SciClaimSeekers at CheckThat! 2026: Retrieving Scientific Sources for Social Media Claims with LLM Reranking

CheckThat! Lab at CLEF 2026

Mohotarema Rashid[1,*], Nansu Baniya[2], Anirban Saha Anik[2], Xiaoying Song[1] and Lingzi Hong[3]

[1]Department of Information Science, University of North Texas, Denton, TX, United States
[2]Department of Computer Science and Engineering, University of North Texas, Denton, TX, United States
[3]Department of Data Science, University of North Texas, Denton, TX, United States

Abstract

Scientific claims often spread on social media faster than they can be verified, while posts rarely link to the original scholarly sources. To tackle this problem this paper presents system called SciClaimSeekers, a retrieval and reranking framework by combining BM25 and zero-shot multilingual E5 retrieval with Reciprocal Rank Fusion (k=60), followed by Qwen2.5-14B-Instruct pointwise reranking. The pipeline reaches 64.36% MRR@5 on the English development set a 13.67-point jump over BM25 and 10.17 points over the unranked hybrid and 64.39% on the official test set, in the CLEF-2026 CheckThat! Task 1 evaluation. Our experiment suggests that large pre-trained models, when combined into a careful pipeline, can be competitive with fine-tuned approaches on this task.

Keywords

Large Language Model, Retrieval, Social Media, CLEF

1. Introduction

Since online misinformation spreads rapidly, tracing social media claims back to their original scientific sources is crucial for automated fact checking and verification based on evidence [1, 2]. Furthermore, users in social media tend to favor information that aligns with their preexisting beliefs and this confirmation bias, combined with the rapid dissemination of scientific information, facilitates the widespread acceptance and propagation of unverified claims as factual information [3, 4]. Because the content of social media is solely user-generated and user-generated posts often paraphrase or loosely reference findings without standardized citations, making reliable identification of the original scientific sources difficult [5]. Furthermore, this task poses challenges due to the linguistic and structural mismatch between informal social media jargon and formal scientific literature [6].

To address this challenge our work proposed a combined reranking foundations integrating hybrid sparse dense retrieval with Large Language Model (LLM); Fig. 1 illustrates our three-stage retrieve-then-rerank architecture: two complementary first-stage retrievers operate in parallel on the indexed corpus, their ranked candidate lists are merged with Reciprocal Rank Fusion, and the fused candidate set is reranked by a LLM used as a pointwise cross-encoder.

In the above context, our contributions are as follows:

1. We present a hybrid retrieve-then-rerank pipeline for the multilingual CheckThat! 2026 setting [7, 8]. Our submitted system ranks 11th out of 37 participating teams on the English test set for Task 1, Source Retrieval for Scientific Web Claims [9], achieving 64.39% MRR@5. The negligible 0.03

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
MohotaremaRashid@my.unt.edu (M. Rashid); NansuBaniya@my.unt.edu (N. Baniya); AnirbanSahaAnik@my.unt.edu (A. S. Anik); XiaoyingSong@my.unt.edu (X. Song); Lingzi.Hong@unt.edu (L. Hong)
0009-0003-2683-7191 (M. Rashid); 0000-0001-7116-9338 (N. Baniya); 0000-0002-7824-3702 (A. S. Anik); 0000-0001-9390-1155 (X. Song); 0000-0001-8412-8180 (L. Hong)

percentage-point gap from the development score of 64.36% suggests that our tuning decisions transfer cleanly to held-out test data.

2. We isolate the contribution of each pipeline stage and identify the LLM cross-encoder as the dominant source of gain approximately +10 MRR@5 points whether applied on top of BM25 or the hybrid first stage accounting for roughly three-quarters of the +13.67-point overall improvement over the BM25 baseline.

