Scientific Claim–Source Retrieval Revisited: A Comparative Study of Style Transfer and Re-Ranking

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

[1] TU Dresden & ScaDS.AI Dresden/Leipzig, Dresden, Germany
tobias.schreieder@tu-dresden.de
[2] Friedrich-Alexander University of Erlangen–Nuremberg, Erlangen, Germany

Abstract. Scientific claims shared on social media are often difficult to verify and may contribute to the spread of misinformation. To address this challenge, automated fact verification systems require scientific claim–source retrieval, the task of identifying the source publication underlying a given claim. However, claims often differ substantially from their source publications in language, style, and specificity, making retrieval challenging. We present a comparative study of scientific claim–source retrieval on the CheckThat! 2026 benchmark across sparse and dense retrieval models. Our results show that translating claims into English outperforms both original and bilingual claim representations, while incorporating publication metadata provides additional retrieval gains by capturing indirect source references. In addition, we analyze four style transfer approaches and find that they improve retrieval performance for most models, although the optimal style depends on the underlying retrieval objective. Finally, we investigate similarity- and signal-based re-ranking approaches, introducing three novel re-ranking models based on attribution, entity overlap, and verification-based reasoning. Verification-based re-ranking yields additional gains beyond semantic similarity and achieves the best overall performance with an MRR@5 of 0.758.

Keywords: Scientific Claim–Source Retrieval · Re-Ranking · Text Style Transfer · Multilinguality · Large Language Models · Fact Verification.

1 Introduction

Scientific communication has undergone a notable shift. While peer-reviewed journals and conferences remain central for communication within expert communities, social media platforms such as X (formerly Twitter) serve as additional channels that enable the rapid dissemination of research findings, the promotion of publications, and direct engagement with broader audiences [6,7]. Despite these benefits, social media also raises concerns about the credibility of scientific information. Scientific claims are often presented without sufficient context, expressed imprecisely, or only loosely connected to the underlying evidence, making

their credibility difficult to assess. In the worst case, such claims may constitute misinformation, that is, information that contradicts current scientific evidence or consensus [27], which is widespread on social media, particularly in areas such as vaccination, opioids, and noncommunicable diseases [25].

Addressing these challenges requires automated fact verification systems that assess the factual correctness of claims with respect to relevant evidence. In this context, we focus on scientific claim–source retrieval, a key component of such systems that aims to identify the scientific publication that serves as the source for a given claim. Scientific claim–source retrieval has recently been introduced in the CheckThat! Lab and extended to multilingual settings [1,24]. However, this task is inherently difficult in the context of social media, where claims are often expressed in colloquial or simplified language. In addition, claims and their corresponding scientific publications may be expressed in different languages, introducing a challenge for retrieval systems given the predominantly English scientific literature. Furthermore, claims often contain only partial and implicit references to their sources, such as mentions of authors, institutions, or venues.

Our previous work on scientific claim–source retrieval demonstrated that zero-shot style transfer, in particular rewriting claims into a more formal style, generally improves retrieval performance [18]. While that study focused on monolingual style transfer, this work revisits style transfer on the multilingual CheckThat! 2026 benchmark by evaluating multiple large language models (LLMs). Beyond reproducing and extending the style-transfer analysis, we investigate re-ranking as a complementary retrieval strategy, enabling the first comparative study of style transfer and re-ranking for scientific claim–source retrieval. Code and resources are publicly available.[3] Our contributions are as follows:

1. We analyze multilingual claim representations and show the impact of original, translated, and bilingual claims on retrieval performance.
2. We investigate the role of publication metadata and demonstrate its value for scientific claim–source retrieval.
3. We extend prior work on zero-shot style transfer to a multilingual setting and provide a comparative analysis across sparse and dense retrieval models.
4. We compare similarity-based re-rankers and propose three novel signal-based re-rankers that leverage attribution, entity overlap, and verification signals.

2 Related Work

The task of evidence retrieval aims to identify relevant sources for a given claim, often as part of claim verification. Scientific claim–source retrieval represents a specific instance of this problem, where the goal is to identify the referenced scientific paper. Early work relies on traditional retrieval and learning-to-rank methods [9], followed by a shift toward neural and transformer-based approaches such as BERT-based retrieval models [21,16]. Hybrid retrieval has emerged as

