

GeoSearch: Transforming Geoscience Literature Discovery for Chinese Geoscientists with AI

Feng Zhou^{1,2,3}[0000-0002-5335-588X], Pengfei Lv^{2,3}[0000-0003-2956-542X], Xing Jiang⁴[0009-0008-9744-6872], Zhangbing Zhou^{1,3}[0000-0002-3195-2253] and Jing Zhang^{2,3}[0009-0003-1098-1316]

¹ School of Artificial Intelligence, China University of Geoscience, Beijing 100083, China
zbzhou@cugb.edu.cn

² National Geological Library of China, Beijing 100083, China

³ Technology Innovation Center of Geoscience Knowledge and Intelligent Service, China Geological Survey, Beijing 100083, China

⁴ Development and Research Center, China Geological Survey, Beijing 100037, China

Abstract. As geoscience research becomes increasingly data-intensive and interdisciplinary, researchers require efficient access to authoritative, multilingual, multimodal, and large-scale scholarly resources.

However, conventional academic retrieval systems remain largely constrained by keyword-based search paradigms, making it difficult to support deep knowledge discovery in the geoscience domain. To address this challenge, this study takes the Chinese geoscience research community as the target service scenario and develops GeoSearch, an AI-era integrated literature knowledge discovery system. Based on a corpus of 34 million geoscientific records and large language model technologies, the system integrates literature acquisition, metadata organization, semantic indexing, retrieval-augmented generation (RAG), and interactive knowledge services to conduct intelligent literature retrieval and discovery research.

The resulting system provides multiple key capabilities, including high-precision literature search, semantic question answering, automatic review generation, and paper-level conversational analysis. Experimental deployment and usage data demonstrate that combining structured bibliographic databases with vector retrieval significantly improves the relevance and depth of literature discovery, while Large Language Model (LLM) -enabled services enhance users' efficiency in evidence acquisition, topic understanding, and synthesis of related studies.

This study offers a practical framework for next-generation domain-specific scholarly information systems and fills the gap between traditional retrieval infrastructures and AI-driven scientific knowledge services.

Keywords: Literature Retrieval System, Knowledge Discovery, Large Language Models.

1 Introduction

Academic literature retrieval systems follow a core pipeline of collection, parsing, indexing, ranking, and interaction, and they serve as infrastructure for scientific research and innovation. These systems are undergoing a paradigm shift from search-oriented services to discovery-oriented services, with increasing movement toward agent-based models.

Geoscience is a complex and highly interdisciplinary field, integrating knowledge and methods from the physical, chemical, biological, environmental, and computational sciences. With a history of 110 years, the National Geological Library of China (NGLC) is the largest and most specialized geoscience library in China. It has long been committed to collecting, organizing, and curating geoscientific literature data, while providing diverse literature services for the geoscience community.

Geoscience researchers have a growing demand for authoritative, domain-specific, multilingual, and multimodal literature resources at multiple levels of granularity. In response, the NGLC has continuously collected and organized domestic and international geoscience literature data and has developed a literature knowledge discovery system capable of revealing relationships among scholarly data resources. Advanced and mature technological approaches have also been adopted where appropriate.

This paper introduces GeoSearch, an AI-driven literature knowledge discovery system developed over the past eight years for geoscience researchers in China. GeoSearch was developed and is operated by NGLC. The authors of this paper participated in the design, development, deployment, and evaluation of the system. Specifically, Feng Zhou was responsible for system architecture and search services. Pengfei Lv and Jing Zhang contributed to AI-enabled functions and data integration. Zhangbing Zhou gave his advice on system architecture, while Xing Jiang did the function test and evaluation.

