for Computational Linguistics, 62–84. DOI: https://doi.org/10.18653/v1/2023.eval4nlp-1.6

[32] Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. 2024. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. arXiv:2410.05080. Retrieved from https://arxiv.org/abs/2410.05080

[33] Huzi Cheng, Bin Sheng, Aaron Lee, Varun Chaudhary, Atanas G. Atanasov, Nan Liu, Yue Qiu, Tien Yin Wong, Yih-Chung Tham, and Ying-Feng Zheng. 2024. Have AI-Generated Texts from LLM Infiltrated the Realm of Scientific Writing? A Large-Scale Analysis of Preprint Platforms. bioRxiv:2024.03.25.586710

[34] Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. APE: Argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 7000–7011.

[35] Yizhou Chi, Yizhang Lin, Sirui Hong, Duyi Pan, Yaying Fei, Guanghao Mei, Bangbang Liu, Tianqi Pang, Jacky Kwok, Ceyao Zhang, et al. 2024. SELA: Tree-Search enhanced LLM agents for automated machine learning. arXiv:2410.17238. Retrieved from https://arxiv.org/abs/2410.17238

[36] Gautam Choudhary, Natwar Modani, and Nitish Maurya. 2021. ReAct: A review comment dataset for actionability (and more). In *Proceedings of the International Conference on Web Information Systems Engineering – WISE 2021*. 336–343.

[37] Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Daniel S. Weld. 2019. Pretrained language models for sequential sentence classification. arXiv:1909.04054. Retrieved from https://arxiv.org/abs/1909.04054

[38] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, 2270–2282.

[39] Alessandro Conti, Enrico Fini, Paolo Rota, Yiming Wang, Massimiliano Mancini, and Elisa Ricci. 2024. Automatic benchmarking of large multimodal models via iterative experiment programming. arXiv:2406.12321. Retrieved from https://arxiv.org/abs/2406.12321

[40] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for YouTube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*. 191–198.

[41] Zhiqing Cui, Jiahao Yuan, Hanqing Wang, Yanshu Li, Chenxu Du, and Zhenglong Ding. 2025. Draw with thought: Unleashing multimodal reasoning for scientific diagram generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 5050–5059.

[42] Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. MARG: Multi-Agent review generation for scientific papers. arXiv:2401.04259. Retrieved from https://arxiv.org/abs/2401.04259

[43] Mike D'Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. 2023. ARIES: A corpus of scientific paper edits made in response to peer reviews. arXiv:2306.12587. Retrieved from https://arxiv.org/abs/2306.12587

[44] Zekun Deng, Zixin Zeng, Weiye Gu, Jiawen Ji, and Bolin Hua. 2021. Automatic related work section generation by sentence extraction and reordering. In *Proceedings of the 1st Workshop on AI + Informetrics*.

[45] Ismail Dergaa, Feten Fekih-Romdhane, Jordan M. Glenn, Mohamed Saifeddin Fessi, Karim Chamari, Wissem Dhahbi, Makram Zghibi, Nicola Luigi Bragazzi, Mohamed Ben Aissa, Noomen Guelmami, et al. 2023. Moving beyond the stigma: Understanding and overcoming the resistance to the acceptance and adoption of artificial intelligence chatbots. *New Asian Journal of Medicine* 1, 2 (2023), 29–36.

[46] Danilo Dessí, Francesco Osborne, Diego Reforgiato Recupero, Davide Buscaldi, and Enrico Motta. 2022. CS-KG: A large-scale knowledge graph of research entities and claims in computer science. In *Proceedings of the International Semantic Web Conference*. 678–696.

[47] Iddo Drori and Dov Te'eni. 2024. Human-in-the-Loop AI reviewing: Feasibility, opportunities, and risks. *Journal of the Association for Information Systems* 25, 1 (2024), 98–109.

[48] Jennifer D'Souza, Sören Auer, and Ted Pedersen. 2021. SemEval-2021 Task 11: NLPContributionGraph - Structuring scholarly NLP contributions for a research knowledge graph. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, Alexis Palmer, Nathan Schneider, Natalie Schluter, Guy Emerson, Aurelie Herbelot, and Xiaodan Zhu (Eds.). Association for Computational Linguistics, 364–376.

[49] 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*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Vol. 1, Association for Computational Linguistics, 5049–5073.

[50] Daniel Nygård Ege, Mark Goudswaard, James Gopsill, Ben Hicks, and Martin Steinert. 2024. IDEA Challenge 2022 Dataset: Prototypes from a Design Hackathon. *Data in Brief* 54 (2024), 110363. DOI: https://doi.org/10.1016/j.dib.2024.110363

