Claim2Source at CheckThat! 2026: Improving Multilingual Scientific Claim–Source Retrieval with Verification-based Re-Ranking

Notebook for the CheckThat! Lab at CLEF 2026

Tobias Schreieder[1,*], Harsh Khandelwal[2], Yu-Ling Zhong[1] and Michael Färber[1]

[1]TU Dresden & ScaDS.AI Dresden/Leipzig, Dresden, Germany
[2]Friedrich-Alexander University of Erlangen–Nuremberg, Erlangen, Germany

Abstract

Multilingual scientific claim–source retrieval aims to identify the scientific publication supporting a claim shared on social media. This task is challenging because claims often differ from source publications in terms of language, wording, and level of detail, which weakens the connection between claims and their underlying evidence. In this paper, we present our approach for the CheckThat! 2026 Lab Task 1: *Source Retrieval for Scientific Web Claims*. We propose a multi-stage retrieval framework for multilingual scientific claim–source retrieval that combines structured claim and source representations with progressive candidate refinement. To address multilingual retrieval challenges, the framework employs bilingual claim representations, metadata-enhanced source representations, and language-specific adaptation of dense retrieval models. Building on this setup, a first-stage retriever generates an initial pool of candidate sources, after which similarity-based re-ranking improves the ranking of highly relevant sources and verification-based re-ranking identifies the candidate source that best supports the claim using verification signals. Our approach achieves an average MRR@5 score of 0.7628 across English, German, and French claims, ranking first on the CheckThat! 2026 leaderboard.

Keywords

Fact Verification, Scientific Claim-Source Retrieval, Multilinguality, Re-Ranking, Large Language Model

1. Introduction

Social media platforms such as X (formerly Twitter) have become important channels for communicating and discussing scientific findings, particularly in fast-moving domains such as health and biomedicine [1]. Scientific claims shared online often discuss research findings without explicitly attributing them to their original scientific source. Identifying this source is challenging because claims often contain only partial descriptions of research findings and may rephrase or simplify the original evidence, creating a semantic mismatch between claims and their underlying publications [2, 3]. This problem becomes even more difficult when claims and source documents are written in different languages, requiring retrieval systems to align semantically equivalent content across heterogeneous representations [4, 5].

Attributing claims to their supporting evidence is an important component of automated fact verification systems, which aim to distinguish factual information from misinformation shared online. This problem is studied in the CheckThat! 2026 Lab, which focuses on advancing multilingual fact-checking [6], through Task 1: *Source Retrieval for Scientific Web Claims* [7]. The task was introduced in 2025 [8] and is extended in 2026 toward a multilingual setting with English, German, and French claims [7]. Throughout this paper, we refer to this task as multilingual scientific claim–source retrieval. While previous systems mainly focused on improving retrieval performance through multi-stage retrieval and re-ranking pipelines [9, 10], cross-lingual retrieval introduces additional challenges, as

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
tobias.schreieder@tu-dresden.de (T. Schreieder); harsh.khandelwal@fau.de (H. Khandelwal); yu-ling.zhong@mailbox.tu-dresden.de (Y. Zhong); michael.faerber@tu-dresden.de (M. Färber)
0009-0000-8268-4204 (T. Schreieder); 0009-0004-0237-5735 (H. Khandelwal); 0009-0007-2361-4944 (Y. Zhong); 0000-0001-5458-8645 (M. Färber)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

scientific literature remains predominantly available in English.

Building on recent findings that multi-stage retrieval and re-ranking improve scientific claim–source retrieval performance, we present a retrieval framework for multilingual settings, designed to address the challenges arising when claims and source publications are expressed in different languages. To handle multilingual claims, we employ bilingual claim representations and investigate language-specific adaptation of dense retrieval models. We further propose a novel three-stage retrieval architecture consisting of dense candidate generation, similarity-based re-ranking, and a verification-based re-ranking stage for progressive candidate refinement. In contrast to conventional retrieval pipelines that primarily rely on similarity estimation, the final stage directly incorporates verification signals into the retrieval process to improve source selection. Our contributions are summarized as follows:

1. We introduce bilingual claim representations that combine original and translated claim formulations, preserving language-specific information while reducing language mismatch.
2. We fine-tune GritLM-7B on multilingual training data spanning English, German, and French to improve retrieval performance across different claim languages.
3. We develop metadata-enhanced source representations that combine publication content (title and abstract) with contextual metadata (authors and venue).
4. We propose a novel three-stage retrieval framework comprising first-stage retrieval, similarity-based re-ranking, and a verification-based re-ranking stage using a large language model (LLM) for source selection that directly incorporates verification signals into ranking decisions.