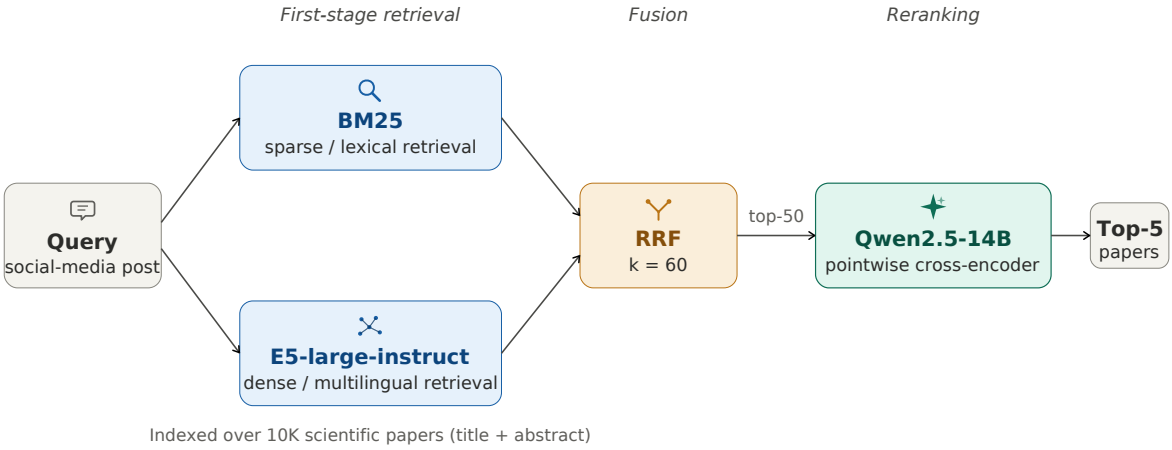

Figure 1: SCiClaimSeekers System Architecture

2. Related Work

Our study is situated in intersection four major research streams I) automated fact checking II) scientific claim verification III) social media based scientific discourse retrieval and IV) LLM based reranking systems.

Early factchecking pipeline was shaped by FEVER benchmark [10] while SciFact extended the framework dedicated to scientific claims [11]. Later in this research stream ARSJoint, and RerrFact improved retrieval and verdict using transformer-based architectures [12, 13].

Though the rise of social media as a channel for scientific communication has created challenges for retrieving scholarly sources from implicit and noisy posts, early work in scientific claim verification rarely focused on noisy and implicit social media discourse. To shed light on this, SciTweets was developed for detecting scientific discourse from social media posts [14]. Building on this line of work, Hafid et al. [15] formalized the task of disambiguating implicit scientific references on X, retrieving the original publications that social media posts implicitly refer to, and contributed the first ground-truth corpus and baselines for the task. Most prior systems adopted a retrieve-then-rerank architecture by combining sparse and dense retrieval [16, 17, 18, 19].

Traditional retrieval system like BM25 which based on lexical matching remains a strong classical sparse retrieval baseline due to its probabilistic ranking formulation [20]. In addition transformer-based dense retrievers such as DPR improved semantic matching for retrieval tasks [21]. Though both of the methods can combined into hybrid retrieval has become dominant as sparse retrievers effectively capture lexical signals such as named entities and numerical values, whereas dense retrievers better model semantic similarity and paraphrastic variation [22].

In recent, reranking approaches increasingly use LLM model as cross encoders. Though early transformer based rankers such as monoBERT utilized reranking as relevance classification task [23], later studies use list wise strategy to improve ranking performance [12].

Recently LLM based closed rerankers such as RankGpt showed strong performance through instruction tuned model [24]. However, open source alternatives also improved the retrieval task while

maintaining competitive effectiveness [25].

In sum, prior studies showed substantial progress in scientific claim verification and retrieval, however challenges remain in linking noisy social media discourse with scholarly evidence. Our work addresses this gap by combining reranking foundations integrating hybrid sparse dense retrieval with LLM-based pointwise reranking for scientific source attribution in social-media environments. Because Pointwise remains promising due to its efficiency, scalability and lower sensitivity to input ordering bias [26].

3. Methods

3.1. Pipeline overview

We formulate the retrieval problem as follows; Given a social media post q written in English the system must return a ranked list of five publications extracted from a fixed candidate pool D of $|D| = 10{,}000$ scientific papers. Each candidate paper is represented by key, a title, an abstract and metadata fields like authors, venue, publication year.

Our pipeline decomposes the task into three computational sequential Phases.

In Phase 1, a sparse retriever based on BM25 [20] and a dense retriever based on multilingual-e5-large-instruct [27] operate in parallel and each independently returns its top-100 candidates per query.

In Phase 2, the two returned lists were merged using Reciprocal Rank Fusion [28] that returned an unified top-50 candidates set against each query.

In phase 3, each of the 50 candidates were scored by Qwen2.5-14B-Instruct [29]. The scoring set up was conducted in a zero shot pointwise cross-encoder and the reranked settings produced top 5 candidates as final predictions.