[51] Aryaz Eghbali and Michael Pradel. 2023. CrystalBLEU: Precisely and efficiently measuring the similarity of code. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*. Article 28.
[52] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. 2019. Neural architecture search: A survey. *Journal of Machine Learning Research* 20, 55 (2019), 1–21.
[53] Faiza Farhat, Shahab Saquib Sohail, and Dag Øivind Madsen. 2023. How trustworthy is ChatGPT? The case of bibliometric analyses. *Cogent Engineering* 10, 1 (2023), 2222988.
[54] Michael Fromm, Evgeniy Faerman, Max Berrendorf, Siddharth Bhargava, Ruoxia Qi, Yao Zhang, Lukas Dennert, Sophia Selle, Yang Mao, and Thomas Seidl. 2021. Argument mining driven analysis of peer-reviews. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 6 (2021), 4758–4766.
[55] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. 2023. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*. Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). Vol. 36, Curran Associates, Inc.
[56] Tsu-Jui Fu, William Yang Wang, Daniel J. McDuff, and Yale Song. 2021. DOC2PPT: Automatic presentation slides generation from scientific documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
[57] 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*. 7337–7353.
[58] Catherine A. Gao, Frederick M. Howard, Nikolay S. Markov, Emma C. Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T. Pearson. 2023. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ Digital Medicine* 6, 1 (2023), 75.
[59] Mingqi Gao, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. LLM-based NLG evaluation: Current status and challenges. *Computational Linguistics* 51, 2 (2025), 661–687.
[60] Andres Garcia-Silva, Cristian Berrio, and Jose Manuel Gomez-Perez. 2024. SPACE-IDEAS: A dataset for salient information detection in space innovation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. 15087–15092.
[61] Jiaxin Ge, Zora Zhiruo Wang, Xuhui Zhou, Yi-Hao Peng, Sanjay Subramanian, Qinyue Tan, Maarten Sap, Alane Suhr, Daniel Fried, Graham Neubig, et al. 2025. Autopresent: Designing structured visuals from scratch. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2902–2911.
[62] Alireza Ghafarollahi and Markus J. Buehler. 2025. SciAgents: Automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials* 37, 22 (2025), 2413523.
[63] Tirthankar Ghosal, Sandeep Kumar, Prabhat Kumar Bharti, and Asif Ekbal. 2022. Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. *PLOS One* 17, 1 (2022), e0259238.
[64] Tirthankar Ghosal, Kamal Kaushik Varanasi, and Valia Kordoni. 2022. HedgePeer: A dataset for uncertainty detection in peer reviews. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*. 1–5.
[65] Tirthankar Ghosal, Rajeev Verma, Asif Ekbal, and Pushpak Bhattacharyya. 2019. DeepSentiPeer: Harnessing sentiment in review texts to recommend peer review decisions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1120–1130.
[66] Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer. 2024. LLMs4Synthesis: Leveraging large language models for scientific synthesis. In *Proceedings of the 2024 ACM/IEEE Joint Conference on Digital Libraries*. 1–12.
[67] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. 2025. Towards an AI co-scientist. arXiv:2502.18864. Retrieved from https://arxiv.org/abs/2502.18864
[68] Andrew V. Gougherty and Hannah L. Clipp. 2024. Testing the reliability of an AI-based large language model to extract ecological information from the scientific literature. *npj Biodiversity* 3, 1 (2024), 13.
[69] Andrew Gray. 2024. ChatGPT "contamination": Estimating the prevalence of LLMs in the scholarly literature. arXiv:2403.16887. Retrieved from https://arxiv.org/abs/2403.16887
[70] Christian Greisinger and Steffen Eger. 2026. TikZilla: Scaling text-to-tikz with high-quality data and reinforcement learning. In *Proceedings of the 14th International Conference on Learning Representations*. In press.
[71] Yanzhu Guo, Guokan Shang, Virgile Rennard, Michalis Vazirgiannis, and Chloé Clavel. 2023. Automatic analysis of substantiation in scientific peer reviews. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 10198–10216.
[72] Thilo Hagendorff. 2024. Mapping the ethics of generative AI: A comprehensive scoping review. arXiv:2402.08323. Retrieved from https://arxiv.org/abs/2402.08323
[73] Janna Hastings. 2023. *AI for Scientific Discovery*. CRC Press.
[74] Xin He, Kaiyong Zhao, and Xiaowen Chu. 2021. AutoML: A survey of the state-of-the-art. *Knowledge-based Systems* 212 (2021), 106622.

[75] Yichen He, Guanhua Huang, Peiyuan Feng, Yuan Lin, Yuchen Zhang, Hang Li, and Weinan E. 2025. PaSa: An LLM agent for comprehensive academic paper search. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, 11663–11679. DOI: https://doi.org/10.18653/v1/2025.acl-long.572

[76] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 7514–7528. DOI: https://doi.org/10.18653/v1/2021.emnlp-main.595

[77] Tony Hey, Stewart Tansley, and Kristin Tolle. 2009. Jim Gray on escience: A transformed scientific method. In *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Tony Hey, Stewart Tansley, and Kristin Tolle (Eds.). Microsoft Research.

[78] Cong Duy Vu Hoang and Min-Yen Kan. 2010. Towards automated related work summarization. In *COLING 2010: Posters*, Chu-Ren Huang and Dan Jurafsky (Eds.). COLING 2010 Organizing Committee, 427–435.

[79] Sally Hopewell, Mike Clarke, David Moher, Elizabeth Wager, Philippa Middleton, Douglas G. Altman, Kenneth F. Schulz, and Consort Group. 2008. CONSORT for reporting randomized controlled trials in journal and conference abstracts: explanation and elaboration. *PLoS Medicine* 5, 1 (2008), e20.

[80] Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. 2024. Nova: An iterative planning and search approach to enhance novelty and diversity of LLM generated ideas. arXiv:2410.14255. Retrieved from https://arxiv.org/abs/2410.14255

[81] Yue Hu and Xiaojun Wan. 2013. PPSGen: Learning to generate presentation slides for academic papers. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*. 2099–2105.

[82] Yue Hu and Xiaojun Wan. 2014. Automatic generation of related work sections in scientific papers: An optimization approach. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. 1624–1633.

[83] 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*. 2131–2137.

[84] Jie Huang and Kevin Chen-Chuan Chang. 2023. Citation: A key to building responsible and accountable large language models. arXiv:2307.02185. Retrieved from https://arxiv.org/abs/2307.02185

[85] Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2024. MLAgentBench: Evaluating language agents on machine learning experimentation. In *Proceedings of the 41st International Conference on Machine Learning*.

[86] Shengzhi Huang, Qicong Wang, Wei Lu, Lingyu Liu, Zhenzhen Xu, and Yong Huang. 2025. PaperEval: A universal, quantitative, and explainable paper evaluation method powered by a multi-agent system. *Information Processing & Management* 62, 6 (2025), 104225.

[87] Taesoon Hwang, Nishant Aggarwal, Pir Zarak Khan, Thomas Roberts, Amir Mahmood, Madlen M. Griffiths, Nick Parsons, and Saboor Khan. 2024. Can ChatGPT assist authors with abstract writing in medical journals? Evaluating the quality of scientific abstracts generated by ChatGPT and original abstracts. *PLOS One* 19, 2 (2024), e0297701.

