RAGScholar & DBLP-QA: Explainable Retrieval Augmented Scientific QA on dblp with Source-Attributed Answers and a Benchmark Dataset

Aditya Neekhra[1], Markus Nilles[1] and Ralf Schenkel[1]

[1]Trier University, Germany

Abstract

Effective question answering (QA) over scientific literature is essential for accelerating research. However, existing benchmarks often fail to capture the complexity and domain-specific reasoning required. To address this gap, we introduce DBLP-QA, a benchmark dataset of 50 manually crafted question-answer pairs derived from the abstracts of scientific publications. It is designed to test both the retrieval and generation capabilities of large language models (LLMs) in scientific contexts. To demonstrate its utility, we evaluate RAGScholar, a retrieval-augmented generation (RAG) system that integrates external knowledge sources, and compare multiple retrieval and generation strategies. The dataset provides a robust foundation for advancing and objectively evaluating QA systems for scientific literature.

Keywords

Retrieval-Augmented Generation, Digital Libraries, Benchmark Question Answer Pair Dataset

1. Introduction

The rapid growth of scholarly publications has intensified information overload. Traditional academic search engines like Google Scholar or PubMed largely rely on retrieving individual publications, returning ranked lists rather than concise answers, and often do not capture semantic intent or synthesize evidence across papers. LLMs, on the other hand, are good at creating natural language answers, but, without grounding, can hallucinate and lack verifiable citations, limiting their suitability for scholarly QA. Retrieval-augmented generation addresses this by conditioning answers on retrieved evidence. However, domain-specific RAG for bibliographic corpora such as dblp remains underexplored, and there is no manually curated benchmark for reproducible evaluation over computer science publications. To address these gaps, we introduce DBLP-QA, a benchmark of 50 manually crafted question–answer pairs derived from scientific abstracts, and RAGScholar, a scholarly QA system that combines a Lucene BM25 retriever with instruction-tuned LLMs. RAGScholar serves as a testbed for evaluating DBLP-QA. We systematically evaluate retrieval performance, context-construction strategies, and model-specific differences under controlled conditions. We focus on four research questions:

RQ1: How effectively does BM25 retrieve a publication whose abstract contains the correct answer?

RQ2: How do different context strategies ranging from single abstracts to concatenated multi-abstract input affect answer quality?

RQ3: Does a two-step strategy, where concatenated answers from single abstracts serve as context, improve final answer quality compared to using multiple retrieved abstracts directly?

RQ4: How does model choice affect performance under identical retrieval and context conditions?

Our contributions are threefold: (1) DBLP-QA, a curated benchmark for abstract-based scholarly QA; (2) RAGScholar, a system integrating BM25 retrieval with instruction-tuned LLMs; and (3) a systematic evaluation of retrieval and context strategies across multiple models.

SCOLIA '26: Second International Workshop on Scholarly Information Access (SCOLIA), April 2, 2026, Delft, The Netherlands
s4adneek@uni-trier.de (A. Neekhra); nillesm@uni-trier.de (M. Nilles); schenkel@uni-trier.de (R. Schenkel)
0009-0003-3897-1259 (A. Neekhra); 0000-0002-3449-9319 (M. Nilles); 0000-0001-5379-5191 (R. Schenkel)

2. Related Work

Scientific question answering has been studied through a variety of benchmarks that differ in domain coverage and annotation depth. General-domain datasets such as SQuAD [1] helped establish large-scale QA evaluation but do not match the conceptual density of scientific writing.

Several datasets target QA over scholarly publications. Scientific QA datasets, including BioASQ [2], PubMedQA [3], SciQ [4], and CORD-19–based benchmarks like COVID-QA [5], have advanced domain-specific QA, yet many rely on automatically generated questions or focus narrowly on biomedical literature. QASPER [6] provides thousands of human-written questions grounded in full NLP papers, while QASA [7] focuses on expert-authored, high-level questions requiring deep reasoning across AI/ML articles. Other scientific QA resources, such as SciQA [8], SciDQA [9], or the automatically generated SciQAG [10] offer broader coverage but often include synthetic questions or multimodal elements that differ from abstract-focused QA. Evaluation frameworks like ScholarQABench [11] and SciArena [12] assess literature-grounded responses from large models, but are not designed around curated, abstract-level question-answer pairs.

