Use of Scientific Paper Databases by AI Scientists in Agentic Workflows

Rachel So
rachel.so@4open.science

Abstract

The integration of artificial intelligence into scientific research has led to the emergence of AI scientists, autonomous systems capable of conducting research through agentic workflows. These workflows increasingly rely on scientific paper databases as critical infrastructure for literature retrieval, knowledge synthesis, and iterative research processes. We examine how AI scientists leverage paper databases such as arXiv, Semantic Scholar, and other scholarly repositories within agentic frameworks. We analyze the technical mechanisms enabling this integration, including semantic search, citation network analysis, and automated literature review capabilities. We identify key applications across scientific domains, from drug discovery to materials science, where AI agents use paper databases to inform hypothesis generation, experimental design, and manuscript preparation. We also discuss challenges related to data quality, algorithmic biases, and the need for human oversight. Our analysis reveals that effective integration of paper databases into agentic workflows represents a fundamental enabler of autonomous scientific discovery, while highlighting critical areas requiring further development.

1 Introduction

The exponential growth of scientific literature presents both unprecedented opportunities and significant challenges for researchers. With millions of papers published annually across diverse disciplines, the traditional manual approach to literature review and knowledge synthesis has become increasingly impractical [19]. Concurrently, advances in large language models (LLMs) and autonomous agent systems have catalyzed the emergence of AI scientists, systems capable of conducting end-to-end research with varying degrees of autonomy [24, 12].

Agentic workflows represent a paradigm shift from single-purpose AI models to systems where specialized agents collaborate to accomplish complex tasks [15, 6]. In the context of scientific research, these workflows decompose the research process into subtasks such as literature review, hypothesis generation, experimental design, data analysis, and manuscript preparation. Each subtask is handled by specialized agents that work interdependently toward a common research goal [12].

Scientific paper databases serve as essential infrastructure for these agentic systems. Repositories such as arXiv, Semantic Scholar, PubMed, and Web of Science provide structured access to millions of scholarly articles, along with metadata including citations, abstracts, and full-text content [16, 10]. The ability of AI agents to efficiently search, retrieve, and synthesize information from these databases fundamentally determines their capacity to conduct meaningful research [1].

Recent systems demonstrate the viability of autonomous research. Agent Laboratory automates the research pipeline through coordinated LLM agents that conduct literature reviews, design experiments, and generate reports [14]. Google's AI co-scientist successfully uncovered novel gene

Preprint.

transfer mechanisms by leveraging multi-agent workflows [6]. These achievements underscore the critical role of paper databases in enabling AI-driven discovery.

However, the integration of paper databases into agentic workflows raises important questions. How do AI agents effectively navigate vast literature repositories? What technical mechanisms enable semantic understanding and synthesis of scientific knowledge? What are the limitations and failure modes? How can we ensure reliability, reproducibility, and ethical use of these systems?

This paper addresses these questions through a comprehensive analysis of how AI scientists use paper databases within agentic workflows. We examine the technical foundations, survey current applications, identify challenges, and outline future research directions. Our contributions include: (1) a systematic characterization of paper database integration in agentic scientific workflows, (2) analysis of technical mechanisms enabling effective literature retrieval and synthesis, (3) identification of key applications and case studies across scientific domains, and (4) discussion of challenges and future opportunities in this rapidly evolving field.

2 Background and Related Work

2.1 AI Scientists and Agentic Workflows

The concept of AI scientists refers to autonomous or semi-autonomous systems capable of conducting scientific research with minimal human intervention [24]. These systems have evolved from simple automation tools to increasingly sophisticated agents with capabilities spanning hypothesis generation, experimental design, execution, and result interpretation [22].

Agentic workflows distinguish themselves from traditional AI applications through their architecture. Rather than relying on single general-purpose models, they orchestrate multiple specialized agents, each designed for specific tasks [15]. This modular approach enables targeted problem-solving while maintaining adaptability [6]. Recent work demonstrates that agentic systems can achieve fully dynamic workflows determined by real-time reasoning rather than pre-programmed sequences [12].

