MeVer at CheckThat! 2026: Cluster-Aware Hard-Negative Mining for Multilingual Scientific-Source Retrieval

Juli Bakagianni[1], Symeon Papadopoulos[1]

[1]*Information Technologies Institute, Centre for Research and Technology Hellas (CERTH), Thessaloniki, Greece*

Abstract

Identifying the scientific source behind a social media claim requires matching short, informal, and often multilingual claims against large collections of scientific publications, where semantically related papers may act as challenging distractors or false negatives during training. We present our submission to CheckThat! 2026 Task 1 on multilingual scientific-source retrieval, focusing on how hard-negative mining should be adapted to multi-stage retrieval pipelines for scientific-source retrieval. We propose cluster-aware hard-negative mining strategies that exploit the semantic structure of retrieved candidate pools in order to construct more informative training negatives for dense retrieval and reranking. Our experiments show that different hard-negative structures induce different retrieval behaviors. Localized cluster negatives tend to favor precision-oriented retrieval, whereas broader non-gold semantic negatives provide stronger candidate coverage and more consistent reranking performance across languages. We further study multiple LLM-based evidence-selection formulations, including direct classification, pairwise comparison, and listwise reranking prompts, and find that constrained classification prompts provide the most reliable final document selection. The final system combines a dense retriever, a multilingual cross-encoder reranker, and a selective LLM-based disagreement resolver, ranking 6th among 37 submissions in the shared task evaluation. Overall, our results suggest that hard-negative mining should be treated as a stage-aware design problem rather than as a single retrieval optimization strategy.

Keywords

information retrieval, social media retrieval, hard-negative mining, scientific-source retrieval, cross-encoder reranking, large language models

1. Introduction

Scientific claims are frequently discussed on social media through informal references to studies, reports, or "research findings" without explicit citations or links to the underlying publication. Automatically identifying the referenced scientific paper is therefore an important problem for scientific fact-checking, evidence tracing, and misinformation analysis [1]. However, this retrieval setting is particularly challenging because short and loosely phrased social media claims must be matched against large collections of scientific articles written in formal academic language. The problem becomes even more difficult in multilingual settings, where claims may appear in different languages while the scientific collection remains primarily English. Moreover, since the dataset provides only a single annotated source paper per claim, semantically related but unlabeled papers may act as false negatives during retrieval and reranking training.

CheckThat! Task 1 formulates this problem as a multilingual scientific-source retrieval task. Given a social media claim in English, German, or French, the goal is to retrieve the referenced scientific paper from a shared collection of scientific publications [2]. The task combines known-item retrieval with semantic evidence matching: the correct paper may differ from competing candidates only in subtle methodological or semantic details, requiring systems to balance broad candidate coverage with fine-grained semantic discrimination.

Modern retrieval systems for such tasks are commonly organized as multi-stage pipelines. First-stage dense retrievers retrieve candidate documents using learned embedding similarity representations [3, 4]. Cross-encoder rerankers then refine the ranking through full query–document interaction modeling [5, 6]. More recently, large language models (LLMs) have also been explored as reasoning-

CheckThat! Lab, 2026

julibak@iti.gr (J. Bakagianni); papadop@iti.gr (S. Papadopoulos)

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

based rerankers or judges capable of comparing a small set of candidate documents and selecting the best supporting evidence [7, 8]. These stages address complementary objectives: retrieval emphasizes candidate-set coverage, reranking emphasizes local semantic discrimination, and LLM-based judgment targets difficult final selection cases where multiple candidates remain plausible.

In this work, we focus on how hard negative (HN) mining during model training should be adapted to such a multi-stage retrieval pipeline. Existing HN approaches for contrastive retriever and reranker training typically select negatives based on retrieval rank or nearest-neighbor similarity [4, 9]. However, retrieved candidate pools often exhibit internal semantic structure: some documents form highly localized neighborhoods around the gold paper, while others act as broader topical distractors. We therefore investigate whether clustering retrieved candidates can provide a more principled mechanism for selecting informative training negatives for both retriever and reranker fine-tuning.

Our central hypothesis is that different HN structures induce different retrieval behaviors across multi-stage pipelines. Broader semantic negatives may improve downstream candidate coverage, whereas highly localized negatives may sharpen fine-grained semantic discrimination between closely related scientific papers. We therefore study how different cluster-aware HN strategies affect both retrieval and reranking stages. To study this question, we introduce cluster-aware HN mining strategies that sample negatives from the gold-document cluster, neighboring clusters, or non-gold semantic regions of the retrieved space.

