PEI at CheckThat! 2026: Investigating Query Normalization and Retrieve-and-Rerank Strategies for Linking Scientific Web Claims to Publications

CheckThat! at CLEF 2026

Farnoush Zeidi[1,*], Farnaz Zeidi[1], Roman Christof[1], Renate König[1] and Liam Childs[1]

[1]Host-Pathogen Interactions, Paul-Ehrlich-Institute, 63225 Langen, Germany

Abstract

This paper presents our approach for Task 1 of the CheckThat! Lab at CLEF 2026, which focuses on multilingual scientific source retrieval from noisy social media claims. The task requires retrieving the corresponding scientific publication for claims written in English, German, and French from a shared collection of scientific papers. We propose a retrieve-and-rerank baseline consisting of the Harrier-27b dense retriever and the Jina cross-encoder reranker. We also explore several extensions to improve the baseline, including multilingual query normalization through translation and noise removal, dense retriever and reranker fine-tuning, and large language model (LLM)-based reranking using summarized scientific abstracts. While the additional enhancement strategies did not consistently outperform the baseline system, query normalization improved retrieval performance for both French and German on the development set. Our best-performing system, based on the baseline retrieve-and-rerank pipeline with query normalization, achieved mean reciprocal rank at 5 (MRR@5) scores of 0.698 for English, 0.702 for French, and 0.645 for German on the official test set. The system ranked 6th for English and 7th for both German and French, placing 7th among 37 participating teams and within the top 20%. To support reproducibility, we also publicly release our baseline implementation.

Keywords

scientific claim linking, retrieve-and-rerank, dense retrieval, cross-encoder reranking, LLM-based reranking

1. Introduction

Linking implicit scientific claims shared on social media to their original publications is essential for evidence-based fact-checking, scholarly analysis, and informed public discourse [1]. The task is challenging because social media posts often paraphrase scientific findings without explicit identifiers such as DOIs or URLs [2] and frequently contain informal and noisy language, including hashtags, emojis, abbreviations, and slang [3].

To address this issue, Task 1 of the CLEF 2026 CheckThat! Lab [4, 5] focuses on multilingual scientific source retrieval from social media claims. Given a short social media post (claim) in English, French, or German, systems must retrieve the corresponding scientific publication from a collection of 10,000 candidate papers. The task presents several challenges, including bridging the gap between the informal and often noisy language used in social media and the formal language of scientific publications. Furthermore, multilingual retrieval increases the difficulty, as the availability of pretrained models and annotated datasets is generally lower for non-English languages than for English, making the development of robust Natural Language Processing (NLP) systems more challenging [6, 7]. Overall, the task requires systems to identify the correct scientific publications from multilingual social media claims despite the linguistic differences between informal social media language and formal scientific literature.

To tackle these retrieval challenges, some studies have adopted dense retrieval models for semantic matching and reported reasonable results [8], while others have turned to cross-encoders to achieve

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

*Corresponding author.

farnoush.zeidikolehparcheh@pei.de (F. Zeidi); farnaz.zeidi@pei.de (F. Zeidi); roman.christof@pei.de (R. Christof); renate.koenig@pei.de (R. König); liam.childs@pei.de (L. Childs)

more accurate retrieval through explicit query-document interaction modeling [9]. Recent studies have also explored large language model (LLM)-based rerankers, showing that they can be competitive with cross-encoders in some settings [10]. Other work has further leveraged large language models for query reformulation and summarization as a way to augment the input to these retrieval pipelines [11]. Previous studies have highlighted the effectiveness of query rewriting and translation strategies for improving retrieval performance in noisy and multilingual settings [12, 13]. These observations motivate the exploration of retrieval, reranking, and multilingual query normalization strategies for scientific source retrieval.

In our participation as the PEI team[1] in Task 1 of the CheckThat! Lab, we employ the Harrier dense retriever [14] together with the Jina cross-encoder reranker [15] as our baseline retrieval framework. Since multilingual social media claims frequently contain noisy and informal language, we first translate and reformulate all claims into clean English using Qwen3-4B [16], removing hashtags, emojis, URLs, and other noisy social media elements. We additionally investigate several strategies for improving different stages of the retrieval pipeline, including dense retriever fine-tuning with hard-negative examples, reranker fine-tuning with different sampling strategies, and reasoning-based LLM reranking using Qwen3-8B with summarized scientific abstracts to reduce prompt complexity. We experimentally evaluate these approaches on the development dataset. Although the explored approaches did not consistently outperform the baseline system, query normalization provided slight improvements for both French and German retrieval. Our best-performing system, the baseline retrieve-and-rerank pipeline with query normalization, ranked 7th among 37 participating teams on the official test-set leaderboard. To support reproducibility, the implementation of our baseline pipeline is publicly available online[2].

2. Related Work

Automated fact-checking has long been studied as a multi-step pipeline covering check-worthiness detection, evidence retrieval, and claim verification [17, 18]. Scientific claim retrieval becomes challenging because social media users frequently discuss scientific findings without citing the original publication, instead paraphrasing results using informal and noisy language [2, 3]. Sager et al. [19] further highlighted the gap between informal claim language and formal scientific text as a major challenge for automated fact-checking systems.

