MG_CheckThat2026 at CheckThat! 2026: A Multilingual Cascade Pipeline for Scientific Claim Source Retrieval

Notebook for the CheckThat! Lab at CLEF 2026

Junyi Li, Zhenyu Peng and Leilei Kong*

Foshan University, Foshan, Guangdong, China

Abstract

Scientific Claim Source Retrieval aims to automatically retrieve the original scientific papers cited in tweets from a large-scale corpus. To address this task, we propose a three-stage cascaded retrieval pipeline. In the first stage, a hybrid retrieval module generates a candidate pool via merging and deduplication. In the second stage, we design a language-specific routing mechanism for initial ranking. This component is fine-tuned on multilingual data distributions to mitigate cross-lingual interference. In the third stage, we apply LLM-based listwise reranking to the candidate documents. This step replaces independent pointwise scoring with a joint multi-document input and supports cross-document comparison among semantically similar candidates. Experimental results show that dense retrieval provides a cross-lingual recall baseline, while adding LLM listwise reranking improves MRR@5 in the development results. Overall, our proposed system achieves an MRR@5 of 64.77% on the development set (dev set) and 63.94% on the hidden test set.

Keywords

LLM, listwise reranking, Scientific Claim Source Retrieval, Multilingual, Cascade Retrieval Pipeline

1. Introduction

Scientific Claim Source Retrieval is the task of retrieving the original source documents cited in tweets from a large-scale scientific corpus [1]. Within NLP and IR, this task falls under open-domain evidence retrieval [2, 3].

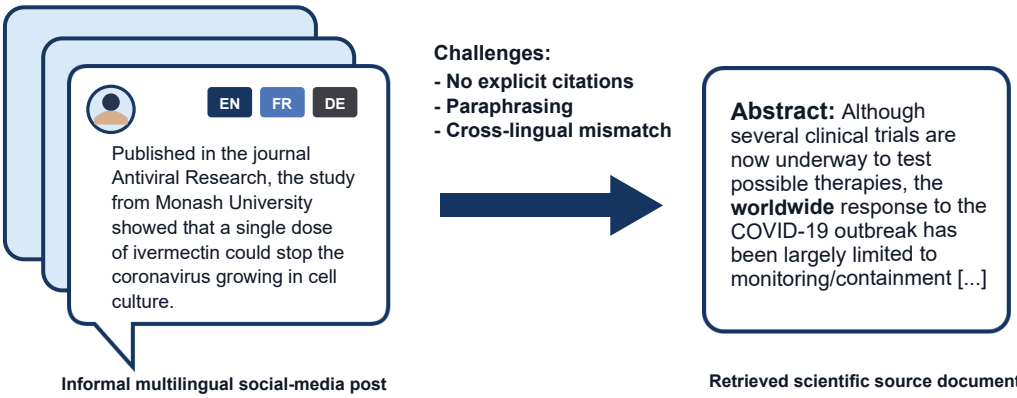

Figure 1: Illustration of the Multilingual Scientific Source Attribution Task. The system takes an informal social media post (EN, FR, DE) as input and retrieves the originally referenced document from a scientific collection, overcoming challenges such as paraphrasing and cross-lingual semantic mismatches.

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

*Corresponding author.

2112571052@stu.fosu.edu.cn (J. Li); 2112571078@stu.fosu.edu.cn (Z. Peng); kongleilei@fosu.edu.cn (L. Kong)

Existing methods fall into three main categories: sparse retrieval based on lexical overlap [4], dense retrieval based on semantic representations [3], and hybrid pipelines combining both with cross-encoder reranking [5, 6]. While these pipelines perform well on homogeneous texts, applying them to our task reveals three key challenges: 1) A universal multilingual model might struggle to capture fine-grained academic paraphrases across languages. 2) Standard Reciprocal Rank Fusion (RRF) [7] relies solely on relative rankings and ignores score distributions, often diluting the long-tail entity signals [8] captured by sparse retrieval. 3) Pointwise reranking lacks cross-document comparison, making it difficult to distinguish semantically similar hard negatives.

