Infraglyph at CheckThat! 2026: Deep Hybrid Multi-Stage Scientific Source Retrieval from Implicit Social Media Claims*

Harsh Patel[1,*,†], Urvi Kava[1,*,†], Thomas Mandl[2,1] and Sandip Modha[1]

[1]Dhirubhai Ambani University, Gandhinagar, Gujarat, India
[2]Universität Hildesheim, Germany

Abstract

The proliferation of scientific claims on social media, often stated without explicit citations and across multiple languages, poses a significant challenge for automated fact-checking systems. This paper presents a system developed by Team Infraglyph for Task 1 of the CheckThat! Lab at CLEF 2026, which requires retrieving the correct scientific publication from a collection of approximately 10,000 papers given a multilingual social media post that implicitly references it. We propose a deep hybrid multi-stage retrieval pipeline that integrates lexical retrieval (BM25), dense semantic retrieval (multilingual-E5), neural sparse retrieval (SPLADE), neural machine translation (MarianMT and NLLB-200), Reciprocal Rank Fusion (RRF), and cross-encoder reranking. Comprehensive ablation studies suggest that translation is among the most influential components of the pipeline for cross-lingual retrieval. Without translation, BM25-based hybrid retrieval performs poorly on German (MRR@5 = 0.1035) and French (MRR@5 = 0.1968) queries. MarianMT increases German MRR@5 from 0.1035 obtained by the Hybrid BM25+Dense system without translation to 0.4032 in the Deep Hybrid + MarianMT configuration, corresponding to approximately a **3.9× improvement**. The proposed deep hybrid pipeline consistently outperforms all individual retrieval strategies across multilingual evaluation settings, achieving a 14.7% relative gain over the dense-only baseline. On the official CLEF 2026 leaderboard, our best submission achieves an average MRR@5 of **0.49** in the development phase (Rank 30) and **0.4825** in the evaluation phase (Rank 33), with per-language scores of EN 0.5056, DE 0.4319, and FR 0.5099.

Keywords

scientific source retrieval, hybrid retrieval, dense retrieval, BM25, cross-encoder reranking, multilingual retrieval, reciprocal rank fusion, machine translation, CheckThat! 2026, CLEF 2026

1. Introduction

Social media platforms have become a primary channel through which scientific information is communicated to the general public. While this enables rapid dissemination of research findings, it also introduces critical challenges in verification and traceability. Scientific claims shared on Twitter/X or similar platforms are typically expressed in simplified, paraphrased, or informal forms, and crucially lack explicit references such as URLs, DOIs, or formal citations. This creates a fundamental problem for automated fact-checking systems: identifying the original scientific publication behind an implicitly referenced claim.

Task 1 of the CLEF-2026 CheckThat! Lab on Source Retrieval for Scientific Web Claims [1] formalises this problem as a retrieval task. Given a social media post that implicitly references a scientific paper, the system must retrieve the correct publication from a collection of approximately 10,000 candidate

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

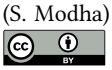
*Code and configuration available at: https://github.com/harshpatel080503/Deep-Hybrid-Multi-Stage-Scientific-Source-Retrieval-from-Implicit-Social-Media-Claims

*Corresponding author.

†These authors contributed equally.

harshpatel080503@gmail.com (H. Patel); urvikawa2004@gmail.com (U. Kava); mandl@uni-hildesheim.de (T. Mandl); sandip_modha@dau.ac.in (S. Modha)

0009-0007-5694-6486 (H. Patel); 0009-0009-5418-0768 (U. Kava); 0000-0002-8398-9699 (T. Mandl); 0000-0003-2427-2433 (S. Modha)



papers. The task is evaluated using Mean Reciprocal Rank at rank 5 (MRR@5), separately for English (EN), German (DE), and French (FR) queries, making multilingual alignment a central challenge.

This task presents several unique challenges. First, scientific claims are often expressed through implicit references, where no explicit identifiers link a social media post to its corresponding publication. Second, there exists a substantial stylistic mismatch between informal and conversational social media language and the structured technical language used in scientific articles. Third, the multilingual nature of the task introduces cross-lingual retrieval difficulties because queries are provided in English, German, and French, while the document collection is predominantly English. In addition, scientific findings are frequently paraphrased, simplified, or exaggerated on social media, creating semantic variability that limits the effectiveness of exact keyword matching. Finally, social media posts often contain hashtags, emojis, abbreviations, and informal grammar, introducing noise that can negatively affect retrieval performance.