[88] Peter Jansen, Oyvind Tafjord, Marissa Radensky, Pao Siangliulue, Tom Hope, Bhavana Dalvi, Bodhisattwa Prasad Majumder, Daniel S. Weld, and Peter Clark. 2025. Codescientist: End-to-end semi-automated scientific discovery with code-based experimentation. In *Findings of the Association for Computational Linguistics: ACL 2025*. 13370–13467.

[89] Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. 2024. SWE-bench: Can language models resolve real-world github issues?. In *Proceedings of the 12th International Conference on Learning Representations*.

[90] Xia Jing, James J. Cimino, Vimla L. Patel, Yuchun Zhou, Jay H. Shubrook, Chang Liu, and Sonsoles De Lacalle. 2024. Data-driven hypothesis generation in clinical research: What we learned from a human subject study? *Medical Research Archives* 12, 2 (2024).

[91] Léane Jourdan, Florian Boudin, Richard Dufour, Nicolas Hernandez, and Akiko Aizawa. 2025. ParaRev: Building a dataset for scientific paragraph revision annotated with revision instruction. In *Proceedings of the 1st Workshop on Writing Aids at the Crossroads of AI, Cognitive Science and NLP*. 35–44.

[92] 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*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA Language Resources Association and the International Committee on Computational Linguistics, 2883–2892.

[93] Samira Ebrahimi Kahou, Adam Atkinson, Vincent Michalski, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2018. FigureQA: An annotated figure dataset for visual reasoning. arXiv:1710.07300. Retrieved from https://arxiv.org/abs/1710.07300

[94] Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard 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*. 1647–1661.

[95] SeongKu Kang, Yunyi Zhang, Pengcheng Jiang, Dongha Lee, Jiawei Han, and Hwanjo Yu. 2024. Taxonomy-guided semantic indexing for academic paper search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 7169–7184.

[96] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision: ECCV 2016*. 235–251.

[97] Neha Nayak Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. 2022. DISAPERE: A dataset for discourse structure in peer review discussions. In *Proceedings of NAACL*. 1234–1249.

[98] M. M. Kessler. 1963. Bibliographic coupling between scientific papers. *American Documentation* 14, 1 (1963), 10–25. DOI: https://doi.org/10.1002/asi.5090140103

[99] Seong-Gon Kim. 2023. Using ChatGPT for language editing in scientific articles. *Maxillofacial Plastic and Reconstructive Surgery* 45, 1 (2023), 13.

[100] William R. King and Jun He. 2005. Understanding the role and methods of meta-analysis in IS research. *Communications of the Association for Information Systems* 16 (2005), 665–686.

[101] Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. 2024. Delving into ChatGPT usage in academic writing through excess vocabulary. arXiv:2406.07016. Retrieved from https://arxiv.org/abs/2406.07016

[102] Jing Yu Koh, Stephen McAleer, Daniel Fried, and Ruslan Salakhutdinov. 2024. Tree search for language model agents. arXiv:2407.01476. Retrieved from https://arxiv.org/abs/2407.01476

[103] Kayvan Kousha and Mike Thelwall. 2024. Artificial intelligence to support publishing and peer review: A summary and review. *Learned Publishing* 37, 1 (2024), 4–12.

[104] Asheesh Kumar, Tirthankar Ghosal, and Asif Ekbal. 2021. A deep neural architecture for decision-aware meta-review generation. In *Proceedings of the 2021 ACM/IEEE Joint Conference on Digital Libraries*. 222–225.

[105] Shrinidhi Kumbhar, Venkatesh Mishra, Kevin Coutinho, Divij Handa, Ashif Iquebal, and Chitta Baral. 2025. Hypothesis generation for materials discovery and design using goal-driven and constraint-guided LLM agents. arXiv:2501.13299. Retrieved from https://arxiv.org/abs/2501.13299

[106] Nino Künzli, Anke Berger, Katarzyna Czabanowska, Raquel Lucas, Andrea Madarasova Geckova, Sarah Mantwill, and Olaf von Dem Knesebeck. 2022. «I do not have time»—Is this the end of peer review in public health sciences? *Public Health Reviews* 43, Article 1605407 (2022).

[107] Ilia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K. Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, et al. 2024. What can natural language processing do for peer review? arXiv:2405.06563. Retrieved from https://arxiv.org/abs/2405.06563

[108] Ilia Kuznetsov, Jan Buchmann, Max Eichler, and Iryna Gurevych. 2022. Revise and resubmit: An intertextual model of text-based collaboration in peer review. *Computational Linguistics* 48, 4 (12 2022), 949–986.

[109] 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. arXiv:2407.10362. Retrieved from https://arxiv.org/abs/2407.10362

[110] Po-Shen Lee, Jevin D. West, and Bill Howe. 2016. Viziometrics: Analyzing visual information in the scientific literature. *IEEE Transactions on Big Data* 4, 1 (2016), 117–129.

[111] Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. 2025. CheckEval: A reliable LLM-as-a-judge framework for evaluating text generation using checklists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 15782–15809.

[112] Christoph Leiter, Jonas Belouadi, Yanran Chen, Ran Zhang, Daniil Larionov, Aida Kostikova, and Steffen Eger. 2024. NLLG quarterly arXiv report 09/24: What are the most influential current AI papers? arXiv:2412.12121. Retrieved from https://arxiv.org/abs/2412.12121

[113] Adrian Letchford, Helen Susannah Moat, and Tobias Preis. 2015. The advantage of short paper titles. *Royal Society Open Science* 2, 8 (2015).

[114] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of NeurIPS* 33 (2020), 9459–9474.

[115] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. 2025. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 2757–2791.