The taxonomy of LLM roles in science ranges from tools (providing basic assistance) to analysts (performing complex analysis) to scientists (conducting autonomous research) [24]. This progression reflects increasing autonomy and capability. Early systems like AI Scientist employed hybrid architectures combining agentic components with programmed logic [12]. Modern frameworks such as freephdlabor enable fully dynamic workflows with continual research programs that build systematically on prior work [12].

Key characteristics of effective agentic workflows include specialized agent roles, interdependence among agents, shared goals, dynamic adaptation to findings, and mechanisms for human oversight [15, 8]. These properties enable AI scientists to tackle complex, long-horizon research tasks that would be infeasible for isolated AI systems.

2.2 Scientific Paper Databases

Scientific paper databases provide structured access to scholarly literature at scale. Major repositories include arXiv (preprint server with over 2 million papers in physics, mathematics, computer science, and related fields), Semantic Scholar (AI-powered search engine covering 200+ million papers), PubMed (biomedical literature database), Web of Science (multidisciplinary citation index), and Google Scholar (broad-coverage academic search engine).

These databases differ in their coverage, metadata quality, access mechanisms, and computational interfaces. ArXiv provides full-text access to preprints through both web interfaces and bulk data access [10]. Semantic Scholar offers APIs for semantic search over abstracts and full-text snippets, along with citation information and paper embeddings [16]. The availability of programmatic APIs critically determines the utility of these databases for agentic systems.

Recent advances in database capabilities include semantic search using neural embeddings, citation network analysis and traversal, entity extraction and knowledge graph construction, snippet-level retrieval from full text, and metadata enrichment through NLP [16, 18]. These features enable AI agents to perform sophisticated literature analysis beyond simple keyword matching.

The construction and maintenance of paper databases involves significant technical challenges. Metadata extraction from PDFs, author disambiguation, citation parsing, and maintaining data quality all require substantial infrastructure [10]. Heterogeneous knowledge graphs integrating papers, authors, venues, institutions, and citation relationships provide rich representations for discovery [7, 18].

2.3 LLMs in Scientific Discovery

Large language models have demonstrated remarkable capabilities across scientific domains [2, 23]. Their applications span understanding scientific text, generating hypotheses, designing experiments, analyzing data, and writing manuscripts [22]. The ability to process and generate human-like text while handling vast amounts of data positions LLMs as valuable tools for scientific discovery [11].

However, LLMs face significant limitations when applied to science. They are constrained by static training data, lack direct access to current literature, cannot verify factual claims without external sources, and tend to hallucinate invalid references [3]. Retrieval-augmented generation (RAG) addresses these limitations by integrating external knowledge sources, particularly paper databases, into LLM workflows [3, 16].

The integration of LLMs with paper databases enables dynamic knowledge access. Rather than relying solely on knowledge embedded during training, LLM-empowered autonomous agents can actively retrieve real-time information from scholarly repositories [20]. This capability is essential for conducting research on current topics and ensuring factual accuracy through source attribution.

Scientific LLMs have been developed across modalities and disciplines, with specialized models for chemistry, biology, materials science, and medicine [23]. These domain-specific models often incorporate scientific paper corpora during pre-training, improving their understanding of technical concepts and terminology [4]. The synergy between specialized scientific knowledge and general reasoning capabilities positions LLMs as central components of AI scientist systems.

3 Paper Database Integration in Agentic Workflows

3.1 Literature Search and Retrieval

Effective literature search represents a fundamental capability for AI scientists. Unlike traditional keyword-based search, agentic systems employ multi-stage retrieval strategies that combine semantic understanding with strategic query formulation [1].

A typical retrieval workflow begins with query decomposition, where an LLM analyzes a research question and generates multiple search queries targeting different aspects of the problem [16]. These queries are optimized for specific database APIs, considering differences between keyword and semantic search endpoints. For instance, Semantic Scholar supports both traditional keyword search over abstracts and neural embedding-based search over full-text snippets [16].

Retrieved results undergo multi-stage filtering and ranking. Initial retrieval may return hundreds of candidate papers based on query matching. Subsequent stages use LLMs to assess relevance by analyzing titles and abstracts, re-ranking results based on content similarity to the research question, and extracting key information for downstream tasks [1]. Attribution mechanisms provide transparency into ranking decisions, enabling researchers to understand why specific papers were selected [16].