Dense retrieval and reranking approaches are widely used to address retrieval tasks involving semantic matching between queries and documents. Thakur et al. [20] evaluated dense retrieval models across heterogeneous domains and reported that domain-specific fine-tuning can substantially improve retrieval performance. Izacard et al. [8] proposed Contriever, a dense retriever trained with unsupervised contrastive learning, and reported strong zero-shot retrieval performance across domains without requiring labelled data. Prior work has also shown that hard-negative sampling can substantially improve dense retriever training [21]. Beyond dense retrieval, Nogueira and Cho [9] showed that combining retrieval with cross-encoder reranking leads to large improvements in ranking quality, while Gao et al. [22] demonstrated that reranker performance strongly depends on the quality of negative samples used during training. More recently, large language models have also been explored for reranking tasks. Sun et al. [23] investigated using ChatGPT for document reranking and reported competitive zero-shot performance compared to supervised rerankers, while Ma et al. [24] showed that LLM-based relevance reasoning can outperform BM25 in several retrieval settings. In related biomedical matching tasks, Christof et al. [25] explored LLM-based reranking with reasoning for linking biomedical text mentions to structured medical terminology entries. In this context, Feng et al. [26] noted that applying LLMs directly to long retrieval documents is often limited by context length constraints, motivating summarization-based reranking approaches.

[1]The PEI team is part of the Artificial Intelligence Working Group at the Paul Ehrlich Institute (PEI), an agency of the German Federal Ministry of Health.

[2]https://github.com/paul-ehrlich-institute/claim-linking-rag

Large language models have also been explored for improving the query side of retrieval pipelines. As discussed by Gao et al. [11], LLMs can be used for query reformulation to generate more retrieval-friendly queries. In multilingual retrieval settings, Zhang et al. [27] showed that translate-then-retrieve approaches provide strong and practical baselines for cross-lingual retrieval tasks. Ma et al. [28] further showed that rewriting noisy queries into cleaner and more retrieval-oriented text before retrieval can improve retrieval performance across multiple settings. DS@GT [29] also explored augmentation strategies to improve retrieval robustness.

The closest shared task to ours is Subtask 4b of the CLEF 2025 CheckThat! Lab [1], where systems were given English tweets implicitly referencing scientific papers and had to retrieve the correct paper from a candidate pool evaluated using Mean Reciprocal Rank at 5 (MRR@5). Several participating systems explored dense retrieval and reranking strategies. AIRwaves [30] fine-tuned a dual-encoder retriever using hard negatives and combined it with a SciBERT reranker, while Deep Retrieval [19] combined BM25 with dense retrieval and contextual reranking. DS@GT [29] further explored retrieval augmentation and reranking variants using augmented training data. However, all of these systems focused only on English retrieval. In CLEF 2026, the task was extended to multilingual retrieval involving English, French, and German [4], introducing additional challenges related to cross-lingual retrieval and multilingual query normalization. Building on prior work in dense retrieval, cross-encoder reranking, query reformulation, and LLM-based reranking, we investigate these strategies for scientific source retrieval from noisy multilingual social media claims.

3. Task and Data

3.1. Task Definition

Task 1 of the CLEF 2026 CheckThat! Lab focuses on multilingual scientific source retrieval from social media claims [4]. Given a social media post containing a scientific claim and a reference to a scientific publication, the goal is to retrieve the corresponding paper from a shared pool of candidate studies. The challenge arises because social media users often refer to scientific findings informally rather than citing them directly. System performance is evaluated using MRR@5 (see Section 4.6 for more details).

3.2. Dataset

The dataset, publicly available through the official Hugging Face repository[3], consists of social media claims in English, German, and French, each paired with its corresponding scientific publication. For each language, the organizers provide query sets together with a shared collection of 10,000 scientific papers used for retrieval across all languages. Each paper in the collection includes metadata such as the title, abstract, authors, and publication venue.

To better understand the characteristics of the scientific publication collection, we additionally analyzed the abstract lengths using the Qwen3 tokenizer. The abstracts contained an average of 365 tokens, with a standard deviation of 545 tokens and a maximum length of 10,001 tokens.

The dataset is divided into training, development, and test splits. All experiments, model selection, and hyperparameter tuning in this work were performed on the development set, since the test labels were not available during the official competition period. After the competition, the test labels were also officially released through the same Hugging Face repository. Table 1 summarizes the statistics of all dataset splits.

4. Methodology

To address the Source Retrieval for Scientific Web Claims task, we developed a baseline retrieve-and-rerank pipeline consisting of two main components: dense retrieval and reranking. In addition

[3] https://huggingface.co/datasets/sschellhammer/CT26_Task1_SourceRetrievalForScientificWebClaims

Table 1
Statistics of the dataset splits across all languages.

Language	Train	Development	Test	Total
English	14,977	3,905	6,076	24,958
French	2,807	702	1,220	4,729
German	1,460	386	876	2,722

to the baseline system, we explored several approaches for improving both retrieval and reranking performance. Since the social media queries are multilingual and include English, German, and French posts, and frequently contain informal language, abbreviations, hashtags, emojis, mentions, and URLs, we also investigated a query normalization strategy to improve retrieval effectiveness.

