

Large Language Model Evaluation for Domain-Specific Scientific Information Retrieval

Dr. Amélie De Smet

Associate Professor, Ardennes Institute of Knowledge Engineering, Belgium

Abstract

Scientific information retrieval plays an essential role in supporting academic research, evidence discovery, knowledge synthesis, and the identification of relevant scholarly publications. Conventional search systems frequently employ lexical matching and statistical ranking methods to identify documents associated with user queries. Although these approaches provide established mechanisms for locating scientific information, they may encounter difficulties when researchers use specialized terminology, interdisciplinary expressions, abbreviations, or conceptual questions that differ from the wording of indexed publications. Advances in natural language processing and large language models have introduced additional opportunities for improving query interpretation, semantic document representation, and relevance ranking. The present paper develops a hypothetical methodological framework for evaluating large language model-assisted scientific information retrieval across multiple research domains. Four retrieval configurations are considered: BM25 lexical retrieval, scientific-domain dense retrieval, large language model-based query expansion with BM25, and hybrid retrieval with large language model reranking. A fictional corpus containing 3,200 scientific document abstracts and 240 research queries is introduced to demonstrate the proposed comparative design. The represented disciplines include biomedical sciences, materials science, environmental sciences, and computer science.

The methodology incorporates scientific corpus preparation, query construction, relevance assessment, model configuration, ranking evaluation, and evidence-grounding analysis. Normalized Discounted Cumulative Gain, Recall, and Mean Reciprocal Rank are selected as principal retrieval measures. All numerical values are manually constructed and do not represent executed retrieval experiments or verified model outputs. The methodological discussion examines domain-specific terminology, semantic relevance, cross-disciplinary retrieval, computational costs, model generalization, and the reliability of generated scientific explanations. Particular attention is given to distinguishing document retrieval effectiveness from the factual accuracy of language-model-generated responses. The paper concludes that large language models offer a relevant methodological direction for scientific information retrieval, although their practical effectiveness requires reproducible experiments, independent relevance judgments, transparent evaluation, and careful verification of cited scientific evidence.

Keywords: Large Language Models, Scientific Information Retrieval, Domain-Specific Search, Semantic Retrieval, Natural Language Processing, Scientific Document Ranking, Retrieval-Augmented Generation, Query Expansion, Relevance Evaluation, Scholarly Information Systems.

1. Introduction

Scientific information retrieval has become an important component of contemporary research because scholars increasingly depend on digital databases, institutional repositories, and online publication platforms to identify relevant academic information. Scientific documents contain specialized terminology, methodological descriptions, empirical findings, and disciplinary concepts that require careful interpretation during information searching. Conventional retrieval systems commonly rank documents according to relationships between query terms and indexed content. Although these methods provide useful access to scholarly materials, their effectiveness can be limited when relevant publications use terminology that differs from the words selected by researchers. Consequently, improving the interpretation of scientific queries has become an important objective within information retrieval research.

Domain-specific scientific information presents challenges that differ from those commonly encountered in general web search. Biomedical publications may contain technical abbreviations and specialized clinical terminology, while materials science articles frequently describe chemical structures, experimental procedures, and physical properties. Environmental research may employ geographical, statistical, and climatic terminology, whereas computer science literature often contains mathematical expressions and rapidly changing technical concepts. These differences complicate the representation of scientific meaning through isolated keywords. A retrieval system may identify documents containing matching terms without adequately recognizing their methodological relevance or the relationships among scientific concepts. Effective scientific retrieval therefore requires attention to both terminology and contextual meaning.

Large language models provide additional computational capabilities for interpreting natural language queries, representing semantic relationships, and generating text associated with retrieved information. Transformer-based architectures have contributed to advances in contextual representation learning, while scientific-domain language models have introduced specialized approaches for processing scholarly terminology. Large language models may also support query expansion, document reranking, and evidence-based response generation. However, these capabilities do not establish that model-generated interpretations are scientifically reliable. A language model may produce plausible explanations that are unsupported by the retrieved documents or introduce technical information that was not present within the original sources.

