Jesters at CheckThat! 2026: Multilingual Scientific Source Retrieval with Hybrid Dense-Sparse Ranking

CheckThat! Task 1 – Source Retrieval for Scientific Web Claims at CLEF 2026

Daksh Vaghasiya[1], Vasu Goti[1]

[1]Dept. of Computer Science and Engineering, Indian Institute of Information Technology Surat, Surat, India

Abstract

We present the Jesters submission to the CheckThat! 2026 Task 1 shared task on multilingual scientific source retrieval. Our approach combines BM25 sparse retrieval and dense retrieval using the `intfloat/multilingual-e5-large` embedding model. Dense and sparse retrieval scores are fused using Reciprocal Rank Fusion (RRF). To improve retrieval quality, we fine-tuned the multilingual E5 model using contrastive learning with hard negatives mined from BM25 and pretrained E5 retrieval outputs. To improve multilingual performance, we augmented the training data by translating English claims into French and German using the `facebook/nllb-200-3.3B` translation model. Our final system achieved an average MRR@5 of 0.4881 on the test set.

Keywords

Scientific Source Retrieval, Multilingual Retrieval, Dense Retrieval, Hybrid Retrieval, BM25, Multilingual-E5, Reciprocal Rank Fusion, Scientific Information Retrieval

1. Introduction

Scientific information is increasingly disseminated through online articles, blogs, social media platforms, and fact-checking websites. Many of these sources summarize findings from scientific publications without explicitly citing the original research papers. As a result, automatically identifying the scientific source behind a claim has become an important problem for evidence retrieval, scientific verification, citation discovery, and misinformation analysis.

Scientific source retrieval differs from traditional document retrieval in several important ways. Claims found online are often short, paraphrased, and written in non-technical language, while scientific papers use domain-specific terminology and formal writing styles. This creates a significant semantic gap between the query claim and the original source document. Furthermore, multilingual retrieval introduces an additional challenge when claims are expressed in languages different from the document collection language.

Traditional sparse retrieval methods such as BM25 [1] rely heavily on lexical overlap and therefore struggle in multilingual and semantically complex retrieval settings. Dense retrieval approaches based on transformer encoders have demonstrated strong performance by learning semantic representations capable of capturing conceptual similarity beyond exact keyword matching [2, 3, 4]. Prior work has also shown that hybrid retrieval systems combining sparse and dense retrieval signals can improve retrieval robustness and ranking quality [5].

In this work, we develop a multilingual scientific source retrieval system for the CheckThat! 2026 Task 1 benchmark. Our approach combines BM25 sparse retrieval with dense retrieval using the `intfloat/multilingual-e5-large` embedding model. We further improve retrieval quality through hard negative mining [6], multilingual query augmentation using machine translation [7], retrieval-oriented fine-tuning, and Reciprocal Rank Fusion (RRF).

Our final hybrid system achieves an average MRR@5 score of 0.4881 on the test set across English, German, and French queries.

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany

vaghasiyadaksh59@gmail.com (D. Vaghasiya); vasugoti2005@gmail.com (V. Goti)

2. Related Work

Scientific claim verification and evidence retrieval have received significant attention in recent years. Early work such as FEVER [8] introduced large-scale fact verification pipelines combining evidence retrieval and claim verification stages. Later work including SciFact [9] extended these ideas to scientific literature retrieval and verification tasks.

Traditional sparse retrieval approaches such as BM25 [1] remain strong lexical baselines for information retrieval tasks. However, sparse retrieval relies heavily on exact term overlap and therefore struggles with semantic paraphrasing and multilingual retrieval settings.

Dense retrieval methods based on transformer encoders have shown strong improvements in semantic retrieval tasks. Dense Passage Retrieval (DPR) [2] demonstrated the effectiveness of bi-encoder architectures for semantic search, while Sentence-BERT [3] enabled efficient sentence-level embedding generation using Siamese transformer networks. More recent embedding models such as E5 [4] and multilingual-e5 use contrastive learning objectives to produce retrieval-oriented representations suitable for multilingual retrieval tasks.

Cross-lingual retrieval introduces additional challenges because queries and documents may exist in different languages. Models such as LaBSE [10] and multilingual-e5 address this through multilingual representation learning and alignment across languages. However, multilingual scientific retrieval remains challenging due to domain-specific terminology and semantic mismatch between claims and scientific documents.

