SINAI-UGPLN at CheckThat! 2026: Dense Retrieval and Neural Reranking for Scientific Source Retrieval from Social Media Claims

CheckThat! Lab at CLEF 2026

Mariuxi del Carmen Toapanta-Bernabé[1,2,*,†], Miguel Ángel García-Cumbreras[1,†] and Luis Alfonso Ureña-López[1,†]

[1]*Computer Science Department, SINAI, CEATIC, Universidad de Jaén, 23071, Jaén, Spain*
[2]*Universidad de Guayaquil, 090514, Guayas, Ecuador*

Abstract

This paper describes the SINAI-UGPLN team's participation in CheckThat! 2026 Task 1, Source Retrieval for Scientific Web Claims. The task consists of retrieving the scientific publication implicitly referenced in a social media post from a closed official collection of candidate papers. We approached the problem as a multilingual document-ranking task constrained to the official collection and to valid `pubkey` identifiers. Our system evaluated lexical retrieval with BM25, dense retrieval with `intfloat/multilingual-e5-large`, hybrid retrieval with Reciprocal Rank Fusion, and neural reranking with `BAAI/bge-reranker-v2-m3`. The official Final Evaluation submission was based on multilingual E5 dense retrieval, which provided the most stable validated configuration before the deadline. In Development, neural reranking over RRF top-60 candidates achieved the strongest score, improving the average MRR@5 from 0.4389 with E5 to 0.4832. These results show that multilingual dense retrieval is a robust baseline for scientific source retrieval, while neural reranking can further improve fine-grained claim–paper matching when the correct source is present among the retrieved candidates.

Keywords

Scientific source retrieval, fact-checking, dense retrieval, neural reranking, multilingual retrieval, CheckThat! 2026

1. Introduction

Scientific claims circulating on social media often refer to research findings without providing complete bibliographic information. Posts may cite studies indirectly, paraphrase results, omit authors or venues, or use non-technical language. In this setting, fact-checking requires not only assessing a claim but also identifying the scientific source on which the claim is based. This is particularly relevant for public health, climate, technology, social science, and biomedical claims, where a short post may compress, simplify, or distort the contribution of a full scientific publication.

CheckThat! 2026 Task 1, Source Retrieval for Scientific Web Claims, addresses this problem by requiring systems to retrieve the scientific publication implicitly referred to by a social media post [1]. Given a claim-like post and a fixed official collection of candidate papers, each system must return a ranked list of five candidate publications identified by their official `pubkey`. The task covers English, German, and French queries, while the candidate collection provides metadata for scientific papers, including titles, abstracts, venues, and authors.

The SINAI-UGPLN team approached the task as a closed-set, multilingual scientific document-ranking problem. All experiments were restricted to the official collection provided by the organizers, and all

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
†These authors contributed equally.
mctb0005@red.ujaen.es,mariuxi.toapantab@ug.edu.ec (M. d. C. Toapanta-Bernabé); magc@ujaen.es (M. : García-Cumbreras); laurena@ujaen.es (L. A. Ureña-López)
0000-0002-4839-7452 (M. d. C. Toapanta-Bernabé); 0000-0003-1867-9587 (M. : García-Cumbreras); 0000-0001-7540-4059 (L. A. Ureña-López)

submitted predictions used valid pubkey identifiers. Our participation builds on previous CheckThat! work on numerical claim verification [2], while the present task focuses on scientific source retrieval using lexical retrieval [3], multilingual dense retrieval [4], hybrid ranking [5], and neural reranking based on the BGE-M3 family [6].

The main contributions of this paper are threefold. First, we present a constrained retrieval pipeline that operates exclusively over the official candidate collection and validates all predicted pubkey identifiers before submission. Second, we compare lexical, dense, hybrid, and reranking-based retrieval strategies under the same multilingual evaluation setting. Third, we analyze the language-dependent behavior of retrieval methods, showing that lexical–dense fusion benefits English but can degrade German and French unless controlled by language or followed by neural reranking.

The official Final Evaluation submission was based on multilingual E5 dense retrieval, which provided the most stable validated configuration before the deadline. In addition, we report Development experiments with language-aware fusion and BGE reranking to analyze the potential of stronger candidate reranking architectures for scientific source retrieval.