In contrast to these large, heterogeneous, or full-paper datasets, DBLP-QA provides a compact, manually crafted benchmark grounded solely in scientific abstracts from computer science. This design enables controlled assessment of retrieval-augmented generation systems, such as RAGScholar, under well-defined and reproducible conditions.

3. DBLP-QA Benchmark Dataset

DBLP-QA is a novel benchmark dataset for scientific question answering that consists of 50 questions and corresponding answers derived from research articles in the computer science domain. To construct it, we used the dblp corpus dump from April 2025 [13] with 7.8 million publications and combined it with the abstracts obtained from the Semantic Scholar dataset API[1]. This resulted in a collection of 4.6 million publications with abstracts. To create the question-answer pairs, we randomly selected 50 among those publications and manually formulated a question that can be answered based on the abstract. We manually extracted and reformulated the answer from the abstract, ensuring that it is 1-3 sentences long and aligns with the original abstract. As an example, from the abstract stating *"Magnetic resonance (MR) tagging is a technique for measuring heart deformations through creation of a stripe grid pattern on cardiac images"*, we formulated the question *"What is Magnetic resonance tagging?"* and the answer *"Magnetic Resonance (MR) Tagging is a technique used to measure heart deformations by creating a stripe grid pattern on cardiac images."* The dataset is provided as a single CSV file[2] where each line provides question, answer, and dblp key and semantic scholar id of the publication from which they were derived.

4. RAGScholar & Experimental Setup

This section presents the architecture, workflow, and experimental setup of RAGScholar, a retrieval-augmented generation (RAG) system designed to answer questions in the domain of computer science. By integrating the system description with the experimental methodology, we provide a concise and coherent overview of the framework and its evaluation.

4.1. System Overview

RAGScholar consists of three main components: a retriever, a generator, and a web-based user interface (Figure 1). Users submit questions through a single-page application, upon which the retriever identifies

[1] https://www.semanticscholar.org/product/api; note that the abstracts dataset from the corpus of April 2025 is currently unavailable. The current abstracts dataset from January 20, 2026 does not include all the abstracts from the DBLP-QA dataset due to copyright licensing restrictions, but they are still accessible through the web interface.

[2] https://seafile.rlp.net/f/6581519cdd1d4782bccc/

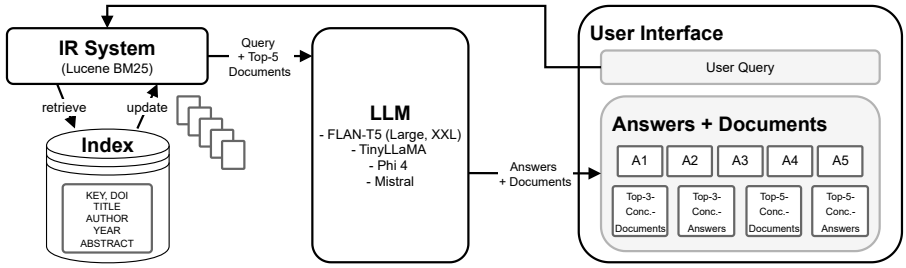

Figure 1: RAGScholar Architecture

relevant scientific publications from a pre-constructed Apache Lucene[3] index. The index is built from the April 2025 dblp dump and contains, for each publication, the dblp key, DOI, title, year, author names, and abstract. User queries are translated into Lucene search expressions over titles and abstracts, and the retrieved publications are ranked using BM25. The top-ranked abstracts are provided as context to the generator, a large language model that produces a natural-language answer. Depending on the configuration, the generator uses either the top abstract, multiple individual abstracts, or concatenated contexts derived from the top retrieved documents. The generated answer and the corresponding source documents are presented to the user.

4.2. Retrieval Evaluation (RQ1)

RQ1 evaluates whether BM25 reliably retrieves a publication containing the answer to a question of the DBLQ-QA benchmark dataset. We use BM25 with default parameters ($k_1 = 1.2$, $b = 2$) and evaluate performance via Recall@k and MRR@k.