[116] Da-Wei Li, Danqing Huang, Tingting Ma, and Chin-Yew Lin. 2021. Towards topic-aware slide generation for academic papers with unsupervised mutual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[117] Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 6449–6464. DOI: https://doi.org/10.18653/v1/2023.emnlp-main.397

[118] Jiyi Li, Ayaka Sato, Kazuya Shimura, and Fumiyo Fukumoto. 2020. Multi-task peer-review score prediction. In *Proceedings of the 1st Workshop on Scholarly Document Processing*. 121–126.

[119] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* 1 (2024), 14369–14387.

[120] Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, et al. 2024. Chain of ideas: Revolutionizing research via novel idea development with LLM agents. arXiv:2410.13185. Retrieved from https://arxiv.org/abs/2410.13185

[121] Miao Li, Eduard Hovy, and Jey Lau. 2023. Summarizing multiple documents with conversational structure for meta-review generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 7089–7112.

[122] Xiangci Li, Biswadip Mandal, and Jessica Ouyang. 2022. CORWA: A citation-oriented related work annotation dataset. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, 5426–5440. DOI: https://doi.org/10.18653/v1/2022.naacl-main.397

[123] 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*. 13846–13864.

[124] Xiangci Li and Jessica Ouyang. 2025. Explaining relationships among research papers. In *Proceedings of the 31st International Conference on Computational Linguistics*. 1080–1105.

[125] Yutong Li, Lu Chen, Aiwei Liu, Kai Yu, and Lijie Wen. 2025. ChatCite: LLM agent with human workflow guidance for comparative literature summary. In *Proceedings of the 31st International Conference on Computational Linguistics*. 3613–3630.

[126] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. arXiv:2308.03281. Retrieved from https://arxiv.org/abs/2308.03281

[127] Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, et al. 2023. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. arXiv:2310.01783. Retrieved from https://arxiv.org/abs/2310.01783

[128] Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, et al. 2025. Quantifying large language model usage in scientific papers. *Nature Human Behaviour* 9 (2025), 2599–2609.

[129] Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, et al. 2024. Mapping the increasing use of LLMs in scientific papers. In *Proceedings of COLM*.

[130] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. 74–81.

[131] Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. 2023. Automated scholarly paper review: Concepts, technologies, and challenges. *Information Fusion* 98 (2023), 101830.

[132] Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. 2023. MOPRD: A multidisciplinary open peer review dataset. *Neural Computing and Applications* 35, 34 (Sept. 2023), 24191–24206.

[133] Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. TruthfulQA: Measuring how models mimic human falsehoods. arXiv:2109.07958. Retrieved from https://arxiv.org/abs/2109.07958

[134] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Jian Su, Kevin Duh, and Xavier Carreras (Eds.). Association for Computational Linguistics, 2122–2132. DOI: https://doi.org/10.18653/v1/D16-1230

[135] Haokun Liu, Sicong Huang, Jingyu Hu, Yangqiaoyu Zhou, and Chenhao Tan. 2025. HypoBench: Towards systematic and principled benchmarking for hypothesis generation. arXiv:2504.11524. Retrieved from https://arxiv.org/abs/2504.11524

[136] Quanliang Liu, Maciej P. Polak, So Yeon Kim, M. D. Shuvo, Hrishikesh Shridhar Deodhar, Jeongsoo Han, Dane Morgan, and Hyunseok Oh. 2024. Beyond designer's knowledge: Generating materials design hypotheses via large language models. arXiv:2409.06756. Retrieved from https://arxiv.org/abs/2409.06756

[137] Ryan Liu and Nihar B. Shah. 2023. ReviewerGPT? An exploratory study on using large language models for paper reviewing. arXiv:2306.00622. Retrieved from https://arxiv.org/abs/2306.00622

[138] Sizhe Liu, Yizhou Lu, Siyu Chen, Xiyang Hu, Jieyu Zhao, Tianfan Fu, and Yue Zhao. 2024. DrugAgent: Automating AI-aided drug discovery programming through LLM multi-agent collaboration. arXiv:2411.15692. Retrieved from https://arxiv.org/abs/2411.15692

[139] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 2511–2522. DOI: https://doi.org/10.18653/v1/2023.emnlp-main.153

[140] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S. Weld. 2019. S2ORC: The semantic scholar open research corpus. arXiv:1911.02782. Retrieved from https://arxiv.org/abs/1911.02782

[141] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI scientist: Towards fully automated open-ended scientific discovery. arXiv:2408.06292. Retrieved from https://arxiv.org/abs/2408.06292

[142] Yuyu Luo, Jiawei Tang, and Guoliang Li. 2021. nvBench: A large-scale synthesized dataset for cross-domain natural language to visualization task. arXiv:2112.12926. Retrieved from https://arxiv.org/abs/2112.12926

[143] Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025. LLM4SR: A survey on large language models for scientific research. arXiv:2501.04306. Retrieved from https://arxiv.org/abs/2501.04306

[144] Kangyong Ma. 2025. AI agents in chemical research: GVIM – An intelligent research assistant system. *Digital Discovery* 4, 2 (2025), 355–375.

[145] Himanshu Maheshwari, Sambaran Bandyopadhyay, Aparna Garimella, and Anandhavelu Natarajan. 2024. Presentations are not always linear! GNN meets LLM for text document-to-presentation transformation with attribution. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 15948–15962.

[146] 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. arXiv:2402.12255. Retrieved from https://arxiv.org/abs/2402.12255

[147] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 2263–2279. DOI: https://doi.org/10.18653/v1/2022.findings-acl.177

[148] Prakhar Mishra, Chaitali Diwan, Srinath Srinivasa, and Gopalakrishnan Srinivasaraghavan. 2021. Automatic title generation for text with pre-trained transformer language model. In *Proceedings of the 2021 IEEE 15th International Conference on Semantic Computing*. 17–24.

[149] Ishani Mondal, Zongxia Li, Yufang Hou, Anandhavelu Natarajan, Aparna Garimella, and Jordan Lee Boyd-Graber. 2024. SciDoc2Diagrammer-MAF: Towards generation of scientific diagrams from documents guided by multi-aspect feedback refinement. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 13342–13375.