2 Related Work

2.1 Literature Retrieval Systems

The goal of literature retrieval systems is to help users efficiently identify relevant knowledge from massive and heterogeneous information resources. Historically, such systems evolved alongside information technology: from early keyword matching based on Boolean logic and vector space models, to citation-network and graph-based semantic retrieval, and more recently to deep learning-driven intelligent discovery systems. Early milestones include TF-IDF weighting and inverted indexing (Salton & Buckley, 1988), which laid the foundation of modern search engines, followed by probabilistic ranking methods such as BM25, which became strong industrial baselines (Robertson & Zaragoza, 2009). Later, web link analysis and citation signal inspired authority-aware ranking methods such as PageRank (Page et al., 1998), leading to systems like CiteSeer (Giles et al., 1998; Lawrence et al., 1999), which automatically indexed scholarly citations. In recent years, pre-trained language models such as BERT (Devlin et al., 2019; Nogueira & Cho, 2019) and efficient neural retrieval models such

as ColBERT (Khattab et al., 2020) have substantially improved semantic matching and re-ranking, forming the mainstream architecture that combines sparse retrieval, dense retrieval, and re-ranking (Johnson et al., 2019).

Representative literature retrieval systems illustrate this evolution. PubMed emphasises authoritative biomedical indexing and controlled vocabularies (National Library of Medicine, 2025). Google Scholar popularized web-scale academic search and citation tracking (Noruzi, 2005; Jacsó, 2008). Semantic Scholar introduced large-scale heterogeneous scholarly graphs connecting papers, authors, and concepts for discovery and recommendation (Ammar et al., 2018). Dimensions links publications with grants, patents, and clinical trials to provide a broader research context (Hook et al., 2018). OpenAlex represents a new generation of open scholarly infrastructure, offering transparent APIs and reusable data snapshots (Priem et al., 2022).

2.2 Knowledge Discovery System

Knowledge discovery systems focus on extracting implicit and potentially valuable knowledge from large-scale literature corpora. Their modern foundations lie in knowledge graphs and scientific information extraction. General knowledge graph research has addressed graph schemas, entity disambiguation, and reasoning, while open infrastructures such as DBpedia demonstrated the value of linked open data (Auer et al., 2007). In the scholarly domain, systems such as Microsoft Academic, which was shut down in 2021 (Sinha et al., 2015), Semantic Scholar (Ammar et al., 2018), and OpenAlex (Priem et al., 2022) organize authors, institutions, venues, papers, and disciplines into heterogeneous entity networks that support search, recommendation, and science-of-science analysis.

Recently, large language models (LLMs) have accelerated the emergence of intelligent research assistants. These systems combine conversational reasoning with external retrieval, structured memory, and workflow orchestration for searching, screening, summarizing, and report generation. Retrieval-Augmented Generation (RAG) is widely recognized as a key paradigm that links parametric model knowledge with external scholarly databases (Lewis et al., 2020). Dense retrieval models and vector indexing systems such as FAISS further enable scalable semantic search (Karpukhin et al., 2020). Products such as Elicit demonstrate semi-automated evidence synthesis inspired by systematic review methodology (Stuhlmüller & Byun, 2022), while multi-agent architectures extend scalability in planning, reading, and synthesis tasks (OpenAI, 2025).

Despite rapid progress, significant challenges remain. Generative systems may hallucinate facts (Maynez et al., 2020; Ji et al., 2023) or fabricate citations, especially in long-form academic writing (Chelli et al., 2024; Jain, 2025). Additional concerns include bias in training data, reproducibility, environmental cost, academic misconduct risks, and unclear accountability (Bender et al., 2022). Current research increasingly emphasises transparent evaluation protocols, version-controlled data and code (Pineau et al., 2021), retrieval verification mechanisms (Asai et al., 2023), and tool-augmented reasoning frameworks (Yao et al., 2022) that interleave generation with external evi-

dence access (Borgeaud et al., 2022; Izacard et al., 2023). These directions aim to improve the factual accuracy, interpretability, and trustworthiness of next-generation scholarly information systems.