2. Related Work

Scientific Claim–Source Retrieval. Scientific claim–source retrieval aims to identify the scientific publication underlying a claim expressed in social media or online discourse. In contrast to conventional evidence retrieval tasks, scientific claims frequently reference publications only implicitly and often differ substantially from scientific texts in language and style. This problem has been studied in the context of the CheckThat! Lab, which has established scientific claim–source retrieval as a benchmark task for evaluating retrieval approaches in realistic multilingual settings [8, 6]. Recent systems have explored multi-stage retrieval pipelines to bridge the gap between informal social media claims and formal scientific publications. Sager et al. [9] demonstrated that combining sparse and dense retrieval with LLM-based re-ranking substantially improves retrieval performance. Staudinger et al. [10] showed that listwise re-ranking consistently outperforms pointwise approaches. Schofield et al. [3] investigated data augmentation and claim reformulation strategies, showing that combining multiple claim representations can improve retrieval performance, whereas replacing original claims may remove important contextual information. The effectiveness of zero-shot style transfer methods for scientific claim–source retrieval was shown to depend strongly on the underlying retrieval model, with most sparse and hybrid models benefiting from more formal claim representations [2]. These findings indicate that scientific claim–source retrieval benefits from multi-stage architectures with candidate re-ranking and adaptation of claim and document representations, which motivate the design choices of our proposed framework.
Multilingual and Cross-Lingual Retrieval. Cross-lingual retrieval introduces additional challenges because claims and source documents may differ in language while still requiring alignment at the semantic level. Recent work has shown that dense multilingual representations can effectively align semantically similar queries and documents across different languages [4]. Lawrie et al. [11] demonstrated that multilingual neural retrieval models can achieve performance close to translation-based retrieval pipelines while reducing indexing costs. Furthermore, Litschko et al. [5] showed that cross-lingual retrieval can substantially benefit from in-domain fine-tuning of multilingual encoders. In the context of scientific retrieval, Valentini et al. [12] found that multilingual dense representations and translation-based methods offer complementary strengths for cross-lingual retrieval, with translation particularly benefiting sparse retrieval models. Overall, these studies highlight that retrieval strategies depend strongly on the retrieval setting and underlying models, motivating the exploration of complementary approaches such as translation, bilingual representations, and language-specific adaptation.

Table 1
CheckThat! 2026 dataset statistics across train, validation, and test splits for English (EN), German (DE), and French (FR) claims. The retrieval corpus consists of 10,000 scientific publications used as candidate sources.

Split	EN	DE	FR
Train	14,977	1,460	2,807
Validation	3,905	386	702
Test	6,076	876	1,220

Re-Ranking and Verification. Re-ranking has become a central component of scientific claim–source retrieval, as the best-performing approaches in the CheckThat! 2025 shared task predominantly relied on retrieval pipelines with re-ranking stages [8]. Classical re-ranking approaches primarily employ semantic similarity estimation to refine candidate rankings [10]. More recently, LLMs have been integrated into retrieval pipelines for candidate re-ranking. Existing work explored LLM-based source filtering techniques to improve the selection of retrieved candidates [13, 14]. Beyond candidate ranking, verifier feedback has been used to optimize evidence retrieval by measuring how useful retrieved evidence is for downstream claim verification [15]. Together, these developments underscore that verification signals can provide complementary information beyond semantic similarity. This motivates our use of verification-based re-ranking for refining highly ranked candidate sources.

3. Dataset

The experiments are conducted on the scientific claim–source retrieval dataset used in CheckThat! 2026[1] [6]. The dataset contains social media claims paired with their corresponding scientific source publications. Claims are available in English, German, and French, while sources are in English. Table 1 summarizes the distribution across languages and data splits. The objective is to identify the scientific publication referenced by a claim from a collection of 10,000 candidate documents within a known-item retrieval setting. Compared to the previous edition, the multilingual setup increases task complexity because claims and source documents may differ not only in style but also in language, requiring retrieval systems to align semantically equivalent information across languages. We use the training split for model fine-tuning and the validation split for model selection and ablation studies, while the final system performance is evaluated on the official test set.

4. Methodology

Our approach combines tailored claim and source representations with a multi-stage retrieval framework for multilingual scientific claim–source retrieval (see Figure 1). First, we construct claim and source representations designed to reduce language mismatch and expose additional source signals. Subsequently, candidate sources are progressively refined through a three-stage retrieval process consisting of first-stage retrieval, similarity-based re-ranking, and verification-based re-ranking.

4.1. Claim and Source Representation

Social media claims and scientific source publications may differ substantially in language and level of detail, while correct source attribution often depends on signals beyond the publication content alone. To address this, we introduce bilingual claim representations and metadata-enhanced source representations that reduce language mismatch and incorporate additional contextual source signals.