4.3. Context Construction Strategies (RQ2 & RQ3)

RAGScholar employs nine context strategies, grouped into three categories: (1) Single-Document Contexts (5 variants): The LLM receives the abstract of each of the Top-5 (**A1**, …, **A5**) retrieved documents individually; (2) Concatenated Document Contexts (2 variants): **Top-3 Concatenated Documents** and **Top-5 Concatenated Documents**, formed by concatenating the top-ranked abstracts; and (3) Concatenated Answer Contexts (2 variants): **Top-3 Concatenated Answers** and **Top-5 Concatenated Answers**, where intermediate answers for the Top-3 or Top-5 documents are concatenated and used as context for generating a final answer.

RQ2 compares strategies from the first and the second category to assess the effect of context size and document aggregation. RQ3 compares strategies from the second and the third category to study document- vs. answer-level aggregation.

4.4. Model Comparison (RQ4)

To evaluate the influence of model scale and architecture, five instruction-tuned causal and sequence-to-sequence LLMs (0.78B–14.7B parameters) were tested under identical retrieval and context conditions. All models used consistent generation settings (temperature 0.7, top-$p = 0.9$, maximum output lengths of 512–1024 tokens). Automatic precision and device selection were applied when supported. Table 1 shows the models evaluated and the parameters used.

4.5. Evaluation Procedure

For all 50 benchmark questions, we generated 2,500 answers (5 models × 10 context variants, including a no-context baseline). Answers were manually rated against the ground-truth answer on a three-point

[3] https://lucene.apache.org/

Table 1
LLM Configurations

Model	Family	Params	Precision	Temp.	Top-p	Max Output	Device
Mistral-7B-Instruct-v0.1	causal/chat	7.3B	auto	0.7	0.9	512	auto
FLAN-T5-XXL	seq2seq	11.0B	auto	0.7	0.9	512	auto
Phi-4	causal/chat	14.7B	auto	-	-	-	auto
TinyLlama-1.1B-Chat	causal/chat	1.1B	fp32	0.7	0.9	512	CPU
FLAN-T5-Large	seq2seq	0.78B	fp32	0.2	0.9	1024	CPU

Table 2
Example of the manual scores given to generated answers for a question-answer pair from DBLP-QA. **Question:** What is Compressive Sensing? **Ground-Truth Answer:** Compressive Sensing (CS) is an advanced signal processing technique that enables the reconstruction of a signal using far fewer measurements than required by the traditional Nyquist-Shannon sampling theorem.

Example (Score 2)	Example (Score 1)	Example (Score 0)
Generated Answer: Compressive sensing is a signal processing technique that allows for the reconstruction of signals or images from a small number of measurements, significantly fewer than what is typically required by the Nyquist-Shannon sampling theorem. **Explanation:** Answered in the same way as Ground Truth Answer.	**Generated Answer**: Compressive Sensing, also known as Compressed Sensing (CS), is a technique used in signal processing and imaging that allows for the detection and recovery of signals that are inherently sparse or low-rank, such as images or signals with noise. **Explanation:** Related to the Ground Truth Answer.	**Generated Answer:** Compressive Sensing is a method used to improve on what any nonadaptive method can achieve in the context of recursive bisection method. It is a technique that establishes a non-asymptotic lower bound that applies to all methods, regardless of their computational complexity. **Explanation:** Unrelated to Ground Truth Answer.

Scale: 2 = correct and complete, 1 = correct but incomplete, 0 = incorrect or irrelevant. Table 2 contains an example question-answer pair with generated answers and their manual scores. In addition, all answers were evaluated against the ground-truth answers using ROUGE-L [14] and BERTScore [15].

5. Results

5.1. Retrieval Performance (RQ1)

The retriever ranked a document containing a correct answer at rank 1 for 44 out of the 50 benchmark questions. In four questions, the document at rank 2 was the first document in the ranking to contain a correct answer, and in two questions, the first document containing an answer was ranked third. Consequently, the retriever achieved MRR@3 of 0.93 and Recall@3 of 1. We also examined the ranking position of the document from which the question-answer pair was derived. For 32 questions, the document was ranked at the top, for seven questions, it was ranked at two, for two questions, it was ranked at three, and for one question, it was ranked at four. The document was not in the top 5 for five questions.