Contributions.

1. A multilingual deep hybrid retrieval pipeline integrating BM25, multilingual-E5 dense retrieval, MarianMT/NLLB-200 translation, RRF fusion, and cross-encoder reranking for implicit scientific source retrieval across EN/DE/FR query sets.
2. A comprehensive ablation study quantifying the contribution of each pipeline component, demonstrating that translation is associated with an approximately 3.9× improvement in German MRR@5 within our experimental setting.
3. An evaluation of five pipeline variants (baseline dense, Dense+CE, Hybrid, SPLADE, and Field-Aware BM25), providing a systematic comparison that identifies the best configuration for each language.
4. Evaluation on the official CLEF 2026 CheckThat! Task 1 benchmark achieving an average MRR@5 of 0.49 on the development phase and 0.4825 on the evaluation phase, providing insights into multilingual scientific source retrieval under a zero-shot setting.

The remainder of this paper is organised as follows. Section 2 reviews related work. Section 3 describes the task, dataset, and problem formulation. Section 4 presents the proposed methodology. Section 5 details the experimental setup. Section 6 reports results, analysis, and error discussion. Section 7 concludes the paper.

2. Related Work

Information retrieval has evolved from traditional lexical matching approaches toward semantic and hybrid neural retrieval architectures. Sparse retrieval models such as BM25 [2] remain strong baselines due to their effectiveness in matching exact terminology, named entities, and numerical expressions, but they struggle with paraphrasing and cross-lingual retrieval because of their dependence on lexical overlap. To address these limitations, dense retrieval models based on transformer architectures [3] introduced shared semantic embedding spaces with efficient nearest-neighbour search using FAISS [4]. Multilingual dense encoders such as multilingual-E5 [5] further improved cross-lingual semantic alignment, while neural sparse retrieval models like SPLADE [6] combined sparse indexing with contextual semantic expansion. Recent retrieval systems increasingly employ hybrid pipelines that integrate sparse and dense retrieval using Reciprocal Rank Fusion (RRF) [7], followed by cross-encoder rerankers [8] for fine-grained relevance estimation. In multilingual retrieval settings, neural machine translation models such as MarianMT [9] and NLLB-200 [10] have become important for reducing lexical mismatch between multilingual queries and predominantly English document collections.

The broader CheckThat! Lab addresses multiple multilingual fact-checking challenges, including source retrieval, claim verification, claim normalisation, and fact-checking article generation [11]. Within this framework, recent CheckThat! Lab systems have explored increasingly sophisticated multi-stage retrieval pipelines for scientific claim-source retrieval. AIRwaves [12] combined BM25, fine-tuned

E5 dual encoders, and SciBERT reranking for tweet-to-paper matching. Deep Retrieval [13] employed hybrid retrieval with FAISS indexing and LLM-based reranking, demonstrating the effectiveness of combining lexical and semantic retrieval. ATOM [14] investigated pointwise and listwise reranking strategies with enriched scientific document representations, while Claim2Source [15] explored stylistic divergence between social media claims and scientific publications using zero-shot style transfer. SeRRa [16] further extended this direction through sequential multi-stage reranking pipelines with challenging negative sampling. Collectively, these studies demonstrate a transition from purely lexical retrieval toward hybrid multilingual retrieval and neural reranking architectures for implicit scientific source retrieval.

Despite the effectiveness of recent retrieval systems, most prior approaches either rely on supervised fine-tuning, focus primarily on English retrieval, or do not systematically investigate the interaction between translation, sparse retrieval, dense retrieval, and reranking within a unified multilingual framework. This motivates our investigation of a modular hybrid retrieval architecture designed for multilingual scientific source retrieval under a zero-shot setting.