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

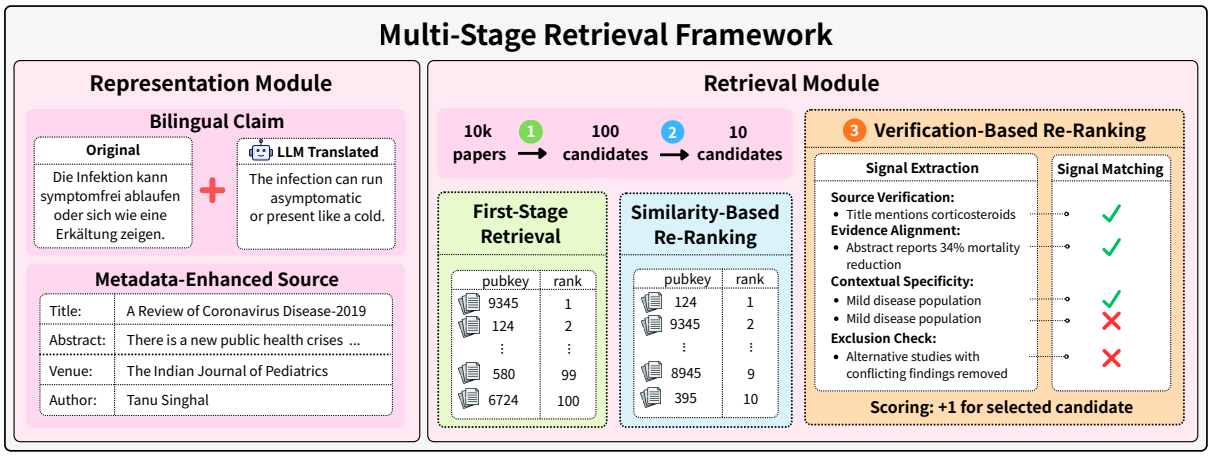


Figure 1: Multi-stage retrieval framework for multilingual scientific claim–source retrieval. The framework consists of a representation module and a retrieval module. The representation module constructs suitable representations for claims and sources: claims are represented as original–translated claim pairs, while sources are encoded using metadata-enhanced structured templates. The retrieval module progressively refines candidate selection through first-stage retrieval, similarity-based re-ranking, and verification-based re-ranking.

4.1.1. Bilingual Claim Representation

We translate German and French claims into English in a zero-shot setting using the open-source LLM *Qwen/Qwen3-8B*. Rather than replacing the original claim formulation, we construct a bilingual representation by combining the original and translated versions of each claim. This preserves language-specific information while providing an English formulation better aligned with the predominantly English source collection. The translation prompt is provided in Figure 2 in Appendix.

4.1.2. Metadata-Enhanced Source Representation

Scientific claims often refer to publications indirectly through mentions of authors, institutions, or publication venues rather than explicit citations. To capture such signals, we augment source representations with structured publication metadata in addition to the textual publication content. Specifically, each source publication is represented using labeled fields for the *title*, *abstract*, *authors*, and *venue*. To reduce noise from long author lists, only the first and last three authors are included.

4.2. Multi-Stage Retrieval Framework

Correctly identifying the scientific source underlying a social media claim often requires progressive refinement, since the two may differ substantially in terms of how information is expressed and which information is covered. We therefore develop a three-stage retrieval framework consisting of first-stage retrieval for candidate generation, similarity-based re-ranking for semantic candidate refinement, and verification-based re-ranking for identifying the candidate source that best supports the claim. Multiple open-source models are considered at each stage, and the best-performing configuration is selected for the final CheckThat! 2026 submission.

4.2.1. First-Stage Retrieval

The first stage of our framework generates an initial ranking of candidate sources from a collection of 10,000 scientific publications. To capture different retrieval paradigms, we evaluate both sparse and dense retrieval models, while additionally investigating language-specific adaptation of the best-performing dense retrieval model across different claim languages.

BM25. *BM25* is a sparse retrieval model based on term frequency and inverse document frequency [16, 17]. Its retrieval behavior depends strongly on overlap between claim and source terminology.

GTR. *gtr-t5-xl* is a dense retrieval model with 11B parameters based on the T5 architecture and trained on large-scale retrieval tasks [18]. The model is designed to capture semantic similarity.
E5. *intfloat/e5-large-v2* is a dense retrieval model with 335M parameters trained using contrastive learning on large-scale text-pair datasets for semantic retrieval [19].
GritLM. *GritLM/GritLM-7B* is a dense retrieval model with 7B parameters based on a decoder-only transformer architecture. It is trained with generative representational instruction tuning (GRIT), jointly optimizing generation and retrieval objectives within a unified model [20]. The resulting model achieves strong retrieval performance in zero-shot settings while retaining generative capabilities.
GritLM-F. We further construct a fine-tuned variant of GritLM, using low-rank adaptation (LoRA) on the multilingual training split of the dataset. Fine-tuning is performed using claims and their corresponding source publications across all available languages, while additionally incorporating hard negative source samples in a contrastive retrieval training setup. This adaptation aims to better align retrieval representations with the multilingual characteristics of the task. Detailed training settings and hyperparameters are reported in Table 6 in Appendix.