Beyond retrieval and reranking, we also investigate how LLMs should operate as selective final-stage judges. Rather than applying the LLM universally, we use it only on disagreement cases between the retriever and reranker. We further study multiple prompting and decision formulations for LLM-based evidence selection, including direct classification, pairwise comparison, permutation-style ranking inspired by RankGPT [7], shuffled candidate ordering, and committee-style voting. Our experiments show that simple constrained classification prompts are more reliable than more elaborate ranking-style formulations in this retrieval setting.

Although multilingual handling is not the primary focus of this work, our experiments also reveal an interesting interaction between translation strategy and judge model family. Translation into English consistently improves first-stage retrieval and reranking for non-English queries, likely because the scientific collection itself is English. However, at the LLM judgment stage, GPT-family models benefit from the original-language claim formulation, whereas Llama-based judges perform better with translated claims. This suggests that optimal query representation depends not only on the pipeline stage but also on the downstream judge model and the candidate distribution produced by earlier retrieval stages.

The paper is organized around the following research questions:

- Can cluster-aware HN mining improve retrieval candidate coverage and reranking effectiveness in multilingual scientific-source retrieval?
- Do retrieval and reranking stages benefit from different forms of semantic HN?
- How should LLMs be used as final-stage judges in retrieve-then-rerank pipelines, and which prompting formulations are most effective for evidence selection?

Our final system combines a dense retriever and a cross-encoder reranker trained with cluster-aware HN and a selective GPT-5.5 disagreement resolver. Across experiments, we find that non-gold-cluster negatives provide the strongest retrieval backbone for reranking-oriented pipelines, while localized cluster negatives are more effective for sharpening retrieval-only precision. We further show that selective LLM intervention is most effective when restricted to disagreement cases and formulated as constrained classification rather than free-form ranking. Together, these results suggest that HN mining should be treated as a stage-aware design problem rather than a single retrieval optimization strategy.

2. Related Work

Dense retrieval has become a standard approach for semantic document retrieval using contrastive representation learning [3, 4]. A central component of dense retriever training is HN mining, where models are trained to distinguish relevant documents from semantically similar distractors. Prior work has explored Approximate Nearest Neighbor Negative Contrastive Learning (ANCE) [4], BM25-based negatives [3], denoised HN [9], and curriculum-style training strategies. However, most existing approaches define negative difficulty primarily through retrieval rank or embedding similarity, without explicitly modeling the semantic structure of the retrieved candidate space.

Retrieve-then-rerank pipelines further improve retrieval quality by combining dense retrievers with cross-encoder rerankers [5, 6]. More recently, LLMs have also been studied for ranking or reranking tasks [7, 8], including permutation-based ranking and pairwise prompting strategies. Existing work has shown that LLMs can act as effective rerankers over small candidate sets, although their reliability and prompting sensitivity remain active research questions.

3. Dataset

We use the official *CheckThat! 2026 Task 1* dataset for multilingual scientific-source retrieval,[1] The dataset consists of social media claims from X in English, German, and French paired with a shared scientific document collection containing 10,000 English-language scientific papers. Each query contains a scientific claim together with an implicit reference to a scientific paper, and the task is to retrieve the referenced paper from the candidate collection [2]. Each query is associated with a single annotated gold scientific paper. Each paper includes title, abstract, venue, and author metadata. Table 1 presents examples of social media posts with implicit references to scientific articles, while Table 2 shows example papers from the scientific collection.

Table 1
Example social media claims with implicit references to scientific publications from the CheckThat! 2026 Task 1 dataset.

Lang	Tweet Text	Pub ID
EN	Published in the journal *Antiviral Research*, the study from Monash University showed that a single dose of Ivermectin could stop the coronavirus growing in cell culture – effectively eradicating all genetic material of the virus within two days.	234567
DE	Die in der Fachzeitschrift *Antiviral Research* veröffentlichte Studie der Monash University zeigte, dass eine einzige Dosis Ivermectin das Wachstum des Coronavirus in Zellkulturen stoppen kann und das gesamte genetische Material des Virus innerhalb von zwei Tagen effektiv zerstört.	234567
FR	L'outil génétique CRISPR-Cas9 a été utilisé pour permettre à la provitamine D3 de s'accumuler dans les feuilles et fruits de la tomate. Une fois exposée aux UV, elle s'est bien convertie en vitamine D (dont 1 milliard de personnes est carencé dans le monde).	345678

In the collection, paper titles are short (13 words on average), whereas abstracts are much longer (232 words on average). Table 3 summarizes the query distribution across train, development, and test splits. The dataset is heavily English-dominant, while German and French constitute substantially smaller portions of the benchmark.