3. Task Description

3.1. Task

Task 1 of the CheckThat! Lab at CLEF 2026 [1] addresses the problem of *Source Retrieval for Scientific Web Claims*. Given a multilingual social media post containing a scientific claim and an implicit reference to a scientific publication, the objective is to retrieve the correct paper from a collection of candidate scientific documents. Unlike traditional citation retrieval, the task does not provide explicit identifiers such as URLs, DOIs, titles, or author names, requiring systems to bridge the semantic gap between informal social media language and formal scientific writing. The task is multilingual, comprising English (EN), German (DE), and French (FR) queries, while the scientific document collection is predominantly in English. This introduces additional challenges related to paraphrasing, lexical mismatch, noisy social media text, and cross-lingual retrieval.

3.2. Dataset

The dataset for Task 1 is publicly available on Hugging Face [17]. It contains a shared scientific document collection together with multilingual query sets for English, German, and French. The collection consists of approximately 10,000 scientific publications spanning multiple research domains and includes four structured metadata fields: **title**, **abstract**, **venue**, and **authors**. Following the official dataset structure, we construct a unified document representation by concatenating all fields:

$$d = \text{title} \parallel \text{abstract} \parallel \text{venue} \parallel \text{authors} \tag{1}$$

The query sets contain multilingual social media posts collected from platforms such as Twitter/X and are divided into training, development, and test splits for each language. Each query is paired with the corresponding publication identifier (`pubkey`) of the referenced scientific paper. The posts are typically short, informal, noisy, and highly paraphrased, often containing hashtags, emojis, abbreviations, URLs, and conversational language, making retrieval substantially more challenging than conventional document search tasks.

3.3. Problem Formulation

Given a query q in language $\ell \in \{\text{EN}, \text{DE}, \text{FR}\}$ and a scientific document collection $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$, the goal is to retrieve the correct publication $d^* \in \mathcal{D}$ such that:

$$d^* = \arg\max_{d \in \mathcal{D}} \text{score}(q, d) \tag{2}$$

The system returns a ranked list of top-5 candidate documents for each query. System performance is evaluated using Mean Reciprocal Rank at rank 5 (MRR@5):

$$\text{MRR@5} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \tag{3}$$

where $\text{rank}_i \in \{1, \ldots, 5\}$ denotes the rank position of the correct document for query i, and contributes 0 if the correct document does not appear within the top-5 retrieved results. The evaluation metric is computed separately for English, German, and French query sets.

4. Methodology

Figure 1 illustrates the proposed deep hybrid multi-stage retrieval pipeline. The architecture consists of three major stages: (1) preprocessing and multilingual translation, (2) hybrid document retrieval using lexical and semantic retrieval models, and (3) cross-encoder reranking for final ranking refinement. The overall pipeline is designed to address lexical mismatch, multilingual variation, and semantic divergence between informal social media claims and formal scientific publications. Semantic divergence is addressed primarily through multilingual-E5 dense retrieval and cross-encoder reranking, which leverage contextual semantic representations rather than exact lexical overlap. In addition, hashtag expansion and text normalisation reduce informal social-media-specific variations before retrieval.