Figure 1 presents an overview of all retrieve-and-rerank pipelines explored in this work, including the baseline system, query normalization, dense retriever fine-tuning, cross-encoder fine-tuning, and LLM-based reranking.

4.1. Baseline

Our baseline system consists of a dense retriever followed by a cross-encoder reranker. For dense retrieval, we used the microsoft/harrier-oss-v1-27b [14] model to retrieve the top-50 candidate documents from the shared scientific publication collection. Each document representation combined multiple metadata fields, including title, abstract, authors, and venue information. The retrieved candidates were then reranked using the jinaai/jina-reranker-v3 [15] model, and the final top-5 ranked documents were used as the output predictions. Additional implementation details, preprocessing scripts, and retrieval pipeline configurations are available in our public repository.

Running the baseline retrieve-and-rerank pipeline on the development set required approximately 3 hours on a single NVIDIA A100 GPU.

4.2. Baseline with Query Normalization

To improve query quality, we introduced a query normalization pipeline before retrieval and reranking. Building on our previous work on a paraphrasing-based data augmentation strategy [31, 32], we adapted the same paraphrasing approach for query normalization to remove noisy social media elements and reformulate the original claims into cleaner, more formal, and retrieval-oriented queries. In addition to normalization, the German and French queries were translated into English, while the original English queries were kept unchanged. Therefore, all queries were normalized through noise removal and reformulation, but only the non-English queries were translated into English.

For this step, we used Qwen/Qwen3-4B [16]. The model was instructed to remove hashtags, emojis, mentions, and URLs, and rewrite the input into fluent and natural English while preserving the original semantic meaning of the claim. The normalization prompt is provided in Appendix A. For inference, we used a batch size of 8, temperature 0.6, and max_new_tokens=32768. We additionally enabled the reasoning capability of the model during normalization, which substantially increased generation length and inference time.

Processing the multilingual development sets required several hours on a single NVIDIA A100 GPU, ranging from approximately 3 hours for the German development set (386 queries) to approximately 6 hours for the French development set (702 queries), and approximately 31 hours for the substantially larger English development set (3,905 queries), with runtime increasing proportionally to dataset size.

4.3. Dense Retriever Fine-Tuning

To improve the retrieval quality of the dense retriever, we fine-tuned the microsoft/harrier-oss-v1-27b model using hard-negative mining. First, we retrieved the top-100 candidate documents for each training

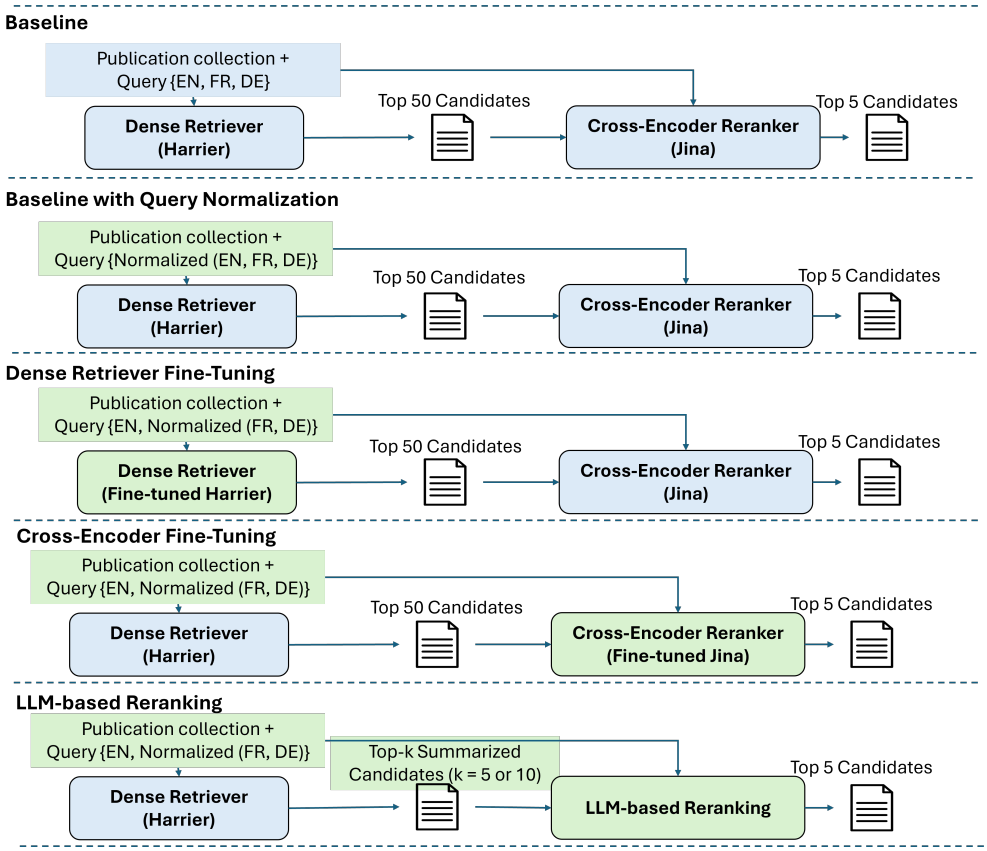

Figure 1: Overview of the retrieve-and-rerank pipelines explored in our experiments. Green components indicate modified or fine-tuned modules compared to the baseline system. Abbreviations: EN = English; FR = French; DE = German.

query using the baseline retriever (harrier-oss-v1-27b). We focused on difficult examples where the gold publication was not ranked within the first or second retrieved results.