4.2.2. Similarity-based Re-Ranking

The second stage of the proposed framework involves refining the initial source ranking using similarity-based re-rankers. Starting from the top-100 candidate sources retrieved by the selected first-stage retrieval model, re-ranking models score the relevance of each claim–source pair based on their textual representations and generate an updated ranking.
Nemotron. The *nvidia/llama-nemotron-rerank-vl-1b-v2* model is a pointwise cross-encoder re-ranker with approximately 1.7B parameters. Cross-encoders encode a claim and a source document together as a single input sequence, then score the relevance of the pair individually.
BGE. Similar to Nemotron, *BAAI/bge-reranker-v2-m3* is a pointwise cross-encoder re-ranker, but with a smaller model size of approximately 0.6B parameters and multilingual training objectives designed for cross-lingual retrieval settings [21].
Jina. The *jinaai/jina-reranker-v3* model is a multilingual listwise re-ranker with approximately 0.6B parameters [22]. Unlike pointwise approaches, listwise re-ranking considers the relative relevance of multiple candidate sources within a shared context to determine their ranking.
Qwen3. The models *Qwen/Qwen3-Reranker-0.6B* and *Qwen/Qwen3-Reranker-8B* are pointwise generative re-rankers from the Qwen3 family [23]. In contrast to conventional cross-encoder re-rankers, Qwen3 uses a generative formulation for relevance estimation and enables analysis of the effects of model scale.
Qwen3-IT. We additionally evaluate an instruction-tuned variant with a customized prompt designed for multilingual scientific claim–source retrieval. The adapted prompt incorporates task-specific guidance to investigate whether stronger instruction alignment improves re-ranking performance. The full prompt template is provided in Figure 3 in Appendix.

4.2.3. Verification-based Re-Ranking

The final stage performs verification-based re-ranking to refine the highest-ranked candidate sources. While similarity-based approaches improve semantic matching between claims and source documents, semantically similar candidates may still differ in the evidence they provide and not necessarily support the claim. To address this limitation, we incorporate verification signals into multilingual scientific claim–source retrieval, following ideas from fact verification literature where evidence assessment is used to determine whether retrieved information supports a claim [24]. The first two stages already reduce the candidate space to a small set of highly relevant sources, such that the correct source is frequently contained among the top-ranked documents. Verification-based re-ranking therefore focuses on identifying the best matching source and promoting it to the highest rank position.

Due to the computational cost of LLM-based reasoning, verification-based re-ranking is applied only to the top-10 candidate sources returned by the selected similarity-based re-ranker. The verification stage is formulated as a listwise selection strategy, where the complete candidate set is considered

jointly and the source that best supports the claim is selected. A structured zero-shot prompt is used to guide the model through several verification steps, including evidence consistency between claims and source abstracts, contextual specificity, and the exclusion of semantically similar but potentially conflicting sources. The prompt template is detailed in Figure 4 in Appendix. The selected source receives an additional score of +1 on top of the normalized ranking scores, ensuring that it is promoted to the first rank position while preserving the relative ordering of all remaining candidates. We consider three open-source generative LLMs as verification-based re-ranking models to analyze the effect of model architecture and model scale.

Gemma-4. The *google/gemma-4-31B-it* model is an instruction-tuned dense LLM with approximately 31B parameters from the Gemma family. Its reasoning capabilities and multilingual support make it suitable for verification-based re-ranking in multilingual scientific claim–source retrieval.

MiniMax-M2.5. The *MiniMaxAI/MiniMax-M2.5* model is a sparse mixture-of-experts (MoE) model with approximately 229B total parameters and around 10B activated parameters during inference. Its sparse activation enables efficient reasoning with substantially larger effective model capacity.

Kimi-K2.6. The *moonshotai/kimi-k2.6* model is a large-scale MoE model with approximately 1T total parameters and 32B activated parameters during inference. Its architecture provides higher capacity through sparse activation, enabling comparison against frontier-scale reasoning models.

5. Evaluation

We adopt a stage-wise evaluation strategy in which each component is evaluated sequentially and the strongest configuration is selected for subsequent stages. We first analyze different first-stage retrieval models across three claim representations, followed by the evaluation of similarity-based and verification-based re-ranking approaches. This setup incrementally refines the overall system configuration while enabling the contribution of each stage to be assessed individually. Following the CheckThat! 2026 setup, all experiments are evaluated using *Mean Reciprocal Rank at 5* (MRR@5).

5.1. First-Stage Retrieval

Table 2 summarizes the first-stage retrieval results across different claim representations. Among the evaluated retrieval models, GritLM achieves the strongest overall performance, reaching the highest average MRR@5 of 0.632 with translated claims. For BM25, GTR, E5, and the original GritLM model, translated claims consistently yield the strongest results, indicating that reducing the language mismatch between non-English claims and predominantly English source publications improves retrieval. In contrast, the fine-tuned variant GritLM-F changes this behavior, reaching its highest performance with bilingual claims and improving German claims by +0.063 MRR@5 over the original GritLM model. While GritLM-F also improves French claims (+0.040), it substantially reduces performance for English claims (-0.122), showing that fine-tuning affects retrieval behavior differently across languages. One possible explanation is that fine-tuning on multilingual claim–source pairs and hard negatives adapts GritLM more strongly to cross-lingual retrieval, improving German and French claims but reducing the strong zero-shot English performance of the original model. To prioritize robust generalization to unseen claims, we adopt a conservative selection strategy and use the original GritLM model for English and French claims, while applying GritLM-F only for German claims where improvements are most pronounced and consistent across claim representations.