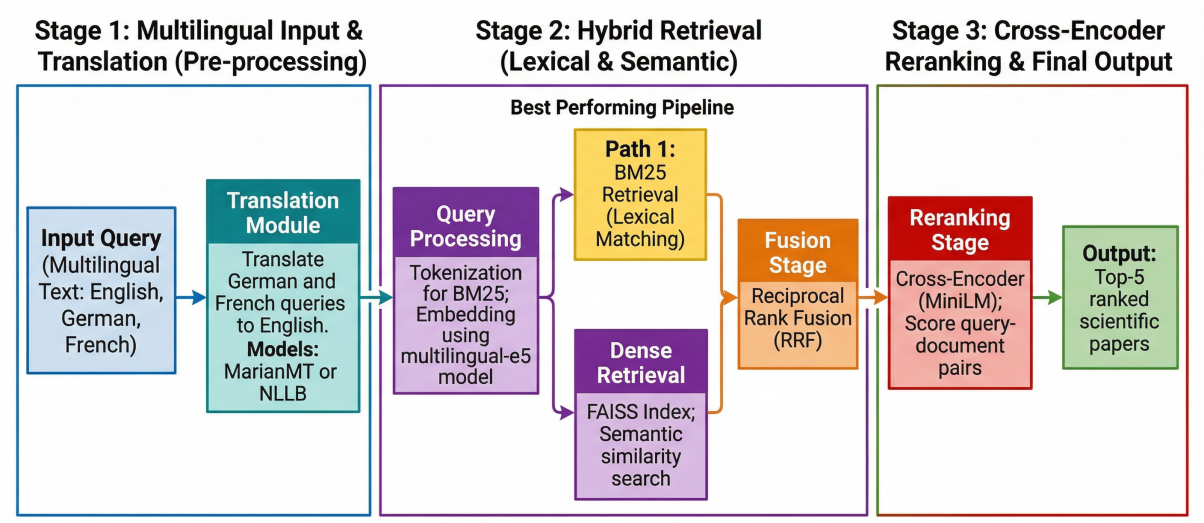


Figure 1: Proposed Deep Hybrid Multi-Stage Retrieval Pipeline. Stage 1 normalises and translates non-English queries (DE/FR→EN). Stage 2 runs parallel BM25 and dense retrieval. Stage 3 merges ranked lists via RRF and refines results using a cross-encoder.
Source: Workflow conceptualised by the authors; figure visually generated using Google Gemini AI.

4.1. Query Preprocessing and Translation

All queries undergo preprocessing before retrieval. This includes Unicode NFKC normalisation, lowercasing, hashtag expansion (removing # while retaining the associated word), URL removal, and whitespace normalisation. For German queries, Twitter mentions and non-semantic special characters are additionally removed to reduce retrieval noise. Since the scientific document collection is predominantly in English, all non-English queries are translated into English prior to retrieval:

$$q_{\text{en}} = \mathcal{T}(q, \ell \rightarrow \text{en}) \tag{4}$$

We evaluate two neural machine translation systems: MarianMT [9], using the Helsinki-NLP `opus-mt-de-en` and `opus-mt-fr-en` models, and NLLB-200 [10], using Facebook's `nllb-200-distilled-600M` multilingual translation model. MarianMT was selected for the primary submission due to its stronger development-set performance and more consistent translation of scientific terminology.

4.2. Hybrid Document Retrieval

The retrieval stage combines lexical, dense semantic, and neural sparse retrieval methods to exploit complementary retrieval signals. For lexical retrieval, we employ BM25Okapi [2] with default parameters ($k_1 = 1.5, b = 0.75$):

$$\text{score}_{\text{BM25}}(q, d) = \sum_{t \in q} \text{IDF}(t) \cdot \frac{f(t, d)(k_1 + 1)}{f(t, d) + k_1(1 - b + b \cdot |d|/\text{avgdl})} \tag{5}$$

BM25 is highly effective for matching named entities, numerical expressions, and domain-specific terminology, but without translation it performs poorly for German and French queries due to the absence of lexical overlap with English documents. Tokenisation includes Unicode normalisation, lowercasing, ASCII transliteration using `unidecode`, and punctuation removal.

For semantic retrieval, we use `intfloat/multilingual-e5-base` [5] with E5-style query and passage prefixes and L2-normalised embeddings:

$$\text{score}_{\text{dense}}(q, d) = \hat{q} \cdot \hat{d}, \quad \hat{q} = \frac{f(q_{\text{en}})}{\|f(q_{\text{en}})\|}, \quad \hat{d} = \frac{f(d)}{\|f(d)\|} \tag{6}$$

Dense retrieval is implemented using FAISS `IndexFlatIP` [4] for exact inner-product nearest-neighbour search. Unlike BM25, multilingual-E5 provides stronger robustness to cross-lingual semantic variation through its shared multilingual embedding space.

As an additional ablation, we evaluate SPLADE [6] (`naver/splade-cocondenser-ensembledistil`) as a neural sparse retrieval alternative to BM25. SPLADE generates vocabulary-space sparse representations through masked language modelling:

$$\text{sparse}(x) = \max_{\text{tokens}} \log(1 + \text{ReLU}(\text{MLM}(x))) \tag{7}$$

To compensate for the increased morphological complexity of German queries, we apply language-adaptive retrieval depth. English and French queries use BM25 top-30 and Dense top-150 retrieval, while German queries use deeper retrieval with BM25 top-50 and Dense top-300 candidates. We additionally evaluate a field-aware retrieval variant that maintains separate BM25 indexes for document titles and abstracts, producing three retrieval signals (BM25-Title, BM25-Abstract, and Dense retrieval) fused using three-way Reciprocal Rank Fusion.