For each selected query, we created one positive pair consisting of the query and the gold document, together with five hard-negative pairs consisting of highly ranked non-gold documents retrieved by the baseline model. This process produced a dataset of 5,522 training instances, which was divided into training and development subsets.

We fine-tuned the retriever using PEFT/LoRA [33] for two epochs with a batch size of 2, a learning rate of 2e-5, and the AdamW optimizer. We used a LoRA rank of $r = 32$, alpha of 64, and dropout of 0.05, applied to the query, key, value, and output projection layers. Training the dense retriever required approximately 18.6 hours on a single NVIDIA A100 GPU.

4.4. Cross-Encoder Fine-Tuning

Since reranking quality plays a critical role in retrieve-and-rerank systems, we also fine-tuned the jina-reranker-v3 cross-encoder model. Similar to the retriever experiments, we first extracted the top-100 retrieved documents for each query using the baseline retriever.

We explored three strategies for constructing reranker training data. In the first strategy, each query was paired with one positive example (gold document) and five negative examples sampled from highly ranked non-gold documents. In the second strategy, queries were divided into easy and hard examples according to the ranking position of the gold document. Easy queries were defined as cases where the gold document appeared within the top-10 retrieved results, while hard queries corresponded to lower-ranked gold documents. For easy queries, we created one positive pair and five negative pairs, while for hard queries, we created one positive pair and ten negative pairs. The third strategy followed the same setup as the second strategy, but defined easy queries as those where the gold document

appeared within the top-5 retrieved results. These strategies produced datasets of 115,464, 117,763, and 122,203 training pairs, respectively, which were divided into training and validation subsets.

We trained the reranker using the Hugging Face Transformers library with a batch size of 16, three epochs, a learning rate of 2e-5, and the AdamW optimizer. Training the reranker for the first, second, and third strategies required approximately 4.4, 4.5, and 4.7 hours, respectively, on a single NVIDIA A100 GPU.

4.5. LLM-based Reranking

We further explored LLM-based reranking using a reasoning-based language model. The main idea of this approach was to use an LLM to rerank candidate scientific publications generated by the dense retrieval model. Candidate documents were first retrieved using the baseline retriever (harrier-oss-v1-27b). We then used Qwen3-8B to rerank the top-5 and top-10 retrieved candidates. This approach consisted of two main stages: abstract summarization and LLM-based reranking.

Abstract Summarization. Because scientific abstracts are often long and substantially increase prompt length, we first summarized the abstracts before reranking. The summarization step generated a short plain-language summary of each paper together with key points describing findings, methods, datasets, or named entities likely to appear in social media discussions. The summarization prompt is provided in Appendix B.

For abstract summarization, we used `Qwen3-4B` with vLLM[4]. The model was configured with max_new_tokens=256, temperature=0.0, and batch_size=16. Processing all abstracts required approximately 26 minutes on a single NVIDIA A100 GPU.

LLM-based Reranking. After generating summarized representations for all abstracts, we used the LLM as a reranker in thinking mode. The reranking prompt included the publication title, authors, venue information, generated summary, and key points for each candidate document. The model was instructed to rank the candidate publications from most to least likely to be the publication implicitly referenced in the social media post. The reranking prompt is provided in Appendix C. If parsing the generated ranking failed due to issues such as non-terminating reasoning generation or malformed output, we used the original encoder-based ranking as a fallback result.

The reranking model (`Qwen3-8B`) was executed using vLLM with `max_new_tokens=2000`, `temperature=0.7`, `batch_size=8` and in thinking mode. Processing the reranking stage required approximately 4.5 hours on a single NVIDIA A100 GPU.

4.6. Evaluation Metric

System performance is evaluated using Mean Reciprocal Rank at 5 (MRR@5), a ranking-based evaluation metric commonly used in information retrieval tasks [34, 35]. The metric emphasizes the position of the first relevant retrieved document by giving higher importance to rankings where the correct scientific publication appears near the top of the retrieved candidates. If the relevant publication is not retrieved within the top five results, the contribution to the score is zero.

In our experiments, we conducted end-to-end evaluation for all explored pipelines using MRR@5, with evaluation performed separately for English, German, and French. Additionally, we analyzed retrieval performance using Hit Accuracy@k, which measures the proportion of queries for which the gold publication appears within the top-k retrieved results. We report Hit Accuracy at $k \in \mathbf{5}, \mathbf{20}, \mathbf{50}, \mathbf{100}$ to better understand the recall behavior of the retrieval component.