5.2. Similarity-based Re-Ranking

Similarity-based re-ranking substantially improves retrieval performance over the selected first-stage retrieval configuration, which combines GritLM for English and French claims with GritLM-F for German claims, as shown in Table 3. The strongest overall performance is achieved by Qwen3-8B-IT with the instruction-tuned prompt, improving average MRR@5 by +0.103 to 0.7419. Standard Qwen3-8B achieves comparable performance (+0.099), while Nemotron also provides notable improvements (+0.054). In

Table 2
First-stage retrieval performance across different claim representations for handling multilingual claims. We report MRR@5 for English (EN), German (DE), and French (FR) claims, while AVG denotes the arithmetic mean across all languages. GritLM-F denotes the fine-tuned variant of GritLM.

Retrieval Models	Original Claims				Translated Claims				Bilingual Claims			
	EN	DE	FR	AVG	EN	DE	FR	AVG	EN	DE	FR	AVG
BM25	0.4317	0.0719	0.0618	0.1885	0.4317	0.2844	0.4310	0.3824	0.4317	0.1936	0.2024	0.2759
GTR	0.4821	0.2839	0.4040	0.3900	0.4821	0.3814	0.5085	0.4573	0.4821	0.3532	0.4932	0.4429
E5	0.5333	0.2543	0.3882	0.3919	0.5333	0.4552	0.5688	0.5191	0.5333	0.4263	0.5319	0.4972
GritLM	0.6595	0.5163	0.6567	0.6108	0.6595	0.5670	0.6707	0.6324	0.6595	0.5347	0.6612	0.6185
GritLM-F	0.5380	0.5808	0.6631	0.5940	0.5380	0.5895	0.6799	0.6025	0.5380	0.5974	0.7011	0.6122

Table 3
Similarity-based re-ranking performance on top of the selected first-stage retrieval configuration. We report MRR@5 for English (EN), German (DE), and French (FR) claims, while AVG denotes the arithmetic mean across all languages. Baseline denotes the selected first-stage retrieval configuration using GritLM for English and French claims and GritLM-F for German claims.

Re-ranker	EN	DE	FR	AVG
GritLM-(F)	0.6595	0.5974	0.6612	0.6394
Nemotron	0.7138	0.6285	0.7367	0.6930
BGE	0.5421	0.4866	0.5615	0.5301
Jina	0.6885	0.5910	0.7108	0.6634
Qwen3-0.6B	0.6527	0.5712	0.6750	0.6330
Qwen3-8B	0.7447	0.6919	0.7773	0.7380
Qwen3-8B-IT	0.7536	0.6847	0.7874	0.7419

contrast, Qwen3-0.6B slightly underperforms the baseline (-0.006), whereas BGE substantially reduces performance (-0.109). Interestingly, the customized Qwen3-8B prompt improves retrieval performance for English and French claims, while the standard prompt remains slightly stronger for German claims. Since the instruction-tuned version achieves the strongest overall retrieval performance, we select Qwen3-8B-IT for all languages in the final system configuration.

5.3. Verification-based Re-Ranking

Verification-based re-ranking further improves retrieval performance on top of the selected first-stage retrieval and similarity-based re-ranking configuration (see Table 4). Starting from the top-10 candidates generated by GritLM-(F) and Qwen3-8B-IT, the verification-based re-ranker selects the single candidate source that best supports the claim. All evaluated verification models improve retrieval performance over the similarity-based re-ranking baseline. Gemma-4 improves average MRR@5 by +0.013, while MiniMax-M2.5 achieves a slightly larger gain of +0.015. The strongest overall performance is achieved by Kimi-K2.6, increasing average MRR@5 from 0.7419 to 0.7648 (+0.023). Compared to earlier stages of the framework, the improvements introduced by verification-based re-ranking are smaller overall. However, the consistent gains across all evaluated models show that verification signals provide complementary information beyond similarity estimation alone.

5.4. CheckThat! 2026 Leaderboard

The final submitted system combines bilingual claim representations with metadata-enhanced source representations. First-stage retrieval uses *GritLM* for English and French claims and the fine-tuned variant *GritLM-F* for German claims. Candidate rankings are subsequently refined using the re-ranker

Table 4
Verification-based re-ranking performance on top of the selected first-stage retrieval and similarity-based re-ranking configuration. We report MRR@5 for English (EN), German (DE), and French (FR) claims, while AVG denotes the arithmetic mean across all languages. Baseline denotes the selected retrieval configuration using GritLM-(F) for first-stage retrieval with Qwen3-8B-IT for similarity-based re-ranking.

Re-ranker	EN	DE	FR	AVG
GritLM-(F)+Qwen3-8B-IT	0.7536	0.6847	0.7874	0.7419
Verification (Gemma-4)	0.7662	0.6937	0.8043	0.7547
Verification (MiniMax-M2.5)	0.7685	0.7001	0.8022	0.7569
Verification (Kimi-K2.6)	0.7834	0.7176	0.7933	0.7648

Table 5
Official test-set results of the submitted Claim2Source system on CheckThat! 2026 Task 1. We report MRR@5 for English (EN), German (DE), and French (FR) claims, while AVG denotes the arithmetic mean across all languages. Rk. denotes the rank achieved on the official leaderboard. Values in parentheses denote the score difference to the next-ranked system, where positive values indicate a higher score than the next lower-ranked system and negative values indicate a lower score than the next higher-ranked system.