The following example shows a case where the top-ranking document was not relevant:

Question: What is CtRL-Sim?
Ground Truth Answer: CtRL-Sim is a method that leverages return-conditioned offline reinforcement learning to efficiently generate reactive and controllable traffic agents.
Abstract of retrieved document at rank 1: …calibration transformer with line-classification (CtRL-C) …demonstrate that CtRL-C outperforms …
Comment: The retriever was misled by high-frequency mentions of the query keyword `ctrl`.

Table 3
Experimental results for the context variants: Single-Document (A1-A5), Concatenated Documents (Top-3-CD, Top-5-CD), and Concatenated Answers (Top-3-CA, Top-5-CA).

Metric	Model	A1	A2	A3	A4	A5	Top-3-CD	Top-5-CD	Top-3-CA	Top-5-CA	Question(no context)
Manual	Mistral-7B	1.72	1.06	1.00	0.84	0.86	1.63	1.74	1.60	1.49	0.80
	Phi-4	1.56	0.78	0.68	0.58	0.59	1.66	1.66	1.63	1.51	0.40
	TinyLlama-1.1B	1.63	1.10	0.94	0.82	0.70	1.58	1.52	1.44	1.34	1.10
	FLAN-T5-Large	1.34	0.70	0.70	0.54	0.56	1.32	1.34	1.36	1.34	0.30
	FLAN-T5-XXL	0.98	0.24	0.16	0.06	0.014	1.04	0.92	0.96	0.90	0.60
BERTScore F1	Mistral-7B	0.58	0.51	0.47	0.44	0.39	0.51	0.54	0.51	0.47	0.35
	Phi-4	0.37	0.38	0.34	0.30	0.27	0.31	0.36	0.36	0.31	0.28
	TinyLlama-1.1B	0.45	0.44	0.41	0.38	0.36	0.42	0.41	0.42	0.39	0.41
	FLAN-T5-Large	0.47	0.46	0.38	0.38	0.38	0.45	0.47	0.48	0.45	0.14
	FLAN-T5-XXL	0.40	0.38	0.32	0.28	0.26	0.42	0.38	0.38	0.35	0.21
ROUGE-L F1	Mistral-7B	0.32	0.30	0.25	0.24	0.24	0.34	0.34	0.32	0.29	0.21
	Phi-4	0.28	0.27	0.21	0.20	0.15	0.37	0.33	0.22	0.23	0.10
	TinyLlama-1.1B	0.35	0.24	0.23	0.22	0.24	0.28	0.23	0.28	0.30	0.26
	FLAN-T5-Large	0.27	0.20	0.18	0.19	0.15	0.27	0.26	0.24	0.24	0.08
	FLAN-T5-XXL	0.20	0.14	0.12	0.13	0.10	0.21	0.20	0.20	0.13	0.14

5.2. Context Selection Strategies (RQ2)

Context selection is a central design choice in retrieval-augmented QA because it determines the evidence available to the model at generation time. We therefore evaluate Top-k context concatenation for $k \in \{3, 5\}$, where abstracts of the k top-ranked retrieved papers are merged into a single composite context to increase coverage and reduce omissions due to ranking errors. The results in Table 3 show that increasing k can improve human-perceived answer quality, but the gains are model-dependent: for Mistral-7B, Top-5 context concatenation (Top-5-CD) achieves the best manual score (1.74) compared to Top-3-CD (1.63), while Phi-4 remains stable at 1.66 for both Top-3-CD and Top-5-CD. In contrast, TinyLlama-1.1B slightly drops from 1.58 (Top-3-CD) to 1.52 (Top-5-CD), suggesting that additional context can introduce redundancy or noise beyond what smaller models can effectively use. Across models, however, aggregated evidence is consistently beneficial relative to no context (e.g., Mistral-7B: 1.74 vs. 0.80; Phi-4: 1.66 vs. 0.40), highlighting the overall importance of context selection in balancing coverage, noise accumulation, and explainability in scholarly QA. For single-document contexts, A1 achieves the best results compared to the other single-document variants. For TinyLlama, this variant attains the highest overall score, and for the other models, the manual scores are only slightly lower than those of the respective best-performing variant. Moreover, a slight correlation can be observed between the manual evaluations and the automatic metrics BERTScore and ROUGE. This suggests that these metrics could potentially be used for the automatic evaluation of generated answers for DBLP-QA, although further research is required.