Recent work has also explored hybrid retrieval systems that combine sparse and dense retrieval signals. Reciprocal Rank Fusion (RRF) [5] is widely used to combine lexical and semantic retrieval rankings while remaining simple and computationally efficient. In addition, hard negative mining [6] has been shown to substantially improve dense retrieval models by exposing them to semantically similar but incorrect retrieval candidates during training.

3. Dataset and Task Setup

The CheckThat! 2026 (CT26) Task 1 dataset [11] consists of scientific claims extracted from web content and paired with supporting scientific publications. Queries are provided in three languages: English (EN), German (DE), and French (FR). The document collection is a fixed set of 10,000 English-language scientific papers, where each paper contains metadata including title, abstract, venue, and author information. Ground truth annotations map each query to its corresponding source publication using a unique publication identifier.

The dataset exhibits a strong imbalance across languages, with English queries forming the majority of the training data. German and French queries are comparatively limited, making multilingual generalization more challenging. This imbalance motivated the multilingual augmentation strategy described later in Section 4.

Table 1
Dataset Statistics for CT26 Task 1

Split / Language	Train	Dev
English (EN)	14,977	3,905
German (DE)	1,460	386
French (FR)	2,807	702
Total Queries	19,244	4,993
Collection (EN)	–	10,000 papers

The retrieval task is formulated as ranking scientific documents for a given multilingual query. Given a query q in language L and a fixed English document collection D, the retrieval system must return a ranked list of documents from D ordered by estimated relevance.

Performance is primarily evaluated using Mean Reciprocal Rank at 5 (MRR@5), which measures the rank quality of the first correct retrieved document:

$$MRR@5 = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{rank_q} \tag{1}$$

where $rank_q$ denotes the rank position of the correct document for query q. If the correct document does not appear within the top 5 retrieved results, the contribution for that query is set to zero.

In addition to MRR@5, Recall@100 is used to evaluate whether the correct source document appears within the top 100 retrieved candidates. This metric is particularly useful for measuring the upper-bound effectiveness of downstream reranking approaches.

4. Methodology

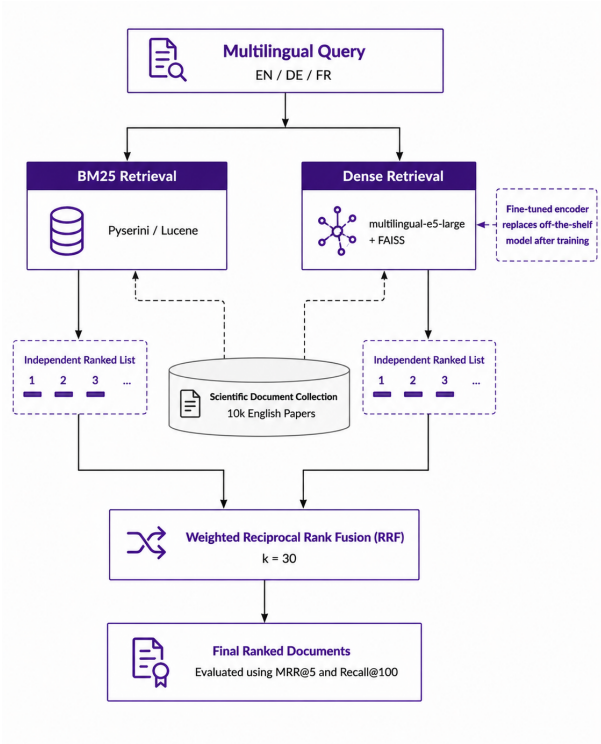

Figure 1: Overall retrieval pipeline combining BM25 sparse retrieval, dense retrieval using multilingual-E5, weighted RRF fusion, hard negative mining, multilingual augmentation, and retrieval-oriented fine-tuning.

4.1. BM25 Sparse Retrieval

The sparse retrieval baseline uses BM25 [1] implemented through Pyserini [12] with a Lucene backend. Documents are indexed using a boosted text representation in which the title is repeated three times and concatenated with the abstract, increasing the importance of title-term matches during retrieval. Indexing is performed over the JsonCollection format, with each document identified by its `pubkey`. A single BM25 index is constructed over the English scientific document collection, and multilingual queries are independently submitted to the same index during retrieval.