4.3. Cross-Encoder Reranking

The ranked candidate lists generated by BM25 and dense retrieval are merged using Reciprocal Rank Fusion (RRF) [7] with smoothing constant $k = 60$:

$$\text{RRF}(d) = \sum_{r \in \mathcal{R}} \frac{1}{k + \text{rank}_r(d)} \tag{8}$$

RRF effectively combines lexical and semantic retrieval signals without requiring score normalisation and is robust to score distribution differences between retrieval models.

The fused candidate documents are subsequently reranked using the cross-encoder model `cross-encoder/ms-marco-MiniLM-L-6-v2` [8], which jointly encodes query-document pairs:

$$\text{score}_{\text{CE}}(q, d) = g([q_{\text{en}}; d]) \tag{9}$$

Unlike bi-encoder retrieval models, cross-encoders capture token-level interactions between queries and documents, enabling more precise relevance estimation. The reranking stage operates on the top-50 candidates for English and French queries and the top-100 candidates for German queries due to the higher retrieval ambiguity observed in German.

5. Experimental Setup

This section describes the experimental configuration used to evaluate the proposed multilingual hybrid retrieval pipeline. We first present the retrieval and translation models used across different experimental variants, followed by implementation details and language-adaptive retrieval settings employed throughout the experiments. The overall setup is designed to evaluate the contribution of lexical retrieval, dense semantic retrieval, neural sparse retrieval, translation strategies, and cross-encoder reranking for multilingual scientific source retrieval.

5.1. Models and Components

Table 1 summarises the primary models and components used across all experimental variants, including the main hybrid pipeline, translation ablations, SPLADE-based retrieval, and field-aware retrieval experiments.

Table 1
Models and components across experiments.

Component	Model	Used In
Dense (baseline)	`multilingual-e5-large`	Baseline, Hybrid (no-trans.), Dense+CE
Dense (main)	`multilingual-e5-base`	Main, SPLADE, Field-Aware
Translation (primary)	`opus-mt-{de,fr}-en`	Main, SPLADE, Field-Aware
Translation (ablation)	`nllb-200-distilled-600M`	Translation ablation
Sparse	BM25Okapi	Main, Hybrid, Field-Aware
Neural sparse	`splade-cocondenser-ensembledistil`	SPLADE variant
Reranker (main)	`ms-marco-MiniLM-L-6-v2`	All reranking systems
Reranker (ablation)	`mmarco-mMiniLMv2-L12-H384-v1`	Dense+CE variant
Vector search	FAISS `IndexFlatIP`	All dense systems

5.2. Implementation Details and Retrieval Configuration

All experiments were conducted using a CUDA-enabled GPU with CPU fallback when GPU resources were unavailable. Document embeddings were computed offline and indexed using FAISS `IndexFlatIP` for efficient dense retrieval. BM25 preprocessing applied Unicode NFKC normalisation, lowercasing, ASCII transliteration using `unidecode`, and punctuation removal. Neural machine translation was executed on CPU to avoid GPU memory conflicts during retrieval and reranking. Cross-encoder inference used a batch size of 32.

To improve multilingual retrieval robustness, we employed language-adaptive retrieval depth based on the observed complexity of each language. English and French queries used BM25 top-30 retrieval, Dense top-150 retrieval, and reranking over the top-50 candidates. German queries required deeper retrieval due to higher morphological complexity and lexical variation. Preliminary development-set experiments indicated that increasing retrieval depth improved German recall more consistently than for English and French, motivating the language-specific configuration adopted in this work. Therefore, we used BM25 top-50 retrieval, Dense top-300 retrieval, and reranking over the top-100 candidates. The complete retrieval configuration used in the main pipeline is summarised in Table 2.

Table 2
Language-adaptive retrieval depth for the main pipeline.