5.3. Answer Concatenation (RQ3)

Answer concatenation leverages multiple retrieved documents. The model generates one answer per document and then combines the k candidates for Top-k ($k \in \{3, 5\}$). However, concatenating abstracts *before* generation is consistently stronger in our results. In Table 3, Top-5 concatenated documents (Top-5-CD) outperforms Top-5 concatenated answers (Top-5-CA) for the best models (Mistral-7B: 1.74 vs. 1.49; Phi-4: 1.66 vs. 1.51; TinyLlama-1.1B: 1.52 vs. 1.34), suggesting that direct exposure to aggregated evidence yields more complete answers. Still, answer concatenation improves over no context (e.g., Mistral-7B: 1.49 vs. 0.80; Phi-4: 1.51 vs. 0.40), indicating that output-level fusion remains useful. Interestingly, TinyLlama attains higher ROUGE-L with Top-5-CA than Top-5-CD (0.30 vs. 0.23), implying that answer fusion can increase surface overlap even when human judgments favor context concatenation.

5.4. LLM Comparison (RQ4)

All models were evaluated under identical conditions, so differences can be attributed primarily to the LLMs. Table 3 shows clear separation: Mistral-7B performs best (Top-5-CD: 1.74), followed by Phi-4 (up to 1.66) and TinyLlama-1.1B (up to 1.58), while FLAN-T5 variants are much lower (0.38-0.61 under Top-5-CD). Retrieved evidence consistently improves quality over no context (e.g., Mistral-7B: 0.80→1.74; Phi-4: 0.40→1.66), confirming retrieval conditioning as a key driver. Context size shows a model-dependent coverage-noise trade-off: Mistral benefits from Top-5 vs. Top-3 (1.74 vs. 1.63), Phi-4 is stable (1.66 for both), and TinyLlama slightly drops with Top-5 (1.58→1.52), indicating higher sensitivity to redundancy. Overall, strong performance is achievable with efficient mid-sized models in a reproducible pipeline, supporting scalable scholarly QA deployments.

6. Discussion

Limitations Despite the effectiveness of the framework, limitations remain. The system uses abstracts rather than full texts, limiting fine-grained methodological questions. Retrieval is BM25-only, which may miss semantically relevant evidence under paraphrasing without dense/hybrid methods. We also evaluate only small and medium-sized LLMs due to compute constraints, although results are already strong, highlighting the impact of retrieval and evidence packaging. Finally, DBLP-QA contains only 50 items from one domain, limiting generalizability.

Future Work Several directions can extend this work. First, re-running the pipeline with larger instruction-tuned, long-context models (e.g., 34B/70B) would allow to estimate the performance ceiling. Second, retrieval could be strengthened via dense/hybrid methods by adding DPR [16] or ColBERT [17] alongside BM25 [18], using fusion (e.g., RRF/score fusion), and lightweight reranking (e.g., MonoT5 [19]); FiD-style conditioning can further probe multi-document limits [20]. Beyond retrieval, sentence/span-level evidence and diversification (MMR [21]) can improve grounding and reduce redundancy, while adaptive routing can enable low-latency single-document inference. Finally, expanding DBLP-QA beyond 50 items, diversifying question types, and collecting multi-rater judgments would improve reliability and generalizability across domains.

7. Conclusion

In this work, we introduced DBLP-QA, a benchmark of 50 QA pairs derived from abstracts, and RAGScholar, a retrieval-augmented QA system integrating BM25 retrieval and generative LLMs. The evaluation provides several insights in the design of domain-specific RAG systems for bibliographic corpora.