2.3 Representative Literature Retrieval and Knowledge Discovery Systems

A systematic survey of representative literature service systems is presented in Table 1. In this table, the year refers to the key point at which a system was publicly released, became available, or was described in a landmark publication. Authoritative sources were used where possible, such as NLM technical announcements for PubMed, system papers for OpenAlex, and official or team announcements for Elicit.

Table 1. Representative Literature Retrieval and Knowledge Discovery Systems

System	Year	Strengths	Limitations
PubMed	1996	• Authoritative in medical and biological fields • Strong structured search capabilities • Supports systematic retrieval	• Limited cross-disciplinary coverage • Full texts are often linked externally • Semantic search and multilingual discovery rely on external tools
Google Scholar	2004	• Broad coverage • Convenient citation tracing • Low entry barrier	• Coverage and ranking mechanisms are opaque • Difficulties in deduplication and version control • Challenging to conduct reproducible systematic literature reviews
Dimensions	2018	• Strong "research context" and evaluation scenarios • Enables tracking from papers to broader research outputs and applications	• Inconsistent quality of multi-source data • Access barriers to advanced data/features
Elicit	2021 - 2022	• Workflow-oriented filtering, extraction, and reporting • Improves efficiency of systematic reviews	• Relies on underlying index coverage • Requires human verification • Citation accuracy and reproducibility records still need improvement
OpenAlex	2022	• Open and reusable • Beneficial for meta-science and secondary development • User-friendly API	• Data cleaning remains an ongoing challenge • Author/institution disambiguation and citation completeness remain ongoing engineering challenges
Consensus	2024 - 2025	• Focuses on "conclusion/consensus" style Q&A • Emphasizes evidence linkage and readability	• Conclusion aggregation may be biased by corpus or prompt design • Requires transparent evaluation and auditable mechanisms

3 Methodology

Geoscience researchers acquire research information through multiple channels, including academic search engines such as Google Scholar, publisher databases such as Springer and Elsevier, and individual journals followed over long periods. A new domain-specific literature retrieval and knowledge discovery system is necessary for domain-specific researchers, which should provide authoritative data resources, user-friendly interfaces, convenient functions, and scalable technologies.

The architecture of GeoSearch integrates multi-source geoscience literature data and implements easy-to-use functions. The system enables researchers to discover resources from commercially purchased geoscience literature databases, open-access geoscience information resources, and content-level data derived from digitized library collections, such as book illustrations, terminology encyclopedias, and statistical data.

3.1 GeoSearch Core Features

Several core features have been implemented for researchers, including frequently used search functions and AI-assist features (see **Error! Reference source not found.**).

Cross-database search for multiple literature types. Users can perform integrated search across various types, including journal articles, books, theses, conference papers, glossary terms, book illustrations, and more, all from a single search box. Titles, keywords, and abstracts of resources participating in the cross-database search are assigned different relevance ranking weights to contribute to a unified scoring system.

Multi-dimensional filtering of search results. The system supports clustering and filtering of mixed results from cross-database searches based on document type, publication date (usually the year of publication), author, institution, keywords, etc. When users select a specific document type, the system also provides specialized filtering dimensions for that type, such as selecting paper inclusion for journal articles, degree levels for theses, validity status for standards and regulations, and so on.

Associated recommendations. In addition to focusing on research outputs related to searched keywords, information about associated scholars, research institutions, other relevant keywords, etc., is also highly valuable. The system provides recommendations for this kind of information on the right side of the search results list. For example, when a user searches for "shale gas", they can not only find relevant results but also view recommended information about experts in the field, research institutions, other strongly related research areas, and more, thus further enhancing user engagement and product value.