[3] https://github.com/faerber-lab/CheckThat2026

a strong paradigm that combines sparse and dense signals to leverage complementary strengths [26]. Recent work shows that hybrid approaches, especially when combined with re-ranking, consistently outperform single-model retrieval setups [15]. To address heterogeneous sources, cross-genre retrieval frameworks were proposed to align scientific and journalistic evidence [33]. Data augmentation strategies that modify or reformulate claims have been investigated to improve retrieval performance [17]. Subsequent work moved beyond relevance estimation by incorporating verifier feedback [31] and modeling multi-step reasoning over interdependent evidence [8]. In addition, re-ranking strategies are critical, with listwise re-rankers outperforming pointwise approaches in the experiments on the CheckThat! 2025 dataset [23]. A growing line of work leverages LLMs to enhance retrieval and reasoning, including question decomposition [4], contrastive re-ranking [22], and sub-question enrichment strategies [5]. Finally, RAG frameworks have been used to jointly assess factual accuracy and relevance by generating evidence summaries from retrieved sources, which provides a more interpretable basis for claim verification [28]. LLM-based source filtering has been proposed to refine retrieved candidates [2]. While prior work has advanced retrieval through dense methods, hybrid approaches, LLM-based reasoning, and style transfer, it largely relies on textual similarity signals. We extend this line of work by systematically comparing style transfer and introducing re-ranking signals based on attribution, entity overlap, and verification reasoning.

3 Experimental Setup

This section introduces the dataset and retrieval models used in our study.

3.1 The CheckThat! 2026 Dataset

We conduct our experiments on the CheckThat! 2026 *Source Retrieval for Scientific Web Claims* dataset [24]. The dataset consists of social media claims and their corresponding source publications. Claims are provided in English (EN), German (DE), and French (FR), while 10,000 English scientific publications form the candidate set. Since our study does not involve training, we restrict our experiments to the validation set, which contains 4,993 claims (3,905 EN, 386 DE, and 702 FR). The task is formulated as a known-item retrieval problem, where the goal is to retrieve the source publication corresponding to a claim from the candidate set. This is challenging because claims often refer to their sources only implicitly and may differ substantially in language and style.

3.2 Retrieval Models

We examine one sparse retrieval model and three dense models to provide a diverse set of retrieval paradigms for our experiments.

BM25. *BM25* is a widely used sparse retrieval model based on term frequency and inverse document frequency, relying on exact term matching to rank source documents [13,14]. Its lexical nature makes it sensitive to claim formulation.

GTR. Based on the T5 architecture, *gtr-t5-xl* is an 11B-parameter dense model trained on large-scale datasets for complex retrieval tasks [12].
E5. The dense model *intfloat/e5-large-v2*, with 335M parameters, is based on a transformer architecture. It is trained with a contrastive objective on large-scale retrieval data for semantic retrieval tasks [30].
GritLM. *GritLM/GritLM-7B* is a dense model with 7B parameters based on a decoder-only transformer architecture. It is trained with instruction tuning on large-scale corpora and performs well in zero-shot retrieval settings [11].

4 Facets of Scientific Claim–Source Retrieval

We conduct a comparative study of different approaches for scientific claim–source retrieval in multilingual settings. We analyze five key facets of the retrieval pipeline: (1) handling multilingual claims, (2) leveraging publication metadata, (3) query reformulation via style transfer, (4) similarity-based re-ranking, and (5) signal-based re-ranking. Each facet captures a distinct aspect of the problem, ranging from linguistic and stylistic variation in claims and incomplete or implicit source references to modeling choices in retrieval and re-ranking, while building on the best-performing configuration from preceding experiments.

4.1 Handling Multilingual Claims

To address multilinguality, we evaluate three claim representations: original claims, translated claims, and bilingual claims. **Original claims** correspond to the claims in their respective languages (English, German, and French). For **translated claims**, we translate all claims into English using zero-shot translation with the LLM *Qwen/Qwen3-8B* to analyze the impact of language on downstream retrieval performance. **Bilingual claims** are constructed by jointly encoding the original and translated claims, while for English claims only the original version is used. This setup allows us to compare original, translated, and bilingual claim representations for scientific claim–source retrieval.

4.2 Leveraging Publication Metadata

As discussed in Section 3.1, claims may include only indirect references to their sources, for example through mentions of authors, institutions, or venues. To investigate whether such signals can improve retrieval performance, we compare two strategies for encoding sources. In the **no metadata** setting, source representations consist only of the title and abstract provided in the dataset, which are concatenated into a single input sequence. In the **metadata** setting, we augment the source representation with additional structured information in the form of labeled fields: *title*, *abstract*, *authors*, and *venue*. To limit noise from long author lists, we include only the first and last three authors. This comparison allows us to evaluate whether incorporating publication metadata improves retrieval performance in scenarios where claims refer to sources only implicitly.