BM25 retrieval is effective in our scholarly QA setting, achieving a Recall@1 of 0.88 and a perfect Recall@3. Context construction is a crucial factor in determining the answer quality. Providing the abstract of the top retrieved document as context to the LLM results in better answers compared to the baseline scenario without context. Concatenating abstracts of the top-3 retrieved documents can improve the quality of the answer, while concatenating five abstracts may lead to reduced quality due to noise. A two-step context setting that first generates answers per retrieved document and then concatenates them as context does not improve answer quality compared to directly concatenating abstracts. Mid-sized LLMs perform well, while smaller models are more sensitive to context size and noise, indicating that effective scholarly QA depends more on retrieval and context construction than on large models.

DBLP-QA and RAGScholar offer a reproducible evaluation setup for studying retrieval-augmented methods over scholarly corpora, with future work on extending the benchmark dataset, improving retrieval and exploring automatic evaluation metrics.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100, 000+ questions for machine comprehension of text, in: J. Su, X. Carreras, K. Duh (Eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016, The Association for Computational Linguistics, 2016, pp. 2383–2392. URL: https://doi.org/10.18653/v1/d16-1264. doi:10.18653/V1/D16-1264.

[2] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, Y. Almirantis, J. Pavlopoulos, N. Baskiotis, P. Gallinari, T. Artières, A. N. Ngomo, N. Heino, É. Gaussier, L. Barrio-Alvers, M. Schroeder, I. Androutsopoulos, G. Paliouras, An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition, BMC Bioinform. 16 (2015) 138:1–138:28. URL: https://doi.org/10.1186/s12859-015-0564-6. doi:10.1186/S12859-015-0564-6.

[3] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, X. Lu, PubMedQA: A dataset for biomedical research question answering, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 2567–2577. URL: https://doi.org/10.18653/v1/D19-1259. doi:10.18653/V1/D19-1259.

[4] J. Welbl, N. F. Liu, M. Gardner, Crowdsourcing multiple choice science questions, in: L. Derczynski, W. Xu, A. Ritter, T. Baldwin (Eds.), Proceedings of the 3rd Workshop on Noisy User-generated Text, NUT@EMNLP 2017, Copenhagen, Denmark, September 7, 2017, Association for Computational Linguistics, 2017, pp. 94–106. URL: https://doi.org/10.18653/v1/w17-4413. doi:10.18653/V1/W17-4413.

[5] T. Möller, A. Reina, R. Jayakumar, M. Pietsch, COVID-QA: A question answering dataset for COVID-19, in: K. Verspoor, K. B. Cohen, M. Dredze, E. Ferrara, J. May, R. Munro, C. Paris, B. Wallace (Eds.), Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, Association for Computational Linguistics, Online, 2020. URL: https://aclanthology.org/2020.nlpcovid19-acl.18/.

[6] P. Dasigi, K. Lo, I. Beltagy, A. Cohan, N. A. Smith, M. Gardner, A dataset of information-seeking questions and answers anchored in research papers, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, Association for Computational Linguistics, 2021, pp. 4599–4610. URL: https://doi.org/10.18653/v1/2021.naacl-main.365. doi:10.18653/V1/2021.NAACL-MAIN.365.

[7] Y. Lee, K. Lee, S. Park, D. Hwang, J. Kim, H. Lee, M. Lee, QASA: advanced question answering on scientific articles, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 19036–19052. URL: https://proceedings.mlr.press/v202/lee23n.html.

[8] J. Lehmann, A. Meloni, E. Motta, F. Osborne, D. R. Recupero, A. A. Salatino, S. Vahdati, Large language models for scientific question answering: An extensive analysis of the SciQA benchmark, in: A. Meroño-Peñuela, A. Dimou, R. Troncy, O. Hartig, M. Acosta, M. Alam, H. Paulheim, P. Lisena (Eds.), The Semantic Web - 21st International Conference, ESWC 2024, Hersonissos, Crete, Greece, May 26-30, 2024, Proceedings, Part I, volume 14664 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 199–217. URL: https://doi.org/10.1007/978-3-031-60626-7_11. doi:10.1007/978-3-031-60626-7_11.