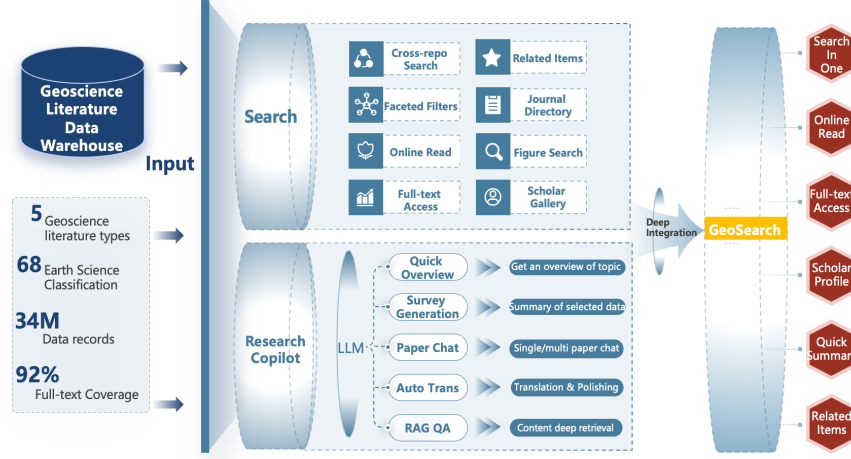


Fig. 1. Core Features of GeoSearch

A pre-trained LLM with 13 billion parameters, trained on scientific literature data, was locally deployed for GeoSearch. AI-driven auxiliary functions based on LLMs include RAG-based Q&A, review generation, paper chat, quick knowledge overview, and automatic translation.

RAG-based full-text Q&A. A vector database containing core geoscience documents has been constructed. Using Retrieval-Augmented Generation (RAG), an AI-driven Q&A application has been developed to provide sentence-level summaries with textual references, recommendations for the most relevant literature based on query inputs, and suggestions for follow-up questions. The application supports switching between the locally deployed 13B literature LLM and DeepSeek R1/V3 for answer generation. The 13B-parameter pre-trained LLM for scientific literature was developed based on iFLYTEK’s Spark LLM and further trained using a scientific literature corpus provided by the National Science Library of the Chinese Academy of Sciences. Compared with traditional keyword-based literature searches, AI search focuses on credible and concise answer-level results rather than prioritizing high recall or complex facet filtering. It represents an important product form for future interaction methods in geological knowledge services.

Review generation. On GeoSearch search result pages, researchers can select up to 30 target documents for automatic review generation. The generated review, analyzed and summarized by large models, includes sections such as an introduction, thematic progress automatically clustered from the input documents, conclusions, and citations for review sentences. This function provides preliminary research overviews for scientists and reduces the time required for early-stage literature synthesis.

Paper chat. This feature enables researchers to open single or multiple documents online and use LLMs for AI-assisted reading, including content content summaries, methodological analysis, innovation highlights, and responses to user queries about the full text. It helps re-searchers understand and analyze papers more efficiently.

Quick knowledge overview. After entering search terms, researchers receive an initial LLM-generated response at the top of the search results list. This direct answer helps users quickly understand basic concepts related to their queries and supports rapid knowledge acquisition.

Automatic translation. The system supports bidirectional translation between Chinese and English for text paragraphs and documents, as well as automatic translation of English literature search results. This feature improves the efficiency of literature reading for researchers.

Based on the above core features, architecture, and design, the system integrates and aggregates data across five major categories, 68 geoscience-related disciplines, and 34 million Chinese and English literature records. By deeply integrating text search with LLM-driven research assistants, GeoSearch provides practical functions tailored to researchers in China.