[4] https://github.com/vllm-project/vllm

Table 2
Development-set performance of all explored retrieval and reranking pipelines evaluated using MRR@5.

Pipeline	Query Setting	English	French	German
DR-only	Original {EN, FR, DE}	0.6897	0.6886	0.5851
Baseline	Original {EN, FR, DE}			
DR → CE-RR		**0.7163**	0.7188	0.6189
Baseline + Query Normalization	Normalized {EN, FR, DE}			
DR → CE-RR		0.6871	**0.7209**	**0.6286**
Dense Retriever Fine-Tuning	Original {EN}, Normalized {FR, DE}			
FT-DR → CE-RR		0.6989	0.7183	0.5933
Cross-Encoder Fine-Tuning	Original {EN}, Normalized {FR, DE}			
DR → FT-CE-RR				
Training Data: Strategy 1		0.6307	0.6751	0.5862
Training Data: Strategy 2		0.6462	0.6967	0.5937
Training Data: Strategy 3		0.6547	0.7083	0.5881
LLM-based Reranking	Original {EN}, Normalized {FR, DE}			
DR → LLM-RR				
Top-5 Summarized Candidates		0.7103	0.7170	0.6054
Top-10 Summarized Candidates		0.7044	0.7139	0.5962

Abbreviations: EN = English; FR = French; DE = German; DR = Dense Retriever; CE-RR = Cross-Encoder Reranker; FT = Fine-Tuned; LLM-RR = LLM-based Reranker.
* All reranking experiments used the top-50 retrieved candidates, except for the LLM-based reranking experiments, which used summarized top-5 and top-10 candidates.

5. Results and Discussion

The results of the different explored approaches on the development set for each language are presented in Table 2. Detailed analyzes and discussions of each pipeline are provided in the following subsections.

5.1. Retriever Recall Analysis

Before applying reranking, we first analyzed the retrieval quality of the dense retriever alone using Hit Accuracy@k. This metric measures the proportion of queries for which the gold publication appears within the top-k retrieved results. Table 3 presents the retrieval recall results for different values of k.

As shown in Table 3, the dense retriever already achieved strong recall performance across all languages. The gold publication appeared within the top-50 retrieved candidates for approximately 87.8% of German queries, 90.6% of English queries, and 89.3% of French queries. These results indicate that the retriever was generally successful at retrieving the relevant publication within the candidate pool, motivating the use of reranking models to improve the ordering of the retrieved documents rather than focusing only on retrieval recall.

Table 3
Hit Accuracy@k results of the dense retriever on the development set.

Language	@5	@10	@20	@50	@100
German	0.6865	0.7358	0.7979	0.8782	0.9067
English	0.7877	0.8305	0.8681	0.9058	0.9296
French	0.7849	0.8319	0.8604	0.8932	0.9145

This observation is further supported by the retriever-only results shown in Table 2. Although the retriever alone achieved relatively strong MRR@5 scores, adding the reranker consistently improved

performance across all languages. The baseline retrieve-and-rerank pipeline improved MRR@5 from 0.6897 to 0.7163 for English, from 0.6886 to 0.7188 for French, and from 0.5851 to 0.6189 for German.

5.2. Comparison of Retrieval and Reranking Approaches

The baseline retrieve-and-rerank pipeline achieved strong performance across all languages, obtaining MRR@5 scores of 0.7163 for English, 0.7188 for French, and 0.6189 for German. Compared to the retriever-only setting, the reranker consistently improved the ranking quality of the retrieved candidate publications.

Query normalization improved performance for both German and French queries, increasing MRR@5 from 0.7188 to 0.7209 for French and from 0.6189 to 0.6286 for German. In contrast, normalization reduced English performance from 0.7163 to 0.6871. These results suggest that translation and reformulation mainly benefited the non-English queries by converting multilingual and noisy social media posts into cleaner English retrieval-oriented queries. Based on this observation, the remaining experiments used normalized German and French queries while keeping the original English queries unchanged.

Fine-tuning the dense retriever using hard-negative mining did not improve performance over the baseline system, likely because the original retriever already achieved strong recall performance, leaving limited room for further improvement.

Similarly, none of the explored cross-encoder fine-tuning strategies outperformed the original reranker. Among the three strategies, Strategy 3 achieved the best performance, where hard queries received more negative training examples based on a top-5 retrieval threshold, obtaining MRR@5 scores of 0.6547 for English, 0.7083 for French, and 0.5881 for German. These results suggest that reranker fine-tuning on automatically generated hard negatives may have reduced the generalization capability of the reranker.

5.3. LLM-based Reranking Analysis

The LLM-based reranking approach improved upon the retriever-only results but did not outperform the cross-encoder reranker. Using summarized top-5 candidates achieved the best performance, while increasing the number of reranked candidates from top-5 to top-10 slightly reduced performance across all languages. The reduced performance with a larger candidate pool may be attributed to the inclusion of less relevant candidates, which introduces additional noise, strains the model's context and reasoning capacity, and increases the overall complexity of the ranking task.