The main contributions of this paper are as follows:

- We construct a hybrid retrieval module that performs direct merging and deduplication on retrieval results during the multi-channel retrieval stage, preserving the long-tail hard entities of tweets captured by sparse retrieval.
- We design a language-specific routing initial ranking mechanism and propose a dynamic routing strategy that combines language-specific expert fine-tuning and unified multilingual fine-tuning to reduce cross-lingual negative interference and few-shot overfitting.
- We introduce an LLM-based listwise reranking module that transforms single-document scoring into a joint multi-document input, separating semantically similar hard negatives through global cross-comparison.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 details the components of the cascade pipeline. Section 4 describes the task, dataset, and evaluation metric. Section 5 presents the experiments and analysis. Section 6 concludes the paper.

2. Related Work

Fact-Checking and Scientific Literature Retrieval. Automated fact-checking initially targeted structured knowledge bases and later extended to open-domain texts. FEVER [2] proposed an evidence retrieval and claim verification pipeline based on Wikipedia. SciFact [9] and MultiVerS [10] extended this task to the scientific and medical domains to handle specialized terminology. Social media source retrieval is a more recent direction. Wright et al. [11] investigated the identification and retrieval of scientific claims in tweets. CLEF CheckThat! [1] introduced the task of cross-register implicit scientific claim source retrieval. However, most existing works focus on monolingual English retrieval. Semantic gaps and insufficient multilingual alignment remain when mapping informal social media texts to formal academic corpora.

Hybrid Retrieval. Sparse retrieval methods like BM25 [4] rely on lexical overlap and can capture exact entities such as proper nouns, abbreviations, and numerical values. Dense Passage Retrieval (DPR) [3] maps texts into a continuous vector space, overcoming the limitations of keyword matching. Common approaches for fusing multiple retrieval signals include Reciprocal Rank Fusion (RRF) [7] and linear interpolation. However, Bruch et al. [8] noted that rank-based fusion ignores absolute confidence scores when the distributions of heterogeneous models differ. In social media matching, this can downweight the long-tail entity signals captured by sparse retrieval.

Cross-Encoder and LLM Reranking. During reranking, cross-encoders [6] enable deep token-level interactions between query-document pairs, yielding higher ranking accuracy than dual-encoder models. However, they score each document independently, lacking cross-document comparison. Sun et al. [12] and Ma et al. [13] explored LLM-based listwise reranking, which simultaneously processes the query and multiple candidate documents to jointly evaluate and compare candidates. Inspired by this paradigm, we design a bilingual prompt mechanism for multilingual scientific source retrieval.

3. Methodology

This section details the three-stage cascaded retrieval pipeline designed for the multilingual scientific source retrieval task. As shown in Figure 2, the pipeline sequentially narrows down the candidate scope through hybrid retrieval, language-specific initial ranking, and LLM-based listwise reranking. The first stage constructs a candidate pool with increased recall by merging results from the sparse and dense retrieval paths. The second stage performs initial ranking by applying independently fine-tuned cross-encoders based on the query language to reduce cross-lingual interference. The third stage inputs the candidate list into an LLM for listwise reranking to support cross-document comparison among hard negatives.

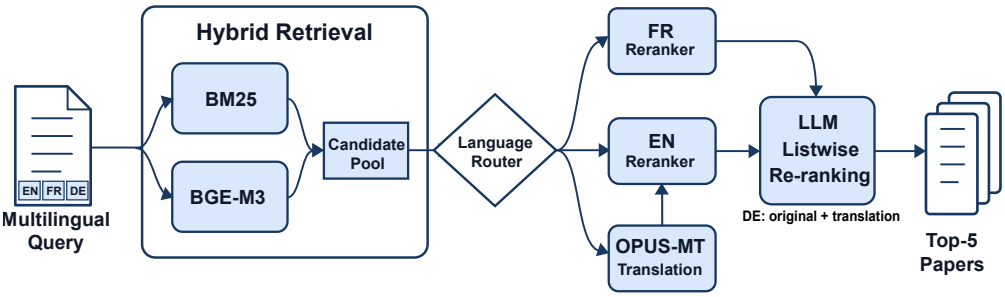

Figure 2: Overview of the three-stage hybrid retrieval pipeline.