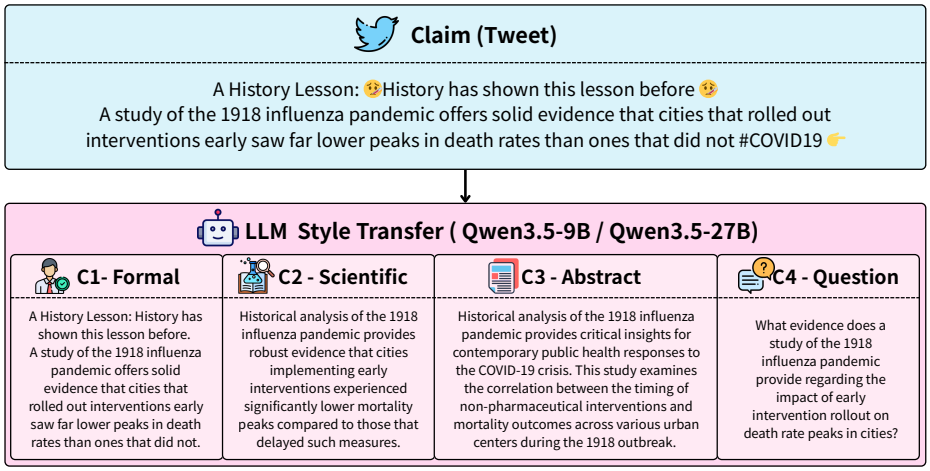


Fig. 1. Illustration of the style transfer process. Given a claim, the LLM generates stylistic variants to create diverse representations while preserving the original meaning.

4.3 Style Transfer for Scientific Claims

As illustrated in Figure 1, we apply style transfer to claims to improve their alignment with scientific publications. Our prior work [18] showed that formalizing claims improves retrieval performance, especially for sparse retrieval models, while modifying source documents typically introduces noise and degrades performance. Therefore, we restrict style transfer to claims and keep sources unchanged. Building on this setup, we extend the analysis to a multilingual setting and adopt LLMs (*Qwen/Qwen3.5-9B* and *Qwen/Qwen3.5-27B*) for zero-shot style transfer. To ensure comparability with our previous study [18], we reuse its prompt templates. This enables a controlled comparison of model sizes while providing a more efficient and scalable setup for retrieval-oriented applications. Due to the non-deterministic behavior of LLM-based style transfer, all experiments are repeated three times and results are reported as averages across runs. We evaluate the following four claim styles:

Formal. Claims are rewritten in a more formal style by removing informal elements such as hashtags and emojis while preserving the original meaning.

Scientific. Claims are reformulated as scientific statements using domain-specific terminology, aiming to better match the language of scientific publications.

Abstract. Claims are expanded into short abstract-like texts that summarize the underlying statement in a structured and coherent form, increasing stylistic similarity to scientific documents.

Question. Claims are converted into concise scientific questions, aligning with the input format of retrieval models optimized for question-based queries.

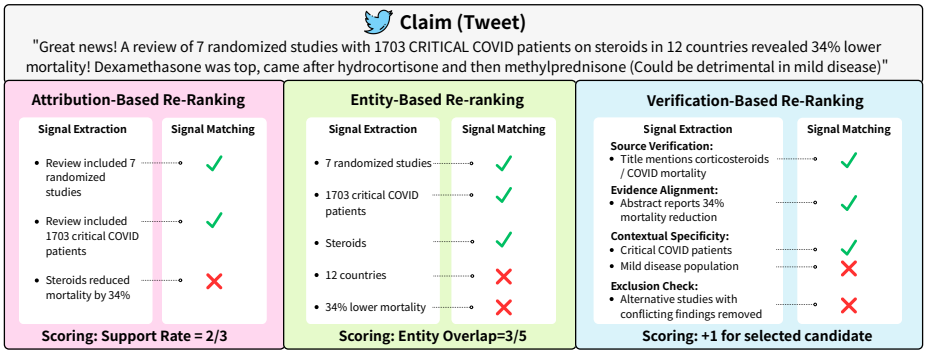


Fig. 2. Overview of the signal-based re-ranking process. Signals extracted from a claim are used to re-rank top-k sources, where sources are scored according to signal matching.

4.4 Similarity-based Re-Ranking

Re-ranking has proven to be an effective component in scientific claim–source retrieval, as top-performing systems in the CheckThat! 2025 shared task incorporated re-ranking approaches to improve retrieval performance [1]. Motivated by these findings, we analyze the impact of similarity-based re-ranking in our setting. Similarity-based re-ranking estimates the relevance of a claim–source pair from their textual content, in contrast to the signal-based approach described in Section 4.5. For all experiments, we use the top-100 sources retrieved by GritLM as candidates. The re-rankers assign new relevance scores to these candidates and generate a new ranking independently of the original retrieval scores.

Nemotron. The *nvidia/llama-nemotron-rerank-vl-1b-v2* model is a pointwise cross-encoder re-ranker with approximately 1.7B parameters.

BGE. The *BAAI/bge-reranker-v2-m3* model is a multilingual pointwise cross-encoder re-ranker with approximately 0.6B parameters from the BGE family [3].

Jina. The *jinaai/jina-reranker-v3* model is a multilingual listwise re-ranker with approximately 0.6B parameters that jointly ranks candidate sources by considering their relative relevance within a shared context [29].

Qwen3. The models *Qwen/Qwen3-Reranker-0.6B* and *Qwen/Qwen3-Reranker-8B* are pointwise generative re-rankers from the Qwen3 family [32]. Their different model sizes additionally enable comparison of scale effects in re-ranking.