4. Methodology

Our system follows a three-stage pipeline. First, a dense retriever selects a candidate pool of potentially relevant scientific papers. Second, a cross-encoder reranker refines this candidate set through finer

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

Table 2
Example scientific publications from the shared document collection of CheckThat! 2026 Task 1.

Pub ID	Study Title	Study Venue + Authors	Study Abstract
123456	*Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant*	[Venue]: *New England Journal of Medicine* [Authors]: Jamie Lopez Bernal, Nick Andrews, Charlotte Gower, Eileen Gallagher, Ruth Simmons, Simon Thelwall, Julia Stowe, Elise Tessier, …	BACKGROUND: The B.1.617.2 (delta) variant of the severe acute respiratory syndrome coronavirus 2 …
234567	*The FDA-approved drug ivermectin inhibits the replication of SARS-CoV-2 in vitro*	[Venue]: *Antiviral Research* [Authors]: Caly, Leon; Druce, Julian D.; Catton, Mike G.; Jans, David A.; Wagstaff, Kylie M.	Although several clinical trials are now underway to test possible therapies, the worldwide response to the COVID-19 outbreak has been largely limited to monitoring/containment …

Table 3
Query distribution across dataset splits.

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

query–document interactions. Finally, an LLM judge is applied selectively to disagreement cases where the retriever and reranker produce different top predictions.

The main methodological focus of this work is cluster-aware HN mining during retriever and reranker training. We study whether the semantic structure of retrieved candidate pools can be exploited through clustering to construct more informative training negatives for different pipeline stages. In addition, we investigate multilingual query handling and multiple LLM-based evidence-selection formulations for final-stage document selection.

4.1. Multilingual Query Handling

The scientific document collection used in the task is entirely English, while claims are provided in English, German, and French. We examine multilingual handling at two stages of the pipeline. The first is the retrieval stage, where the query is matched against English scientific documents. Translation may help by aligning the claim language with the document collection and simplifying semantic matching. At the same time, translation may also remove linguistic nuances or domain-specific phrasing that could remain useful for later evidence selection.

The second stage is the LLM-based judgment component, where the model receives the claim together with a small set of candidate papers and must select the strongest supporting source. Here, we study whether the judge should operate on the original language claim or on its English translation. We therefore treat multilingual query handling as a methodological design choice rather than a fixed preprocessing step.

4.2. Dense Retrieval Stage

The first stage of the pipeline is a dense retriever trained to retrieve a candidate pool of relevant scientific papers for each claim. Retriever training follows the standard contrastive learning setting, where the model is optimized to assign higher similarity to the gold paper than to negative documents. Since

retrieval effectiveness depends strongly on the quality and difficulty of the training negatives, we investigate multiple HN mining strategies.

All training settings use standard in-batch negatives, where negatives are formed from the positive documents of other examples within the same training batch. In-batch negatives provide a large number of inexpensive and diverse easy negatives that stabilize contrastive retriever training [3, 4].

As a stronger HN baseline, we study ANCE HN [4], where the current retriever periodically retrieves top-ranked candidate documents that are then used as HNs for subsequent contrastive training. Our cluster-aware approaches build directly on this setting: rather than sampling HN only according to retrieval similarity, we additionally exploit the semantic structure of the retrieved candidate pool through clustering.

Starting from the retrieved candidate pool of the current retriever, we cluster candidate document embeddings and construct different types of training negatives. Let the *gold cluster* denote the cluster containing the gold document. We study three cluster-aware strategies:

- **Gold-cluster negatives**: negatives are sampled from the same cluster as the gold document, forcing the retriever to distinguish between highly similar local distractors.
- **Nearest-cluster negatives**: negatives are sampled from the non-gold cluster whose centroid is nearest to the gold cluster centroid, targeting semantically neighboring but distinct documents.
- **Non-gold-cluster negatives**: negatives are sampled from retrieved documents outside the gold cluster, providing broader semantic diversity while remaining retrieval-relevant.