[150] Ishani Mondal, Shwetha S, Anandhavelu Natarajan, Aparna Garimella, Sambaran Bandyopadhyay, and Jordan Boyd-Graber. 2024. Presentations by the humans and for the humans: Harnessing LLMs for generating persona-aware slides from documents. In *Proceedings of EACL*. 2664–2684.

[151] Rajiv Movva, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson. 2025. Sparse autoencoders for hypothesis generation. arXiv:2502.04382. Retrieved from https://arxiv.org/abs/2502.04382

[152] Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 2014–2037.

[153] Deepak Nathani, Lovish Madaan, Nicholas Roberts, Nikolay Bashlykov, Ajay Menon, Vincent Moens, Amar Budhiraja, Despoina Magka, Vladislav Vorotilov, Gaurav Chaurasia, et al. 2025. MLGym: A new framework and benchmark for advancing AI research agents. arXiv:2502.14499. Retrieved from https://arxiv.org/abs/2502.14499

[154] Yuansheng Ni, Ping Nie, Kai Zou, Xiang Yue, and Wenhu Chen. 2025. VisCoder: Fine-Tuning LLMs for executable Python visualization code generation. arXiv:2506.03930. Retrieved from https://arxiv.org/abs/2506.03930

[155] Allard Oelen, Mohamad Yaser Jaradeh, and Sören Auer. 2025. Introducing ORKG ASK: An AI-driven scholarly literature search and exploration system taking a Neuro-Symbolic approach. In *Proceedings of the International Conference on Web Engineering*. 11–25.

[156] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Pierre Isabelle, Eugene Charniak, and Dekang Lin (Eds.). Association for Computational Linguistics, 311–318. DOI: https://doi.org/10.3115/1073083.1073135

[157] Yang Jeong Park, Daniel Kaplan, Zhichu Ren, Chia-Wei Hsu, Changhao Li, Haowei Xu, Sipei Li, and Ju Li. 2024. Can ChatGPT be used to generate scientific hypotheses? *Journal of Materiomics* 10, 3 (2024), 578–584.

[158] Guy Paré, Marie-Claude Trudel, Mirou Jaana, and Spyros Kitsiou. 2015. Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management* 52 (2015), 183–199.

[159] Barbara Plank and Reinard van Dalen. 2019. CiteTracked: A longitudinal dataset of peer reviews and citations. In *Proceedings of the 4th Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries*. 116–122.

[160] Aniket Pramanick, Yufang Hou, Saif M. Mohammad, and Iryna Gurevych. 2024. The nature of NLP: Analyzing contributions in NLP papers. arXiv:2409.19505. Retrieved from https://arxiv.org/abs/2409.19505

[161] Aniket Pramanick, Yufang Hou, Saif M. Mohammad, and Iryna Gurevych. 2024. Transforming scholarly landscapes: Influence of large language models on academic fields beyond computer science. arXiv:2409.19508. Retrieved from https://arxiv.org/abs/2409.19508

[162] Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. SPIQA: A dataset for multimodal question answering on scientific papers. In *Advances in Neural Information Processing Systems*, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.). Vol. 38, Curran Associates, Inc.

[163] Kevin Pu, K. J. 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. arXiv:2410.04025. Retrieved from https://arxiv.org/abs/2410.04025

[164] Zhongji Pu, Chun-Lin Shi, Che Ok Jeon, Jingyuan Fu, Shuang-Jiang Liu, Canhui Lan, Yanlai Yao, Yong-Xin Liu, and Baolei Jia. 2024. ChatGPT and generative AI are revolutionizing the scientific community: A Janus-faced conundrum. *iMeta* 3, 2 (2024), e178.

[165] Sukannya Purkayastha, Nils Dycke, Anne Lauscher, and Iryna Gurevych. 2025. Decision-Making with deliberation: Meta-reviewing as a document-grounded dialogue. arXiv:2508.05283. Retrieved from https://arxiv.org/abs/2508.05283

[166] Sukannya Purkayastha, Anne Lauscher, and Iryna Gurevych. 2023. Exploring Jiu-Jitsu argumentation for writing peer review rebuttals. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 14479–14495.

[167] Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. 2023. Large language models are zero shot hypothesis proposers. arXiv:2311.05965. Retrieved from https://arxiv.org/abs/2311.05965

[168] Biqing Qi, Kaiyan Zhang, Kai Tian, Haoxiang Li, Zhang-Ren Chen, Sihang Zeng, Ermo Hua, Hu Jinfang, and Bowen Zhou. 2024. Large language models as biomedical hypothesis generators: A comprehensive evaluation. arXiv:2407.08940. Retrieved from https://arxiv.org/abs/2407.08940

[169] Yuting Qiang, Yanwei Fu, Yanwen Guo, Zhi-Hua Zhou, and Leonid Sigal. 2016. Learning to generate posters of scientific papers. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[170] Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S. Weld. 2024. Scideator: Human-LLM scientific idea generation grounded in research-paper facet recombination. arXiv:2409.14634. Retrieved from https://arxiv.org/abs/2409.14634

[171] Raian Rahman, Rizvi Hasan, Abdullah Al Farhad, Md Tahmid Rahman Laskar, Md. Hamjajul Ashmafee, and Abu Raihan Mostofa Kamal. 2023. ChartSumm: A comprehensive benchmark for automatic chart summarization of long and short summaries. arXiv:2304.13620. Retrieved from https://arxiv.org/abs/2304.13620

[172] Chandan K. Reddy and Parshin Shojaee. 2024. Towards scientific discovery with generative AI: Progress, opportunities, and challenges. arXiv:2412.11427. Retrieved from https://arxiv.org/abs/2412.11427

[173] Isabelle Régner, Catherine Thinus-Blanc, Agnès Netter, Toni Schmader, and Pascal Huguet. 2019. Committees with implicit biases promote fewer women when they do not believe gender bias exists. *Nature Human Behaviour* 3, 11 (2019), 1171–1179.