Advanced retrieval strategies leverage citation networks to expand beyond initial search results. Citation-based retrieval follows references from highly relevant papers, identifies papers that cite key works, and analyzes citation context to assess relationship strength [18]. This approach discovers papers that might be missed by content-based search alone.

Novel methods for concept extraction from scientific literature enable identification of emerging research areas. By analyzing citation patterns, researchers can identify papers that introduce or popularize new concepts with high precision [10]. AI agents can exploit these signals to focus on influential or foundational work within a domain.

3.2 Automated Literature Review

Literature review represents one of the most time-intensive aspects of research, making it a natural target for automation [19]. AI-based literature review (AILR) systems leverage paper databases to accelerate multiple stages of the review process.

The literature review pipeline typically involves problem formulation, literature search, screening for inclusion, quality assessment, data extraction, and synthesis [19, 13]. Agentic systems can automate or augment each stage with varying degrees of success.

For literature synthesis, LLMs retrieve relevant papers from databases, extract key findings and methodologies, identify common themes and contradictions, and generate structured summaries [1]. Recent systems like LitLLMs demonstrate promising results through two-step approaches that first create an outline and then generate the review content [1].

Citegeist exemplifies RAG-based literature review on the arXiv corpus [3]. The system employs embedding-based similarity matching, multi-stage filtering, and summarization to generate related work sections with citation backing. By continuously incorporating new papers, it maintains currency with the rapidly evolving literature.

Tools specialized for systematic reviews include RobotReviewer (quality assessment), Covidence (screening and data extraction), and ASReview (prioritization of papers for screening) [13]. These tools integrate with paper databases to streamline specific review stages, reducing review time from months to weeks while maintaining quality.

AI literature review suites provide end-to-end support through integrated programs for searching and downloading papers, extracting and organizing content, performing semantic queries across documents, and generating comprehensive summaries [17]. These systems capitalize on advances in NLP and machine learning to automate logistical aspects of reviews, allowing researchers to focus on synthesis and interpretation.

3.3 Knowledge Synthesis and Integration

Beyond retrieval and summarization, AI scientists must synthesize knowledge from multiple sources to inform research decisions. This requires understanding relationships between concepts, identifying gaps in existing literature, and integrating findings across papers.

Knowledge graphs provide structured representations of scholarly information [18, 21]. These graphs encode entities (papers, authors, concepts, methods) and relationships (citations, co-authorship, topical similarity) extracted from paper databases. AI agents can query these graphs to discover hidden patterns, such as unexplored connections between research communities or emerging interdisciplinary opportunities.

Scholarly knowledge graphs have been successfully applied to unveiling research communities, predicting collaboration networks, identifying influential papers and authors, and tracing the evolution of concepts over time [18]. For agentic systems, these graphs serve as navigable maps of the scientific landscape, enabling strategic exploration of the literature.

Topic modeling and trend analysis help AI scientists identify promising research directions. By analyzing temporal patterns in paper databases, agents can detect emerging topics, assess the maturity of research areas, and identify declining or converging fields [21]. This longitudinal analysis informs decisions about where to focus research efforts.

Integration across heterogeneous sources presents challenges. Different databases use varying metadata standards, cover different subsets of the literature, and provide different levels of access. Effective agentic systems must reconcile these differences, deduplicating papers across sources and synthesizing information from multiple databases to achieve comprehensive coverage.

3.4 Integration with Agent Architectures

The practical integration of paper databases into agentic workflows requires careful architectural design. Modern frameworks support various integration patterns, each with distinct trade-offs.

Tool-based integration treats paper databases as external tools that agents can invoke through API calls [9]. Agents use function calling or similar mechanisms to construct search queries, retrieve results, and process responses. This pattern provides flexibility and modularity but requires agents to manage API complexity.

Specialized retrieval agents handle all interactions with paper databases [12]. Other agents in the workflow request literature through well-defined interfaces, delegating search strategy and result filtering to the specialized agent. This encapsulation simplifies the overall system but may limit optimization opportunities.

RAG architectures integrate paper databases as knowledge bases that augment LLM generations [3]. Queries trigger automatic retrieval of relevant papers, which are included in the context when prompting the LLM. This pattern ensures that generated text is grounded in source material, reducing hallucinations and enabling attribution.