3.1. First Stage: Hybrid Retrieval and Direct Deduplication

A single retrieval paradigm struggles to balance lexical precision and semantic generalization due to the register differences between informal tweets and formal academic texts. To address this, we construct a dual-path hybrid retrieval module.

Sparse Lexical Path: An inverted index is implemented based on the Okapi BM25 algorithm [4]. Tweets often contain unnormalized abbreviations and specific numerical values. BM25 can capture these exact entities via term frequency matching. To control long-tail noise caused by short text matching, this path retrieves the top-30 candidates.

Dense Semantic Path: The BGE-M3 multilingual dual-encoder [14] is deployed to capture contextual semantics. Dense retrieval fetches the top-100 candidates by calculating the cosine similarity between query and document vectors. For low-resource German queries, we use an OPUS-MT [15] translation module to translate queries into English before encoding, completing the matching within the English-dominated scientific space.

Merging and Deduplication Mechanism: Existing works often adopt Reciprocal Rank Fusion (RRF) [7] to fuse multiple retrieval signals. However, in "short tweet to long abstract" matching, RRF often downweights the entity signals captured by the sparse path by ignoring absolute score distributions [8]. Therefore, we bypass rank-based fusion and perform direct merging and deduplication on the candidate documents from both paths. By expanding retrieval coverage, this mechanism defers precise ranking to the subsequent reranking components.

3.2. Second Stage: Language-Specific Routing Initial Ranking

Directly inputting the deduplicated candidate pool into an LLM incurs high computational costs and exceeds context limits. Therefore, we introduce a cross-encoder [6] in the second stage for initial filtering. Unlike independent dual-encoder models, the cross-encoder concatenates the tweet query and candidate document as a unified input. It uses a multi-layer self-attention mechanism for token-level interactions and outputs relevance scores. This stage extracts a smaller candidate subset to reduce the inference burden on the subsequent LLM reranking step.

For the multilingual environment, we design a language-specific routing strategy. For French (FR), a medium-resource language, we route queries to a language-specific model fine-tuned on the corresponding corpus. This is intended to reduce cross-lingual interference [16] from high-resource languages. For English (EN), a high-resource language, we route queries to a unified multilingual fine-tuned model, using the larger English training split to support generalization. For German (DE), a low-resource language, few-shot task-specific fine-tuning might degrade the base model's alignment due to prepended machine translation and scarce original training samples. Therefore, translated German tweets are scored directly using the English unified multilingual model. All fine-tuned models are initialized from the same multilingual pre-trained base (BAAI/bge-reranker-v2-m3).

To improve the cross-encoder's ability to distinguish hard negatives, we introduce a hard negative mining mechanism [17] during fine-tuning. Following standard contrastive learning paradigms [3], we use the first-stage dense retriever to sample top-ranked incorrect citations as hard negatives. The resulting training tuples contain positive samples, hard negatives, and random negatives, which provide supervision for factual detail differences. After initial ranking, the system selects the top-scoring documents and inputs them into the final LLM listwise reranking stage.

3.3. Third Stage: LLM Listwise Reranking

To address the lack of cross-document comparison in pointwise cross-encoders, we apply LLM-based listwise reranking [12]. This module formats the original tweet query and the top-10 candidates from the second stage into a single multi-document prompt. In our submitted system, the LLM reranker uses DeepSeek Chat (`deepseek-chat`) through the OpenAI-compatible DeepSeek API. This allows the LLM to perform cross-document comparison and logical reasoning to filter out hard negatives.