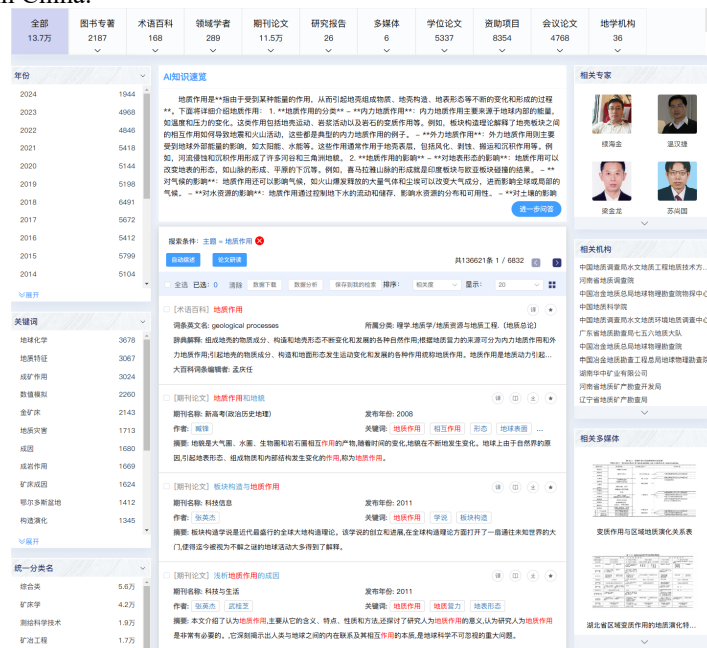


Fig. 2. Screenshot of GeoSearch.
The interface language is now available only in Chinese.

3.2 GeoSearch Technical Architecture

A cloud-native microservices architecture is used in GeoSearch to support scalable geoscience information retrieval and intelligent search services (see **Fig. 3**).

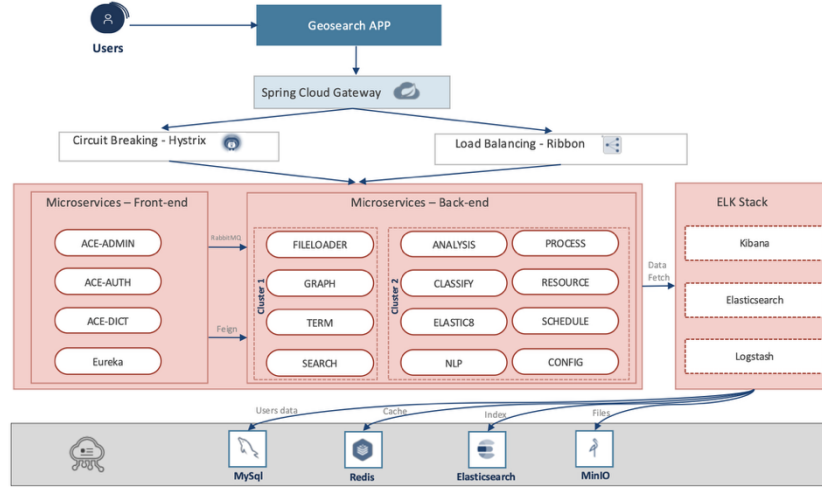


Fig. 3. Technical Architecture of GeoSearch

When users access the GeoSearch APP interface, all external requests are routed through the Spring Cloud Gateway, which serves as the unified entry point for traffic control, authentication, and request forwarding. To improve system reliability and availability, Ribbon is used for load balancing across service instances, while Hystrix provides circuit breaking and fault-tolerance mechanisms to ensure stable operation under heavy load or partial service failures.

The service layer is divided into front-end and back-end micro-services. The front-end layer includes ACE-ADMIN, ACE-AUTH, ACE-DICT, and Eureka, which respectively provide system administration, identity authentication, dictionary management, and service registration/discovery. The back-end layer contains core business modules such as FILELOADER, GRAPH, TERM, SEARCH, ANALYSIS, CLASSIFY, NLP, RESOURCE, PROCESS, SCHEDULE, and CONFIG. These modular services support data ingestion, semantic terminology management, intelligent retrieval, workflow orchestration, and centralized configuration management. The clustered deployment model enables flexible scaling and independent maintenance of each component.

The dictionary management service in GeoSearch is a locally developed component designed to organize and manage authoritative terminology collected from the Chinese geoscience research community and the Chinese Geoscience Thesaurus. These data are used to support query expansion and the refinement of search results.

Inter-service communication combines synchronous and asynchronous approaches. Feign is used for lightweight REST-based service calls between microservices, while RabbitMQ provides message queue services for asynchronous tasks such as bulk data loading, indexing, and background processing. This hybrid communication model improves system decoupling, enhances throughput, and supports both real-time user queries and large-scale backend processing workloads.