Evaluating scientific information retrieval systems requires more than measuring whether documents share vocabulary with a user query. A retrieved publication may be relevant because it addresses a particular research method, investigates a related mechanism, or provides evidence concerning a scientific question. Relevance may therefore vary in degree rather than belonging exclusively to a binary category. Evaluation procedures must also distinguish document discovery from the accuracy of any subsequent generated explanation. A retrieval system can identify appropriate sources while an associated language model misrepresents their findings. Conversely, a fluent response may conceal the failure to retrieve the most relevant scientific evidence.

The present paper proposes a hypothetical comparative investigation of large language model-assisted information retrieval across four scientific domains. BM25, scientific-domain dense retrieval, language model-based query expansion, and hybrid retrieval with language model reranking are examined within a common methodological framework. A fictional collection of scientific abstracts and research queries provides the basis for illustrating document relevance evaluation. The proposed design

emphasizes query interpretation, ranking performance, evidence grounding, and computational limitations. No scientific documents were experimentally indexed, and no retrieval models were executed. The study instead establishes a methodological foundation for future empirical evaluation of domain-specific scientific search systems.

2. Objectives of the Study

The primary objective of this study is to examine the methodological application of large language models in domain-specific scientific information retrieval. The investigation concentrates on scientific query interpretation, comparative document ranking, evidence relevance, and evaluation reliability. Three specific objectives establish the analytical direction of the proposed research framework.

2.1 Examination of Scientific Query Interpretation and Semantic Retrieval

The first objective is to examine how scientific queries can be represented and interpreted using conventional lexical techniques and language-based semantic approaches. Scientific questions frequently contain specialized terminology, methodological expressions, and conceptual relationships that cannot always be captured through exact word matching. The proposed study considers how lexical representations, contextual embeddings, and language model-generated query reformulations may influence document retrieval. Particular attention is given to identifying relevant scientific abstracts that use different terminology while addressing comparable research concepts.

The investigation further aims to examine the potential limitations of semantic query interpretation. A language model may generate expressions that appear relevant but alter the original scientific question or introduce unsupported technical assumptions. Such changes can affect the documents retrieved and complicate the assessment of relevance. Consequently, the proposed framework emphasizes the preservation of query meaning and the need to distinguish semantic expansion from speculative interpretation. This objective seeks to establish appropriate conditions for examining language model assistance in scholarly information searching.

2.2 Comparative Evaluation of Scientific Retrieval Configurations

The second objective focuses on establishing a comparative framework for four scientific information retrieval configurations. BM25 provides a conventional lexical baseline, while a scientific-domain dense encoder represents semantic retrieval. Large language model-based query expansion introduces an additional query interpretation procedure, and hybrid retrieval with neural reranking combines complementary ranking mechanisms. The proposed comparison examines Normalized Discounted Cumulative Gain, Recall, and Mean Reciprocal Rank. These measures provide different perspectives on ranking effectiveness and the identification of relevant scientific documents.

A further objective involves examining whether additional computational complexity produces meaningful improvements in retrieval performance. Language model-based systems may require substantial processing resources for query reformulation or document reranking. However, more extensive language processing does not necessarily improve relevance across every scientific discipline. The proposed framework therefore emphasizes consistent evaluation conditions, transparent model configuration, and independent testing. This objective encourages comparison based on reproducible retrieval evidence rather than assumptions concerning the superiority of increasingly complex language models.

2.3 Assessment of Evidence Reliability and Scientific Retrieval Limitations

The third objective is to identify methodological challenges that may influence the reliability of large language model-assisted scientific retrieval. These include specialized terminology, ambiguous research questions, incomplete abstracts, inconsistent relevance judgments, and differences among scientific disciplines. Such characteristics may affect both lexical and semantic retrieval performance. The proposed study considers how carefully prepared queries, independent relevance assessment, and appropriate ranking metrics could address selected difficulties while preserving the scientific meaning of retrieved information.