The motivation behind these variants is that different retrieval objectives may benefit from different forms of semantic supervision. Gold-cluster negatives emphasize fine-grained discrimination between closely related papers, whereas non-gold-cluster negatives encourage broader candidate coverage by exposing the retriever to more diverse semantic distractors. Since the downstream reranker can only operate on retrieved candidates, candidate coverage remains particularly important in our retrieve–then–rerank pipeline.

4.3. Reranking Stage

The second stage of the pipeline is a cross-encoder reranker that reorders the candidate documents returned by the dense retriever. Unlike dense retrieval, where queries and documents are compared through independent vector representations, the reranker jointly processes the query and candidate document, enabling finer-grained semantic matching between claim wording and scientific abstracts.

Reranker training operates under a different setting from retrieval training. While the retriever must separate the gold paper from the full document collection, the reranker only observes the restricted candidate pool produced by the retriever. Starting from the retrieved candidate pool, we apply the same cluster-aware HN framework used in retrieval training. We compare negatives sampled from the gold cluster, the nearest non-gold cluster, and non-gold semantic regions of the retrieved space. In this setting, gold-cluster negatives emphasize fine-grained discrimination between highly similar scientific papers, whereas broader non-gold negatives expose the reranker to more diverse retrieval distractors.

4.4. LLM-as-Judge Stage

The final stage of the pipeline uses an LLM as a selective evidence-selection judge. Rather than applying the LLM over all reranked candidates, we restrict its use to disagreement cases where the dense retriever and reranker produce different top-ranked predictions. This setting focuses the LLM on difficult cases where multiple plausible candidate papers survive the earlier pipeline stages.

The judge operates over a small candidate set constructed from the top reranked documents together with the retriever top-ranked document whenever it differs from the reranker prediction. In this way, the LLM acts as a targeted resolver between competing retrieval and reranking hypotheses rather than as a full reranker over the entire candidate pool.

Our main focus at this stage is the formulation of the evidence-selection task itself. We investigate multiple prompting strategies:

- **Direct classification**: the LLM directly selects the best supporting document from the candidate set.
- **Pairwise comparison**: the reranker top prediction is treated as a baseline candidate, and the LLM evaluates whether another candidate provides stronger evidence.
- **Listwise reranking**: the LLM produces a complete ordering over the candidate documents using permutation-style ranking prompts [7].

These formulations allow us to study whether LLMs operate more reliably as constrained classifiers, pairwise evaluators, or free-form ranking agents in scientific-source retrieval.

5. Experimental Setup

Our retrieval experiments use `BAAI/bge-large-en-v1.5`[2] as the main dense retriever. Preliminary retrieval experiments additionally included `multilingual-e5-large`[3] and `e5-large-v2`[4] (see Appendix B). For reranking, we evaluated several cross-encoder families, including SciBERT and `BAAI/bge-reranker-base`, while the strongest reranking results were obtained with `jinaai/jina-reranker-v2-base-multilingual` (see Appendix C).[5] The LLM-based judgment stage was evaluated using `Llama 3 70B`, `gpt-4.1-mini`, and `gpt-5.5` (see Appendix D).

For clustering, we use the K-Means clustering algorithm. The number of clusters is selected by maximizing the macro silhouette score over a small search range ($k \in [3, 6]$) [10]. For retrieval training, we use a batch size of 16, yielding 15 in-batch negatives per example. In all HN settings, a single additional HN is sampled per query. For reranking experiments, we use a top-200 retriever pool for HN mining, exclude the gold document and the highest-ranked retrieved documents from negative sampling, and sample 10 HN per query. Hyperparameters, including learning rate, training epochs, and HN counts, are selected on the development set.

The official shared-task metric is MRR@5. Since our system follows a retrieve–then–rerank pipeline, we additionally track recall-oriented metrics during retrieval evaluation, particularly Recall@20, to ensure that the gold document remains available to downstream reranking and LLM-based selection stages.

6. Results

We first examine dense retrieval and reranking experiments with cluster-aware HN mining, and finally LLM-based evidence-selection ablations.

6.1. Dense Retrieval Results

Table 4 compares the negative mining strategies for dense retrieval. We evaluate in-batch negatives, ANCE HN, and the proposed cluster-aware variants using Recall@20 as the primary metric, and then Recall@10 and MRR@5. Detailed multilingual query-handling experiments are provided in Appendix A.