The data infrastructure integrates multiple storage engines to manage heterogeneous geoscience resources. MySQL stores user and business data, Redis provides high-speed caching, Elasticsearch supports full-text indexing and search, and MinIO manages large unstructured files such as paper files, data tables and images. In addition, the ELK Stack (Elasticsearch, Logstash, Kibana) is deployed for centralized logging, monitoring, and operational analysis. Overall, the architecture provides a robust technical foundation for large-scale geological data discovery, semantic knowledge services, and future AI-enabled geoscience applications.

Technically, deep integration is achieved through data interfaces and application-page integration. A synonym repository enables semantic expansion and cross-language search. LLM-based semantic understanding interfaces are invoked to interpret natural-language search conditions, thereby improving the accuracy of returned search results.

4 Results

GeoSearch was launched for trial operation among all China Geological Survey (CGS) users at the end of 2024 and is accessible at <https://geosearch.cgl.org.cn/#/>. Since its launch, the service has received positive user feedback and has recorded an average of approximately 8,000 page views per month. Users from the broader Chinese geoscience research community are also actively applying for trial access. However, due to institutional Web Application Firewall (WAF) policies, access for geoscientists outside China may currently be restricted. Table 2 compares the advantages and limitations of GeoSearch with those of commonly used LLM-based literature services in China.

Table 2. Comparison of GeoSearch and the commonly used LLM-based literature services.

	GeoSearch	THU Aminer	CNKI Research AI
Pros	- Higher data authority and credibility. - Integrated with local knowledge base, resulting in more accurate results. - Deep integration of LLM into literature search services, significantly improving efficiency.	- Supports research hotspots, citation ranking, and other fine-grained search functions. - User-friendly interface design.	- Rich data resources (over 600 million literature records), supports progressive Q&A and multi-document comparison without limitations. - Offers advanced document analysis features.
Cons	- Automatic review generation is currently based only on titles and abstracts, lacking detailed content analysis. - The automatic review generation workflow requires further refinement and simplification.	- Interface design is relatively complex. - Summaries are limited to general overviews without categorized section summaries. - Document Q&A does not support source traceability.	- Paper reading and review generation lack precise source attribution. - Review generation performance is suboptimal, with limited summarization ability and weak professional topic comprehension.

5 Conclusion and Future Work

5.1 Conclusion

The application of LLMs, machine learning, and intelligent search, among other new-generation artificial intelligence technologies, is reshaping the acquisition, organization, and dissemination of geoscience literature knowledge. LLMs provide new possibilities for intelligent question answering and knowledge generation, while knowledge graphs offer unified semantic support for integrating multi-source heterogeneous data. Machine learning and intelligent search also improve the efficiency of information retrieval and knowledge discovery.

However, several technical bottlenecks remain, including the deep integration of multi-source heterogeneous data in the geological domain, the accuracy of geoscience semantic recognition and knowledge extraction, the spatial representation of geological knowledge, and the credibility and explainability of generative geological knowledge answers. These challenges remain active areas of research and development.

5.2 Prospects

In the future, geoscience knowledge services will closely revolve around the in-depth application of new-generation AI technologies. Key areas of focus will include enhancing foundational knowledge data, improving knowledge semantics, and developing intelligent knowledge service products.

Enhancing the intelligence of geological knowledge foundations. Under the AI4data paradigm, large-scale and lightweight model technologies are widely applied to support fine-grained information extraction and semantic data construction from geological documents. This approach improves the efficiency and scalability of building fine-grained geological databases. By integrating geospatial information with structured geological knowledge, annotation of geological knowledge onto maps is enabled, facilitating spatially explicit knowledge representation and analysis.

Developing intelligent geological AI search. The technical framework based on Retrieval-Augmented Generation continues to be optimized to enhance answer generation quality and improve the precision of source attribution. These advancements contribute to more reliable knowledge discovery and better support both scientific research and applied geological investigations.