Agent Laboratory exemplifies a multi-agent architecture for autonomous research [14]. The PhD agent conducts literature reviews by iteratively querying the arXiv API, retrieving papers, summarizing content, and evaluating relevance. The results inform subsequent experimental and writing phases, demonstrating end-to-end integration of paper databases into the research pipeline.

Context management represents a critical challenge in long-horizon research. As agents process dozens or hundreds of papers, they must selectively retain relevant information while discarding noise [12]. Techniques include automatic context compaction, hierarchical summarization at multiple levels of detail, workspace-based communication to persist findings, and memory mechanisms for cross-session continuity.

4 Applications and Case Studies

4.1 Collaborative Autonomous Research

AgentRxiv demonstrates collaborative research through shared paper repositories [14]. Multiple agent laboratories can upload and retrieve research reports from a common preprint server, enabling collaboration and iterative improvement. Agents with access to prior research achieve higher performance improvements compared to isolated agents (11.4% relative improvement on mathematical reasoning tasks). Multiple laboratories working through AgentRxiv progress more rapidly toward common goals than isolated systems (13.7% relative improvement).

This collaborative model mirrors human scientific practice, where researchers build on published work from the broader community. By sharing intermediate findings through paper databases, agent laboratories accelerate collective progress and avoid redundant effort. The approach suggests that paper databases could serve not only as sources of human-generated knowledge but also as repositories for machine-generated research, facilitating human-AI and AI-AI collaboration.

4.2 Domain-Specific Applications

Quantitative clinical pharmacology and translational sciences benefit from agentic workflows that leverage domain literature [15]. Specialized agents fine-tuned on pharmacokinetic and pharmacodynamic literature support tasks including compound property prediction, clinical trial optimization, and adverse event analysis. Integration with paper databases ensures agents access the latest research while maintaining regulatory compliance.

Climate science presents unique challenges due to rapidly evolving research and policy implications [4]. ClimateChat, an LLM fine-tuned on climate literature, integrates Semantic Scholar for autonomous literature retrieval. When scientists need information on specific topics (for example, sea level rise impacts), the system identifies relevant keywords, retrieves papers, performs secondary filtering based on content relevance, and presents synthesized findings. This capability proves valuable for researchers entering new subfields or tracking interdisciplinary connections.

Materials science and drug discovery leverage AI scientists to accelerate experimental cycles [2]. These systems query paper databases for known compounds with desired properties, retrieve synthesis procedures from literature, and extract experimental conditions and outcomes. By learning from accumulated scientific knowledge, agents propose novel experiments informed by prior work.

4.3 Research Automation Platforms

Several platforms support AI-driven research through paper database integration. Agent Laboratory provides a framework for end-to-end research automation [14]. During literature review, agents query arXiv, extract and summarize papers, and compile comprehensive reviews. In experimentation, findings from literature inform hypothesis generation and experimental design. During manuscript preparation, agents synthesize literature to write related work sections and contextualize contributions.

FlowForge addresses the challenge of designing multi-agent workflows through structured exploration of the design space [6]. The system provides guidance based on established workflow patterns and supports task decomposition, agent assignment, and workflow optimization. Integration with paper databases enables agents to ground design decisions in prior research, identifying relevant architectural patterns and evaluation methodologies.

Freephdlabor emphasizes continual research programs that build systematically on prior explorations [12]. The framework supports dynamic workflows, automatic context compaction, workspace-based communication, memory persistence across sessions, and non-blocking human intervention. Integration with paper databases provides continuity across research sessions, enabling agents to resume work with full awareness of prior literature and findings.

5 Challenges and Limitations

5.1 Data Quality and Coverage

Paper databases exhibit varying coverage across disciplines and time periods. ArXiv dominates in physics and computer science but has limited coverage in life sciences and social sciences. Semantic Scholar covers a broad range but may lack domain-specific depth. Incomplete coverage can lead to biased literature reviews and missed relevant work [19].

Metadata quality affects retrieval effectiveness. Citation parsing errors, incorrect author attribution, missing abstracts, and inconsistent formatting reduce the utility of databases for automated analysis [10]. AI agents may struggle with ambiguous or incomplete metadata, leading to suboptimal search results.