[174] Georg Rehm, Stefan Dietze, Sonja Schimmler, and Frank Krüger. 2024. *Natural Scientific Language Processing and Research Knowledge Graphs: 1st International Workshop, NSLP 2024*. Springer.

[175] Zachary Robertson. 2023. GPT-4 is slightly helpful for peer-review assistance: A pilot study. arXiv:2307.05492. Retrieved from https://arxiv.org/abs/2307.05492

[176] Furkan Şahinuç, Subhabrata Dutta, and Iryna Gurevych. 2025. Expert preference-based evaluation of automated related work generation. arXiv:2508.07955. Retrieved from https://arxiv.org/abs/2508.07955

[177] Furkan Şahinuç, Ilia Kuznetsov, Yufang Hou, and Iryna Gurevych. 2024. Systematic task exploration with LLMs: A study in citation text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* 1 (2024), 4832–4855.

[178] Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2026. HalluCitation matters: Revealing the impact of hallucinated references with 300 hallucinated papers in ACL conferences. arXiv:2601.18724. Retrieved from https://arxiv.org/abs/2601.18724

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

[180] Shubhra Kanti Karmaker Santu, Sanjeev Kumar Sinha, Naman Bansal, Alex Knipper, Souvika Sarkar, John Salvador, Yash Mahajan, Sri Guttikonda, Mousumi Akter, Matthew Freestone, et al. 2024. Prompting LLMs to compose

meta-review drafts from peer-review narratives of scholarly manuscripts. arXiv:2402.15589. Retrieved from https://arxiv.org/abs/2402.15589

[181] Oliver Schilke and Martin Reimann. 2025. The transparency dilemma: How AI disclosure erodes trust. *Organizational Behavior and Human Decision Processes* 188 (2025), 104405.

[182] Daniel Schlagwein and Leslie Willcocks. 2023. 'ChatGPT et al.': The ethics of using (generative) artificial intelligence in research and science. *Journal of Information Technology* 38, 3 (2023), 232–238.

[183] 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:2501.04227. Retrieved from https://arxiv.org/abs/2501.04227

[184] Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. AgentClinic: A multimodal agent benchmark to evaluate AI in simulated clinical environments. arXiv:2405.07960. Retrieved from https://arxiv.org/abs/2405.07960

[185] Dominik Schmidt, Zhengyao Jiang, and Yuxiang Wu. 2024. AIDE: Human-level performance in data science competitions. Retrieved November 24, 2024 from https://www.weco.ai/blog/technical-report

[186] Paul Sebo, Bing Nie, and Ting Wang. 2025. Can ChatGPT write better scientific titles? A comparative evaluation of human-written and AI-generated titles. *F1000Research* 14 (2025), 1470.

[187] 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*. 2521–2535.

[188] Pranav Shetty, Arunkumar Chitteth Rajan, Chris Kuenneth, Sonakshi Gupta, Lakshmi Prerana Panchumarti, Lauren Holm, Chao Zhang, and Rampi Ramprasad. 2023. A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing. *npj Computational Materials* 9, 1 (2023), 52.

[189] Chufan Shi, Cheng Yang, Yaxin Liu, Bo Shui, Junjie Wang, Mohan Jing, Linran Xu, Xinyu Zhu, Siheng Li, Yuxiang Zhang, et al. 2024. ChartMimic: Evaluating LMM's cross-modal reasoning capability via chart-to-code generation. arXiv:2406.09961. Retrieved from https://arxiv.org/abs/2406.09961

[190] Chenglei Si, Tatsunori Hashimoto, and Diyi Yang. 2025. The ideation-execution gap: Execution outcomes of LLM-generated versus human research ideas. arXiv:2506.20803. Retrieved from https://arxiv.org/abs/2506.20803

[191] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. arXiv:2409.04109. Retrieved from https://arxiv.org/abs/2409.04109

[192] Shruti Singh, Mayank Singh, and Pawan Goyal. 2021. COMPARE: A taxonomy and dataset of comparison discussions in peer reviews. arXiv:2108.04366. Retrieved from https://arxiv.org/abs/2108.04366

[193] Henry Small. 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science* 24, 4 (1973), 265–269. DOI: https://doi.org/10.1002/asi.4630240406

[194] Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28, 1 (1972), 11–21.

[195] M. Sravanthi, C. Ravindranath Chowdary, and P. Sreenivasa Kumar. 2009. SlidesGen: Automatic generation of presentation slides for a technical paper using summarization. In *Proceedings of the International Florida AI Research Society Conference (FLAIRS 2009)*.

[196] Lukas Stappen, Georgios Rizos, Madina Hasan, Thomas Hain, and Björn W. Schuller. 2020. Uncertainty-aware machine support for paper reviewing on the Interspeech 2019 submission corpus. In *Proceedings of Interspeech 2020*. 1808–1812.

[197] Andrew Starnes and Clayton Webster. 2024. Mamba for scalable and efficient personalized recommendations. arXiv:2409.17165. Retrieved from https://arxiv.org/abs/2409.17165

[198] Moritz Staudinger, Wojciech Kusa, Florina Piroi, and Allan Hanbury. 2024. An analysis of tasks and datasets in peer reviewing. In *Proceedings of the 4th Workshop on Scholarly Document Processing*. 257–268.

[199] Dana Strauss, Sophia Gran-Ruaz, Muna Osman, Monnica T. Williams, and Sonya C. Faber. 2023. Racism and censorship in the editorial and peer review process. *Frontiers in Psychology* 14, Article 1120938 (2023).

[200] 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. arXiv:2410.09403. Retrieved from https://arxiv.org/abs/2410.09403

[201] Edward Sun, Yufang Hou, Dakuo Wang, Yunfeng Zhang, and Nancy X. R. Wang. 2021. D2S: Document-to-slide generation via query-based text summarization. In *Proceedings of NAACL*. 1405–1418.