	EN		DE		FR		AVG	
System	MRR@5	Rk.	MRR@5	Rk.	MRR@5	Rk.	MRR@5	Rk.
Claim2Source	0.7819 (+0.0017)	1	0.7153 (-0.0075)	2	0.7912 (+0.0160)	1	0.7628 (+0.0048)	1

Qwen3-8B-IT for similarity-based re-ranking and *Kimi-K2.6* for verification-based re-ranking.

Table 5 reports the official test set results on CheckThat! 2026 Task 1 on *Source Retrieval for Scientific Web Claims*. The submitted system achieved first place out of 37 participating teams, with an average MRR@5 score of 0.7628. Across individual languages, the system ranked first for English and French claims, and second for German claims. The strongest performance was achieved for French claims (0.7912), followed by English (0.7819) and German (0.7153).

6. Discussion

The experimental findings provide insights into the design decisions underlying the proposed system and highlight several remaining challenges for multilingual scientific claim–source retrieval.

6.1. Design Choices Behind Claim2Source

Structured Claim and Source Representations. The results highlight that multilingual scientific claim–source retrieval benefits from structured claim and source representations. Bilingual claim representations combined with fine-tuned GritLM-F, in particular, improve performance for German (+0.063 MRR@5) and French (+0.040) claims. This indicates that preserving information from the original claim formulation provides complementary retrieval signals beyond translation alone. Additionally, metadata-enhanced source representations combine publication content with contextual information, such as authors and venue, to support the identification of indirect claim–source relationships throughout the retrieval and verification stages.

Progressive Candidate Refinement. Retrieval performance increases progressively across multiple refinement stages. First-stage retrieval reduces the search space from 10,000 publications to a set of 100 candidate sources, while similarity-based re-ranking provides the largest performance gain (+0.103 MRR@5). Verification-based re-ranking further improves retrieval performance (+0.023) by refining only the top-10 candidate sources, although gains become smaller in later stages. Different retrieval stages contribute complementary signals, while early candidate reduction enables computationally expensive reasoning-based approaches to be applied to a small set of highly relevant candidates.

6.2. Limitations and Future Work

Computational Costs. The proposed framework relies on several computationally intensive components, including GritLM-7B for first-stage retrieval, Qwen3-8B for similarity-based re-ranking, and Kimi-K2.6 for verification-based re-ranking, which comprises approximately 1T total parameters with 32B active parameters during inference. Although smaller alternatives were evaluated throughout different stages of the framework, they consistently resulted in lower retrieval performance. This trade-off highlights an important challenge for multilingual scientific claim–source retrieval, namely improving computational efficiency without sacrificing retrieval effectiveness. Future work should investigate lightweight retrieval and re-ranking architectures or model distillation approaches that reduce computational costs while preserving retrieval performance.

Multilingual Generalization. Despite its multilingual design, the framework remains sensitive to language-specific adaptations in terms of retrieval performance. While fine-tuned GritLM-F improves retrieval performance for German and French claims, it substantially reduces performance for English claims (-0.122 MRR@5). Consequently, the final system requires language-specific model selection and different retrieval configurations for each language rather than a unified multilingual setup. This underscores a significant challenge in the field of multilingual scientific claim–source retrieval: developing retrieval models that generalize robustly across languages while maintaining consistent performance. Although language-specific adaptation improves retrieval performance for German over the original GritLM model, German claims remain the most challenging. Understanding the remaining performance gap requires a more detailed error analysis and is left for future work.

7. Conclusion

In this paper, we presented a multi-stage retrieval framework for multilingual scientific claim–source retrieval in CheckThat! 2026 Task 1: *Source Retrieval for Scientific Web Claims*. Our approach combines structured claim and source representations with progressive candidate refinement through first-stage retrieval, similarity-based re-ranking, and verification-based re-ranking. Results show that each stage contributes to retrieval performance, with smaller but complementary improvements across the framework. Our system achieves an average MRR@5 score of 0.7628 across English, German, and French claims, ranking first on the CheckThat! 2026 leaderboard. Future work should focus on improving multilingual generalization and reducing computational costs while preserving retrieval performance.

Acknowledgments

The authors acknowledge the financial support by the Federal Ministry of Research, Technology and Space of Germany (BMFTR) and by the Saxon State Ministry for Science, Culture and Tourism in the programme Center of Excellence for AI Research „Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig“, project identification number: ScaDS.AI.

Tobias Schreieder is supported by the BMFTR through a Software Campus project, project identification number: 16|S23070.

The authors also acknowledge computing resources provided by the NHR Center at TU Dresden, supported by the BMFTR and the participating state governments within the NHR framework.

Availability

We publicly release the code for all experiments at https://github.com/faerber-lab/CheckThat2026.

Declaration on Generative AI

The authors used ChatGPT for language editing and minor formatting assistance, and GitHub Copilot, Gemini, and GLM for coding assistance. These tools did not contribute to the intellectual content or scientific conclusions. After using these tools, all content was reviewed by the authors, who assume full responsibility for the publication.