The investigation also considers the reliability of scientific explanations generated from retrieved documents. A model may identify relevant publications yet produce statements that misrepresent experimental findings, overlook limitations, or attribute unsupported conclusions to cited sources. Consequently, retrieval quality and response factuality require separate evaluation. This objective emphasizes source traceability, evidence-supported interpretation, and transparent handling of uncertainty. These considerations provide a foundation for assessing large language models as components of scientific information systems rather than independent authorities on scientific correctness.

3. Review of Literature

3.1 Foundations of Scientific Information Retrieval and Ranking Evaluation

Järvelin and Kekäläinen (2002) investigated cumulative gain-based evaluation methods and established an important foundation for assessing ranked retrieval results using graded relevance. Manning et al. (2008) provided a systematic treatment of information retrieval concepts, including indexing, vector representations, ranking, and evaluation. Robertson and Zaragoza (2009) examined the probabilistic relevance framework and discussed the mathematical foundations of BM25 and related ranking methods. These contributions established essential principles for understanding how conventional retrieval systems identify and rank documents according to their relationships with user queries.

The development of scientific information repositories introduced additional requirements involving document metadata, technical terminology, and relationships among scholarly publications. Lo et al. (2020) developed S2ORC as a resource containing academic text and structured bibliographic information. Wadden et al. (2020) introduced SciFact for retrieving scientific evidence and evaluating claims against published research. These contributions demonstrated that scholarly information retrieval requires more than generic keyword matching. Scientific relevance often depends on whether a document addresses a particular claim, method, finding, or research context.

3.2 Scientific Language Models and Neural Retrieval Architectures

Vaswani et al. (2017) introduced the Transformer architecture, establishing an influential foundation for contextual representation learning. Devlin et al. (2019) developed BERT using bidirectional language representations, while Beltagy et al. (2019) introduced SciBERT to address specialized scientific language processing requirements. Reimers and Gurevych (2019) examined sentence embeddings for semantic similarity tasks. These studies demonstrated how contextual language models can represent textual relationships that extend beyond individual keywords, providing important foundations for scientific semantic retrieval.

Karpukhin et al. (2020) investigated dense passage retrieval through dual-encoder representations. Khattab and Zaharia (2020) introduced ColBERT for efficient retrieval using contextualized late interaction. Cohan et al. (2020) developed SPECTER using citation-informed scientific document representations. Thakur et al. (2021) introduced BEIR to assess information retrieval models across heterogeneous evaluation datasets. These contributions established different neural retrieval strategies and emphasized that model performance can vary substantially across domains and tasks. Their findings support domain-sensitive evaluation rather than universal assumptions about retrieval effectiveness.

3.3 Large Language Models, Retrieval Augmentation, and Evidence Assessment

Lewis et al. (2020) developed retrieval-augmented generation by combining neural text generation with access to external information. Gao et al. (2023) investigated hypothetical document embeddings for zero-shot dense retrieval. Wang et al. (2023) developed Query2doc, demonstrating how large language models can generate pseudo-documents for query expansion. Sun et al. (2023) examined language models as document reranking agents. These studies established important approaches for integrating generative language models with conventional and neural information retrieval procedures.

Min et al. (2023) introduced FActScore for examining whether generated factual claims are supported by reliable information sources. Liu et al. (2023) investigated language model-based evaluation and identified concerns associated with automated evaluators. Es et al. (2024) developed RAGAs for assessing retrieval-augmented generation pipelines, while Asai et al. (2024) investigated adaptive retrieval and self-reflective generation. Liu et al. (2024) examined difficulties in utilizing relevant information positioned within long contexts. Together, these contributions demonstrate the importance of evaluating retrieval relevance, generated factual accuracy, and evidence utilization as distinct dimensions of language model performance.

4. Research Methodology

4.1 Research Design and Retrieval Model Configuration

The present paper adopts a hypothetical comparative quantitative research design for evaluating language model-assisted scientific information retrieval. Four system configurations are considered: BM25 lexical retrieval, scientific-domain dense retrieval, BM25 with language model-based query expansion, and hybrid lexical–dense retrieval followed by large language model reranking. The proposed investigation involves four scientific disciplines and predefined research queries. Each configuration would produce ranked lists of scientific documents for the same evaluation queries. No actual retrieval experiments were conducted for this illustrative manuscript.