2. Task Description

CheckThat! 2026 Task 1, Source Retrieval for Scientific Web Claims, is formulated as a scientific source retrieval task according to the official task description [7]. Given a social media post containing a scientific claim and an implicit reference to a scientific paper, without a URL, the system must retrieve the mentioned paper from a pool of candidate papers. Unlike veracity classification, the goal is not to determine whether the claim is true or false, but to recover the source paper that the post implicitly references.

The task is multilingual and includes English, German, and French queries. The candidate collection is fixed and closed: systems must retrieve papers only from the official collection provided by the organizers. Each candidate paper is identified by a unique pubkey and described through metadata fields such as `title`, `abstract`, `venue`, and `authors`. For each query, the expected output is a ranked list of five pubkey values.

The official metric is Mean Reciprocal Rank at 5 (MRR@5). For each query, the score is determined by the reciprocal rank of the correct publication if it appears among the first five predictions; otherwise, the query receives a score of zero. This metric rewards systems that not only retrieve the correct paper but place it near the top of the ranked list.

From a retrieval perspective, the task combines several challenges. Social media posts are short, informal, and often incomplete, while candidate papers are represented by formal scientific metadata. Moreover, the multilingual setting introduces lexical and semantic mismatches between the query and the candidate abstract. These characteristics motivated our use of multilingual dense retrieval and reranking over a closed official collection.

3. Background and Motivation

Scientific source retrieval is related to evidence retrieval, scientific fact-checking, and claim verification, but it has a distinctive objective: the system must identify the specific publication associated with a claim-like social media post. The relevant item is not merely a supporting passage or a related document, but the exact scientific source indexed in the official collection. This makes the task sensitive to fine-grained differences between papers that may share similar terminology, topics, or abstracts.

Previous CheckThat! tasks have addressed different parts of the fact-checking pipeline, including claim detection, evidence retrieval, and numerical claim verification [1]. Our previous participation in CheckThat! 2025 focused on numerical claim verification using a meta-ensemble strategy [2]. In contrast, the 2026 Task 1 setting requires retrieving a source publication rather than assigning a veracity label. Nevertheless, both tasks share a common requirement: connecting short, noisy claims with reliable evidence under the constraints of an official benchmark.

A related direction is scientific claim-source retrieval. In CheckThat! 2025, the Claim2Source system studied scientific claim-source retrieval using sparse, dense, and hybrid retrieval approaches [8]. This line of work is relevant because social media claims often lack formal bibliographic cues and may express scientific findings in simplified language. These characteristics motivate the use of semantic retrieval models and reranking strategies.

Lexical retrieval models such as BM25 remain useful baselines because they capture exact term overlap between queries and candidate documents [3]. However, their effectiveness decreases when posts paraphrase scientific findings, use informal wording, or appear in a language different from the dominant language of the candidate abstract. Dense retrieval methods address this limitation by representing queries and documents in a shared semantic space. In our system, multilingual E5 embeddings were used to connect English, German, and French posts with candidate scientific papers [4].

Hybrid retrieval and neural reranking provide complementary mechanisms. Reciprocal Rank Fusion can combine sparse and dense rankings without requiring score calibration [5], while rerankers such as BGE-M3 can directly score claim–document pairs [6]. These methods motivated our architecture: first retrieve candidates efficiently from the official collection, then analyze whether fusion or reranking improves the final top-five ranking.

4. System Architecture

The proposed system follows a closed-set retrieval architecture designed for the official constraints of Task 1. The pipeline does not search the open Web or query external scientific databases. Instead, all stages operate over the candidate papers provided in the official collection. The architecture is composed of four main layers: candidate representation, first-stage retrieval, optional fusion and reranking, and submission validation.

Figure 1 summarizes the complete workflow. A social media post is used as the query, while each scientific paper is represented using its title and abstract. First-stage retrieval is performed with BM25 and multilingual E5. The resulting rankings can be submitted directly, fused with Reciprocal Rank Fusion, or passed to a neural reranker. The final output is a top-five list of valid `pubkey` identifiers.