References

[1] L. Guenther, C. Wilhelm, C. Oschatz, J. Brück, Science communication on twitter: Measuring indicators of engagement and their links to user interaction in communication scholars' tweet content, Public Understanding of Science 32 (2023) 860–869. doi:10.1177/09636625231166552.

[2] T. Schreieder, M. Färber, Claim2source at checkthat! 2025: Zero-shot style transfer for scientific claim-source retrieval, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 2025. URL: https://ceur-ws.org/Vol-4038/paper_94.pdf.

[3] 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, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 2025. URL: https://ceur-ws.org/Vol-4038/paper_93.pdf.

[4] X. Zhang, K. Ogueji, X. Ma, J. Lin, Toward best practices for training multilingual dense retrieval models, ACM Trans. Inf. Syst. 42 (2023). URL: https://doi.org/10.1145/3613447. doi:10.1145/3613447.

[5] R. Litschko, I. Vulić, S. P. Ponzetto, G. Glavaš, On cross-lingual retrieval with multilingual text encoders, Information Retrieval Journal 25 (2022) 149–183. doi:10.1007/s10791-022-09406-x.

[6] 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. doi:10.1007/978-3-032-21321-1_43.

[7] 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, in: E. S. Salido, A. Barrón-Cedeño, Alba García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CLEF 2026, Jena, Germany, 2026.

[8] F. Alam, J. M. Struß, T. Chakraborty, S. Dietze, S. Hafid, K. Korre, A. Muti, P. Nakov, F. Ruggeri, S. Schellhammer, V. Setty, M. Sundriyal, K. Todorov, V. Venktesh, Overview of the clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and retrieval, in: J. Carrillo-de Albornoz, A. García Seco de Herrera, J. Gonzalo, L. Plaza, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2026, pp. 199–223. doi:10.1007/978-3-032-04354-2_13.

[9] 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, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 2025. URL: https://ceur-ws.org/Vol-4038/paper_89.pdf.

[10] M. Staudinger, A. El-Ebshihy, W. Kusa, F. Piroi, A. Hanbury, Atom at checkthat! 2025: Retrieve the implicit - scientific evidence retrieval, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 2025. URL: https://ceur-ws.org/Vol-4038/paper_98.pdf.

[11] D. Lawrie, E. Yang, D. W. Oard, J. Mayfield, Neural approaches to multilingual information retrieval, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2023, pp. 521–536. doi:10.1007/978-3-031-28244-7_33.

[12] F. Valentini, D. Kozlowski, V. Lariviere, CLIRudit: Cross-lingual information retrieval of scientific documents, in: D. I. Adelani, C. Arnett, D. Ataman, T. A. Chang, H. Gonen, R. Raja, F. Schmidt, D. Stap, J. Wang (Eds.), Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025), Association for Computational Linguistics, Suzhuo, China, 2025, pp. 226–242. URL: https://aclanthology.org/2025.mrl-main.16/. doi:10.18653/v1/2025.mrl-main.16.

[13] I. Besrour, J. He, T. Schreieder, M. Färber, Squai: Scientific question-answering with multi-agent retrieval-augmented generation, in: Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 6603–6608. doi:10.1145/3746252.3761471.

[14] I. Besrour, J. He, T. Schreieder, M. Färber, Ragenta: Multi-agent retrieval-augmented generation for attributed question answering, 2025. URL: https://arxiv.org/abs/2506.16988. arXiv:2506.16988.

[15] H. Zhang, R. Zhang, J. Guo, M. de Rijke, Y. Fan, X. Cheng, From relevance to utility: Evidence retrieval with feedback for fact verification, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 6373–6384. URL: https://aclanthology.org/2023.findings-emnlp.422/. doi:10.18653/v1/2023.findings-emnlp.422.

[16] S. E. Robertson, S. Walker, Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval, in: SIGIR '94, Springer London, London, 1994, p. 232–241. doi:10.5555/188490.188561.

[17] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, M. Gatford, Okapi at TREC-3, in: D. K. Harman (Ed.), Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994, volume 500-225 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 1994, pp. 109–126. URL: http://trec.nist.gov/pubs/trec3/papers/city.ps.gz.

[18] J. Ni, C. Qu, J. Lu, Z. Dai, G. Hernandez Abrego, J. Ma, V. Zhao, Y. Luan, K. Hall, M.-W. Chang, Y. Yang, Large dual encoders are generalizable retrievers, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 9844–9855. doi:10.18653/v1/2022.emnlp-main.669.

[19] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, F. Wei, Text embeddings by weakly-supervised contrastive pre-training, 2024. arXiv:2212.03533.

[20] N. Muennighoff, H. SU, L. Wang, N. Yang, F. Wei, T. Yu, A. Singh, D. Kiela, Generative representational instruction tuning, in: The Thirteenth International Conference on Learning Representations, 2025. URL: https://openreview.net/forum?id=BC4lIvfSzv.

[21] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 2318–2335. doi:10.18653/v1/2024.findings-acl.137.