BM25 retrieval relies on exact lexical matching and inverse document frequency weighting. While effective for English queries where claim vocabulary overlaps with paper titles or abstracts, BM25 performance degrades substantially for German and French queries due to limited lexical overlap with the English document collection. This limitation motivates the use of semantic dense retrieval.

4.2. Dense Retrieval using multilingual-E5

Dense retrieval is implemented using the `intfloat/multilingual-e5-large` encoder [4], a multilingual transformer model with a 1024-dimensional embedding space. Queries are formatted as:

```
query: find relevant scientific paper: <claim text>
```

Documents are formatted as:

```
passage: <title>. <venue>. <abstract>. <authors>
```

Query and document embeddings are computed using mean pooling over the final transformer layer followed by L2 normalization. All 10,000 collection documents are encoded offline and stored within a FAISS FlatIP index [13] for efficient similarity search. Because embeddings are normalized prior to indexing, inner-product retrieval is equivalent to cosine similarity retrieval. Maximum sequence length is set to 512 tokens during offline document encoding.

4.3. Hybrid Retrieval via Reciprocal Rank Fusion

RRF [5] is used to combine the ranked outputs of BM25 and dense retrieval. The weighted RRF score for a document d is defined as:

$$score(d) = \sum_i w_i \cdot \frac{1}{k + rank_i(d)} \tag{2}$$

where $rank_i(d)$ denotes the rank of document d in retrieval system i, k is a smoothing constant, and w_i is a retrieval-specific weight.

We use $k = 30$, assigning a higher weight to dense retrieval ($w_{dense} = 1.2$) and unit weight to BM25 ($w_{BM25} = 1.0$). Lower values of k increase the influence of higher-ranked documents in the fused ranking, which aligns well with the MRR@5 evaluation objective. The asymmetric weighting reflects the greater importance of semantic retrieval using dense embeddings for multilingual retrieval, while still preserving useful lexical matching signals from BM25 for English queries.

4.4. Hard Negative Mining

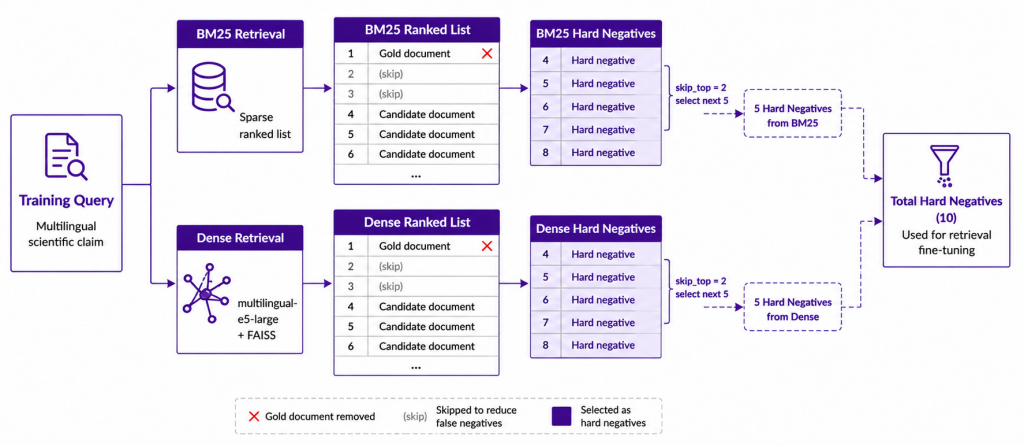

Figure 2: Hard Negative Mining Pipeline: Queries are submitted to both BM25 and dense retrieval; top non-gold results from the fused ranked list (skipping top-2 to avoid false negatives) serve as hard negatives.

Standard in-batch negatives used during contrastive training are often trivially distinguishable from the query and therefore provide limited retrieval supervision. Hard negative mining instead selects semantically challenging retrieval candidates that are highly ranked but incorrect.

Our hard negative mining procedure submits each training query to both the BM25 retriever and the dense FAISS index, retrieving the top-150 candidates from each system. The two ranked lists are fused using weighted RRF. The gold document is removed from the fused ranking, and the highest-ranked remaining candidates are skipped to reduce the likelihood of selecting unlabeled but potentially relevant documents as negatives. The next five candidates are selected as hard negatives, yielding a training structure consisting of one positive document and five hard negatives per query.

Using RRF-fused hard negatives combines two complementary failure modes. BM25 hard negatives tend to be lexically similar to the query, while dense retrieval negatives are semantically proximate but referentially distinct. Training with both types of negatives improves the model's ability to distinguish the correct source document from semantically similar distractors.