Overall, the differences between HN strategies remain marginal to modest. This is partly explained by the training setup itself: each query receives 15 in-batch negatives due to the batch size of 16, while only a single explicit HN is added in the HN settings. However, the later scaling experiments in Section 6.1.1

[2] https://huggingface.co/BAAI/bge-large-en-v1.5
[3] https://huggingface.co/intfloat/multilingual-e5-large
[4] https://huggingface.co/intfloat/e5-large-v2
[5] https://huggingface.co/jinaai/jina-reranker-v2-base-multilingual

Table 4
Dense retrieval results for different HN mining strategies on the development set. Metrics include Recall@20, Recall@10, and MRR@5 across languages and macro averages.

	EN			DE			FR			Macro		
Family	R@20	R@10	MRR@5	R@20	R@10	MRR@5	R@20	R@10	MRR@5	R@20	R@10	MRR@5
Only in-batch	**0.842**	0.794	**0.644**	0.780	0.707	0.559	0.869	0.832	0.670	0.830	0.778	0.624
ANCE	0.838	0.792	0.635	**0.824**	0.764	0.589	0.877	0.846	**0.715**	0.847	0.801	0.646
Cluster gold	0.834	0.784	0.629	0.808	0.756	**0.601**	0.882	0.848	0.709	0.841	0.796	0.646
Cluster nearest	0.840	0.787	0.632	0.816	0.759	0.593	0.882	0.858	0.702	0.846	0.801	0.642
Cluster non-gold	0.839	**0.800**	**0.644**	0.819	**0.769**	0.590	**0.886**	**0.865**	0.707	**0.848**	**0.811**	**0.647**

show that increasing the number of explicit HN mainly sharpens precision-oriented retrieval rather than improving Recall@20, which remains the primary objective for the downstream reranking pipeline.

Despite these small margins, several consistent patterns emerge. First, standard in-batch negatives already provide a strong retrieval baseline, achieving the best English Recall@20. Since English also constitutes the largest portion of the training data, this suggests that large numbers of diverse in-batch negatives may already provide sufficiently strong supervision in higher-resource settings.

Second, the impact of HN structure differs across the translated German and French subsets. While ANCE and cluster-aware approaches consistently improve over the in-batch baseline for these subsets, localized cluster negatives tend to produce stronger precision-oriented retrieval. In particular, gold-cluster negatives achieve the highest German MRR@5 and outperform the nearest and non-gold cluster variants on both German and French MRR@5. This suggests that localized semantic distractors may improve fine-grained ranking precision when retrieval supervision is weaker or noisier.

Finally, non-gold-cluster negatives achieve the strongest macro Recall@20 and Recall@10 scores, indicating that broader semantic negatives improve retrieval coverage across languages. Since downstream reranking and LLM-based selection can only operate on retrieved candidates, we therefore use the non-gold-cluster strategy as the retrieval backbone for subsequent reranking experiments.

6.1.1. Scaling

Table 5 examines the interaction between in-batch negatives and explicit HN for the tuned non-gold-cluster retriever. We vary both the effective number of in-batch negatives and the number of explicit HN per query. Overall, the results reveal a consistent precision–coverage trade-off. Increasing the number of in-batch negatives slightly improves recall-oriented metrics, with the 63 in-batch and 1 HN setting achieving the strongest Recall@20 and Recall@50 scores. In contrast, increasing the number of explicit HN improves precision-oriented retrieval, with the 31 in-batch and 4 HN setting producing the strongest MRR@5.

Table 5
Scaling results for the tuned cluster non-gold retriever. We report macro-average retrieval metrics over English, German, and French. Settings are described by the effective number of in-batch negatives and the number of explicit HN per query.

In-batch neg.	HN	R@20	R@10	R@50	MRR@5
15	1	0.848	**0.811**	0.889	0.647
31	1	0.841	0.807	0.887	0.655
31	2	0.845	0.806	0.885	0.644
31	4	0.847	0.807	0.885	**0.660**
63	1	**0.849**	0.806	**0.891**	0.653

These findings align with the earlier retrieval results. While HN are known to play a critical role in modern dense retrieval training [4, 9], their marginal impact in our setting is moderated by the already strong in-batch supervision signal, where each query is paired with 15–63 in-batch negatives.

Within this setting, broader negative diversity tends to improve candidate coverage, whereas stronger HN pressure sharpens early ranking precision.

6.2. Reranking Results

In the reranking stage, the dense retriever first returns the top-20 candidate papers for each query, and the cross-encoder reranker then reorders these candidates. Since reranking operates within a fixed candidate pool, we report MRR@5.