3.2. Data representation and indexing

We represent each candidate paper concatenated by title and abstract, we ignored meta data field as it could introduce noise to the pipeline followed by prior literature [2]. We indexed full corpus once at preprocessing time for both set of retrievals discussed later in this section.

3.3. Sparse Retrieval

We use Okapi BM25 [20] with its standard parameter to work on retrieval that is based on lexical overlap between social media claim and title and abstract of the candidates. Though we note that lexical matching is insufficient this type of task since social media post contains online jargon and most of the cases the post paraphrase the scientific languages, yet BM25 remains a robust first-stage retriever for queries that include named entities, numerical values, and domain-specific terminology, which keeps the pipeline easy to reproduce [30]. To make this phase effective we preprocessed data by lowering the alphabetic characters, removing mentions and URL, normalizing hashtag, removing of punctuation, tokenizing of subwords, removing stopwords [2, 31].

3.4. Dense Retrieval

For dense retrieval we use `intfloat/multilingual-e5-large-instruct` retriever [27] encoder which is based on built on XLM-RoBERTa. By following this model's convention, we differentiated queries from documents via instruction prefix, each post was tagged `"Instruct: Given a social media post that paraphrases a scientific finding, retrieve the scientific paper it refers to. \nQuery: "`. In this setting we tokenized the input into 512 tokens and mean pooled over the final hidden states and we normalize L2 so that cosine similarity could be reduced to inner product [32]. We match query embeddings against FAISS index that returns top-100 documents per query.

3.5. Reciprocal Rank Fusion

In this stage we merge the two top-100 candidate lists using Reciprocal Rank Fusion [28], which assigns scores to each document solely by its rank position in each input ranking.

$$\mathrm{RRF}(d) = \sum_{r \in R} \frac{1}{k + \mathrm{rank}_r(d)} \tag{1}$$

where R is the set of retrievers (BM25 and dense), $\mathrm{rank}_r(d)$ is the rank of d in retriever r's list (∞ if absent), and k is a smoothing constant. We use $k = 60$ recommended by prior research [28]. We chose this depth based on the findings from the development set (discussed later in result section).

3.6. LLM based point-wise reranking

In this stage we use Qwen2.5-14B-Instruct [29] as a setting of zero-shot pointwise cross-encode. This LLM is open-source serving, and 32K-token context window. We score each candidate independently rather than listwise [24].

For each (post, candidate) pair, we build a prompt instruct. The system message tells the model to act as a scientific source identifier and return an integer between 0 and 10, and the user message provides the post along with the candidate's title and abstract (Table 1). The relevance score were set to 0 to 10, then the Candidates are sorted by score, with ties broken using the Stage 2 RRF score, and the top 5 are returned.

Table 1
Prompt for LLM to rerank pipeline.

SYSTEM: You are an expert scientific source identifier. You will be given a social media post that paraphrases a scientific finding, and a candidate scientific paper. Your job is to judge how likely the post is referring to this specific paper. Output ONLY a single integer score from 0 to 10 (no explanation, no other text). 10 = the post is almost certainly about this paper. 0 = the paper is unrelated.
USER: Social media post: {post} Candidate paper: Title: {title} Abstract: {abstract} Score (0-10):

3.7. Implementation Details

We ran all experiments on a single node with two NVIDIA A40 GPUs (46 GB each) and 128 GB of memory. Models are loaded via Hugging Face transformers, dense encoding uses sentence-transformers, and BM25 runs on a pre-tokenized corpus. We serve the LLM reranker with vLLM for high-throughput inference. End-to-end inference on the English test set (303,800 prompts) takes roughly 11 hours.

4. Result and Analysis

We follow the official CLEF 2026 CheckThat! Lab Task 1 protocol and use **Mean Reciprocal Rank (MRR)** as our primary evaluation metric:

$$\mathrm{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\mathrm{rank}_i} \tag{2}$$

where $|Q|$ is the number of queries in the evaluation set, and rank_i is the rank position of the gold-standard paper for the i-th post within the top-5 predictions (taken as 0 if the gold paper is not

Table 2
English dev and test set results. All scores against the gold pubkey labels; best score per column in **bold**.