4.5. Machine Translation Augmentation

The training set exhibits a strong imbalance between English, German, and French queries. To improve multilingual retrieval performance, we augment the training data using machine translation. English training queries are translated into German and French using the `facebook/nllb-200-3.3B` model [7].

Synthetic German and French translations are added to the training pool using the same positive document labels as the original English queries. This augmentation increases the effective amount of multilingual supervision and improves cross-lingual retrieval alignment during fine-tuning.

4.6. Fine-Tuning Strategy

Fine-tuning is performed using the SentenceTransformers framework [3]. Training examples are represented using query-positive-negative triplets consisting of a query, its corresponding positive document, and up to five hard negatives.

The training objective uses `MultipleNegativesRankingLoss`, which treats other examples within the same batch as additional negatives during optimization. During training, positive and hard negative examples are jointly encoded within the same batch, enabling efficient contrastive learning with large numbers of implicit negatives.

This objective encourages the model to assign higher similarity scores to relevant query-document pairs than to irrelevant candidates, directly optimizing retrieval ranking quality.

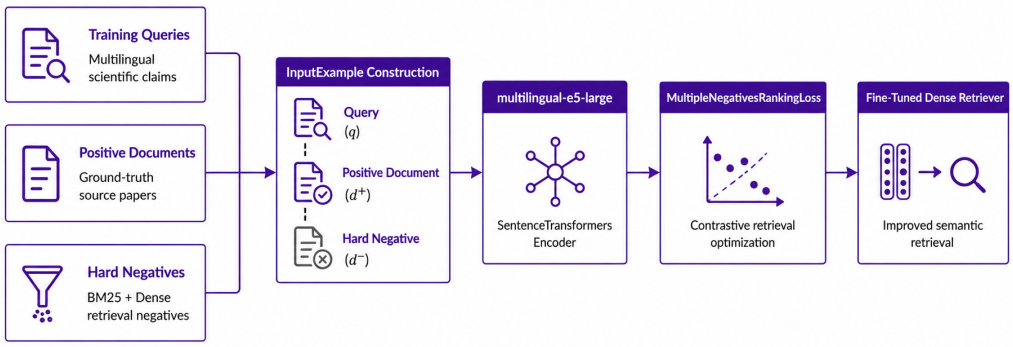

Figure 3: Fine-Tuning Workflow: Positive pairs and RRF hard negatives are assembled into InputExamples and trained using MultipleNegativesRankingLoss with SentenceTransformers.

4.7. Training Configuration

Table 2 summarizes the training configuration used for fine-tuning. Training is performed for a single epoch on an NVIDIA A100 GPU using mixed precision training and gradient checkpointing to reduce memory consumption. The maximum sequence length during training is reduced to 96 tokens to improve throughput efficiency, while full-length document encoding is retained during evaluation.

Table 2
Training Hyperparameters

Hyperparameter	Value
Base Model	multilingual-e5-large
Batch Size	64
Epochs	1
Learning Rate	2e-5
Max Sequence Length	96
Warmup Ratio	5%
Loss Function	MultipleNegativesRankingLoss
Mixed Precision	FP16
Gradient Checkpointing	Enabled
Hardware	NVIDIA A100

5. Experiments and Results

5.1. BM25 Baseline

Table 3 presents BM25 retrieval results on the development set. English retrieval achieves a Recall@100 of 0.7846 and an MRR@5 of 0.5298, confirming that lexical retrieval provides a strong baseline when query terminology overlaps with scientific document vocabulary. However, retrieval quality degrades substantially for German and French queries, where lexical overlap with the English document collection is limited. German MRR@5 drops to 0.1484 and French MRR@5 to 0.2481, demonstrating the limitations of sparse retrieval in multilingual retrieval settings.

Table 3
BM25 Baseline Results (Development Set)

Language	Recall@100	MRR@5
English	0.7846	0.5298
German	0.3938	0.1484
French	0.5342	0.2481

5.2. Dense Retrieval Results

Dense retrieval using the off-the-shelf `multilingual-e5-large` model substantially improves multilingual retrieval performance. German MRR@5 improves from 0.1484 to 0.3788, while French MRR@5 improves from 0.2481 to 0.4579. English retrieval quality remains competitive, with Recall@100 increasing from 0.7846 to 0.8184.