As noted above, candidates were summarized prior to reranking to reduce prompt length and computational cost. The generated summaries and key points were intended to provide compact representations of the abstracts while aiming to preserve the most relevant information for reranking. While manual inspection indicated that key information was generally well preserved, we did not conduct a systematic evaluation of information coverage across all samples. An example of the generated summaries is provided in the Appendix D.

We additionally observed practical limitations related to reasoning-based generation. Approximately 29% of prompts produced outputs that could not be parsed correctly, mainly because the generation exceeded the token limit or the reasoning process did not terminate properly. The failure rates were approximately 28% for English, 35% for German, and 32% for French queries. To mitigate this issue, we applied a fallback strategy that reused the original encoder-based ranking whenever parsing failed.

5.4. Official Test-Set Results

For the official test submission, we selected our best-performing pipeline, namely the baseline retrieve-and-rerank system with query normalization applied only to German and French queries. According to the official competition leaderboard,[5] our system ranked 7th among 37 participating teams, placing

[5] https://checkthat.gitlab.io/clef2026/task1/

Table 4
Top-ranked systems on the official CLEF CheckThat! Task 1 test-set leaderboard evaluated using MRR@5.

Team	Average MRR@5	German	English	French
Claim2Source	**0.7628**	0.7153	**0.7819**	**0.7912**
Outsider	0.7580	**0.7228**	0.7802	0.7709
Boyes Boys	0.7464	0.7070	0.7571	0.7752
CheckMate	0.7194	0.6861	0.7343	0.7378
RETRIX	0.7108	0.6806	0.7117	0.7400
MeVer@CERTH	0.6839	0.6615	0.6829	0.7074
PEI (Ours)	0.6819	0.6455	0.6980	0.7021

within the top 20% overall. The official leaderboard results are presented in Table 4. Our submission ranked 6th for English and 7th for both German and French.

Although our approach relied primarily on retrieval and reranking techniques without ensemble-based optimization, it achieved competitive multilingual retrieval performance across all languages.

6. Limitations

Our work has several limitations that provide directions for future research. First, although multilingual query normalization improved retrieval performance for French and German, we did not perform a systematic quality analysis of the normalized queries or summarized abstracts. Such an analysis could determine whether the original semantic meaning was preserved or whether the generation process introduced noise or omitted retrieval-relevant information. Second, we did not conduct a detailed error analysis for each retrieval pipeline to better understand the strengths and weaknesses of the explored approaches. For example, further investigation could explain why query normalization slightly reduced retrieval performance for English while improving performance for French and German. Finally, fine-tuning both the dense retriever and the cross-encoder resulted in lower retrieval performance than the original pretrained models. Understanding the reasons behind this behavior, including the influence of training data quality, negative sampling strategies, and fine-tuning methodology, remains an important direction for future work.

7. Conclusion

In this paper, we presented the PEI system for CheckThat! 2026 Task 1 on multilingual scientific claim source retrieval. Our approach was based on a retrieve-and-rerank pipeline combining dense retrieval, cross-encoder reranking, query normalization, retriever and reranker fine-tuning, and LLM-based reranking. For query normalization, we translated non-English queries into English, removed social media noise, and reformulated the posts into fluent formal text. We used harrier-oss-v1-27b as the baseline dense retriever and `jina-reranker-v3` as the cross-encoder reranker. Experimental results on the development set showed that the retriever alone successfully retrieved the gold publication within the top-50 candidates for approximately 87%–90% of the queries across all languages, indicating that the main challenge lies in the reranking stage. Adding the reranker substantially improved retrieval quality compared to the retriever-only setting. Query normalization was particularly beneficial for German and French queries, while retriever and reranker fine-tuning did not outperform the original baseline models. We further explored reasoning-based LLM reranking using summarized scientific abstracts. Although this approach produced competitive results, it suffered from generation instability and high computational cost. Our best-performing system, based on the baseline retrieve-and-rerank pipeline with multilingual query normalization, achieved 7th place among 37 participating teams in the official CLEF 2026 leaderboard, demonstrating competitive multilingual retrieval performance. To support reproducibility, we publicly release the implementation of our baseline system. For future work,

we plan to investigate hybrid sparse-dense retrieval methods, more robust reasoning-based reranking approaches, and adaptive retrieval strategies for difficult scientific claims.

Acknowledgments

This work was carried out by the Artificial Intelligence Working Group, led by Dr. Liam Childs within the FoG3 Research Group, under the leadership of Dr. Renate König at the Paul Ehrlich Institute (PEI).

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT[6] for grammar correction, language refinement, and writing assistance. Further, the authors used Grammarly[7] for grammar and spelling correction. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. Hafid, Y. S. Kartal, S. Schellhammer, K. Boland, D. Dimitrov, S. Bringay, K. Todorov, S. Dietze, Overview of the CLEF-2025 CheckThat! lab task 4 on scientific web discourse, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), volume 4038 of *CEUR Workshop Proceedings*, RWTH Aachen, 2025, pp. 722–732. URL: https://ceur-ws.org/Vol-4038/paper_54.pdf.

[2] 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. doi:10.1145/3720553.3746676.