Language	BM25 top-k	Dense top-k	Rerank top-k
English	30	150	50
French	30	150	50
German	50	300	100

6. Results and Analysis

This section presents the experimental results of the proposed multilingual deep hybrid retrieval pipeline on the CheckThat! 2026 Task 1 dataset. We first compare all retrieval variants on the development set to analyse the contribution of translation, hybrid retrieval, reranking, and neural sparse retrieval components. We then report the official leaderboard results from both the development and evaluation phases of the competition, followed by an analysis of multilingual retrieval behaviour, recurring failure modes, and current system limitations.

6.1. Development Set Ablation

Table 3
Development-set MRR@5 for all system variants and relative gain over the dense baseline. **Bold** indicates the best score per column.

System	EN	DE	FR	Avg.	Δ Avg.
Single-stage baselines (no translation)					
Baseline Dense (e5-large)	0.4671	0.3699	0.4524	0.4298	–
Dense + Cross-Encoder	0.4714	0.2031	0.2417	0.3054	−0.1244
Hybrid BM25+Dense (no translation)	0.5311	0.1035	0.1968	0.2771	−0.1527
Full-pipeline variants (with translation)					
Deep Hybrid + MarianMT (main)	**0.5304**	0.4032	0.5449	**0.4928**	**+0.0630**
Deep Hybrid + NLLB-200	**0.5304**	0.3744	**0.5455**	0.4834	+0.0536
Deep Hybrid + SPLADE	0.5221	**0.4047**	0.5249	0.4839	+0.0541
Deep Hybrid + BGE Reranker	0.5153	0.3649	0.5477	0.4760	+0.0462
Field-Aware BM25 + MarianMT	0.5273	0.4046	0.5398	0.4906	+0.0608

The ablation results suggest that multilingual translation is among the most influential components of the proposed pipeline, particularly for German and French retrieval, although the experiments do not isolate translation effects completely from other retrieval interactions. The complete Deep Hybrid + MarianMT system achieves the best overall performance with an average MRR@5 of 0.4928, improving upon the dense-only baseline by 14.7%. Without translation, BM25-based hybrid retrieval fails for German and French queries due to the absence of lexical overlap with English documents. Combining BM25 and dense retrieval consistently improves English retrieval quality, confirming that lexical and semantic signals are complementary for scientific claim retrieval. MarianMT slightly outperforms NLLB-200, while SPLADE and field-aware BM25 provide only marginal gains at higher computational cost.

6.2. Official Leaderboard Results

To evaluate real-world system effectiveness, we report our official leaderboard submissions from both the development and evaluation phases of the CheckThat! 2026 competition. Tables 4 and 5 compare Team Infraglyph against the top-performing systems in each phase.

Table 4
Top development-phase leaderboard results together with Team Infraglyph.

Team	EN	DE	FR	Avg.
Claim2Source	0.79	0.72	0.81	0.77
Outsider	0.78	0.71	0.79	0.76
CheckMate	0.76	0.70	0.80	0.76
Boyes Boys	0.76	0.69	0.79	0.75
MeVer@CERTH	0.72	0.66	0.78	0.72
Infraglyph (Rank 30)	**0.53**	**0.40**	**0.54**	**0.49**

Table 5
Top evaluation-phase leaderboard results together with Team Infraglyph.

Team	EN	DE	FR	Avg.
Claim2Source	0.7819	0.7153	0.7912	0.7628
Outsider	0.7802	0.7228	0.7709	0.7580
Boyes Boys	0.7571	0.7070	0.7752	0.7464
CheckMate	0.7343	0.6861	0.7378	0.7194
RETRIX	0.7117	0.6806	0.7400	0.7108
Infraglyph (Rank 33)	**0.5056**	**0.4319**	**0.5099**	**0.4825**

The relatively lower ranking compared to top-performing systems can largely be attributed to the absence of task-specific retriever fine-tuning, hard-negative mining, and domain-adapted reranking models, all of which have been shown to substantially improve retrieval performance in prior CheckThat! systems.