These results demonstrate that dense semantic embeddings effectively bridge cross-lingual semantic gaps that cannot be handled through lexical matching alone.

Table 4
Dense Retrieval Results using multilingual-E5 (Development Set)

Language	Recall@100	MRR@5
English	0.8184	0.4954
German	0.7150	0.3788
French	0.8091	0.4579

5.3. Hybrid Retrieval Results

Applying Reciprocal Rank Fusion (RRF) improves English retrieval further, increasing English MRR@5 to 0.5371 and Recall@100 to 0.8236. This confirms that lexical BM25 signals remain useful when query vocabulary overlaps with relevant scientific documents.

However, German and French performance decrease slightly compared to dense-only retrieval. BM25 introduces noisy rankings for non-English queries, partially overriding stronger semantic retrieval signals. This observation suggests that fixed-weight retrieval fusion may not optimally balance sparse and dense retrieval signals across languages.

Table 5
Hybrid Retrieval Results using RRF Fusion (Development Set)

Language	Recall@100	MRR@5
English	0.8236	0.5371
German	0.7176	0.2997
French	0.7963	0.4316

5.4. Fine-Tuned Development Results

After retrieval-oriented fine-tuning using hard negatives and multilingual augmentation, the final hybrid system achieves the strongest results across all languages on the internal development split. English Recall@100 improves to 0.9091 with an MRR@5 of 0.5731. German retrieval improves substantially to an MRR@5 of 0.5025, while French achieves the strongest multilingual retrieval performance with an MRR@5 of 0.6517.

These results demonstrate that hard negative mining and multilingual augmentation significantly improve the model's ability to distinguish semantically similar scientific documents during retrieval.

Table 6
Fine-Tuned Hybrid Retrieval Results (Development Set)

Language	Recall@100	MRR@5
English	0.9091	0.5731
German	0.8705	0.5025
French	0.9145	0.6517
Average	0.8980	0.5758

5.5. Official Test Set Results

Table 7
Official Test Set Results

Language	MRR@5
English	0.4798
German	0.4548
French	0.5296
Average	0.4881

Table 7 presents the final results obtained on the official CheckThat! 2026 Task 1 test set. The submitted system corresponds to the fine-tuned hybrid retrieval model combining multilingual-E5

dense retrieval, BM25 sparse retrieval, weighted Reciprocal Rank Fusion, hard negative mining, and multilingual augmentation.

The final system achieves an average MRR@5 score of 0.4881 across all languages. French queries achieve the strongest performance with an MRR@5 of 0.5296, while German remains comparatively more challenging with an MRR@5 of 0.4548.

5.6. Comparative Analysis

Table 8 summarizes retrieval performance across all experimental configurations on the internal development split. Several observations emerge from these experiments.

First, the transition from sparse retrieval to dense retrieval produces the largest multilingual improvement, particularly for German and French queries. This confirms that semantic retrieval is essential for multilingual scientific source retrieval.

Second, naive RRF fusion improves English retrieval quality but slightly degrades non-English retrieval performance due to noisy sparse retrieval rankings. This highlights the difficulty of balancing lexical and semantic retrieval signals across languages.

Finally, retrieval-oriented fine-tuning using RRF-mined hard negatives and multilingual augmentation consistently improves performance across all languages. These improvements suggest that the model learns stronger discrimination between semantically related but referentially distinct scientific documents.

Table 8
Comparative Analysis Across Retrieval Systems (Development Set)

System	EN R@100	DE R@100	FR R@100	EN MRR@5	DE MRR@5	FR MRR@5
BM25	0.7846	0.3938	0.5342	0.5298	0.1484	0.2481
mE5-large	0.8184	0.7150	0.8091	0.4954	0.3788	0.4579
RRF (BM25+mE5)	0.8236	0.7176	0.7963	0.5371	0.2997	0.4316
Fine-tuned + RRF	0.9091	0.8705	0.9145	0.5731	0.5025	0.6517

6. Future Work

6.1. Cross-Encoder Reranking

The current system uses a bi-encoder architecture for efficient retrieval. Future work may extend this to a two-stage retrieval pipeline where a cross-encoder reranker jointly processes query-document pairs to improve ranking precision [14]. Models such as BGE rerankers [15] and MiniLM-based cross-encoders [16] may further improve MRR@5 performance.