4.5 Signal-based Re-Ranking

As illustrated in Figure 2, initial retrieval identifies a broad set of candidate sources, while similarity-based re-ranking refines them based on semantic relevance. However, semantic similarity alone may not fully capture claim–source alignment. We therefore investigate signal-based re-ranking, which incorporates structured signals to improve retrieval performance. Due to the computational

cost of LLM-based reasoning, signal-based re-ranking is applied only to the top-10 candidates from the best-performing retrieval and similarity-based re-ranking setup. We evaluate the proposed re-ranking approaches using three reasoning LLMs. All experiments are repeated three times and averaged across runs.

Qwen3.5. *Qwen/Qwen3.5-27B* is a 27B-parameter reasoning model from the Qwen family with strong instruction-following and multilingual capabilities.

Gemma-4. *google/gemma-4-31B-it* is a 31B-parameter instruction-tuned model from the Gemma family with strong reasoning and multilingual performance.

Kimi-2.6. *moonshotai/Kimi-K2.6* is a large-scale mixture-of-experts model with approximately 1T total parameters and 32B activated parameters during inference, enabling comparison against frontier-scale reasoning models.

We investigate the following three novel signal-based re-rankers.

Attribution-based Re-Ranking. Attribution is the process of linking content to supporting evidence and represents a key component of evidence-based text generation with LLMs [19]. While attribution approaches have been extensively studied in this context, their use as a re-ranking signal for scientific claim–source retrieval remains largely unexplored. We therefore draw inspiration from the atomic fact decomposition introduced in FactScore [10]. Using a zero-shot prompt, claims are decomposed into atomic facts while preserving their content and contextual information. Each atomic fact is then evaluated independently against source documents through natural language inference with the respective reasoning LLM [19]. The resulting fact-level predictions are aggregated into a support rate representing the proportion of supported facts, which is used as an attribution signal for re-ranking candidate sources.

Entity-based Re-Ranking. Salient entities often provide retrieval signals that are not fully captured by semantic similarity [20]. In scientific claim–source retrieval, social media claims frequently contain implicit references through entities such as authors, institutions, datasets, or domain-specific terminology. We therefore analyze whether explicit entity overlap between claims and source documents provides a useful signal for claim–source alignment. Using a structured zero-shot prompt, entities are extracted and compared across claim–source pairs. The prompt considers biomedical, chemical, numerical, methodological, institutional, geographic, and temporal entities, while abbreviations and common synonyms are normalized. The re-ranking score is computed as the proportion of claim entities that also occur in the source representation.

Verification-based Re-Ranking. This approach addresses cases where semantic similarity or entity overlap alone is insufficient to determine whether a source actually underlies a claim. Source documents may share similar terminology while differing in the specific evidence, population, location, or intervention described. We therefore investigate whether verification reasoning provides an additional signal for claim–source alignment. Using a structured zero-shot prompt, the model selects the single most relevant source candidate for a given claim by considering source verification, evidence alignment, contextual specificity, and conflicting findings. The selected source is assigned the highest relevance score and used for re-ranking candidate sources.

Table 1. Retrieval performance (MRR@5) across different strategies for handling multilingual claims, including original, translated, and bilingual claim representations.

Models	Original Claims				Translated Claims				Bilingual Claims			
	EN	DE	FR	AVG	EN	DE	FR	AVG	EN	DE	FR	AVG
BM25	0.420	0.078	0.069	0.189	0.420	0.252	0.391	0.354	0.420	0.162	0.167	0.249
GTR	0.476	0.212	0.300	0.329	0.476	0.381	0.505	0.454	0.476	0.366	0.481	0.441
E5	0.520	0.187	0.336	0.348	0.520	0.468	0.556	0.515	0.520	0.396	0.502	0.473
GritLM	0.648	0.494	0.624	0.589	0.648	0.573	0.668	0.630	0.648	0.527	0.653	0.609

5 Evaluation

We follow a sequential evaluation strategy in which each stage builds upon the strongest results of the previous one. We first identify the most effective strategy for handling multilingual claims and subsequently evaluate the impact of leveraging publication metadata. The resulting setup forms the basis for our comparative study of style transfer and re-ranking approaches. Building on this setup, we compare four claim style transfer approaches and four similarity-based re-rankers across all retrieval models. Finally, we assess the three signal-based re-ranking approaches using the best combination of retrieval model and similarity-based re-ranker. Following the CheckThat! 2026 shared task setup, all experiments are evaluated using *Mean Reciprocal Rank at* 5 (MRR@5).

5.1 Analysis of Strategies for Handling Multilingual Claims

Table 1 shows the retrieval performance of different strategies for handling multilingual claims. Across all retrieval models, translating claims into English consistently achieves the highest performance. Compared to original claims, the largest improvement is observed for E5, where the average MRR@5 increases from 0.348 to 0.515 (+0.167), closely followed by BM25 with an improvement of +0.165. GTR and GritLM also show notable gains of +0.125 and +0.041, respectively. These improvements indicate that reducing the language mismatch between non-English claims and the English sources benefits both sparse and dense retrieval models. Bilingual claims outperform the original multilingual setting but consistently remain below the translated claims. These findings indicate that combining original and translated representations may create competing signals that weaken retrieval performance compared to translated claims. Based on these findings, translated claims are selected for all subsequent experiments.