#	System	Stage	Dev ($N = 3{,}905$)			Test ($N = 6{,}076$)		
			MRR@5	Hit@1	Hit@5	MRR@5	Hit@1	Hit@5
1	BM25	Lexical	0.5069	0.4584	0.5828	0.5021	0.4493	0.5859
2	E5-large-instruct (zero-shot)	Dense	0.5265	0.4638	0.6292	0.5164	0.4536	0.6177
3	BM25 $\oplus$ E5 (RRF, $k = 60$)	Hybrid	0.5419	0.4761	0.6487	0.5577	0.4924	0.6587
4	Hybrid + Cross-encoder rerank (bge-reranker-v2-m3, top-30)	Hybrid + CE	0.5512	0.4912	0.6423	0.5422	0.4724	0.6513
5	BM25 + Qwen2.5-14B rerank	Lexical + LLM	0.6073	0.5649	0.6676	0.6062	0.5566	0.6746
6	Hybrid + Qwen2.5-14B rerank (submitted)	Hybrid + LLM	**0.6436**	**0.5836**	**0.7347**	**0.6439**	**0.5797**	**0.7355**

retrieved within the top 5). MRR is well suited to this task, as it rewards systems that rank correct evidence near the top of the list.

We evaluate all systems on the English development split (3,905 posts) and the official English test split (6,076 posts). In addition to MRR@5; we report Hit@1 and Hit@5 to characterize where in the ranking the correct paper is placed. Table 2 summarizes the contribution of each component of the pipeline.

Our result analysis showed by using lexical retrieval methods like BM25 both dev set and test set achieves MRR@5 of 50%. The near-identical scores across splits indicate that the corpus and query distributions are well matched, and establish a stable reference point against which the subsequent retrieval stages are evaluated. On the other hand semantic retrieval like the multilingual encoder `intfloat/multilingual-e5-large-instruct` in a zero-shot configuration yields an absolute improvement of 1.96 percentage points in MRR@5 over the BM25 baseline, reaching 52.65% on the development set, with a substantially larger 4.64-point gain in Hit@5 (58.28% to 62.92%). This confirms that the instruction-tuned dense encoder captures the implicit semantic links between social-media claims and scientific abstracts more reliably than surface term matching. Combining the BM25 and zero-shot dense rankings with Reciprocal Rank Fusion provides a further +1.54 percentage points in MRR@5 over the dense leg alone (54.19% vs. 52.65%) and +1.95 in Hit@5, confirming that the two signals are complementary.

In subsequent stage we compare two re-rankers operating on the hybrid fused candidates: a cross-encoder over the top-30, and a generative LLM re-ranker over the top-50. The cross-encoder `BAAI/bge-reranker-v2-m3` contributes a modest +0.93-point gain in MRR@5 over the hybrid first stage (55.12% vs. 54.19%), with most of the benefit concentrated at the top of the ranking (Hit@1 improves from 47.61% to 49.12%). However, this gain reverses on the test split, where the cross-encoder loses 1.55 points relative to the hybrid baseline (54.22% vs. 55.77%)—indicating that cross-encoder reranking overfits to the development distribution. The 14B-parameter generative re-ranker Qwen2.5-14B-Instruct delivers a substantially larger improvement: on the development set, applying it on top of the hybrid candidates raises MRR@5 by +10.17 percentage points (54.19% to 64.36%) and Hit@5 by +8.60 points (64.87% to 73.47%). Applied directly to BM25 candidates, the same re-ranker yields an MRR@5 of 60.73%, indicating that approximately three-quarters of its benefit is independent of the first-stage retriever.

Our final pipeline with LLM re-ranking over the hybrid top-50 achieves an MRR@5 of 64.36% on the development set and 64.39% on the test set, with corresponding Hit@1 of 58.36% and Hit@5 of 73.47% and the test (57.97% and 73.55%) respectively. The minimal 0.03-point gap between development and test MRR@5 confirms that the tuning decisions transfer cleanly to held-out data; the LLM re-ranker is the dominant contributor to the +13.67-point overall improvement in MRR@5 over the BM25 baseline.

5. Discussion & Conclusion

Two findings strengthen our empirical contributions to literature.

The LLM reranker drives the gain. Qwen2.5-14B contributes +10 MRR@5 points whether applied to BM25 or hybrid candidates roughly three-quarters of the +13.67-point overall lift over BM25. Pointwise scoring with a strong open-weight LLM is a reliable substitute for training a task-specific

reranker [24].

Hybrid retrieval compounds with reranking. RRF fusion yields a modest +1.54-point first-stage gain but a larger +3.63-point lift after LLM reranking (Hybrid+LLM 64.36% vs. BM25+LLM 60.73%): better first-stage recall translates directly into stronger downstream rankings [33, 19].