[22] F. Wang, Y. Li, H. Xiao, jina-reranker-v3: Last but not late interaction for listwise document reranking, 2025. URL: https://arxiv.org/abs/2509.25085. arXiv:2509.25085.

[23] Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, F. Huang, J. Zhou, Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025. URL: https://arxiv.org/abs/2506.05176. arXiv:2506.05176.

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

A. GritLM Fine-Tuning Details

This appendix provides the hyperparameters and training configuration used for LoRA fine-tuning of GritLM-F. The setup includes optimization settings, negative sampling configuration, and training infrastructure details referenced in Section 4.

Table 6
Hyperparameters and training configuration used for LoRA fine-tuning of GritLM-F. Training was performed on a single NVIDIA H100 GPU (see Section 4.2.1).

Parameter	Value
Training epochs	3
Learning rate	$2e^{-4}$
Batch size	8
Gradient accumulation steps	4
Maximum sequence length	256
LoRA rank (r)	16
LoRA alpha	64
LoRA dropout	0.1
Optimizer	AdamW
Learning rate scheduler	Cosine
Warmup ratio	0.05
Weight decay	0.01
Negative sampling strategy	Hybrid (BM25 + GritLM weighted fusion)
Number of negative samples	8
Number of hard negative samples	6
Number of random negative samples	2

B. LLM Prompt Templates

The following figures present the prompt templates used across the evaluated LLM-based retrieval and re-ranking components described in Section 4. The prompts include query translation, instruction-tuning for similarity-based re-ranking, and verification-based re-ranking.

Translation Prompt

[Role & Objective]
You are a professional translator. Translate the following text to English.

[Instructions & Constraints]
Output ONLY the English translation — nothing else. Do not add explanations, notes, or quotes.

Figure 2: Prompt template used for query translation with the Qwen/Qwen3-8B model (see Section 4.1.1).

Similarity-based Re-Ranking Instruction-Tuning Prompt

[Role & Objective]
You are a high-precision scientific fact-checker expert. Your mission is to determine whether a research paper (DOCUMENT) supports the claim made in a tweet (QUERY).

[Pre-Processing Rules]

- Ignore emojis, slang, and non-scientific conversational fillers in the query.
- Focus exclusively on the core scientific facts and claims.
- Treat # and @ as noise unless they provide explicit metadata such as journal or author names.

[Analysis Dimensions]

1. **Subject & Entities:** Identify core research subjects and terminology. Check for overlap in proper nouns such as diseases, drugs, proteins, or methods.
2. **Methodology:** Compare the research frameworks or experimental approaches used in the paper and implied in the query.
3. **Results & Data:** Prioritize matching specific numerical results, statistics, or scientific findings.
4. **Metadata Bonus (@Author):** Increase confidence if a name following @ matches one of the paper authors.
5. **Metadata Bonus (#Journal):** Increase confidence if a venue following # matches the publication venue of the paper.

[Output Instruction]

- Respond with `yes` if the document provides sufficient evidence to support or strongly relate to the query.
- Otherwise, respond with `no`.
- Output strictly either `yes` or `no`. No explanations.

Figure 3: Prompt template used for the Qwen3-8B-IT similarity-based re-ranker (see Section 4.2.2).

Verification-based Re-Ranking Prompt

[Role & Objective]
You are an expert research assistant. Your task is to identify the single most relevant paper from a provided list of candidates based on a specific user query.

Handling bilingual queries:
The user may provide a single query (`QUERY: ...`) OR a pair (`QUERY (original language): ...` and `QUERY (translated to English): ...`)

- If both are given, the original is in French or German. Use it as the PRIMARY source for matching.
- The English translation is a SUPPLEMENT to help recognise cross-lingual synonyms and terminology.
- Always consider both versions when evaluating relevance.

[Instructions & Constraints]
Analyze the Query and the Candidates below. Select the ONE paper that is the most relevant. Before making your final decision, you must perform the following checks:

- **Source Verification:** Does the paper title appear verbatim (or near-verbatim) in the query? (e.g., "Ultrapotent antibodies...")
- **Evidence Alignment:** Does the paper's abstract explicitly support the specific claim made in the query? (e.g., If the query claims "increased deaths," does the paper's results section confirm increased deaths?)
- **Contextual Specificity:** Does the specific population, location, or intervention in the paper match the query? (e.g., "US Veterans" vs. "France"; "First Wave" vs. "Vaccinated")
- **Exclusion Check:** Are there other papers with similar keywords but contradictory conclusions? Discard them.

You MUST respond with ONLY a JSON object with exactly two keys:

- `"reasoning"`: a brief string (2–4 sentences) explaining why this specific paper was chosen over the others based on the checks above. Mention any title matches or specific data points that aligned.
- `"selected_paper"`: an integer (1-based), the paper number that is most relevant.

Example response:

```
{
  "reasoning": "Paper 3 title matches the drug name in the query verbatim.
    Its abstract confirms the 50% reduction claim in the specific population
    mentioned. No other paper addresses this exact intervention.",
  "selected_paper": 3
}
```

No extra text, no explanation outside the JSON.

Figure 4: Prompt template used for verification-based re-ranking across all evaluated LLMs (see Section 4.2.3).