For prompt design, we set specific instruction constraints for the model. First, the instructions require the model to prioritize documents containing exact entities (e.g., specific numerical values or experimental subjects) mentioned in the tweets. It also explicitly penalizes review articles. Second, for German queries, because the prepended translation used in the first stage may cause the loss of original proper nouns, we design a bilingual prompt template. This template inputs both the original German text and its English translation. This allows the LLM to perform reasoning in the English-dominated scientific space while referencing original German entity clues for auxiliary verification. Finally, the LLM is instructed to output a JSON array containing five 1-based candidate numbers. The exact prompt templates are provided in Appendix A.

4. Task and Dataset

4.1. Datasets

We use the official CLEF 2026 dataset for our experiments. The shared candidate pool includes the titles, abstracts, authors, and journal details of 10,000 scientific papers, with an average length of 520 words. The query set comprises tweets containing implicit citations across three languages: French (FR), English (EN), and German (DE). Table 1 details the dataset splits, proportions, and average lengths.

Table 1
Multilingual distribution proportion and average length of the query set.

Language	Train	Dev	Test	Total & %	Mean Query Length
EN	14977	3905	6076	24,958 (77.9%)	32.87
FR	2807	702	1220	4,729 (14.5%)	34.11
DE	1460	386	876	2,722 (7.6%)	30.05

Table 1 highlights an imbalanced language distribution. English is the largest language split (77.9%), whereas German accounts for 7.6%. This disparity motivates our language-specific routing strategy during reranking. For French and English, we apply a language-specific fine-tuned model and a

unified multilingual model, respectively. For German (a low-resource language), we translate queries into English and reuse the unified multilingual model. This design is intended to reduce the risk of generalization degradation and catastrophic forgetting that may be triggered by few-shot fine-tuning.

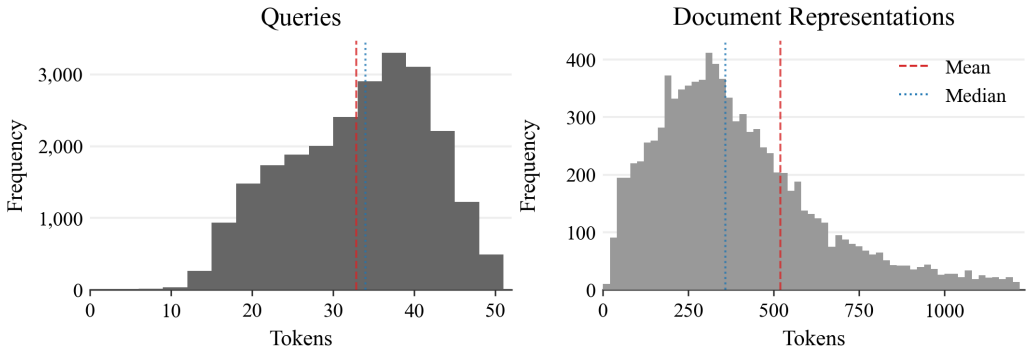

Figure 3: Length distribution comparison between social media tweet queries and academic literature representations.

Beyond resource imbalance, the task features a large length discrepancy between queries and documents. As shown in Table 1 and Figure 3, the average query length is 30–34 words, whereas candidate documents average 520 words. Figure 3 further illustrates a long-tail document length distribution. This imbalance can dilute entity signals within dense retrieval representations. Therefore, our first-stage hybrid retrieval uses direct merging and deduplication to preserve BM25 lexical signals.

4.2. Evaluation Metrics

This paper follows the official setting of the CLEF 2026 task and adopts MRR@5 (Mean Reciprocal Rank at 5) as the evaluation metric. MRR is sensitive to ranking positions, making it suitable for source retrieval tasks with only a single correct source. Its calculation formula is as follows:

$$\text{MRR@5} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \tag{1}$$

where $|Q|$ represents the total number of queries (social media tweets); rank_i represents the ranking position of the correct scientific paper in the top-5 candidate list returned by the system for the i-th query ($\text{rank}_i \in \{1, 2, 3, 4, 5\}$). If the system fails to rank the correct paper within the top 5, the term $\frac{1}{\text{rank}_i}$ for this query is recorded as 0.