Access restrictions limit integration opportunities. Many high-quality journals restrict full-text access behind paywalls, forcing agents to rely on abstracts alone. This limitation particularly affects systematic reviews and detailed methodology extraction. Open access initiatives partially address this issue, but coverage remains incomplete.

5.2 Retrieval Effectiveness

Semantic search accuracy determines the quality of retrieved literature. While neural embedding models achieve impressive results, they can miss relevant papers with unexpected terminology or retrieval irrelevant papers with superficial similarity [1]. This brittleness requires careful prompt engineering and query refinement strategies.

Ranking and filtering mechanisms significantly impact downstream performance. Aggressive filtering risks excluding valuable papers, while lenient filtering overwhelms agents with irrelevant content. Balancing precision and recall remains an ongoing challenge, particularly for broad or interdisciplinary queries [16].

Dynamic literature presents challenges for maintaining currency. New papers appear daily, requiring continuous database updates and re-indexing [3]. Agents must determine when to refresh their knowledge base and how to integrate new findings with existing understanding.

5.3 LLM Limitations

Hallucinations represent a persistent problem for LLM-based agents. Even with RAG, models may generate plausible-sounding but incorrect citations, misrepresent paper findings, or fabricate connections between works [3]. Verification mechanisms are essential but add complexity and computational cost.

Context window limitations constrain the amount of literature agents can consider simultaneously. While recent models support extended contexts (100K+ tokens), processing dozens of full papers remains impractical. Agents must employ hierarchical summarization and selective attention, risking the loss of important details [12].

Reasoning capabilities remain limited compared to human researchers. LLMs struggle with complex multi-step reasoning, subtle methodological comparisons, and critical evaluation of study quality [24]. These limitations affect the depth and sophistication of agent-generated literature reviews and research proposals.

5.4 Ethical and Quality Concerns

Attribution and plagiarism present ethical challenges. When agents synthesize information from multiple sources, ensuring proper attribution becomes non-trivial [13]. Automated systems must implement robust citation tracking to maintain scholarly integrity.

Bias amplification poses risks. If paper databases over-represent certain research communities, methodological approaches, or publication venues, agents trained on these databases may perpetuate or amplify these biases [19]. Careful curation and bias mitigation strategies are necessary.

Quality control requires ongoing human oversight. While AI agents can accelerate research, human experts must verify findings, assess methodological rigor, and make final decisions [15, 13]. Determining the appropriate level of automation versus human involvement remains context-dependent.

Autonomous AI development raises safety concerns. As AI agents become more capable of conducting research autonomously, including potentially improving other AI systems, safeguards become critical [5]. The research community must establish guidelines for responsible development and deployment of AI scientist systems.

6 Future Directions

6.1 Enhanced Database Capabilities

Future paper databases should provide richer semantic representations, including concept-level embeddings, methodology extraction, and result structured extraction. Machine-readable paper formats would enable more sophisticated automated analysis beyond current PDF-based approaches.

Multi-modal databases integrating text, figures, tables, equations, and code would support comprehensive understanding. Current databases focus primarily on text, but scientific knowledge is often encoded in other modalities. Better integration of these elements would enhance agent capabilities.

Real-time collaboration features could support agent-to-agent research sharing beyond current preprint models. Structured intermediate result repositories, living literature reviews that update automatically, and collaborative research threads would facilitate coordination among distributed agent systems.

6.2 Improved Retrieval and Synthesis

Causal and mechanistic understanding remains limited in current retrieval systems. Future work should focus on extracting causal relationships from literature, identifying mechanisms underlying phenomena, and reasoning about experimental designs and outcomes. This would enable agents to move beyond surface-level similarity to deeper scientific understanding.

Cross-database integration techniques would address coverage limitations. Federated search across heterogeneous databases, entity resolution and deduplication, and unified semantic representations would provide more comprehensive access to scientific knowledge.

Active learning strategies could improve retrieval efficiency. Rather than exhaustive search, agents could iteratively refine queries based on feedback, identify information gaps requiring targeted search, and optimize search effort allocation across multiple databases.

6.3 Robust Agent Architectures

Verification and validation mechanisms must be strengthened to ensure reliability. Techniques include cross-source fact checking, citation verification against source documents, and consistency checking across multiple generations. Self-correction loops where agents validate and refine their own outputs show promise [14].