[3] P. Bhargava, N. Spasojevic, G. Hu, Lithium nlp: A system for rich information extraction from noisy user generated text on social media, in: Proceedings of the 3rd Workshop on Noisy User-generated Text, 2017, pp. 131–139. doi:10.18653/v1/W17-4417.

[4] 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.

[5] 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.

[6] F. Zeidi, M. Messelhäusser, R. Christof, X. D. Wang, U. Leser, D. Mentzer, R. König, L. Childs, Medner. de: Medicinal product entity recognition in german-specific contexts, IEEE Access (2025). doi:10.1109/ACCESS.2025.3607266.

[7] J. Frei, F. Kramer, Annotated dataset creation through large language models for non-english medical nlp, Journal of Biomedical Informatics 145 (2023) 104478. doi:10.1016/j.jbi.2023.104478.

[8] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, E. Grave, Unsupervised dense information retrieval with contrastive learning, Transactions on Machine Learning Research (2022). URL: https://openreview.net/forum?id=jKN1pXi7b0.

[9] R. Nogueira, K. Cho, Passage re-ranking with BERT, arXiv preprint arXiv:1901.04085 (2019). doi:10.48550/arXiv.1901.04085.

[6] https://chatgpt.com

[7] https://www.grammarly.com

[10] H. Déjean, S. Clinchant, T. Formal, A thorough comparison of cross-encoders and llms for reranking splade, arXiv preprint arXiv:2403.10407 (2024). URL: https://arxiv.org/abs/2403.10407. doi:10.48550/arXiv.2403.10407.

[11] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, Retrieval-augmented generation for large language models: A survey, arXiv preprint arXiv:2312.10997 (2023). doi:10.48550/arXiv.2312.10997.

[12] Z. Li, J. Wang, Z. Jiang, H. Mao, Z. Chen, J. Du, Y. Zhang, F. Zhang, D. Zhang, Y. Liu, Dmqr-rag: Diverse multi-query rewriting for rag, arXiv preprint arXiv:2411.13154 (2024). URL: https://arxiv.org/abs/2411.13154. doi:10.48550/arXiv.2411.13154.

[13] R. Goworek, O. Macmillan-Scott, E. B. Özyiğit, Bridging language gaps: Advances in cross-lingual information retrieval with multilingual llms, arXiv preprint arXiv:2510.00908 (2025). URL: https://arxiv.org/abs/2510.00908. doi:10.48550/arXiv.2510.00908.

[14] Microsoft Research, Harrier: A large-scale dense retrieval model, 2025. Model available at https://huggingface.co/microsoft/harrier-oss-v1-27b.

[15] Jina AI, Jina reranker v3: A cross-encoder reranking model, 2024. Model available at https://huggingface.co/jinaai/jina-reranker-v3.

[16] Qwen Team, Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025). doi:10.48550/arXiv.2505.09388.

[17] R. Panchendrarajan, A. Zubiaga, Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research, Natural Language Processing Journal 7 (2024) 100066. doi:10.1016/j.nlp.2024.100066.

[18] Z. Guo, M. Schlichtkrull, A. Vlachos, A survey on automated fact-checking, Transactions of the Association for Computational Linguistics 10 (2022) 178–206. doi:10.1162/tacl_a_00454.

[19] 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, arXiv preprint arXiv:2505.23250 (2025). doi:10.48550/arXiv.2505.23250.

[20] 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 Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2021. URL: https://openreview.net/forum?id=wCu6T5xFjeJ.

[21] J. Zhan, J. Mao, Y. Liu, J. Guo, M. Zhang, S. Ma, Optimizing dense retrieval model training with hard negatives, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 1503–1512. doi:10.1145/3404835.3462880.

[22] L. Gao, Z. Dai, J. Callan, Rethink training of bert rerankers in multi-stage retrieval pipeline, in: Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 – April 1, 2021, Proceedings, Part II, Springer-Verlag, Berlin, Heidelberg, 2021, p. 280–286. doi:10.1007/978-3-030-72240-1_26.

[23] 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, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 14918–14937. doi:10.18653/v1/2023.emnlp-main.923.

[24] X. Ma, X. Zhang, R. Pradeep, J. Lin, Zero-shot listwise document reranking with a large language model, arXiv preprint arXiv:2305.02156 (2023). doi:10.48550/arXiv.2305.02156.

[25] R. Christof, F. Zeidi, M. Messelhäußer, D. Mentzer, R. Koenig, L. Childs, A. Mehler, MedLinkDE – MedDRA entity linking for German with guided chain of thought reasoning, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2025, pp. 31581–31593. doi:10.18653/v1/2025.emnlp-main.1609.

[26] J. Feng, W. Liu, Z. Dou, Sumrank: Aligning summarization models for long-document listwise reranking, arXiv preprint arXiv:2603.24204 (2026). doi:doi.org/10.48550/arXiv.2603.