[202] Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. 2024. TrustLLM: Trustworthiness in large language models. arXiv:2401.05561. Retrieved from https://arxiv.org/abs/2401.05561

[203] Jaroslaw Szumega, Lamine Bougueroua, Blerina Gkotse, Pierre Jouvelot, and Federico Ravotti. 2023. The Open Review-Based (ORB) dataset: Towards automatic assessment of scientific papers and experiment proposals in high-energy physics. arXiv:2312.04576. Retrieved from https://arxiv.org/abs/2312.04576

[204] Neset Tan, Trung Nguyen, Josh Bensemann, Alex Peng, Qiming Bao, Yang Chen, Mark Gahegan, and Michael Witbrock. 2023. Multi2Claim: Generating scientific claims from multi-choice questions for scientific fact-checking. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 2652–2664.

[205] Liyan Tang, Yifan Peng, Yanshan Wang, Ying Ding, Greg Durrett, and Justin Rousseau. 2023. Less likely brainstorming: Using language models to generate alternative hypotheses. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 12532–12555. DOI: https://doi.org/10.18653/v1/2023.findings-acl.794

[206] Song Tong, Kai Mao, Zhen Huang, Yukun Zhao, and Kaiping Peng. 2024. Automating psychological hypothesis generation with AI: Large language models meet causal graph. arXiv:2402.14424. Retrieved from https://arxiv.org/abs/2402.14424

[207] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288. Retrieved from https://arxiv.org/abs/2307.09288

[208] Patara Trirat, Wonyong Jeong, and Sung Ju Hwang. 2024. AutoML-Agent: A multi-agent LLM framework for full-pipeline AutoML. arXiv:2410.02958. Retrieved from https://arxiv.org/abs/2410.02958

[209] Yun-Da Tsai, Yu-Che Tsai, Bo-Wei Huang, Chun-Pai Yang, and Shou-De Lin. 2023. AutoML-GPT: Large language model for AutoML. arXiv:2309.01125. Retrieved from https://arxiv.org/abs/2309.01125

[210] Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. 2023. FreshLLMs: Refreshing large language models with search engine augmentation. arXiv:2310.03214. Retrieved from https://arxiv.org/abs/2310.03214

[211] William H. Walters and Esther Isabelle Wilder. 2023. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports* 13, 1 (2023), 14045.

[212] Fuchun Wang, Xian Zhou, Wenpeng Hu, Zhunchen Luo, Wei Luo, and Xiaoying Bai. 2024. LLM assists hypothesis generation and testing for deliberative questions. In *Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 424–436.

[213] Hao Wang and Wu-Jun Li. 2014. Relational collaborative topic regression for recommender systems. *IEEE Transactions on Knowledge and Data Engineering* 27, 5 (2014), 1343–1355.

[214] Pancheng Wang, Shasha Li, Haifang Zhou, Jintao Tang, and Ting Wang. 2019. ToC-RWG: Explore the combination of topic model and citation information for automatic related work generation. *IEEE Access* 8 (2019), 13043–13055.

[215] Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024. SciMON: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. 279–299.

[216] Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. 2019. PaperRobot: Incremental draft generation of scientific ideas. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 1980–1991. DOI: https://doi.org/10.18653/v1/P19-1191

[217] Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. 2020. ReviewRobot: Explainable paper review generation based on knowledge synthesis. In *Proceedings of the 13th International Conference on Natural Language Generation*. 384–397.

[218] Sida Wang, Xiaojun Wan, and Shikang Du. 2017. Phrase-based presentation slides generation for academic papers. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[219] Xingyao Wang, Boxuan Li, Yufan Song, Frank F. Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, et al. 2024. OpenHands: An open platform for AI software developers as generalist agents. arXiv:2407.16741. Retrieved from https://arxiv.org/abs/2407.16741

[220] Yubo Wang, Xueguang Ma, Ping Nie, Huaye Zeng, Zhiheng Lyu, Yuxuan Zhang, Benjamin Schneider, Yi Lu, Xiang Yue, and Wenhu Chen. 2025. ScholarCopilot: Training large language models for academic writing with accurate citations. arXiv:2504.00824. Retrieved from https://arxiv.org/abs/2504.00824

[221] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. 2024. CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. In *Advances in Neural Information Processing Systems*, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.). Vol. 38, Curran Associates, Inc.

[222] Jingxuan Wei, Cheng Tan, Qi Chen, Gaowei Wu, Siyuan Li, Zhangyang Gao, Linzhuang Sun, Bihui Yu, and Ruifeng Guo. 2025. From words to structured visuals: A benchmark and framework for text-to-diagram generation and editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 13315–13325.

[223] Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2025. CycleResearcher: Improving automated research via automated review. In *Proceedings of the 13th International Conference on Learning Representations*.

[224] Leslie P. Willcocks, John Hindle, Matt Stanton, and John Smith. 2023. *Maximizing Value with Automation and Transformation: A Realist's Guide*. Palgrave Macmillan. DOI: https://doi.org/10.1007/978-3-031-46569-7

[225] Siwei Wu, Yizhi Li, Xingwei Qu, Rishi Ravikumar, Yucheng Li, Tyler Loakman, Shanghaoran Quan, Xiaoyong Wei, Riza Batista-Navarro, and Chenghua Lin. 2025. LongEval: A comprehensive analysis of long-text generation through a plan-based paradigm. arXiv:2502.19103. Retrieved from https://arxiv.org/abs/2502.19103

[226] Yuhao Wu, Yushi Bai, Zhiqiang Hu, Roy Ka-Wei Lee, and Juanzi Li. 2025. LongWriter-Zero: Mastering ultra-long text generation via reinforcement learning. arXiv:2506.18841. Retrieved from https://arxiv.org/abs/2506.18841

[227] Guangzhi Xiong, Eric Xie, Amir Hassan Shariatmadari, Sikun Guo, Stefan Bekiranov, and Aidong Zhang. 2024. Improving scientific hypothesis generation with knowledge grounded large language models. arXiv:2411.02382. Retrieved from https://arxiv.org/abs/2411.02382