5.2 Analysis of Strategies for Leveraging Publication Metadata

Table 2 summarizes the impact of leveraging publication metadata on retrieval performance. Across all retrieval models, augmenting the source representation with publication metadata consistently improves retrieval performance, al-

Table 2. Retrieval performance (MRR@5) on translated claims across different strategies for leveraging publication metadata.

Model	No Metadata				Metadata			
	EN	DE	FR	AVG	EN	DE	FR	AVG
BM25	0.420	0.252	0.391	0.354	0.432	0.284	0.431	<u>0.382</u>
GTR	0.476	0.381	0.505	0.454	0.482	0.381	0.509	<u>0.457</u>
E5	0.520	0.468	0.556	0.515	0.533	0.455	0.569	<u>0.519</u>
GritLM	0.648	0.573	0.668	0.630	0.660	0.567	0.672	<u>0.633</u>

though the magnitude of the gains varies between models. The largest improvement is observed for BM25, where the average MRR@5 increases from 0.354 to 0.382 (+0.028). Smaller gains are observed for GTR (+0.003), E5 (+0.004), and GritLM (+0.003). The larger gains observed for BM25 indicate that sparse retrieval benefits more from additional metadata fields. In contrast, dense retrieval models already capture semantic relationships effectively from titles and abstracts, resulting in more limited gains from leveraging metadata. Based on these findings, publication metadata is included in all subsequent experiments.

5.3 Impact of Style Transfer

Table 3 summarizes the impact of different claim style transfer approaches on retrieval performance. The results show that the effectiveness of style transfer strongly depends on the retrieval model. For BM25, the highest performance is achieved with formal claims, improving MRR@5 from 0.382 to 0.401 (+0.019). This observation is consistent with prior findings [18], where formalizing claims also yielded the strongest improvements for sparse retrieval. In contrast, GritLM achieves its highest performance without applying style transfer, reaching an MRR@5 of 0.633. This result suggests that stronger retrieval models may already be robust to stylistic variations in claims. For dense retrieval models, a different pattern emerges. Both GTR and E5 achieve their highest performance with question-style claims, improving from 0.457 to 0.466 (+0.009) and from 0.519 to 0.545 (+0.026), respectively. Unlike prior findings [18], where formal and scientific claim styles showed the strongest improvements, question-based formulations consistently achieve the highest performance in the current multilingual setting. One explanation is that question formulations align more closely with retrieval objectives by expressing information needs more explicitly. Across most style transfer settings, Qwen3.5-27B slightly outperforms Qwen3.5-9B, indicating that larger LLMs produce modest improvements in claim transformations. Results were stable across runs, with standard deviations ≤ 0.004 MRR@5. Overall, style transfer improves retrieval performance for most models, although the optimal claim style depends on the underlying retrieval architecture.

Table 3. Retrieval performance (MRR@5) on translated claims with publication metadata for style transfer using Qwen3.5-9B and Qwen3.5-27B. Original (Orig.) denotes claims without style transfer, while Formal (Form.), Scientific (Sci.), Abstract (Abs.), and Question (Quest.) denote the evaluated claim styles. * marks a statistically significant improvement over Orig. (paired Student's t-test, $p < 0.05$).

		Qwen3.5-9B				**Qwen3.5-27B**			
Model	**Orig.**	**Form.**	**Sci.**	**Abs.**	**Quest.**	**Form.**	**Sci.**	**Abs.**	**Quest.**
BM25	0.382	0.397*	0.394*	0.298	0.383	0.401*	0.397*	0.324	0.385
GTR	0.457	0.452	0.445	0.411	0.458	0.451	0.449	0.420	0.466*
E5	0.519	0.524*	0.530*	0.471	0.544*	0.528*	0.535*	0.486	0.545*
GritLM	0.633	0.629	0.602	0.567	0.601	0.630	0.615	0.576	0.608

Table 4. Retrieval performance (MRR@5) on translated claims with publication metadata using similarity-based re-ranking on top of GritLM retrieval. Baseline denotes GritLM without re-ranking, while Nemotron, BGE, Jina, Qwen3-0.6B, and Qwen3-8B denote the evaluated similarity-based re-rankers.