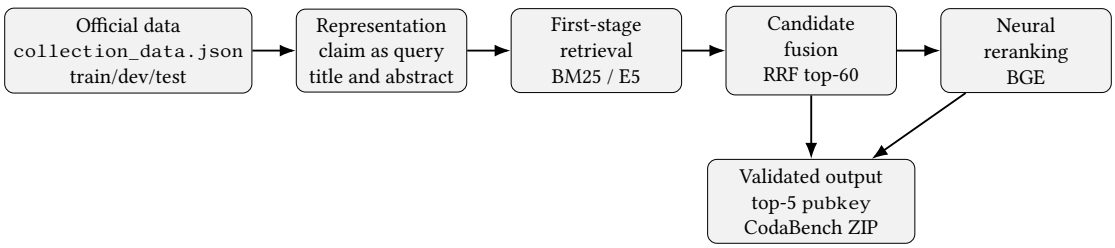

Figure 1: Closed-set retrieval and reranking pipeline for CheckThat! 2026 Task 1.

4.1. Closed-Set Candidate Collection

The system is constrained by the official candidate collection. Each valid answer must correspond to a `pubkey` contained in `collection_data.json`. This constraint is central to the system design because a semantically relevant external paper cannot be used unless it is part of the official collection. Therefore, the collection was treated as a fixed retrieval index.

The official collection contains paper metadata including `title`, `abstract`, `venue`, and `authors`. Although all fields were available, the core retrieval representation was based on title and abstract. This decision was motivated by the nature of the task: social media posts usually refer to a paper's topic or findings rather than its venue or full author list.

4.2. Query and Document Representation

Each social media post was represented as a query. The post may contain informal wording, hashtags, user mentions, incomplete references, or paraphrased scientific findings. Therefore, the query representation had to preserve the original semantic content while avoiding excessive normalization that could remove useful scientific terms.

Each candidate paper was represented by concatenating its title and abstract. The title provides a compact description of the scientific topic, while the abstract provides context about the research question, method, population, and findings. This representation was used consistently across BM25, E5, RRF, and BGE experiments.

For E5-based retrieval, we followed the query–passage convention of the model family. Social media posts were encoded as queries, and scientific papers were encoded as passages. For BGE reranking, the input pair consisted of the original claim-like post and the candidate paper text.

4.3. First-Stage Retrieval

The first-stage retrieval layer was responsible for producing candidate rankings from the full official collection. We implemented two complementary retrieval strategies.

The lexical strategy used BM25 over the candidate paper representation. BM25 captures exact term overlap and is useful when the post mentions distinctive scientific terms, disease names, treatments, or title fragments. However, it is sensitive to lexical mismatch and was expected to be weaker for German and French posts.

The dense strategy used `intfloat/multilingual-e5-large`. This model maps queries and documents into a shared multilingual semantic space. Candidate papers were ranked by similarity between normalized query and passage embeddings. This stage was designed to reduce the gap between informal social media language and formal scientific abstracts.

4.4. Hybrid Fusion

To analyze whether lexical and dense evidence were complementary, BM25 and E5 rankings were combined using Reciprocal Rank Fusion. RRF was selected because it directly combines rankings and does not require calibration of scores between BM25 and embedding-based similarities.

The global fusion setting used the same RRF configuration for all languages. However, the Development results showed that retrieval behavior was language-dependent. Lexical evidence improved English but introduced noise for German and French. For this reason, we also analyzed a language-aware strategy in which RRF was applied to English while E5 was kept for German and French.

4.5. Neural Reranking

The final architecture component was a neural reranker based on `BAAI/bge-reranker-v2-m3`. The reranker was applied only to a compact candidate set instead of the full collection. Specifically, candidates were first retrieved and fused, and the top-60 candidates were then reranked by scoring each claim–paper pair.

This reranking stage was designed to capture fine-grained semantic relations that may not be fully represented by BM25 or embedding similarity alone. It is particularly useful when several candidate papers discuss similar topics and the system must identify the exact publication referenced in the post. Since reranking is computationally more expensive, it was evaluated as an additional Development experiment.

4.6. Submission Validation

Output validation was implemented as part of the pipeline. Before generating the final ZIP files, each prediction list was checked to ensure that it contained exactly five identifiers, that all identifiers were integers, and that every identifier belonged to the official `pubkey` set.

The final submission format consisted of three TSV files, one per language: `predictions_en.tsv`, `predictions_de.tsv`, and `predictions_fr.tsv`. Each file contained the columns `index` and `preds`. The three files were compressed into a single ZIP archive for CodaBench submission.