6.2. Translation-Based Retrieval

Another possible direction is explicit query translation prior to retrieval. German and French queries could first be translated into English using models such as NLLB [7], allowing retrieval to operate entirely within a monolingual English retrieval space. However, retrieval quality would then depend heavily on translation accuracy.

6.3. Advanced Reranking Models

Specialized reranking architectures, including multilingual cross-encoders and domain-specific rerankers, may further improve retrieval quality through scientific retrieval fine-tuning and stronger semantic relevance estimation.

6.4. Chunk-Level Retrieval

The current system performs retrieval using document-level title and abstract representations. Future work could explore chunk-level or sentence-level retrieval [2, 17] to improve fine-grained evidence matching within scientific documents.

7. Conclusion

This work presented a multilingual scientific source retrieval system for the CT26 Task 1 benchmark. Starting from a BM25 sparse retrieval baseline, the system was progressively improved through dense semantic retrieval, hybrid Reciprocal Rank Fusion, hard negative mining, multilingual augmentation, and retrieval-oriented fine-tuning.

Experimental results showed that sparse retrieval alone is insufficient for multilingual scientific retrieval, particularly for German and French queries where lexical overlap with the English document collection is limited. Dense retrieval substantially improved multilingual retrieval quality by capturing semantic similarity beyond exact term matching, while hybrid retrieval combined the complementary strengths of lexical and semantic retrieval signals.

Fine-tuning using RRF-mined hard negatives proved especially effective, producing substantial improvements over off-the-shelf multilingual embedding models across all languages. These results highlight the importance of retrieval-specific fine-tuning even for strong pretrained multilingual retrievers.

The final hybrid system achieved an average MRR@5 score of 0.4881 on the official test set across English, German, and French queries. Overall, the results demonstrate the effectiveness of combining sparse and dense retrieval strategies for multilingual scientific source retrieval and provide a strong foundation for future reranking-based extensions.

Acknowledgments

The authors thank the support received in terms of computational resources and Google Colab credits, which helped a lot in the experimentation and evaluation processes of this work. The availability of these resources made it possible to train and fine-tune the models.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (OpenAI GPT-5.5) to assist with paraphrasing and rewording, improving writing style, grammar and spelling correction, formatting assistance, citation formatting, and content refinement. The paper structure, literature selection, methodology, experiments, and analysis were designed and written by the authors. Generative AI tools were also used to assist in creating workflow and system pipeline figures, which were iteratively refined and verified by the authors to ensure technical correctness. After using these tools, the authors reviewed, edited, and validated all content and take full responsibility for the final publication.

References

[1] S. Robertson, H. Zaragoza, The probabilistic relevance framework: Bm25 and beyond, Foundations and Trends in Information Retrieval (2009).

[2] V. Karpukhin, et al., Dense passage retrieval for open-domain question answering, in: EMNLP, 2020.

[3] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: EMNLP, 2019.

[4] L. Wang, et al., Text embeddings by weakly-supervised contrastive pre-training, arXiv preprint arXiv:2212.03533 (2022).
[5] G. Cormack, C. Clarke, S. Buettcher, Reciprocal rank fusion outperforms condorcet and individual rank learning methods, SIGIR (2009).
[6] L. Xiong, et al., Approximate nearest neighbor negative contrastive learning for dense text retrieval, ICLR (2021).
[7] M. Costa-jussà, et al., No language left behind: Scaling human-centered machine translation, arXiv preprint arXiv:2207.04672 (2022).
[8] J. Thorne, et al., Fever: a large-scale dataset for fact extraction and verification, in: NAACL, 2018.
[9] D. Wadden, et al., Fact or fiction: Verifying scientific claims, in: EMNLP, 2020.
[10] F. Feng, et al., Language-agnostic bert sentence embedding, arXiv preprint arXiv:2007.01852 (2020).
[11] 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.
[12] J. Lin, X. Ma, S.-C. Lin, J.-H. Yang, R. Pradeep, R. Nogueira, Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations, in: SIGIR, 2021.
[13] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with gpus, IEEE Transactions on Big Data (2019).
[14] R. Nogueira, K. Cho, Passage re-ranking with bert, in: arXiv preprint arXiv:1901.04085, 2019.
[15] S. Xiao, et al., C-pack: Packed resources for general chinese embeddings, arXiv preprint arXiv:2309.07597 (2023).
[16] W. Wang, et al., Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, NeurIPS (2020).
[17] P. Lewis, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, NeurIPS (2020).