The intended analytical process consists of scientific corpus preparation, query construction, relevance annotation, retrieval index development, model configuration, ranking generation, and independent performance evaluation. The dense retrieval system would use a scientific-domain encoder adapted for query–document matching. The query expansion configuration would preserve the original query while incorporating generated explanatory terms. The hybrid configuration would combine lexical and dense candidates before reranking a predefined candidate list. Model parameters, prompts, and retrieval thresholds would be selected without access to independent test relevance judgments.

4.2 Hypothetical Scientific Corpus and Relevance Assessment

For illustrative purposes, the study assumes a fictional corpus containing 3,200 scientific document abstracts distributed equally across biomedical sciences, materials science, environmental sciences, and computer science. Each domain contributes 800 abstracts and 60 research queries. The resulting evaluation collection contains 240 scientific queries. These specifications were manually constructed and do not describe publications collected from a scientific database. No actual article texts, bibliographic records, or proprietary scholarly materials were accessed or indexed.

The proposed query preparation procedure would distinguish terminology-based queries, conceptual research questions, methodological comparisons, and evidence-seeking requests. Relevance would be assessed using a three-level scale: not relevant, partially relevant, and highly relevant. Independent reviewers with suitable subject expertise would assess candidate documents using predefined criteria. Disagreements would be resolved through documented adjudication. Because complete relevance assessment across all documents may be impractical, pooled judgments would be required, with the limitations of incomplete annotation explicitly reported during evaluation.

4.3 Retrieval Evaluation and Evidence-Grounding Procedures

The proposed query partition would allocate 120 queries to model development, 60 to validation, and 60 to independent testing. Each scientific domain would contribute 30 development queries, 15 validation queries, and 15 test queries. Retrieval performance would be evaluated using $nDCG@10$, $Recall@20$, and $MRR@10$. Graded relevance would inform $nDCG$, while MRR would identify the rank of the first highly relevant document. Recall would be calculated using the available judged-relevant documents, with the incompleteness of relevance judgments acknowledged.

A separate evidence-grounding assessment would evaluate scientific explanations generated from retrieved documents. This component would examine whether each substantive claim is supported by an identifiable source and whether the source actually contains the attributed information. Generated responses would be assessed through claim-level annotation rather than fluency alone. Citation validity, unsupported assertions, contradiction, and appropriate abstention would be recorded. Human expert review would remain necessary because automated language model evaluation can itself introduce errors and biases.

5. Analysis

5.1 Illustrative Distribution of Scientific Documents and Research Queries

The composition of an evaluation corpus can influence conclusions concerning the effectiveness of scientific information retrieval. Different disciplines may contain specialized terminology, distinct publication conventions, and varying relationships between query language and document content. Table 1 presents a hypothetical distribution of scientific abstracts and research queries across four domains. Each discipline contributes an equal number of documents and queries to simplify comparative evaluation. These values are methodological assumptions rather than observations obtained from an actual scholarly database.

The illustrative distribution demonstrates how scientific query partitions could be organized while maintaining domain representation across development, validation, and independent testing. Equal allocation reduces certain forms of domain imbalance but does not guarantee that the selected research questions represent the full complexity of scientific information needs. A genuine investigation would

require documented sampling procedures, transparent query construction, and subject-specific relevance assessment. The hypothetical dataset therefore provides a reporting example rather than evidence concerning actual scientific retrieval conditions.

Table 1: Hypothetical Scientific Information Retrieval Evaluation Dataset

Scientific Domain	Document Abstracts	Development Queries	Validation Queries	Test Queries	Total Queries
Biomedical Sciences	800	30	15	15	60
Materials Science	800	30	15	15	60
Environmental Sciences	800	30	15	15	60
Computer Science	800	30	15	15	60
Total	3,200	120	60	60	240

The constructed dataset contains 3,200 scientific abstracts and 240 research queries. Its principal purpose is to demonstrate balanced domain allocation and independent evaluation partitions rather than establish an empirical representation of scientific information retrieval workloads.