Our official evaluation submission achieves an average MRR@5 of 0.4825 across English, German, and French queries. French achieves the strongest overall performance, while German remains the most challenging language across all experiments. Notably, German performance improves from 0.4000 on the development phase to 0.4319 on the evaluation phase. While this observation is encouraging, it should be interpreted cautiously because differences between development and evaluation distributions may also contribute to the observed change. English and French scores decrease moderately during evaluation, consistent with the expected distribution shift between development and hidden test sets.

Compared to prior systems from CLEF 2025, our approach operates entirely in a zero-shot multilingual setting without task-specific fine-tuning. The strongest 2025 systems achieved MRR@5 scores between 0.58 and 0.68 on English-only datasets using supervised bi-encoder fine-tuning and hard-negative training [12, 13, 14]. Despite these constraints, the proposed multilingual hybrid pipeline achieves competitive English retrieval performance while simultaneously supporting German and French scientific claim retrieval.

6.3. Impact of Translation

Translation appears to be one of the most influential components in the pipeline for multilingual retrieval. Without translation, BM25-based retrieval produces near-zero lexical overlap for German and very weak retrieval for French queries because the document collection is predominantly English. Although multilingual-E5 embeddings partially mitigate this issue, BM25 negatively affects RRF fusion when lexical overlap is absent. Applying MarianMT translation increases German retrieval performance from 0.1035 in the Hybrid BM25+Dense (no translation) configuration to 0.4032 in the Deep Hybrid + MarianMT configuration, corresponding to approximately 3.9$\times$ improvement, demonstrating that translation is essential for lexically grounded multilingual retrieval in this task.

6.4. German vs. French Performance Gap

German consistently achieves lower MRR@5 than English and French across almost all system variants. One major reason is the morphological complexity of German, including compound words and inflectional variation, which increases translation difficulty for scientific terminology. German social media posts also contain more code-switching and orthographic variation, introducing additional noise into the retrieval pipeline. Although deeper retrieval depth partially compensates for this issue, the gap is not fully eliminated.

6.5. Error Analysis

Inspection of development-set predictions reveals four recurring retrieval failure modes. First, high-degree paraphrasing reduces both lexical and semantic matching quality when claims describe scientific findings without mentioning key terminology. Second, near-duplicate papers discussing highly similar findings often create ambiguity when queries lack distinguishing quantitative or contextual information. Third, translation artefacts occasionally introduce unrecoverable retrieval noise through incorrect rendering of scientific terminology, gene names, numerical values, or German compound nouns. Finally, information-sparse queries containing very little scientific content frequently provide insufficient evidence for reliable retrieval regardless of retrieval strategy. For example, German compound scientific terms were occasionally translated into less specific English expressions, reducing lexical overlap with the corresponding publication. Similarly, highly paraphrased claims sometimes omitted key entities such as disease names, interventions, or study identifiers, making both lexical and semantic retrieval substantially more difficult.

6.6. Limitations

Our system has several limitations compared to the strongest systems from previous CheckThat! editions. First, we do not perform task-specific bi-encoder fine-tuning with hard negatives, which previous work has shown to provide substantial retrieval improvements. Second, the cross-encoder reranker is primarily optimised for English and may underperform when translated queries still contain multilingual artefacts. Third, we do not explore LLM-based rerankers such as `bge-reranker-v2-gemma` or GritLM, which demonstrated strong performance in recent retrieval systems. Finally, the current pipeline does not include query rewriting or style-transfer-based normalisation to bridge the linguistic gap between informal social media language and formal scientific writing.

7. Conclusion

We presented Team Infraglyph's system for Task 1 of the CheckThat! Lab at CLEF 2026, addressing the retrieval of scientific publications implicitly referenced in multilingual social media posts. Our deep hybrid multi-stage pipeline integrates BM25 lexical retrieval, multilingual-E5 dense retrieval, MarianMT machine translation, Reciprocal Rank Fusion, and cross-encoder reranking. Comprehensive ablation studies suggest that translation is among the most influential components for multilingual retrieval, yielding a 3.9× MRR@5 improvement for German and enabling a 14.7% relative improvement over the dense baseline. The full pipeline achieves average MRR@5 = 0.49 on the development leaderboard (Rank 30) and 0.4825 on the evaluation leaderboard (Rank 33).