24204.
[27] J. Lin, D. Alfonso-Hermelo, V. Jeronymo, E. Kamalloo, C. Lassance, R. Nogueira, O. Ogundepo, M. Rezagholizadeh, N. Thakur, J.-H. Yang, et al., Simple yet effective neural ranking and reranking baselines for cross-lingual information retrieval, arXiv preprint arXiv:2304.01019 (2023). doi:10.48550/arXiv.2304.01019.
[28] X. Ma, Y. Gong, P. He, H. Zhao, N. Duan, Query rewriting in retrieval-augmented large language models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 5303–5315. doi:10.18653/v1/2023.emnlp-main.322.
[29] J. Schofield, S. Tian, H. T. T. Truong, M. Heil, Ds@gt at checkthat! 2025: Exploring retrieval and reranking pipelines for scientific claim source retrieval on social media discourse, arXiv preprint arXiv:2507.06563 (2025). doi:10.48550/arXiv.2507.06563.
[30] 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, arXiv preprint arXiv:2509.19509 (2025). doi:10.48550/arXiv.2509.19509.
[31] F. Zeidi, R. Christof, F. Zeidi, R. König, L. Childs, Pei at# smm4h-heard 2026: Enhancing patient metadata detection via hypothesis-conditioned classification and paraphrase-based data augmentation, in: Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, 2026, pp. 146–153. doi:10.18653/v1/2026.smm4h-1.24.
[32] F. Zeidi, R. Christof, R. König, L. Childs, Pei at#smm4h-heard 2025: Enhancing adverse event detection via error-driven data augmentation, in: Proceedings of the 19th International AAAI Conference on Web and Social Media Workshops (ICWSM 2025), 2025. URL: https://doi.org/10.36190/2025.56. doi:10.36190/2025.56.
[33] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021. doi:10.48550/arXiv.2106.09685.
[34] S. Knollmeyer, O. Caymazer, L. Koval, M. U. Akmal, S. Asif, S. George Mathias, D. Grossmann, Benchmarking of retrieval augmented generation: A comprehensive systematic literature review on evaluation dimensions, evaluation metrics and datasets, in: Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2024), SciTePress, 2024, pp. 137–148.
[35] Z. Rackauckas, A. Camara, J. Zavrel, Evaluating rag-fusion with ragelo: An automated elo-based framework, arXiv preprint arXiv:2406.14783 (2024).

A. Query Normalization Prompt

The following prompt was used for query normalization and translation of non-English social media posts into clean retrieval-oriented English queries.

> **Translate the following noisy tweet into clean, natural English.**
>
> **Guidelines:**
>
> - Remove noise such as emojis, hashtags, URLs, mentions, and repeated characters
> - Correct informal or misspelled words
> - Rewrite the sentence in fluent, natural English
> - Preserve the original meaning
> - Output ONLY one clean sentence
> - Do NOT explain anything
>
> **Tweet:** `<input tweet>`
>
> **Clean English sentence:**

B. Abstract Summarization Prompt

You are a scientific summarisation assistant.

Given a publication title, venue, and abstract, produce:

1. A one- or two-sentence plain-language SUMMARY of the paper's main contribution ($\leq$ 40 words)
2. A bullet list of 3–6 KEY POINTS: concrete facts, findings, methods, datasets, or named entities that someone might mention when tweeting about this paper.

Format your response exactly as:

```
SUMMARY: <summary text>

KEY POINTS:
- <point 1>
- <point 2>
```

Be concise and factual. No filler, no opinions, and no caveats. Use the same technical terms as the abstract.

C. LLM Re-ranking Prompt

You are an expert academic assistant.

The user shows you a social-media post (a tweet) that implicitly cites a scientific publication, plus a short list of candidate publications (each described by title, authors, venue, summary, and key points).

Your job is to rank the candidates from MOST to LEAST likely to be the publication the tweet is citing.

Consider:

- topical overlap
- named entities (including author names mentioned in the tweet)
- specific findings or claims
- venue
- timing

Be strict: a candidate is highly relevant only when it plausibly *is* the cited work, not merely related to its topic.

Respond with a single line containing the candidate labels in ranked order, best first, separated by >.

Example for 5 candidates:

```
3 > 1 > 5 > 2 > 4
```

Use only the labels that were given to you. Do not add explanations.

D. Example of Generated Abstract Summarization

Title: "Hypericum perforatum and Its Ingredients Hypericin and Pseudohypericin Demonstrate an Antiviral Activity against SARS-CoV-2"

Abstract:

For almost two years, the COVID-19 pandemic has constituted a major challenge to human health, particularly due to the lack of efficient antivirals to be used against the virus during routine treatment interventions. Multiple treatment options have been investigated for their potential inhibitory effect on SARS-CoV-2. Natural products, such as plant extracts, may be a promising option, as they have shown antiviral activity against other viruses in the past. Here, a quantified extract of …

Generated Summary:

Hypericum perforatum extracts and its components hypericin and pseudohypericin show antiviral activity against SARS-CoV-2, offering a potential natural treatment option for COVID-19.

Generated Key Points:

- Hypericum perforatum extract demonstrates antiviral activity against SARS-CoV-2.
- Hypericin and pseudohypericin are key components with antiviral effects.
- The study investigates natural products as potential treatments for COVID-19.
- The findings suggest possible use of plant-based compounds in antiviral therapy.
- The research addresses the need for effective antiviral treatments during the pandemic.