5.2 Illustrative Comparison of Scientific Retrieval Performance

Scientific information retrieval systems can be compared through their ability to rank relevant publications and identify appropriate evidence among available documents. $nDCG@10$ considers graded relevance within the first ten results, $Recall@20$ describes the proportion of judged-relevant documents retrieved within the first twenty, and $MRR@10$ emphasizes the position of the first highly relevant result. Table 2 presents hypothetical performance values for four retrieval configurations. The values were manually constructed and do not represent outputs generated through actual model execution.

The illustrative comparison assigns the highest ranking performance to hybrid retrieval with language model reranking. The dense retrieval and query expansion configurations receive higher fictional values than the lexical baseline, demonstrating one possible analytical scenario. However, these differences cannot establish actual improvement because the models were not implemented. A genuine comparison would require reproducible indexing, independent relevance judgments, controlled prompts, repeated evaluation, and statistical uncertainty assessment. Differences in processing costs would also require systematic measurement.

Table 2: Hypothetical Comparative Evaluation of Scientific Information Retrieval Configurations

Retrieval Configuration	$nDCG@10$	$Recall@20$	$MRR@10$
BM25 Lexical Retrieval	0.57	0.63	0.54

Scientific-Domain Dense Retrieval	0.64	0.70	0.61
LLM Query Expansion with BM25	0.67	0.74	0.64
Hybrid Retrieval with LLM Reranking	0.73	0.79	0.69

The fictional pattern illustrates how language model assistance could be assessed alongside lexical and dense retrieval methods. However, ranking metrics alone would not establish the factual reliability of any scientific explanations generated from the retrieved documents.

6. Discussion

6.1 Scientific Terminology and Semantic Relevance

Scientific retrieval depends substantially on the relationship between user queries and the terminology used within scholarly publications. Conventional lexical methods can identify exact technical expressions and may perform effectively when scientific questions contain distinctive terms. However, relevant documents may describe similar research concepts using alternative terminology or disciplinary conventions. Dense language representations provide a method for examining semantic relationships beyond direct word overlap. The proposed framework evaluates these approaches as complementary retrieval strategies rather than mutually exclusive alternatives.

The importance of domain-specific representation is evident in scientific literature containing specialized abbreviations, technical methods, and discipline-dependent meanings. A term may represent different concepts across biomedical and engineering publications, while several expressions may refer to related scientific phenomena. Models adapted to scholarly language may help represent these relationships, but semantic similarity does not automatically establish scientific relevance. Consequently, the proposed evaluation emphasizes independently assessed relevance and the need to examine whether retrieved documents genuinely address the question being asked.

6.2 Large Language Models and Scientific Query Expansion

Large language models can contribute to scientific information retrieval by interpreting complex questions and proposing additional terminology for search. Query expansion may help identify publications that employ related technical expressions or describe methods indirectly associated with the original request. However, language-generated terms can also introduce inaccuracies or shift the intended meaning of a scientific question. Consequently, the usefulness of query expansion depends on semantic preservation and the ability to distinguish plausible terminology from unsupported scientific assumptions.

The hypothetical comparison presents one possible scenario in which language model-based query expansion receives favorable retrieval scores. Nevertheless, this numerical pattern does not establish actual improvement. An implemented study would need to investigate which types of scientific queries benefit from expansion and which experience deterioration through topic drift. Researchers should also examine whether generated terminology reflects domain-specific meanings accurately. Such analysis would provide a more informative assessment than simply comparing average retrieval scores across all research questions.

6.3 Retrieval Effectiveness and Evidence-Grounded Generation

Retrieval effectiveness and scientific response reliability represent related but distinct evaluation objectives. A system may identify relevant documents while producing an explanation that omits important qualifications or inaccurately summarizes the evidence. Conversely, an apparently coherent scientific explanation may contain statements unsupported by the retrieved sources. The proposed framework therefore separates document ranking evaluation from evidence-grounding assessment. This distinction is necessary because fluent natural language generation cannot independently establish the validity of scientific information.