5. Experiments

5.1. Experimental Setup

All model training and inference are conducted on a single NVIDIA RTX 5090 (32GB) GPU. During preprocessing, we flatten the provided structured metadata and concatenate it into a single string formatted as `{title}. {venue}. {abstract}. {authors}`. This serves as the standard document input for the entire pipeline.

In the first stage (hybrid retrieval and deduplication), we build the sparse retrieval inverted index using the `rank_bm25` library. After lowercasing the queries, we select the top-30 candidate documents. For dense retrieval, we use the BGE-M3 multilingual model. The maximum document encoding length is fixed at 768 tokens, and the query encoding length is limited to 512 tokens. We compute cosine similarity with FP16 precision and retrieve the top-100 documents. For German queries, we use the `Helsinki-NLP/opus-mt-de-en` translation model to translate them into English before encoding. After merging and deduplicating the retrieval results, the candidate pool size ranges from 110 to 130 documents per query.

In the second stage (language-specific initial ranking), all fine-tuned cross-encoders are initialized from the `BAAI/bge-reranker-v2-m3` base. To address resource imbalance for the unified multilingual model, we apply oversampling data augmentation to the training corpus at a 1:5:10 ratio for English, French, and German, respectively. We construct training samples using a hard negative mining mechanism. For each query in the training set, we use the first-stage dense retriever to fetch the top-50 documents. We select the annotated correct source as the positive sample, the top-ranked incorrect citations as hard negatives (3 per query), and 1 random document from the corpus as a simple negative. This yields 5 query-document pairs per query in a 1:3:1 ratio. During fine-tuning, the maximum sequence length is 768. The physical batch size is 1, with 16 gradient accumulation steps to achieve an effective batch size of 16. The learning rate is 2×10^{-5} with a 15% linear warmup. We train for 2 epochs to reduce the risk of catastrophic forgetting. We enable BF16 mixed precision and gradient checkpointing to reduce memory overhead. During inference, the routing module selects the top-10 highest-scoring documents to pass to the final ranking stage.

In the third stage (LLM listwise reranking), we use DeepSeek Chat (`deepseek-chat`) through the OpenAI-compatible endpoint `https://api.deepseek.com/v1`. Each request contains a single user message constructed from the prompt templates in Appendix A. For French and English queries, the tweet text is truncated to 500 characters. For German queries, both the original German tweet and its English translation are truncated to 500 characters. The LLM receives the top-10 cross-encoder candidates, with each candidate document truncated to 800 characters. We set `temperature=0.0` and `max_tokens=50`. We run the calls with 8 worker threads. For output parsing, we remove Markdown code fences when present, extract the bracketed array, and parse it as JSON. The returned 1-based candidate numbers are mapped back to document identifiers. If the parsed list contains fewer than five valid identifiers, the system returns the cross-encoder top-5. API errors and JSON parsing failures are retried up to three times with a one-second interval; after repeated failure, the system also falls back to the cross-encoder top-5.

5.2. Main Results

Table 2 presents our pipeline's performance on the development set compared to the official baseline. To separate the cascade stages, we evaluate hybrid indexing recall and reranking accuracy (MRR@5) separately. Alongside these metrics, we analyze the cross-lingual limitations of basic retrieval paradigms and our design motivations.

Table 2
Comprehensive performance comparison between each stage of the cascade pipeline and the official baseline on the development set.