5. Experimental Setup

The experiments were conducted on the official CheckThat! 2026 Task 1 data. The objective of the setup was to compare retrieval configurations under the same closed-set constraint and to identify a stable system for the Final Evaluation phase.

5.1. Dataset and Splits

The official Task 1 overview describes the shared task resources, including the official data used for source retrieval experiments [7]. In our experiments, the dataset was accessed via the Hugging Face repository `sschellhammer/CT26_Task1_SourceRetrievalForScientificWebClaims`. The repository includes train and Development query sets for English, German, and French, together with a single shared publication collection used across all languages and splits. Access requires completing the Declaration of Commitment and authenticating the runtime with a Hugging Face account granted access to the dataset.

Table 1
Official data used in Task 1 experiments.

Language	Train	Development	Test
English	14,977	3,905	6,076
German	1,460	386	876
French	2,807	702	1,220

The Development set was used for model selection and ablation analysis. The test labels were not available during the competition, so all decisions about the official submission were based on Development performance and stability across languages.

5.2. Retrieval Configurations

Table 2 summarizes the configurations evaluated during the experimental process. All configurations used the same official candidate collection and the same output validation procedure.

Table 2
Retrieval and reranking configurations evaluated in Development.

Configuration	Strategy	Purpose
BM25	Lexical retrieval over title and abstract	Lexical baseline and output validation
E5 dense retrieval	Multilingual E5 query–passage retrieval	Main multilingual retrieval system and official Final Evaluation run
BM25 and E5 RRF	Reciprocal Rank Fusion over BM25 and E5 rankings	Hybrid retrieval analysis over the official collection
Language-aware selection	RRF for English; E5 for German and French	Language-specific retrieval control
BGE reranking	BGE reranking over RRF top-60 candidates	Development reranking analysis over a compact candidate set

The BM25 configuration was used as a lexical baseline. The E5 configuration was used as the primary dense retrieval system and served as the official Final Evaluation submission. RRF and language-aware selection were used to analyze the complementarity between sparse and dense retrieval. BGE reranking was evaluated as a stronger second-stage ranking model over a compact candidate set.

5.3. Parameters and Implementation Details

For BM25, candidate documents were constructed from the titles and abstracts of each scientific paper. Queries were obtained from the social media posts. Basic text normalization was applied before indexing and retrieval.

For dense retrieval, we used `intfloat/multilingual-e5-large`. Candidate paper embeddings were computed once and reused for Development and test inference. Query embeddings were generated independently for each language split. Documents were ranked using similarity between normalized query and passage embeddings.

For RRF, BM25 and E5 ranked lists were combined at the rank level. A global configuration with $c = 60$ was evaluated first. Additional analysis showed that a lower fusion constant benefited English, so a language-aware variant was also evaluated.

For neural reranking, we used `BAAI/bge-reranker-v2-m3`. The reranker was applied to the top-60 candidates obtained from hybrid retrieval. Each input pair consisted of the social media post and the candidate paper text. The reranker produced a relevance score for each pair, and candidates were sorted according to this score.

5.4. Evaluation Protocol

The official evaluation metric is MRR@5. For each query, the score is the reciprocal rank of the correct `pubkey` if it appears among the first five predictions. If the correct paper is not present in the top five, the score is zero. We report MRR@5 for English, German, and French, as well as the average across the three languages.

5.5. Official Submission Policy

The official Final Evaluation submission was based on multilingual E5 dense retrieval with `intfloat/multilingual-e5-large`. This configuration was selected before the final deadline because it provided the most stable Development performance across English, German, and French. The submitted archive contained one TSV file per language and used only valid integer `pubkey` identifiers from the official collection.

The official leaderboard result is reported in Section 6, while the BGE reranking configuration is reported only as a Development analysis.

6. Results

This section reports the official Final Evaluation run and the Development ablation analysis. The distinction is important: the official submission corresponds to the E5 dense retrieval configuration, whereas BGE reranking is reported as the best Development experiment.

6.1. Official Final Evaluation Run

The official Final Evaluation submission used multilingual E5 dense retrieval with `intfloat/multilingual-e5-large`. The submitted archive was generated with one TSV file per language and contained only valid integer `pubkey` identifiers from the official collection.