Table 6 reports reranking results under different HN sampling strategies. The retrieval-pool approach provides a strong reranking baseline and achieves the strongest English MRR@5 score together with the non-gold-cluster variant. However, cluster-aware negatives consistently improve German and French reranking performance, leading to higher macro-average results. Among the evaluated strategies, non-gold-cluster negatives produce the strongest overall reranking performance, achieving the best German, French, and macro MRR@5 scores.

Although the absolute gains remain small, the improvements are consistent across the non-English subsets, suggesting that broader semantic negatives help the reranker distinguish between difficult surviving retrieval confounders. Earlier reranking experiments with gold-cluster negatives produced consistently weaker results and are reported in Appendix F for completeness.

Table 6
Development-set MRR@5 results for the cross-encoder reranker using different HN sampling strategies.

Negative Strategy	EN	DE	FR	Macro
Retrieval pool	**0.674**	0.590	0.729	0.664
Cluster nearest	0.672	0.592	0.732	0.665
Cluster non-gold	**0.674**	**0.593**	**0.734**	**0.667**

Interestingly, unlike the retrieval experiments where gold-cluster negatives favored precision-oriented retrieval, reranking performance benefits more consistently from broader semantic negatives than from highly localized cluster supervision.

6.3. LLM-as-Judge Results

After fixing the best dense retriever and reranker, we examined whether an external LLM can improve difficult disagreement cases between the two earlier pipeline stages. The intervention is selective: the reranker prediction is retained whenever the retriever and reranker agree on the top-ranked document, while the LLM is only applied when the two stages produce different top-1 predictions.

Retriever–reranker agreement proved to be a strong correctness signal. The two stages selected the same top-ranked document for 60.6% of development queries, and this shared prediction was correct in 84.5% of agreement cases. In contrast, disagreement cases were substantially harder, motivating the use of the LLM as a targeted resolver rather than as a full reranker.

We first studied different evidence-selection formulations for the LLM judge on the German development disagreement subset. The corresponding prompt formulations are provided in Appendix E. To isolate prompt quality from retrieval coverage effects, these experiments were restricted to recoverable disagreement cases where the gold document was already present in the shown candidate set.

Table 7 shows that simpler constrained prompting is more reliable than more complex ranking formulations. Direct classification consistently outperforms both pairwise comparison and listwise reranking, suggesting that LLM-based evidence selection is more effective when formulated as a constrained decision problem rather than as a free-form ranking task.

We finally evaluate the full selective disagreement pipeline with `gpt-5.5` as the final-stage judge. Table 8 reports the end-to-end MRR@5 gains across development languages relative to the plain reranker.

Table 7
German LLM-judge prompt ablations using `Llama 3 70B` on recoverable disagreement cases. The stage-2 row reports reranker performance before LLM intervention. Hit@1 measures whether the gold document is selected as the correct prediction.

Prompt formulation	Hit@1
Stage-2 reranker before LLM	0.4070
Direct classification	**0.523**
Pairwise comparison	0.442
Listwise reranking	0.488

Table 8
Final selective `gpt-5.5` disagreement gains on the development split. Scores are end-to-end MRR@5.

Lang	Plain reranker	With LLM judge	Δ
EN	0.6706	0.7170	+0.0464
DE	0.5983	0.6601	**+0.0618**
FR	0.7284	0.7772	+0.0488

The selective LLM stage improves end-to-end reranking performance across all three development languages, with the largest absolute gain observed on German.

Additional judge-family comparisons, multilingual query formulations, and prompt-engineering ablations are reported in Appendix D.

7. Discussion and Conclusion

In this work, we investigated how HN mining should be adapted across multi-stage scientific-source retrieval pipelines. Our experiments show that the semantic structure of the retrieved candidate space can be exploited to control different retrieval behaviors. Localized cluster negatives improve precision-oriented retrieval, whereas broader non-gold semantic negatives produce stronger retrieval coverage and more consistent reranking performance across languages. Although the absolute improvements remain modest, the results consistently show that different HN structures influence the balance between early ranking precision and downstream candidate coverage.