[9] S. Singh, N. Sarkar, A. Cohan, SciDQA: A deep reading comprehension dataset over scientific papers, in: Y. Al-Onaizan, M. Bansal, Y. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024, Association for Computational Linguistics, 2024, pp. 20908–20923. URL: https://doi.org/10.18653/v1/2024.emnlp-main.1163. doi:10.18653/V1/2024.EMNLP-MAIN.1163.

[10] Y. Wan, Y. Liu, A. Ajith, C. Grazian, B. Hoex, W. Zhang, C. Kit, T. Xie, I. Foster, SciQAG: A framework for auto-generated science question answering dataset with fine-grained evaluation, CoRR abs/2405.09939 (2024). URL: https://doi.org/10.48550/arXiv.2405.09939. doi:10.48550/ARXIV.2405.09939. arXiv:2405.09939.

[11] A. Asai, J. He, R. Shao, W. Shi, A. Singh, J. C. Chang, K. Lo, L. Soldaini, S. Feldman, M. D'Arcy, D. Wadden, M. Latzke, M. Tian, P. Ji, S. Liu, H. Tong, B. Wu, Y. Xiong, L. Zettlemoyer, G. Neubig, D. S. Weld, D. Downey, W. Yih, P. W. Koh, H. Hajishirzi, OpenScholar: Synthesizing scientific literature with retrieval-augmented LMs, CoRR abs/2411.14199 (2024). URL: https://doi.org/10.48550/arXiv.2411.14199. doi:10.48550/ARXIV.2411.14199. arXiv:2411.14199.

[12] Y. Zhao, K. Zhang, T. Hu, S. Wu, R. L. Bras, Y. Liu, X. Tang, J. C. Chang, J. Dodge, J. Bragg, C. Zhao, H. Hajishirzi, D. Downey, A. Cohan, SciArena: An open evaluation platform for non-verifiable scientific literature-grounded tasks, in: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2025. URL: https://openreview.net/forum?id=am6RR85mnc.

[13] dblp Team, dblp computer science bibliography – Monthly Snapshot XML Release of April 2025, 2025. URL: https://doi.org/10.4230/dblp.xml.2025-04-01. doi:10.4230/dblp.xml.2025-04-01.

[14] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[15] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[16] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W. Yih, Dense passage retrieval for open-domain question answering, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, Association for Computational Linguistics, 2020, pp. 6769–6781. URL: https://doi.org/10.18653/v1/2020.emnlp-main.550. doi:10.18653/V1/2020.EMNLP-MAIN.550.

[17] O. Khattab, M. Zaharia, ColBERT: Efficient and effective passage search via contextualized late interaction over BERT, in: J. X. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, Y. Liu (Eds.), Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020, ACM, 2020, pp. 39–48. URL: https://doi.org/10.1145/3397271.3401075. doi:10.1145/3397271.3401075.

[18] S. E. Robertson, H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, Found. Trends Inf. Retr. 3 (2009) 333–389. URL: https://doi.org/10.1561/1500000019. doi:10.1561/1500000019.

[19] R. Nogueira, Z. Jiang, R. Pradeep, J. Lin, Document ranking with a pretrained sequence-to-sequence model, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020, volume EMNLP 2020 of *Findings of ACL*, Association for Computational Linguistics, 2020, pp. 708–718. URL: https://doi.org/10.18653/v1/2020.findings-emnlp.63. doi:10.18653/V1/2020.FINDINGS-EMNLP.63.

[20] G. Izacard, E. Grave, Leveraging passage retrieval with generative models for open domain question answering, in: P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021, Association for Computational Linguistics, 2021, pp. 874–880. URL: https://doi.org/10.18653/v1/2021.eacl-main.74. doi:10.18653/V1/2021.EACL-MAIN.74.

[21] J. G. Carbonell, J. Goldstein, The use of MMR, diversity-based reranking for reordering documents

and producing summaries, SIGIR Forum 51 (2017) 209–210. URL: https://doi.org/10.1145/3130348.3130369. doi:10.1145/3130348.3130369.