Table 3 shows the official Final Evaluation leaderboard excerpt. SINAI-UGPLN ranked 34th with an average MRR@5 of 0.4328, obtaining 0.3734 for German, 0.4943 for English, and 0.4306 for French.

Table 3
Official Final Evaluation leaderboard excerpt.

Rank	Team	Avg.	DE	EN	FR
1	Claim2Source	0.7628	0.7153	0.7819	0.7912
2	Outsider	0.7580	0.7228	0.7802	0.7709
3	Boyes Boys	0.7464	0.7070	0.7571	0.7752
4	CheckMate	0.7194	0.6861	0.7343	0.7378
5	RETRIX	0.7108	0.6806	0.7117	0.7400
	... omitted intermediate ranks ...				
34	SINAI-UGPLN	0.4328	0.3734	0.4943	0.4306

This official score is close to the Development performance observed for the same E5 configuration, which obtained an average MRR@5 of 0.4389. The similarity between Development and Final Evaluation supports the selection of E5 as the official run, since it provided the most stable multilingual configuration before the deadline.

The higher BGE result reported later in this paper corresponds to a Development experiment. Therefore, it is used to analyze reranking potential, but it is not presented as the official Final Evaluation submission.

6.2. Development Ablation Study

Table 4 presents the Development results obtained during the ablation study. The table compares lexical retrieval, dense retrieval, hybrid fusion, language-aware selection, and neural reranking under the same closed-set candidate constraint.

Table 4
Development ablation results for CheckThat! 2026 Task 1.

Configuration	EN	DE	FR	Avg.	Interpretation
BM25	0.4449	0.0941	0.0851	0.2080	Lexical baseline; useful for English but weak cross-lingually
E5 dense retrieval	0.4887	0.3757	0.4523	0.4389	Official Final Evaluation system; stable across languages
BM25 and E5 RRF ($c = 60$)	0.5011	0.1641	0.1956	0.2869	Global fusion; improves English but degrades German and French
Language-aware selection	0.5185	0.3757	0.4523	0.4488	Uses RRF for English and E5 for German and French
BGE reranking over RRF top-60	0.5094	0.4232	0.5168	0.4832	Best Development experiment; strongest reranking result

The BM25 baseline obtained an average MRR@5 of 0.2080. Its behavior was highly uneven across languages: English reached 0.4449, while German and French remained below 0.10. This indicates that exact lexical matching was useful only when the social media post shared terminology with the target publication.

Dense retrieval with E5 produced the largest stable gain among first-stage retrieval methods. The average MRR@5 increased to 0.4389. The improvement was especially important for German and French, confirming that multilingual semantic representations were more robust than lexical overlap in the cross-lingual setting.

Global RRF did not improve the overall score. Although English increased to 0.5011, German and French decreased substantially compared with E5. This result shows that lexical evidence was not equally reliable across languages. The language-aware variant partially addressed this issue by using RRF only for English and E5 for German and French, increasing the Development average to approximately

0.4488.

The best Development result was obtained with BGE reranking over RRF top-60 candidates. This configuration reached an average MRR@5 of 0.4832. It improved German and French compared with E5, while maintaining a competitive English score. This suggests that neural reranking is especially useful when the candidate set already contains semantically plausible papers and the system must identify the exact source.

6.3. Language-Level Analysis

English was the only language where lexical evidence consistently helped. BM25 alone achieved a relatively high English score, and RRF further improved it. This suggests that many English posts contained terms that overlapped with the target paper title or abstract.

German and French showed a different behavior. BM25 performed poorly, and global RRF reduced the gains obtained by E5. A likely explanation is that lexical overlap between non-English social media posts and English scientific abstracts was limited. In these languages, dense retrieval was more reliable because it captured semantic similarity beyond exact word matching.

The BGE reranker improved German and French by evaluating claim–paper pairs directly. Instead of relying only on lexical overlap or embedding similarity, it assigned a relevance score to each candidate pair. This helped distinguish between semantically related papers that discussed similar scientific topics.

6.4. Submission Summary

Table 5 summarizes the status of the main configurations. The official Final Evaluation run is separated from Development-only experiments.

Table 5
Submission and analysis status of the main configurations.