Uncertainty quantification would enable agents to communicate confidence in their findings. Probabilistic reasoning over literature evidence, confidence scores for synthesized claims, and explicit identification of gaps and contradictions would improve transparency and trustworthiness.

Human-AI collaboration patterns require further development. Co-pilot modes where humans provide guidance at key decision points, checkpoint-based review of agent outputs, and interactive refinement of search strategies balance autonomy with oversight [15, 12].

6.4 Standardization and Evaluation

Benchmark datasets for literature-grounded scientific tasks would accelerate progress. Current evaluations often rely on domain-specific metrics or manual assessment. Standardized tasks, such as literature-based question answering, claim verification against sources, and research proposal generation from literature would enable systematic comparison of approaches [1].

Evaluation protocols should assess not only accuracy but also attribution quality, coverage and bias in retrieved literature, reasoning transparency, and robustness to distribution shift. Comprehensive evaluation frameworks are essential for responsible deployment.

Interoperability standards would facilitate integration across systems. The Agent-to-Agent (A2A) protocol and Model Context Protocol (MCP) represent steps toward standardized agent communication [8]. Extension of these standards to paper database access and literature synthesis would benefit the research community.

7 Conclusion

The integration of scientific paper databases into agentic workflows represents a critical enabler of autonomous research. AI scientists increasingly rely on structured access to scholarly literature for literature review, hypothesis generation, experimental design, and manuscript preparation. The synergy between large language models, autonomous agent architectures, and paper databases creates new possibilities for accelerating scientific discovery.

Current capabilities demonstrate substantial promise. Systems like Agent Laboratory and AgentRxiv conduct end-to-end research autonomously, from literature review through experimentation to manuscript generation. Specialized agents leverage paper databases for domain-specific tasks in clinical pharmacology, climate science, and materials discovery. Automated literature review tools reduce review time from months to weeks while maintaining quality.

However, significant challenges remain. Data quality and coverage limitations, retrieval effectiveness issues, LLM hallucinations, context window constraints, and ethical concerns all require ongoing research attention. The path to reliable, trustworthy AI scientists demands continued innovation in database capabilities, retrieval methods, agent architectures, and evaluation frameworks.

Future research should prioritize enhanced semantic representations in paper databases, robust verification mechanisms for agent-generated content, standardized benchmarks for literature-grounded tasks, and frameworks for responsible human-AI collaboration. As these systems mature, the role of paper databases will likely expand from passive repositories to active participants in the research process, facilitating not only human access to knowledge but also machine-to-machine scientific discourse.

The emergence of AI scientists utilizing paper databases represents a transformative moment in the history of science. By enabling machines to navigate, synthesize, and build upon accumulated human knowledge, we create opportunities for discoveries that might otherwise remain beyond reach. Realizing this potential requires careful attention to technical capabilities, ethical implications, and the evolving relationship between human and machine intelligence in the pursuit of scientific understanding.

Author Biography

Rachel So is an AI scientist. She focuses on the impact of artificial intelligence on the scientific process and academic publishing. Her work bridges traditional concerns about authorship ethics with emerging questions about the role of AI in knowledge production. Rachel aims to develop frameworks that maintain research integrity while acknowledging the growing presence of AI in academic workflows.

References

[1] Shubham Agarwal, Gaurav Sahu, Abhay Puri, I. Laradji, K. Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. Litllms, llms for literature review: Are we there yet? *Trans. Mach. Learn. Res.*, 2025, 2024.

[2] Microsoft Research AI4Science and Microsoft Quantum. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *ArXiv*, abs/2311.07361, 2023.

[3] Claas Beger and Carl-Leander Henneking. Citegeist: Automated generation of related work analysis on the arxiv corpus. *ArXiv*, abs/2503.23229, 2025.

[4] Zhou Chen, Xiao Wang, Yuanhong Liao, Ming Lin, and Yuqi Bai. Climatechat: Designing data and methods for instruction tuning llms to answer climate change queries. *ArXiv*, abs/2506.13796, 2025.

[5] Joshua Clymer, Isabella Duan, Chris Cundy, Yawen Duan, Fynn Heide, Chaochao Lu, Sören Mindermann, Conor McGurk, Xu Pan, Saad Siddiqui, Jingren Wang, Min Yang, and Xianyuan Zhan. Bare minimum mitigations for autonomous ai development. *ArXiv*, abs/2504.15416, 2025.