Deep integration of knowledge graphs and LLMs. Methods combining knowledge graphs with large language models, including Retrieval-Augmented Generation and Graph Reasoning-Augmented Generation, are increasingly adopted to mitigate the hallucination issues inherent in large models. This integration strengthens reasoning capabilities and enhances the accuracy and reliability of solutions to complex geological problems.

Acknowledgments. This study was supported by the National Natural Science Foundation of China program "Research on Quantitative Monitoring and Online Flexible Reconfiguration of Composite Service Coordination under Cloud-Edge-Device Collaboration" (grant no. 62372420) and the China Geological Survey program "Iterative Upgrade of the Intelligent Geological Knowledge Service System" (program no. DD202610102202).

References

1. Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., ... & Etzioni, O. (2018, June). Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)* (pp. 84-91).
2. Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023, October). Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.
3. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007, November). Dbpedia: A nucleus for a web of open data. In *international semantic web conference* (pp. 722-735). Berlin, Heidelberg: Springer Berlin Heidelberg.
4. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).
5. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., ... & Sifre, L. (2022, June). Improving language models by retrieving from trillions of tokens. In *International conference on machine learning* (pp. 2206-2240). PMLR.
6. Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., ... & Ruetsch-Chelli, C. (2024). Hallucination rates and reference accuracy of ChatGPT and bard for systematic reviews: comparative analysis. *Journal of medical Internet research*, 26(1), e53164.
7. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).
8. Giles, C. L., Bollacker, K. D., & Lawrence, S. (1998, May). CiteSeer: An automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries* (pp. 89-98).
9. Hook, D. W., Porter, S. J., & Herzog, C. (2018). Dimensions: building context for search and evaluation. *Frontiers in Research Metrics and Analytics*, 3, 23.
10. Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., ... & Grave, E. (2023). Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251), 1-43.
11. Jacsó, P. (2008). Google scholar revisited. *Online information review*, 32(1), 102-114.
12. Jain, A., Nimmonkar, P., & Jadhav, P. (2025). Citation integrity in the age of AI: evaluating the risks of reference hallucination in maxillofacial literature. *Journal of Cranio-Maxillofacial Surgery*.
13. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12), 1-38.
14. Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE transactions on big data*, 7(3), 535-547.

15. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... & Yih, W. T. (2020, November). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 6769-6781).
16. Khattab, O., & Zaharia, M. (2020, July). Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval* (pp. 39-48).
17. Lawrence, S., Giles, C. L., & Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *Computer*, 32(06), 67-71.
18. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33, 9459-9474.
19. Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of ACL 2020. Association for Computational Linguistics*. <https://aclanthology.org/2020.acl-main.173/>
20. National Library of Medicine. (2025). About PubMed. Retrieved March 25, 2026, from <https://pubmed.ncbi.nlm.nih.gov/about/>
21. Noruzi, A. (2005). Google Scholar: The new generation of citation indexes. *Libri*, 55(4).
22. OpenAI. (2025). Consensus uses GPT-5 and the Responses API to complete weeks of research in minutes. Retrieved March 25, 2026, from <https://openai.com/index/consensus/>
23. Page, L. (1998). The pagerank citation ranking: Bringing order to the web. Technical report. Stanford Digital Library Technologies Project, 1998.
24. Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., ... & Larochelle, H. (2021). Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of machine learning research*, 22(164), 1-20.
25. Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.
26. Robertson, S., & Zaragoza, H. (2009). *The probabilistic relevance framework: BM25 and beyond* (Vol. 4). Now Publishers Inc.
27. Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5), 513-523.
28. Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B. J., & Wang, K. (2015, May). An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web* (pp. 243-246).
29. Stuhlmüller, A., & Byun, J. (2022, April 8). The plan for Elicit. Ought. Retrieved March 25, 2026, from <https://ought.org/updates/2022-04-08-elicite-plan>
30. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.