Pipeline Stages	FR Dev	EN Dev	DE Dev	Avg
Hybrid Index Coverage (Recall)				
BM25 (Top-30)	8.68%	54.18%	9.84%	24.23%
BGE-M3 (Top-100)	82.33%	82.94%	75.90%	80.39%
Ours: Hybrid retrieval	82.47%	84.32%	75.90%	80.90%
End-to-End System Performance (MRR@5)				
Official Baseline	45.84%	49.87%	37.67%	44.46%
Ours: Hybrid Routing Initial Ranking	64.65%	57.93%	47.47%	56.68%
Ours: LLM Listwise Reranking	70.33%	66.98%	57.01%	64.77%

As shown in Table 2, BM25 exhibits low recall when processing French (8.68%) and German (9.84%) tweets. This reflects the lexical sparsity when mapping informal tweets to formal English academic corpora. In contrast, BGE-M3 provides higher cross-lingual recall than BM25 across the three languages.

Based on this data distribution, our first-stage hybrid retrieval adopts an asymmetric truncation strategy. We use BGE-M3 (top-100) as the primary retriever for cross-lingual semantic matching, while restricting BM25 to the top-30 to capture exact entities (e.g., specific numerical values or abbreviations)

mentioned in the tweets. By limiting long-tail noise, this direct merging and deduplication mechanism increases the average cross-lingual recall to 80.90%.

The third stage (LLM listwise reranking) achieves an average MRR@5 of 64.77%, which is 20.31 percentage points above the official baseline. Initially, we attempted to deploy an LLM on the query side to rewrite informal tweets into academic-style queries before retrieval. However, our preliminary tests showed that autoregressive generation introduced extraneous vocabulary, diluting the original entity signals and degrading sparse retrieval performance.

Consequently, we bypassed pre-retrieval rewriting, positioning the LLM at the end of the pipeline for listwise reranking. This architecture allows the model to perform cross-document comparison and logical reasoning within a bounded candidate pool, helping filter out semantically similar hard negatives.

Table 3 reports the official hidden test results of our final submitted system. The cascade pipeline achieves an average MRR@5 of 63.94%, with the highest score on French (68.28%), followed by English (63.82%) and German (59.71%).

Table 3
Official hidden test results of the final cascade pipeline.

Language	MRR@5
FR	68.28%
EN	63.82%
DE	59.71%
Avg	63.94%

5.3. Error Analysis

The main remaining error source is candidate recall in the first stage. Although hybrid retrieval improves coverage over single-channel retrieval, the average Recall@130 remains 80.90%, indicating that some correct papers are absent from the candidate pool before cross-encoder and LLM reranking. These cases are unlikely to be corrected by reranking unless the correct paper is included in the candidate pool.

German queries are more challenging in our results. The German hybrid retrieval recall is 75.90%, lower than French (82.47%) and English (84.32%). BM25 provides limited recall for German, reaching 9.84% Recall@30. This indicates that direct lexical matching is weak when German social media posts are matched against an English-dominated scientific corpus.

We attribute the lower German performance to two factors. First, German queries are translated into English before dense retrieval and cross-encoder reranking. Translation may alter or remove disease names, drug names, abbreviations, numerical clues, and informal expressions. Although the LLM reranking stage receives both the original German query and its English translation, it is unlikely to recover cases where the correct paper has already been missed by the candidate generation or initial reranking stages. Second, German has the smallest training split, which limits task-specific adaptation for reranking. These factors help explain why German obtains a lower hidden-test MRR@5 than French and English.

5.4. Ablation Experiment

To validate the cascaded pipeline's core reranking components, we conduct an ablation study on the cross-encoder's fine-tuning strategies (prior to the final LLM reranking stage).

Table 4 reports the independent ranking performance of three strategies: zero-shot base model, language-specific fine-tuning, and unified multilingual fine-tuning. These experiments show performance differences when handling unbalanced multilingual resources, providing the empirical basis for our language-specific routing strategy. For French (a medium-resource language), the language-specific

Table 4
MRR@5 performance comparison of the cross-encoder under different fine-tuning strategies. (Note: To exclude the effect of LLM final ranking, only the Top-5 scores output by the pure cross-encoder are reported here.)