Re-Ranker	**EN**	**DE**	**FR**	**AVG**
GritLM	0.660	0.567	0.673	0.633
Nemotron	0.714	0.606	0.740	0.686
BGE	0.542	0.450	0.563	0.518
Jina	0.688	0.583	0.707	0.659
Qwen3-0.6B	0.653	0.546	0.676	0.625
Qwen3-8B	0.745	0.675	0.778	0.733

5.4 Impact of Similarity-based Re-Ranking

Similarity-based re-ranking improves retrieval performance for most models over the initial GritLM retrieval stage (see Table 4). The largest improvement is achieved by Qwen3-8B, increasing average MRR@5 from 0.633 to 0.733 (+0.100), with consistent gains across all languages. Nemotron and Jina also improve performance, reaching average MRR@5 scores of 0.686 (+0.053) and 0.659 (+0.026), respectively. In contrast, Qwen3-0.6B slightly underperforms the retrieval baseline (-0.008), while BGE substantially reduces performance (-0.115). The large performance difference between Qwen3-0.6B and Qwen3-8B further highlights that model capacity plays an important role in similarity-based re-ranking.

5.5 Impact of Signal-based Re-Ranking

Signal-based re-ranking leads to different retrieval behavior depending on the signal type (see Table 5). Attribution-based re-ranking performs comparably to the GritLM+Qwen3-8B baseline, with only minor performance differences across models and languages. Kimi-2.6 achieves the highest score for attribution-based re-ranking with an average MRR@5 of 0.735 (+0.002). In contrast, entity-based

Table 5. Retrieval performance (MRR@5) of signal-based re-ranking on top of GritLM retrieval with Qwen3-8B similarity-based re-ranking. Baseline denotes the selected retrieval and similarity-based re-ranking setup. * marks a statistically significant improvement in AVG over the baseline (paired Student's t-test, $p < 0.05$).

Model	**Attribution**				**Entity**				**Verification**			
	EN	DE	FR	AVG	EN	DE	FR	AVG	EN	DE	FR	AVG
Baseline	0.745	0.675	0.778	0.733	0.745	0.675	0.778	<u>0.733</u>	0.745	0.675	0.778	0.733
Qwen3.5	0.744	0.671	0.782	0.732	0.700	0.629	0.747	0.692	0.749	0.662	0.789	0.733
Gemma-4	0.743	0.671	0.779	0.731	0.699	0.640	0.738	0.692	0.763	0.673	0.798	0.745*
Kimi-2.6	0.748	0.678	0.779	<u>0.735</u>*	0.726	0.654	0.759	0.713	0.782	0.692	0.801	<u>0.758</u>*

re-ranking consistently reduces performance across all models, decreasing average MRR@5 by 0.020 to 0.041 relative to the baseline. Verification-based re-ranking yields the strongest improvements, with Kimi-2.6 achieving the best performance and increasing average MRR@5 from 0.733 to 0.758 (+0.025). Gemma-4 also improves performance through verification-based re-ranking, reaching an average MRR@5 of 0.745 (+0.012), while Qwen3.5 matches the baseline. Results were stable across runs, with standard deviations ≤ 0.004.

6 Discussion

Our findings reveal several insights into scientific claim–source retrieval.

Reducing Language Mismatch Improves Retrieval Performance. Translating claims into English consistently improves retrieval performance across all tested models, with the largest gain observed for E5 (+0.167 MRR@5). While a bilingual claim representation also improves performance compared to the original multilingual setting, translated claims consistently achieve the strongest results. These findings highlight that language differences between social media claims and source publications remain a major challenge for scientific claim–source retrieval, emphasizing the importance of reducing such mismatch.

Structured Source Representations Improve Retrieval Performance. Representing sources through structured publication information provides useful retrieval signals, particularly for BM25, which shows an improvement of +0.028 MRR@5. The stronger gains for sparse retrieval suggest that structured metadata can provide useful lexical cues beyond titles and abstracts. This underscores the importance of explicitly incorporating such indirect signals into source representations for scientific claim–source retrieval.

Style Transfer Compensates for Retrieval Limitations. The effectiveness of style transfer appears to depend on the retrieval objective underlying the model architecture. BM25 benefits most from formal claim styles (+0.019 MRR@5), as sparse retrieval relies on exact term matching between claims and source documents. In contrast, question-style claims yield the strongest improvements for GTR (+0.009) and E5 (+0.026), suggesting that explicitly reformu-

lating claims as information needs better aligns with the retrieval objectives of these models. Interestingly, GritLM achieves its highest performance without style transfer, while all transformed claim styles reduce retrieval effectiveness. Overall, these findings highlight that style transfer primarily compensates for retrieval limitations rather than providing universal improvements.

Structured Re-Ranking Signals Differ in Effectiveness. Structured re-ranking signals differ substantially in effectiveness. Similarity-based re-ranking provides the largest gains (+0.100 MRR@5 for Qwen3-8B), while verification-based re-ranking yields further improvements (+0.025 for Kimi-2.6). In contrast, attribution-based re-ranking yields only marginal performance gains, whereas entity-based re-ranking consistently reduces retrieval performance. These findings show that not all structured signals are equally useful for scientific claim–source retrieval. In particular, verification signals can refine candidate rankings beyond semantic similarity alone. However, such improvements come at substantially higher computational cost, as LLM-based verification requires expensive reasoning even when applied only to a small set of top-ranked candidates.