[228] Peixin Xu, Yujuan Ding, and Wenqi Fan. 2024. ChartAdapter: Large vision-language model for chart summarization. arXiv:2412.20715. Retrieved from https://arxiv.org/abs/2412.20715

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

[230] Sheng Xu and Xiaojun Wan. 2022. PosterBot: A system for generating posters of scientific papers with neural models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[231] Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. 2025. The AI Scientist-v2: Workshop-level automated scientific discovery via agentic tree search. arXiv:2504.08066. Retrieved from https://arxiv.org/abs/2504.08066

[232] Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2023. Large language models for automated open-domain scientific hypotheses discovery. arXiv:2309.02726. Retrieved from https://arxiv.org/abs/2309.02726

[233] Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. 2024. MOOSE-Chem: Large language models for rediscovering unseen chemistry scientific hypotheses. arXiv:2410.07076. Retrieved from https://arxiv.org/abs/2410.07076

[234] 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 *Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, The 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*. 7386–7393. DOI: https://doi.org/10.1609/aaai.v33i01.33017386

[235] Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2021. Packed levitated marker for entity and relation extraction. arXiv:2109.06067. Retrieved from https://arxiv.org/abs/2109.06067

[236] Xinyang Yi, Ji Yang, Lichan Hong, Derek Zhiyuan Cheng, Lukasz Heldt, Aditee Kumthekar, Zhe Zhao, Li Wei, and Ed Chi. 2019. Sampling-bias-corrected neural modeling for large corpus item recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 269–277.

[237] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know?. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 8653–8665. DOI: https://doi.org/10.18653/v1/2023.findings-acl.551

[238] Yantao Yu, Weipeng Wang, Zhoutian Feng, and Daiyue Xue. 2021. A dual augmented two-tower model for online large-scale recommendation. In *Proceedings of DLP-KDD*.

[239] Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research* 75 (2022).

[240] Abhay Zala, Han Lin, Jaemin Cho, and Mohit Bansal. 2024. DiagrammerGPT: Generating open-domain, open-platform diagrams via LLM planning. In *Proceedings of the COLM*.

[241] Qi Zeng, Mankeerat Sidhu, Hou Pong Chan, Lu Wang, and Heng Ji. 2023. Meta-review generation with checklist-guided iterative introspection. arXiv:2305.14647. Retrieved from https://arxiv.org/abs/2305.14647

[242] Jiajie Zhang, Zhongni Hou, Xin Lv, Shulin Cao, Zhenyu Hou, Yilin Niu, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2024. LongReward: Improving long-context large language models with AI feedback. arXiv:2410.21252. Retrieved from https://arxiv.org/abs/2410.21252

[243] Leixin Zhang, Steffen Eger, Yinjie Cheng, Weihe Zhai, Jonas Belouadi, Christoph Leiter, Simone Paolo Ponzetto, Fahimeh Moafian, and Zhixue Zhao. 2024. SciImage: How good are multimodal large language models at scientific text-to-image generation? arXiv:2412.02368. Retrieved from https://arxiv.org/abs/2412.02368

[244] Lei Zhang, Yuge Zhang, Kan Ren, Dongsheng Li, and Yuqing Yang. 2023. Mlcopilot: Unleashing the power of large language models in solving machine learning tasks. arXiv:2304.14979. Retrieved from https://arxiv.org/abs/2304.14979

[245] Qiang Zhang, Wanyi Chen, Ming Qin, Yuhao Wang, Zhongji Pu, Keyan Ding, Yuyue Liu, Qunfeng Zhang, Dongfang Li, Xinjia Li, et al. 2025. Integrating protein language models and automatic biofoundry for enhanced protein evolution. *Nature Communications* 16, 1 (2025), 1553.

[246] Xuan Zhang, Limei Wang, Jacob Helwig, Youzhi Luo, Cong Fu, Yaochen Xie, Meng Liu, Yu-Ching Lin, Zhao Xu, Keqiang Yan, et al. 2023. Artificial intelligence for science in quantum, atomistic, and continuum systems. arXiv:2307.08423. Retrieved from https://arxiv.org/abs/2307.08423

[247] Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024. A comprehensive survey of scientific large language models and their applications in scientific discovery. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 8783–8817.

[248] Yunyi Zhang, Ruozhen Yang, Siqi Jiao, SeongKu Kang, and Jiawei Han. 2025. Scientific paper retrieval with LLM-guided semantic-based ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, 2049–2060. DOI: https://doi.org/10.18653/v1/2025.findings-emnlp.108

[249] Wei Zhao, Jennifer D'Souza, Steffen Eger, Anne Lauscher, Yufang Hou, Nafise Sadat Moosavi, Tristan Miller, and Chenghua Lin (Eds.). 2025. *Proceedings of the 1st Workshop on Human–LLM Collaboration for Ethical and Responsible Science Production (SciProdLLM)*.

[250] Zhiling Zheng, Oufan Zhang, Christian Borgs, Jennifer T. Chayes, and Omar M. Yaghi. 2023. ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis. *Journal of the American Chemical Society* 145, 32 (2023), 18048–18062.

[251] Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. Hypothesis generation with large language models. arXiv:2404.04326. Retrieved from https://arxiv.org/abs/2404.04326

[252] Dawei Zhu, Rui Meng, Yale Song, Xiyu Wei, Sujian Li, Tomas Pfister, and Jinsung Yoon. 2026. PaperBanana: Automating academic illustration for AI scientists. arXiv:2601.23265. Retrieved from https://arxiv.org/abs/2601.23265

[253] Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, et al. 2024. Agent-as-a-Judge: Evaluate agents with agents. arXiv:2410.10934. Retrieved from https://arxiv.org/abs/2410.10934

Received 4 March 2025; revised 11 April 2026; accepted 23 April 2026