Configuration	Phase	Status
E5 dense retrieval	Final Evaluation	Official submitted system; Avg. MRR@5 = 0.4328
E5 dense retrieval	Development	Validated Development configuration; Avg. MRR@5 = 0.4389
BGE reranking over RRF top-60	Development	Best Development experiment; Avg. MRR@5 = 0.4832

7. Discussion

The experiments provide three main lessons for scientific source retrieval from social media claims. First, multilingual dense retrieval is a strong and stable first-stage strategy when the query and candidate documents differ in language, style, or level of formality. This explains why E5 was selected as the official Final Evaluation system: it provided balanced performance across English, German, and French and generalized better than lexical matching.

Second, hybrid retrieval requires language-aware control. BM25 contributed useful lexical evidence in English, where posts more often shared terminology with the candidate paper title or abstract. However, applying the same fusion strategy to German and French degraded performance. This suggests that sparse and dense signals should not be combined uniformly in multilingual scientific retrieval tasks.

Third, neural reranking is effective when the first-stage retriever provides a sufficiently good set of candidates. The BGE reranker improved the Development score by directly scoring claim–paper pairs, which is valuable when several candidate papers are topically similar. However, the reranker cannot recover a correct paper that was not retrieved among the candidate set. This makes candidate recall a necessary condition for successful reranking.

From an operational perspective, the task also highlights the importance of strict submission validation. Since the benchmark uses a closed set of candidates, generating valid outputs is part of the system design. Each prediction must contain five integer `pubkey` values from the official collection. In our experiments, this validation step was essential for producing valid CodaBench submissions.

8. Error Analysis

Based on the Development behavior and language-level results, the main error sources can be grouped into four conditions: lexical mismatch, cross-lingual mismatch, topical ambiguity, and candidate recall limitations.

8.1. Lexical Mismatch

The first source of error was lexical mismatch between the social media post and the scientific paper representation. Many posts did not reproduce the title or terminology of the target publication. Instead, they summarized a finding, used informal language, or referred to the study indirectly. This explains the behavior of BM25: it achieved a reasonable English score when exact terms were shared, but it performed poorly when the relevant relation was semantic rather than lexical.

8.2. Cross-Lingual Mismatch

The second source of error was cross-lingual mismatch. The evaluated queries were written in English, German, and French, while the candidate papers were represented mainly through scientific titles and abstracts. In German and French, direct lexical overlap with the candidate abstracts was often limited. Dense multilingual retrieval reduced this problem, but some cases remained difficult when the social media post expressed the finding in a highly compressed or paraphrased form.

8.3. Topical Ambiguity

The third source of error was topical ambiguity. Some posts referred to broad scientific areas with many similar candidate papers, such as COVID-19, biomedical treatments, public health, climate, or social science studies. In these cases, the system often retrieved papers that were scientifically plausible and semantically close, but not the exact target publication. This type of error is especially challenging because the retrieved candidates may appear relevant at the topic level yet be incorrect for the official `pubkey`.

8.4. Candidate Recall Limitations

The fourth source of error was candidate recall. The BGE reranker improved the ordering of candidates when the correct paper was among the top 60 retrieved documents. However, if the correct paper was absent from the first-stage candidate set, the reranker could not recover it. This indicates that first-stage retrieval and reranking must be optimized jointly: the retriever must maximize recall, while the reranker must refine the ranking among plausible candidates.

Overall, the error analysis suggests that the most promising direction is not simply to replace one retrieval model with another, but to improve the interaction between query representation, candidate generation, language-aware fusion, and reranking.

9. Limitations

A first limitation is that the official Final Evaluation run prioritized the most stable validated dense retrieval configuration, while the strongest reranking configuration was completed as a Development analysis. Although BGE achieved the best Development score, the official Final Evaluation run was

based on E5 dense retrieval. Therefore, the official result should be interpreted as the performance of the multilingual dense retrieval configuration, while the BGE result should be understood as a Development analysis of the potential gains of reranking.

A second limitation is the absence of explicit claim reformulation. Social media posts often contain informal language, incomplete references, hashtags, rhetorical framing, or shortened descriptions of scientific findings. Reformulating these posts into more explicit scientific queries could improve alignment with candidate paper abstracts, but this step was not included in the official run.