7 Conclusion

In this paper, we presented a comparative study of scientific claim–source retrieval on the CheckThat! 2026 benchmark. Our results show that claim translation improves retrieval performance, while publication metadata provides additional retrieval signals beyond titles and abstracts. Furthermore, the effectiveness of style transfer depends strongly on the retrieval model and its underlying objective. Among the evaluated re-ranking approaches, similarity-based re-ranking yields the largest performance gains, while verification-based re-ranking further improves source selection by explicitly verifying claim–source alignment.

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.

The authors used AI-based assistance tools to support language editing, minor formatting, and coding tasks. These tools did not contribute to the intellectual content or scientific conclusions. All content was reviewed by the authors, who assume full responsibility for the publication.

References

1. Alam, F., Struß, J.M., Chakraborty, T., Dietze, S., Hafid, S., Korre, K., Muti, A., Nakov, P., Ruggeri, F., Schellhammer, S., Setty, V., Sundriyal, M., Todorov, K., Venktesh, V.: Overview of the clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and retrieval. In: Carrillo-de Albornoz, J., García Seco de Herrera, A., Gonzalo, J., Plaza, L., Mothe, J., Piroi, F., Rosso, P., Spina, D., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 199–223. Springer Nature Switzerland, Cham (2026). https://doi.org/10.1007/978-3-032-04354-2_13
2. Besrour, I., He, J., Schreieder, T., Färber, M.: Squai: Scientific question-answering with multi-agent retrieval-augmented generation. In: Proceedings of the 34th ACM International Conference on Information and Knowledge Management. p. 6603–6608. CIKM '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3746252.3761471
3. Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., Liu, Z.: M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics: ACL 2024. pp. 2318–2335. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.findings-acl.137
4. Chen, J., Kim, G., Sriram, A., Durrett, G., Choi, E.: Complex claim verification with evidence retrieved in the wild. In: Duh, K., Gomez, H., Bethard, S. (eds.) Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 3569–3587. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024). https://doi.org/10.18653/v1/2024.naacl-long.196
5. Churina, S., Barik, A.M., Phaye, S.R.: Improving evidence retrieval on claim verification pipeline through question enrichment. In: Schlichtkrull, M., Chen, Y., Whitehouse, C., Deng, Z., Akhtar, M., Aly, R., Guo, Z., Christodoulopoulos, C., Cocarascu, O., Mittal, A., Thorne, J., Vlachos, A. (eds.) Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER). pp. 64–70. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.fever-1.6
6. Guenther, L., Wilhelm, C., Oschatz, C., Brück, J.: Science communication on twitter: Measuring indicators of engagement and their links to user interaction in communication scholars' tweet content. Public Understanding of Science **32**(7), 860–869 (2023). https://doi.org/10.1177/09636625231166552
7. Haunschild, R., Bornmann, L., Potnis, D., Tahamtan, I.: Investigating dissemination of scientific information on twitter: A study of topic networks in opioid publications. Quantitative Science Studies **2**(4), 1486–1510 (2021). https://doi.org/10.1162/qss_a_00168
8. Liao, H., Peng, J., Huang, Z., Zhang, W., Li, G., Shu, K., Xie, X.: Muser: A multi-step evidence retrieval enhancement framework for fake news detection. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. p. 4461–4472. KDD '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3580305.3599873
9. Ma, W., Chao, W., Luo, Z., Jiang, X.: Claim retrieval in twitter. In: Hacid, H., Cellary, W., Wang, H., Paik, H.Y., Zhou, R. (eds.) Web Information Systems Engineering – WISE 2018. pp. 297–307. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-02922-7_20

10. Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.t., Koh, P., Iyyer, M., Zettlemoyer, L., Hajishirzi, H.: FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 12076–12100. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.741
11. Muennighoff, N., SU, H., Wang, L., Yang, N., Wei, F., Yu, T., Singh, A., Kiela, D.: Generative representational instruction tuning. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=BC4lIvfSzv
12. Ni, J., Qu, C., Lu, J., Dai, Z., Hernandez Abrego, G., Ma, J., Zhao, V., Luan, Y., Hall, K., Chang, M.W., Yang, Y.: Large dual encoders are generalizable retrievers. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 9844–9855. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.emnlp-main.669
13. Robertson, S.E., Walker, S.: Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In: SIGIR '94. p. 232–241. Springer London, London (1994). https://doi.org/10.5555/188490.188561
14. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at TREC-3. In: Harman, D.K. (ed.) Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994. NIST Special Publication, vol. 500-225, pp. 109–126. National Institute of Standards and Technology (NIST) (1994), http://trec.nist.gov/pubs/trec3/papers/city.ps.gz
15. Sager, P.J., Kamaraj, A., Grewe, B.F., Stadelmann, T.: 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), https://ceur-ws.org/Vol-4038/paper_89.pdf
16. Samarinas, C., Hsu, W., Lee, M.L.: Improving evidence retrieval for automated explainable fact-checking. In: Sil, A., Lin, X.V. (eds.) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations. pp. 84–91. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.naacl-demos.10
17. Schofield, J., Tian, S., Truong, H.T.T., Heil, M.: 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), https://ceur-ws.org/Vol-4038/paper_93.pdf
18. Schreieder, T., Färber, M.: 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), https://ceur-ws.org/Vol-4038/paper_94.pdf
19. Schreieder, T., Schopf, T., Färber, M.: Attribution, citation, and quotation: A survey of evidence-based text generation with large language models. In: Liakata, M., Moreira, V.P., Zhang, J., Jurgens, D. (eds.) Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