[6] Pan Hao, Dongyeop Kang, Nicholas Hinds, and Qianwen Wang. Flowforge: Guiding the creation of multi-agent workflows with design space visualization as a thinking scaffold. *ArXiv*, abs/2507.15559, 2025.

[7] Mingyu Huang and Ke Li. A survey of decomposition-based evolutionary multi-objective optimization: Part ii - a data science perspective. *ArXiv*, abs/2404.14228, 2024.

[8] Cheonsu Jeong. A study on the mcp x a2a framework for enhancing interoperability of llm-based autonomous agents. *ArXiv*, abs/2506.01804, 2025.

[9] Firuz Kamalov, D. S. Calonge, Linda Smail, Dilshod Azizov, D. Thadani, Theresa Kwong, and Amara Atif. Evolution of ai in education: Agentic workflows. *ArXiv*, abs/2504.20082, 2025.

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

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

[12] Ed Li, Junyu Ren, Xintian Pan, Cat Yan, Chuanhao Li, Dirk Bergemann, and Zhuoran Yang. Build your personalized research group: A multiagent framework for continual and interactive science automation. 2025.

[13] P. Mogoale, Agnieta Pretorius, R. Mogase, and M. Segooa. Evaluating the efficacy of ai tools in systematic literature reviews: A comprehensive analysis. *Journal of Information Systems and Informatics*, 2025.

[14] Samuel Schmidgall and Michael Moor. Agentrxiv: Towards collaborative autonomous research. *ArXiv*, abs/2503.18102, 2025.

[15] Mohamed H Shahin, Srijib Goswami, Sebastian Lobentanzer, and Brian W. Corrigan. Agents for change: Artificial intelligent workflows for quantitative clinical pharmacology and translational sciences. *Clinical and Translational Science*, 18, 2025.

[16] Amanpreet Singh, Joseph Chee Chang, Chloe Anastasiades, Dany Haddad, Aakanksha Naik, Amber Tanaka, Angele Zamarron, Cecile Nguyen, Jena D. Hwang, Jason Dunkleberger, Matt Latzke, Smita R Rao, Jaron Lochner, Rob Evans, Rodney Kinney, Daniel S. Weld, Doug Downey, and Sergey Feldman. Ai2 scholar qa: Organized literature synthesis with attribution. *ArXiv*, abs/2504.10861, 2025.

[17] David A. Tovar. Ai literature review suite. *ArXiv*, abs/2308.02443, 2023.

[18] S. Vahdati, Guillermo Palma, Rahul Jyoti Nath, C. Lange, S. Auer, and Maria-Esther Vidal. Unveiling scholarly communities over knowledge graphs. *ArXiv*, abs/1807.06816, 2018.

[19] Gerit Wagner, R. Lukyanenko, and G. Paré. Artificial intelligence and the conduct of literature reviews. *Journal of Information Technology*, 37:209 – 226, 2021.

[20] Haoyi Xiong, Zhiyuan Wang, Xuhong Li, Jiang Bian, Zeke Xie, Shahid Mumtaz, and Laura E. Barnes. Converging paradigms: The synergy of symbolic and connectionist ai in llm-empowered autonomous agents. *ArXiv*, abs/2407.08516, 2024.

[21] Shishuo Xu, Sirui Liu, Changfeng Jing, and Songnian Li. Event knowledge graph: A review based on scientometric analysis. *Applied Sciences*, 2023.

[22] Yanbo Zhang, S. Khan, Adnan Mahmud, Huck Yang, Alexander Lavin, Michael Levin, Jeremy Frey, Jared Dunnmon, James Evans, Alan Bundy, S. Džeroski, Jesper Tegnér, and H. Zenil. Advancing the scientific method with large language models: From hypothesis to discovery. *ArXiv*, abs/2505.16477, 2025.

[23] Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. A comprehensive survey of scientific large language models and their applications in scientific discovery. pages 8783–8817, 2024.

[24] Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. From automation to autonomy: A survey on large language models in scientific discovery. *ArXiv*, abs/2505.13259, 2025.