Fine-tuning Strategy	FR Dev	EN Dev	DE Dev
Zero-shot Base Model	56.97%	53.88%	45.82%
Language-Specific Fine-tuned Model	64.65%	57.51%	–
Unified Multilingual Fine-tuned Model	61.37%	57.92%	47.47%

model (64.65%) scores higher than the unified multilingual model (61.37%). This result is consistent with the interpretation that language-specific training may reduce cross-lingual interference from high-resource English data. For English (a high-resource language), the oversampled cross-lingual hard negatives may act as a regularizer; the unified multilingual model (57.92%) scores slightly higher than the language-specific model (57.51%). For German (a low-resource language), the unified multilingual model (47.47%) scores higher than the zero-shot base model (45.82%). This suggests that unified fine-tuning can use signals from other languages to partially compensate for the scarcity of German training data.

6. Conclusion

In this paper, we propose a three-stage cascaded retrieval pipeline for the multilingual scientific claim source retrieval task. By integrating hybrid retrieval, language-specific initial ranking, and LLM-based listwise reranking, our pipeline is designed to address lexical sparsity and cross-lingual matching challenges between informal tweets and formal academic corpora. Experiments show improved ranking performance over the official baseline across English, French, and German, while maintaining inference on a single consumer-grade GPU.

Future work will focus on three directions. First, we plan to improve sparse retrieval for French and German queries, where BM25 recall remains low. Possible extensions include entity normalization, abbreviation expansion, scientific terminology mapping, and bilingual query-term expansion. Second, we plan to replace the API-based LLM reranker with a locally deployed open-source model. This could reduce inference cost, latency, and service availability risks, while also enabling task-specific fine-tuning or preference optimization. Third, we plan to improve German processing by exploring entity-preserving translation, bilingual query representations, and German-oriented reranker training. These directions target the recall and reranking errors observed in the current system.

Acknowledgments

This work is supported by the National Social Science Foundation of China (No. 22BTQ101).

Declaration on Generative AI

During the preparation of this work, the authors used Gemini (by Google) for text translation and writing-style improvement. After using this tool and service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. M. Struß, S. Schellhammer, S. Dietze, et al., The CLEF-2026 CheckThat! lab: Advancing multilingual fact-checking, in: European Conference on Information Retrieval, Springer Nature Switzerland, Cham, 2026, pp. 325–335.

[2] J. Thorne, A. Vlachos, C. Christodoulopoulos, et al., 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), 2018, pp. 809–819.

[3] V. Karpukhin, B. Oguz, S. Min, et al., Dense passage retrieval for open-domain question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 6769–6781.

[4] S. Robertson, H. Zaragoza, The Probabilistic Relevance Framework: BM25 and Beyond, Now Publishers Inc, 2009.

[5] P. J. Sager, A. Kamaraj, B. F. Grewe, et al., Deep retrieval at CheckThat! 2025: Identifying scientific papers from implicit social media mentions via hybrid retrieval and re-ranking, arXiv preprint arXiv:2505.23250 (2025).

[6] R. Nogueira, K. Cho, Passage re-ranking with BERT, arXiv preprint arXiv:1901.04085 (2019).

[7] G. V. Cormack, C. L. A. Clarke, S. Buettcher, 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, 2009, pp. 758–759.

[8] S. Bruch, S. Gai, A. Ingber, An analysis of fusion functions for hybrid retrieval, ACM Transactions on Information Systems 42 (2023) 1–35.

[9] D. Wadden, S. Lin, K. Lo, et al., Fact or fiction: Verifying scientific claims, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 7534–7550.

[10] D. Wadden, K. Lo, L. L. Wang, et al., MultiVerS: Improving scientific claim verification with weak supervision and full-document context, in: Findings of the Association for Computational Linguistics: NAACL 2022, 2022, pp. 61–76.

[11] D. Wright, D. Wadden, K. Lo, et al., Generating scientific claims for zero-shot scientific fact checking, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 2448–2460.