Future Work. Several promising directions remain: (i) *bi-encoder fine-tuning* using the available training split with hard negatives, as demonstrated to be highly effective by 2025 participants [12, 13]; (ii) *LLM-based reranking* using models such as bge-reranker-v2-gemma; (iii) *LLM-based query rewriting* to formalise informal social media language before retrieval; (iv) *improved cross-lingual reranking* using multilingual cross-encoders trained on scientific text pairs; and (v) *ensemble reranking* combining multiple cross-encoders trained on diverse datasets.

Acknowledgments

The authors thank the organisers of the CheckThat! Lab 2026 for providing the dataset, evaluation infrastructure, and support throughout the competition. We also acknowledge the developers of the open-source tools used in this work: Hugging Face Transformers, Sentence-Transformers, FAISS, and BM25Okapi.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) to assist with refining and structuring the manuscript. After using this tool, the authors reviewed and edited all content and take full responsibility for the publication's content.

References

[1] 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: CLEF 2026 Working Notes, CLEF 2026, CEUR-WS.org, Jena, Germany, 2026.

[2] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford, Okapi at TREC-3, NIST Special Publication 500 (1994) 109–126. URL: https://trec.nist.gov/pubs/trec3/papers/city.ps.gz.

[3] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6769–6781. doi:10.18653/v1/2020.emnlp-main.550.

[4] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with GPUs, IEEE Transactions on Big Data 7 (2021) 535–547. doi:10.1109/TBDATA.2019.2921572.

[5] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual E5 text embeddings: A technical report, arXiv preprint arXiv:2402.05672 (2024). URL: https://arxiv.org/abs/2402.05672.

[6] T. Formal, B. Piwowarski, S. Clinchant, SPLADE: Sparse lexical and expansion model for first stage ranking, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2021, pp. 2288–2292. doi:10.1145/3404835.3463098.

[7] G. V. Cormack, C. L. A. Clarke, S. Buettcher, Reciprocal rank fusion outperforms Condorcet and individual rank learning methods, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2009, pp. 758–759. doi:10.1145/1571941.1572114.

[8] R. Nogueira, Z. Jiang, R. Pradeep, J. Lin, Document ranking with a pretrained sequence-to-sequence model, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 708–718. doi:10.18653/v1/2020.findings-emnlp.63.

[9] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. Fikri Aji, N. Bogoychev, A. F. T. Martins, A. Birch, Marian: Fast neural machine translation in C++, in: Proceedings of ACL 2018, System Demonstrations, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 116–121. doi:10.18653/v1/P18-4020.

[10] NLLB Team, M. R. Costa-jussà, J. Cross, et al., No language left behind: Scaling human-centered machine translation, arXiv preprint arXiv:2207.04672 (2022). URL: https://arxiv.org/abs/2207.04672.

[11] 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.

[12] C. Ashbaugh, L. Baumgärtner, T. Greß, N. Sidorov, D. Werner, AIRwaves at CheckThat! 2025: Retrieving scientific sources for implicit claims on social media with dual encoders and neural re-ranking, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, Madrid, Spain, 2025, pp. 826–844. URL: https://ceur-ws.org/Vol-4038/paper_63.pdf.

[13] 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, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, Madrid, Spain, 2025, pp. 1141–1155. URL: https://ceur-ws.org/Vol-4038/paper_89.pdf.

[14] 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, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, Madrid, Spain, 2025, pp. 1246–1255. URL: https://ceur-ws.org/Vol-4038/paper_98.pdf.

[15] 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, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, Madrid, Spain, 2025, pp. 1330–1343. URL: https://ceur-ws.org/Vol-4038/paper_94.pdf.

[16] G. A. Marchetti, G. Rocha, H. L. Cardoso, Team serra at CheckThat! clef 2025: Sequential re-ranking in a scientific claim source retrieval pipeline, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, Madrid, Spain, 2025, pp. 1217–1227. URL: https://ceur-ws.org/Vol-4038/paper_81.pdf.

[17] S. Schellhammer, CT26_Task1_SourceRetrievalForScientificWebClaims, Hugging Face Datasets, 2026. URL: https://huggingface.co/datasets/sschellhammer/CT26_Task1_SourceRetrievalForScientificWebClaims.