Our full pipeline is open-source, modular, and runs end-to-end on a single two-GPU node with vLLM serving make our pipeline practical for privacy-sensitive or offline deployment.

6. Limitations & Future Work

Our set up was evaluated on the English split only and represent each paper by title and abstract, as a result the German and French test splits and metadata-aware representations are left for future work. The Qwen2.5-14B reranker is also the runtime bottleneck (∼11 hours per 6,076 test queries on two A40 GPUs). Future directions will include (i) Retrieval Augmented Generation (RAG) along with the metadata feature feeding into the LLM for additional cues for guided ranking (ii) distilling the LLM reranker into a smaller cross-encoder or adopting listwise inference to reduce per-query cost.

Declaration on Generative AI

During the creation of this work, the authors used Claude to refine the pre-written text and to create Figure 1. After using the tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. Muhammed T, S. K. Mathew, The disaster of misinformation: A review of research in social media, International Journal of Data Science and Analytics 13 (2022) 271–285. doi:10.1007/s41060-022-00311-6.

[2] P. J. Sager, A. Kamaraj, B. F. Grewe, T. Stadelmann, Deep retrieval at CheckThat! 2025: Identifying scientific papers from implicit social media mentions via hybrid retrieval and re-ranking, 2025. doi:10.48550/arXiv.2505.23250. arXiv:2505.23250.

[3] S. Altay, A. Acerbi, People believe misinformation is a threat because they assume others are gullible, New Media & Society 26 (2024) 6440–6461. doi:10.1177/14614448231153379.

[4] P. Moravec, R. Minas, A. R. Dennis, Fake news on social media: People believe what they want to believe when it makes no sense at all, SSRN Scholarly Paper No. 3269541, 2018. doi:10.2139/ssrn.3269541.

[5] S. Zhang, F. Ma, Y. Liu, W. Pian, Identifying features of health misinformation on social media sites: An exploratory analysis, Library Hi Tech 40 (2021) 1384–1401. doi:10.1108/LHT-09-2020-0242.

[6] L. Booth, A. Aldaihani, J. Davidson, C. Wilson, C. Lawlor, P. Hong, M. E. Graham, Misinformation and readability of social media content on pediatric ankyloglossia and other oral ties, JAMA Otolaryngology–Head & Neck Surgery 151 (2025) 143–150. doi:10.1001/jamaoto.2024.4211.

[7] J. M. Struß, S. Schellhammer, S. Dietze, V. V., V. Setty, T. Chakraborty, P. Nakov, A. Anand, P. Chungkham, S. Hafid, D. Sahnan, K. Todorov, The clef-2026 checkthat! lab: Advancing multilingual fact-checking, in: R. Campos, A. Jatowt, Y. Lan, M. Aliannejadi, C. Bauer, S. MacAvaney, A. Anand, Z. Ren, S. Verberne, N. Bai, M. Mansoury (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2026, pp. 325–335.

[8] J. M. Struß, S. Schellhammer, S. Dietze, V. V., V. Setty, T. Chakraborty, P. Nakov, A. Anand, P. Chungkham, S. Hafid, D. Sahnan, K. Todorov, Overview of the CLEF-2026 CheckThat! Lab: Advancing multilingual fact-checking, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. S. Salido, A. Barrón-Cedeño, Alba García Seco de Herrera (Eds.), Exper-

imental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer, 2026.

[9] S. Schellhammer, S. Hafid, Y. S. Kartal, K. Boland, D. Dimitrov, K. Todorov, S. Dietze, Overview of the CLEF-2026 CheckThat! lab task 1 on source retrieval for scientific web claims, CLEF 2026, Jena, Germany, 2026.

[10] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: A large-scale dataset for fact extraction and VERification, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, 2018, pp. 809–819. doi:10.18653/v1/N18-1074.

[11] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, H. Hajishirzi, Fact or fiction: Verifying scientific claims, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020, pp. 7534–7550. doi:10.18653/v1/2020.emnlp-main.609.

[12] R. Pradeep, S. Sharifymoghaddam, J. Lin, RankVicuna: Zero-shot listwise document reranking with open-source large language models, 2023. doi:10.48550/arXiv.2309.15088. arXiv:2309.15088.