The literature on retrieval-augmented generation and factual evaluation supports this separation. Lewis et al. (2020) established an approach for combining external retrieval with language generation, while Min et al. (2023) examined claim-level factual support. Es et al. (2024) further addressed evaluation across retrieval and generation components. These contributions indicate that scientific retrieval systems should be assessed through multiple complementary criteria. Future implementation should therefore report ranking accuracy, citation validity, unsupported claims, and appropriate uncertainty rather than relying exclusively on aggregate retrieval performance.

7. Challenges

7.1 Specialized Scientific Language and Domain Generalization

Scientific disciplines employ specialized terminology that may be difficult for general-purpose language models to interpret consistently. Technical abbreviations, chemical terminology, mathematical expressions, and discipline-specific conceptual distinctions can introduce ambiguity during query processing. A model trained primarily on general language may represent some scientific relationships inadequately. Even scientific-domain models may perform differently across disciplines because their training data contain unequal coverage of specialized subjects. Consequently, reliable retrieval requires appropriate evaluation across distinct scientific domains.

Domain generalization introduces further difficulties when a model encounters scientific terminology or research methods that differ from those represented during development. A configuration performing well on computer science abstracts may not achieve comparable performance on materials science or biomedical literature. Query expansion and reranking may also respond differently to domain-specific linguistic characteristics. The proposed methodology therefore emphasizes separate domain-level reporting and independent evaluation. Average performance across multiple scientific fields should not conceal substantial weaknesses in individual disciplines.

7.2 Hallucinated Scientific Information and Citation Reliability

Large language models can generate text that appears scientifically plausible without adequate support from relevant publications. This difficulty becomes particularly significant when a model produces specific numerical findings, methodological explanations, or research conclusions that are not contained in retrieved documents. Query expansion can also introduce nonexistent or inappropriate technical expressions, directing retrieval toward unrelated information. Consequently, scientific retrieval systems must distinguish model-generated suggestions from verified source evidence and prevent unsupported statements from being presented as established findings.

Citation reliability represents an additional challenge because a generated reference may be incomplete, incorrectly attributed, or unrelated to the claim it accompanies. The presence of a valid-

looking citation does not establish that the corresponding publication supports the generated statement. Reliable evaluation therefore requires checking source identity and examining the relationship between each substantive claim and its supporting text. The proposed framework emphasizes claim-level verification, citation traceability, and suitable abstention when the available evidence is insufficient.

7.3 Computational Complexity, Evaluation Bias, and Model Transparency

Large language model-assisted retrieval systems may require substantially more computational resources than conventional lexical approaches. Query generation, dense encoding, candidate merging, and neural reranking introduce processing stages that can affect response latency and operational cost. Such requirements become important when scholarly search systems must process numerous queries or large document collections. Consequently, a method producing higher retrieval scores may not necessarily provide the most suitable operational solution. Computational efficiency should therefore be evaluated alongside ranking effectiveness.

Evaluation bias introduces further difficulties because scientific relevance judgments are inherently contextual and may vary among qualified reviewers. Incomplete annotation can cause relevant documents to be treated as unjudged or incorrectly assessed during metric calculation. Model selection using test queries can also produce optimistic performance estimates. The proposed methodology therefore emphasizes independent relevance assessment, transparent annotation criteria, and controlled model optimization. These procedures are necessary for distinguishing reliable retrieval improvements from artifacts introduced through dataset construction or evaluation decisions.

8. Conclusion

Large language models provide a methodological direction for improving scientific information retrieval through semantic query interpretation, document representation, query expansion, and relevance reranking. The present paper developed a hypothetical comparative framework involving BM25, scientific-domain dense retrieval, large language model-assisted query expansion, and hybrid retrieval with language model reranking. Its principal focus was the evaluation of domain-specific retrieval effectiveness and the distinction between document relevance and the reliability of generated scientific explanations.