[12] W. Sun, L. Yan, X. Ma, et al., Is ChatGPT good at search? investigating large language models as re-ranking agents, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 14918–14937.

[13] X. Ma, X. Zhang, R. Pradeep, et al., Zero-shot listwise document reranking with a large language model, arXiv preprint arXiv:2305.02156 (2023).

[14] J. Chen, S. Xiao, P. Zhang, et al., BGE m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, arXiv preprint arXiv:2402.03216 4 (2024).

[15] J. Tiedemann, Parallel data, tools and interfaces in OPUS, in: LREC, 2012, pp. 2214–2218.

[16] Z. Wang, Z. C. Lipton, Y. Tsvetkov, On negative interference in multilingual models: Findings and a meta-learning treatment, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 4438–4450.

[17] Y. Qu, Y. Ding, J. Liu, et al., RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 5835–5847.

A. LLM Listwise Reranking Prompts

The following templates are used in the final LLM listwise reranking stage. The placeholders `{k}`, `{tweet}`, `{tweet_de}`, `{tweet_en}`, and `{candidates}` are filled at inference time.

A.1. Monolingual Prompt for French and English Queries

```
You are a scientific paper identification reranker.

Task:
Given a social media post and {k} candidate scientific papers, rank the candidates by how likely the
    post is referring to, citing, discussing, or summarizing that exact paper.

Important:
- Do NOT simply choose papers with the same broad topic.
- Prefer the paper that best matches the exact entities, claims, findings, population, intervention,
    method, dataset, author/year clues, acronym, trial name, disease, drug, biomarker, or outcome
    mentioned in the post.
- Penalize candidates that are only topically similar but differ in the specific finding, population,
    disease, intervention, method, or conclusion.
- Penalize review papers if the post appears to refer to an original study, trial, dataset, or
    experiment.
- If the post is informal, sarcastic, or abbreviated, infer the likely scientific reference from the
    concrete clues.
- If multiple candidates are plausible, rank the one with the strongest exact-match evidence first.
- Candidate numbers are fixed. Output 5 distinct candidate numbers from 1 to {k}.

Social media post:
{tweet}

Candidate papers:
{candidates}

Return ONLY a JSON array of the 5 best candidate numbers, ranked from most likely to least likely.
Example: [3, 7, 1, 5, 2]

Output:
```

Listing 1: Prompt template for French and English listwise reranking.

A.2. Bilingual Prompt for German Queries

```
You are a scientific paper identification reranker.

Task:
Given an original German social media post, its English translation, and {k} candidate scientific
     papers, rank the candidates by how likely the post is referring to, citing, discussing, or
     summarizing that exact paper.

Use both the German original and the English translation.
The translation may be imperfect, so preserve clues from the original text such as named entities,
     acronyms, hashtags, author names, disease names, drug names, trial names, numbers, and quoted
     phrases.

Important:
- Do NOT simply choose papers with the same broad topic.
- Prefer the paper that best matches the exact entities, claims, findings, population, intervention,
     method, dataset, author/year clues, acronym, trial name, disease, drug, biomarker, or outcome
     mentioned in the post.
- Penalize candidates that are only topically similar but differ in the specific finding, population,
     disease, intervention, method, or conclusion.
- Penalize review papers if the post appears to refer to an original study, trial, dataset, or
     experiment.
- If the post is informal, sarcastic, or abbreviated, infer the likely scientific reference from the
     concrete clues.
- If multiple candidates are plausible, rank the one with the strongest exact-match evidence first.
- Candidate numbers are fixed. Output 5 distinct candidate numbers from 1 to {k}.

Original German post:
{tweet_de}

English translation:
{tweet_en}

Candidate papers:
{candidates}

Return ONLY a JSON array of the 5 best candidate numbers, ranked from most likely to least likely.
Example: [3, 7, 1, 5, 2]

Output:
```

Listing 2: Bilingual prompt template for German listwise reranking.