[13] A. Rana, D. Khanna, T. Ghosal, M. Singh, H. Singh, P. S. Rana, RerrFact: Reduced evidence retrieval representations for scientific claim verification, 2022. doi:10.48550/arXiv.2202.02646. arXiv:2202.02646.

[14] S. Hafid, S. Schellhammer, S. Bringay, K. Todorov, S. Dietze, SciTweets—a dataset and annotation framework for detecting scientific online discourse, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM '22), 2022, pp. 3988–3992. doi:10.1145/3511808.3557693.

[15] S. Hafid, Y. S. Kartal, S. Schellhammer, V. Jacot, S. Bringay, S. Dietze, K. Todorov, Disambiguation of implicit scientific references on x, in: Proceedings of the 36th ACM Conference on Hypertext and Social Media, 2025, pp. 165–170.

[16] C. Ashbaugh, L. Baumgärtner, T. Gress, N. Sidorov, D. Werner, AIRwaves at CheckThat! 2025: Retrieving scientific sources for implicit claims on social media with dual encoders and neural re-ranking, 2025. doi:10.48550/arXiv.2509.19509. arXiv:2509.19509.

[17] D. Lee, S.-w. Hwang, K. Lee, S. Choi, S. Park, On complementarity objectives for hybrid retrieval, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2023, pp. 13357–13368. doi:10.18653/v1/2023.acl-long.746.

[18] Y. Ren, Y. Cao, P. Guo, F. Fang, W. Ma, Z. Lin, Retrieve-and-sample: Document-level event argument extraction via hybrid retrieval augmentation, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2023, pp. 293–306. doi:10.18653/v1/2023.acl-long.17.

[19] L. Yang, J. Hu, M. Qiu, C. Qu, J. Gao, W. B. Croft, X. Liu, Y. Shen, J. Liu, A hybrid retrieval-generation neural conversation model, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19), 2019, pp. 1341–1350. doi:10.1145/3357384.3357881.

[20] S. Robertson, H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, Foundations and Trends in Information Retrieval 3 (2009) 333–389. doi:10.1561/1500000019.

[21] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. 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), Association for Computational Linguistics, 2020, pp. 6769–6781. doi:10.18653/v1/2020.emnlp-main.550.

[22] O. Khattab, M. Zaharia, ColBERT: Efficient and effective passage search via contextualized late interaction over BERT, 2020. doi:10.48550/arXiv.2004.12832. arXiv:2004.12832.

[23] 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, Association for Computational Linguistics, 2020, pp. 708–718. doi:10.18653/v1/2020.findings-emnlp.63.

[24] W. Sun, L. Yan, X. Ma, S. Wang, P. Ren, Z. Chen, D. Yin, Z. Ren, Is ChatGPT good at search? investigating large language models as re-ranking agents, 2024. doi:10.48550/arXiv.2304.09542. arXiv:2304.09542.

[25] X. Ma, L. Wang, N. Yang, F. Wei, J. Lin, Fine-tuning LLaMA for multi-stage text retrieval, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), 2024, pp. 2421–2425. doi:10.1145/3626772.3657951.

[26] H. Zhuang, Z. Qin, K. Hui, J. Wu, L. Yan, X. Wang, M. Bendersky, Beyond yes and no: Improving zero-shot LLM rankers via scoring fine-grained relevance labels, in: K. Duh, H. Gómez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), Association for Computational Linguistics, 2024, pp. 358–370. doi:10.18653/v1/2024.naacl-short.31.

[27] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual E5 text embeddings: A technical report, 2024. doi:10.48550/arXiv.2402.05672. arXiv:2402.05672.

[28] G. V. Cormack, C. L. A. Clarke, S. Büttcher, Reciprocal rank fusion outperforms Condorcet and individual rank learning methods, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09), 2009, pp. 758–759. doi:10.1145/1571941.1572114.

[29] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, et al., Qwen2.5 technical report, 2024. doi:10.48550/arXiv.2412.15115. arXiv:2412.15115.

[30] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, 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 (NeurIPS 2021), 2021. arXiv:2104.08663.

[31] R. Sennrich, B. Haddow, A. Birch, Neural machine translation of rare words with subword units, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2016, pp. 1715–1725. doi:10.18653/v1/P16-1162.

[32] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410.

[33] S. Bruch, S. Gai, A. Ingber, An analysis of fusion functions for hybrid retrieval, ACM Transactions on Information Systems 42 (2023) 20:1–20:35. doi:10.1145/3596512.