The illustrative tables demonstrate possible procedures for presenting scientific document distributions and comparative ranking metrics. However, all numerical values were manually constructed and do not represent executed retrieval experiments or independently verified model performance. Establishing meaningful comparative findings would require authentic scientific corpora, representative research queries, independently assessed relevance judgments, implemented retrieval configurations, and suitable statistical analysis. Consequently, the hypothetical results cannot establish the superiority of language model-assisted retrieval over conventional lexical or neural retrieval systems.

Overall, the study identifies scientific terminology, domain generalization, evaluation bias, computational complexity, hallucinated information, and citation reliability as central considerations in scientific retrieval research. Future empirical investigations should examine these factors through reproducible experiments and transparent evidence-grounding assessments. Large language models may support scholarly information discovery when integrated with dependable retrieval procedures and appropriate verification safeguards. Nevertheless, scientific credibility must remain grounded in

accessible evidence and independently validated retrieval performance rather than the apparent fluency or sophistication of generated responses.

References

1. Chen, J., Lin, H., Han, X., & Sun, L. (2024). Benchmarking large language models in retrieval-augmented generation. Proceedings of the AAAI Conference on Artificial Intelligence, 38(16), 17754–17762. <https://doi.org/10.1609/aaai.v38i16.29728>
2. Es, S., James, J., Espinosa Anke, L., & Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (pp. 150–158). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.eacl-demo.16>
3. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, M., Wang, M., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. arXiv. <https://doi.org/10.48550/arXiv.2312.10997>
4. Jin, Q., Dhingra, B., Liu, Z., Cohen, W., & Lu, X. (2019). PubMedQA: A dataset for biomedical research question answering. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (pp. 2567–2577). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1259>
5. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-T. (2020). Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 6769–6781). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.550>
6. Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., & Petrov, S. (2019). Natural Questions: A benchmark for question answering research. Transactions of the Association for Computational Linguistics, 7, 453–466. https://doi.org/10.1162/tacl_a_00276
7. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (Vol. 33, pp. 9459–9474).
8. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press. <https://nlp.stanford.edu/IR-book/>
9. Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2023). MTEB: Massive text embedding benchmark. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (pp. 2014–2037). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.eacl-main.148>
10. Nogueira, R., & Cho, K. (2019). Passage re-ranking with BERT. arXiv. <https://doi.org/10.48550/arXiv.1901.04085>
11. 19. Journal of the American Medical Informatics Association, 27(9), 1431– [1] 1436. <https://doi.org/10.1093/jamia/ocaa091>

12. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (Vol. 1).
13. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. arXiv. <https://doi.org/10.48550/arXiv.2104.08663>
14. Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M. R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., Almirantis, Y., Pavlopoulos, J., Baskiotis, N., Gallinari, P., Artières, T., Ngomo, A. C. N., Heino, N., Gaussier, E., Barrio-Alvers, L., ... Paliouras, G. (2015). An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinformatics, 16, Article 138. <https://doi.org/10.1186/s12859-015-0564-6>
15. Voorhees, E. M., & Harman, D. K. (Eds.). (2005). TREC: Experiment and evaluation in information retrieval. MIT Press.
16. Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., & Hajishirzi, H. (2020). Fact or fiction: Verifying scientific claims. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 7534–7550). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.609>
17. Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Improving text embeddings with large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics.
18. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. Transactions on Machine Learning Research.
19. Zhu, K., Luo, Y., Xu, D., Yan, Y., Liu, Z., Yu, S., Wang, R., Wang, S., Li, Y., Zhang, N., Han, X., Liu, Z., & Sun, M. (2025). RAGEval: Scenario specific RAG evaluation dataset generation framework. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 8520–8544). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.acl-long.418>
20. Song, T., Gan, G., Shang, M., & Zhao, Y. (2025). IFIR: A comprehensive benchmark for evaluating instruction-following in expert-domain information retrieval. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 10186–10204). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.naacl-long.511>