pp. 30956–31000. Association for Computational Linguistics, San Diego, California, United States (Jul 2026). https://doi.org/10.18653/v1/2026.acl-long.1430

20. Shavarani, H., Sarkar, A.: Entity retrieval for answering entity-centric questions. In: Shi, W., Yu, W., Asai, A., Jiang, M., Durrett, G., Hajishirzi, H., Zettlemoyer, L. (eds.) Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing. pp. 1–17. Association for Computational Linguistics, Albuquerque, New Mexico, USA (May 2025). https://doi.org/10.18653/v1/2025.knowledgenlp-1.1
21. Soleimani, A., Monz, C., Worring, M.: Bert for evidence retrieval and claim verification. In: Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F. (eds.) Advances in Information Retrieval. pp. 359–366. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-45442-5_45
22. Sriram, A., Xu, F., Choi, E., Durrett, G.: Contrastive learning to improve retrieval for real-world fact checking. In: Schlichtkrull, M., Chen, Y., Whitehouse, C., Deng, Z., Akhtar, M., Aly, R., Guo, Z., Christodoulopoulos, C., Cocarascu, O., Mittal, A., Thorne, J., Vlachos, A. (eds.) Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER). pp. 264–279. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.fever-1.28
23. Staudinger, M., El-Ebshihy, A., Kusa, W., Piroi, F., Hanbury, A.: 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), https://ceur-ws.org/Vol-4038/paper_98.pdf
24. Struß, J.M., Schellhammer, S., Dietze, S., V., V., Setty, V., Chakraborty, T., Nakov, P., Anand, A., Chungkham, P., Hafid, S., Sahnan, D., Todorov, K.: The clef-2026 checkthat! lab: Advancing multilingual fact-checking. In: Campos, R., Jatowt, A., Lan, Y., Aliannejadi, M., Bauer, C., MacAvaney, S., Anand, A., Ren, Z., Verberne, S., Bai, N., Mansoury, M. (eds.) Advances in Information Retrieval. pp. 325–335. Springer Nature Switzerland, Cham (2026). https://doi.org/10.1007/978-3-032-21321-1_43
25. Suarez-Lledo, V., Alvarez-Galvez, J.: Prevalence of health misinformation on social media: Systematic review. J. Med. Internet Res. **23**(1), e17187 (Jan 2021). https://doi.org/10.2196/17187
26. Sundriyal, M., Malhotra, G., Akhtar, M.S., Sengupta, S., Fano, A., Chakraborty, T.: Document retrieval and claim verification to mitigate COVID-19 misinformation. In: Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations. pp. 66–74. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.constraint-1.8
27. Swire-Thompson, B., Lazer, D.: Public health and online misinformation: Challenges and recommendations. Annual Review of Public Health **41**(Volume 41, 2020), 433–451 (2020). https://doi.org/10.1146/annurev-publhealth-040119-094127
28. Upadhyay, R., Viviani, M.: Enhancing health information retrieval with RAG by prioritizing topical relevance and factual accuracy. Discover Computing **28**(1), 27 (2025). https://doi.org/10.1007/s10791-025-09505-5
29. Wang, F., Li, Y., Xiao, H.: jina-reranker-v3: Last but not late interaction for listwise document reranking (2025), https://arxiv.org/abs/2509.25085

30. Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., Wei, F.: Text embeddings by weakly-supervised contrastive pre-training (2024), https://arxiv.org/abs/2212.03533
31. Zhang, H., Zhang, R., Guo, J., de Rijke, M., Fan, Y., Cheng, X.: From relevance to utility: Evidence retrieval with feedback for fact verification. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 6373–6384. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.findings-emnlp.422
32. Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., Zhou, J.: Qwen3 embedding: Advancing text embedding and reranking through foundation models (2025), https://arxiv.org/abs/2506.05176
33. Zuo, C., Wang, C., Banerjee, R.: Cross-genre retrieval for information integrity: A covid-19 case study. In: Yang, X., Suhartanto, H., Wang, G., Wang, B., Jiang, J., Li, B., Zhu, H., Cui, N. (eds.) Advanced Data Mining and Applications. pp. 495–509. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-46677-9_34