At the LLM stage, we found that simpler constrained prompting formulations are more reliable than more complex ranking-based prompting strategies. Direct classification consistently outperformed pairwise comparison and listwise reranking prompts, suggesting that LLM-based evidence selection is most effective when framed as a constrained decision problem over a compact candidate set. The final system combines cluster-aware retrieval training, multilingual reranking, and selective LLM-based disagreement resolution, achieving development MRR@5 scores of 0.6829 for English, 0.6615 for German, 0.7074 for French, and 0.6839 macro-average, ranking 6th among 37 submissions in the CheckThat! 2026 Task 1 challenge.

7.1. Limitations and Future Work

The main limitation of this study is that many retrieval and reranking differences remain relatively small, partly because strong in-batch supervision already dominates the training signal. In addition, several later-stage improvements remain constrained by candidate coverage: if the correct document is absent from the retrieved candidate set, neither reranking nor LLM-based resolution can recover it.

Future work should therefore focus on stronger retrieval coverage and more principled clustering strategies. In particular, it would be valuable to perform larger-scale experiments with stronger statistical validation in order to better evaluate clustering-based HN mining, as well as to study joint optimization of retrieval-side and reranker-side HN sampling.

Acknowledgments

This work was supported by the ELLIOT Grant, funded by the European Union under Grant Agreement No. 10121439.

References

[1] S. Hafid, S. Bringay, K. Todorov, Scientific web claims: A survey of definitions, tasks, datasets and methods (2026).

[2] J. M. Struß, S. Schellhammer, S. Dietze, V. V, V. Setty, T. Chakraborty, P. Nakov, A. Anand, P. Chungkham, S. Hafid, et al., The clef-2026 checkthat! lab: Advancing multilingual fact-checking, in: European Conference on Information Retrieval, Springer, 2026, pp. 325–335.

[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), 2020, pp. 6769–6781.

[4] L. Xiong, C. Xiong, Y. Li, K.-F. Tang, J. Liu, P. Bennett, J. Ahmed, A. Overwijk, Approximate nearest neighbor negative contrastive learning for dense text retrieval, arXiv preprint arXiv:2007.00808 (2020).

[5] R. Nogueira, K. Cho, Passage re-ranking with bert, arXiv preprint arXiv:1901.04085 (2019).

[6] Y. Qu, Y. Ding, J. Liu, K. Liu, R. Ren, W. X. Zhao, D. Dong, H. Wu, H. Wang, Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering, in: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies, 2021, pp. 5835–5847.

[7] W. Sun, L. Yan, X. Ma, S. Wang, P. Ren, Z. Chen, D. Yin, Z. Ren, Is chatgpt good at search? investigating large language models as re-ranking agents, in: Proceedings of the 2023 conference on empirical methods in natural language processing, 2023, pp. 14918–14937.

[8] Z. Qin, R. Jagerman, K. Hui, H. Zhuang, J. Wu, L. Yan, J. Shen, T. Liu, J. Liu, D. Metzler, et al., Large language models are effective text rankers with pairwise ranking prompting, in: Findings of the Association for Computational Linguistics: NAACL 2024, 2024, pp. 1504–1518.

[9] J. Zhan, J. Mao, Y. Liu, J. Guo, M. Zhang, S. Ma, Optimizing dense retrieval model training with hard negatives, in: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval, 2021, pp. 1503–1512.

[10] J. Pavlopoulos, G. Vardakas, A. Likas, Revisiting silhouette aggregation, in: International Conference on Discovery Science, Springer, 2024, pp. 354–368.

[11] N. Muennighoff, N. Tazi, L. Magne, N. Reimers, Mteb: Massive text embedding benchmark, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023, pp. 2014–2037.

A. Multilingual Query Handling Experiments

We evaluate multilingual query handling across retrieval, reranking, and LLM-based judgment stages for German and French claims. Since the scientific document collection is entirely English, we compare using the original-language claims against their provided English translations.

We first examined whether German and French queries should be encoded directly with a multilingual retriever or translated into English before retrieval. These preliminary experiments used the challenge-provided multilingual baseline encoder, `intfloat/multilingual-e5-large`, as the reference system. We then compared this baseline against two translation settings for the non-English queries: `facebook/nllb-200-distilled-600M` and Llama 3 70B.

Table 9 shows that translation quality substantially affects non-English retrieval performance. NLLB translation slightly decreases German retrieval effectiveness relative to the original-language multilingual baseline, although it improves French. In contrast, higher-quality LLM translation with Llama3

Table 9
Preliminary retrieval-only translation study on the dev split with `multilingual-e5-large`.