A third limitation is computational cost. Neural reranking improved Development performance, but it required substantially more GPU time than first-stage retrieval. This limits its practical use under strict submission deadlines unless the candidate set is carefully controlled or the reranking process is optimized.

Finally, the system did not use external scientific databases or external evidence sources. This was consistent with the closed-set nature of the task, but it also means that the system could not use additional bibliographic signals beyond those available in the official collection. Future systems could explore richer query reformulation or metadata-aware reranking while still respecting the official candidate constraints.

10. Conclusions and Future Work

This paper presented the participation of the SINAI-UGPLN team in CheckThat! 2026 Task 1, Source Retrieval for Scientific Web Claims. We approached the task as a closed-set multilingual scientific document ranking problem, where each social media post must be linked to the scientific publication implicitly referenced in the official candidate collection.

The official Final Evaluation submission was based on dense retrieval with `intfloat/multilingual-e5-large`. This configuration was selected because it provided the most stable validated performance across English, German, and French before the final deadline. Compared with BM25, E5 substantially improved German and French, confirming the importance of multilingual semantic representations for this task.

The Development experiments showed that retrieval behavior was language-dependent. BM25 and RRF were useful for English, but global fusion degraded German and French. This suggests that multilingual scientific source retrieval benefits from language-aware retrieval decisions rather than a single global fusion strategy.

The best Development result was obtained with `BAAI/bge-reranker-v2-m3` over RRF top-60 candidates. This result indicates that neural reranking can improve fine-grained matching between informal claims and scientific abstracts when the correct paper is present in the candidate set. However, reranking is computationally more expensive and must be balanced against runtime constraints.

Future work will focus on three directions. First, we plan to explore efficient reranking strategies that preserve ranking quality while reducing computational cost. Second, we will investigate claim reformulation methods to transform informal social media posts into more explicit scientific queries. Third, we will study language-aware retrieval and fusion methods, including weighted sparse–dense fusion for German and French, since our experiments showed clear differences between English, German, and French.

Overall, the results suggest that scientific source retrieval from social media claims should be treated as a constrained multilingual ranking problem. Dense retrieval provides robust candidate generation across languages, while language-aware fusion and neural reranking can further improve precision when computational resources and submission deadlines allow it.

Acknowledgments

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia – Funded by EU – NextGenerationEU within

the framework of the project Desarrollo Modelos ALIA. This work has also been partially supported by Project CONSENSO (PID2021-122263OB-C21), Project MODERATES (TED2021-130145B-I00) and Project SocialTox (PDC2022-133146-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. Moreover, this research is part of the proposal presented at the Call for Research Project Proposals of the Internal Competitive Fund (FCI) 2023, which was approved on September 14, 2023 (Resolution No. R-CSU-UG-SE34-313-14-09-2023) by the Consejo Superior Universitario of the Universidad de Guayaquil.

The authors declare that they have contributed equally and share authorship roles for this publication.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (OpenAI) and Grammarly to check grammar and spelling. After using these tools and services, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References

[1] 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, 2026. URL: https://arxiv.org/abs/2602.09516. arXiv:2602.09516.

[2] M. d. C. Toapanta-Bernabé, M. Á. García-Cumbreras, L. A. Ureña-López, D. D. Mora-Intriago, C. T. Bernal-García, SINAI-UGPLN at CheckThat! 2025: Meta-ensemble strategies for numerical claim verification in english, in: Working Notes of CLEF 2025, volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 1265–1275. URL: https://ceur-ws.org/Vol-4038/paper_100.pdf.

[3] S. Robertson, H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, Foundations and Trends in Information Retrieval 3 (2009) 333–389. doi:10.1561/1500000019.

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

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

[6] 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: Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, 2024, pp. 2318–2335. URL: https://aclanthology.org/2024.findings-acl.137/. doi:10.18653/v1/2024.findings-acl.137.

[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, CLEF 2026, Jena, Germany, 2026.

[8] T. Schreieder, M. Färber, Claim2source at CheckThat! 2025: Zero-shot style transfer for scientific claim-source retrieval, in: Working Notes of CLEF 2025, volume 4038 of *CEUR Workshop Proceedings*, 2025. URL: https://ceur-ws.org/Vol-4038/paper_94.pdf.