System	DE			FR		
	MRR@5	R@10	R@20	MRR@5	R@10	R@20
Original-language query + multilingual-e5-large	0.374	0.580	0.674	0.456	0.654	0.752
NLLB translation + multilingual-e5-large	0.366	0.557	0.658	0.491	0.691	0.779
Llama 3 70B translation + multilingual-e5-large	**0.416**	**0.614**	**0.723**	**0.511**	**0.721**	**0.801**

70B improves both ranking and recall-oriented metrics for both languages. Although these experiments were conducted with a single retrieval model, they suggest that higher-quality English translation may improve semantic alignment between non-English social media claims and the English scientific document collection.

At the LLM judgment stage, however, the optimal formulation depends on the judge model family. GPT-family models perform better when provided with the original language claim, whereas Llama-based judges obtain stronger results using translated English claims. One possible explanation is that retrieval and reranking candidates are themselves generated using translation-based retrieval, making translated claims more aligned with the retrieved candidate distribution for Llama-based judges.

B. Retrieval models

We used the translated-query setup to benchmark several off-the-shelf first-stage retrievers in a retrieval-only setting. Table 10 reports the saved preliminary encoder results, while for German and French we use their translations to English with Llama3 70B.

Table 10
Preliminary retrieval-only encoder comparison on the development set. German and French queries use English translations generated with Llama 3 70B.

Encoder	EN			DE			FR		
	MRR@5	R@10	R@20	MRR@5	R@10	R@20	MRR@5	R@10	R@20
`intfloat/multilingual-e5-large`	0.499	0.661	0.713	0.416	0.565	0.614	0.511	0.663	0.721
`intfloat/e5-large-v2`	0.504	0.661	0.709	**0.435**	0.593	0.655	0.534	0.682	0.735
`BAAI/bge-large-en-v1.5`	**0.545**	**0.700**	**0.747**	**0.435**	**0.604**	**0.661**	**0.578**	**0.739**	**0.779**

We observe that `BAAI/bge-large-en-v1.5` consistently produced the strongest overall retrieval performance across the three languages. These experiments motivated the later use of `bge-large-en-v1.5` as the retrieval backbone for the final retrieve–then–rerank pipeline. Despite being smaller than the multilingual `multilingual-e5-large` encoder, `bge-large-en-v1.5` remained highly competitive under the translated-query setup, which is also consistent with its strong retrieval performance on the MTEB benchmark [11].

C. Cross-Encoder Model Comparison

Table 11 summarizes representative reranker-family experiments on the development split. These experiments were conducted during the exploratory reranker-selection phase and therefore reflect representative family configurations rather than a fully controlled comparison protocol. The later controlled Jina reranker ablations are reported in the main text.

The main pattern from this preliminary family comparison is that the Jina reranker family benefited most consistently from task-specific fine-tuning. While SciBERT and `BAAI/bge-reranker-base` produced competitive results in some languages, they did not match the strongest fine-tuned Jina configuration at the macro level.

Table 11
Representative cross-encoder reranker-family results on the dev split. Scores are MRR@5 after reranking the retrieved candidate pool.

Reranker family	EN	DE	FR	Macro
SciBERT cross-encoder (fine-tuned)	0.635	0.573	0.728	0.645
`BAAI/bge-reranker-base` (fine-tuned)	0.628	0.580	0.698	0.635
`jinaai/jina-reranker-v2-base-multilingual` (off-the-shelf)	0.646	0.541	0.693	0.626
`jinaai/jina-reranker-v2-base-multilingual` (fine-tuned)	**0.674**	**0.601**	**0.736**	**0.670**

D. LLM Judge Comparison

Table 12 reports the LLM-judge comparisons on the German development disagreement subset as described in Section 6.3. All scores correspond to MRR@5 after selective disagreement resolution.

Table 12
LLM judge comparison on the German development disagreement subset. The judge is applied only on retriever–reranker disagreement cases. Rows with † use translated English queries.

Judge configuration	MRR@5
Ollama committee + `Llama3 70B` tie-break	0.605
`Llama 3 70B`	0.618
`gpt-4.1-mini`†	0.628
`gpt-4.1-mini`	0.629
`gpt-5.5`	**0.653**

The comparison suggests that stronger closed-weight judge models consistently improve disagreement resolution performance. The results also support the earlier multilingual observations: GPT-family judges perform slightly better with the original-language German claim than with the translated query formulation.

The Ollama committee was an exploratory attempt to reduce dependence on a single large judge by combining `llama3:8b`, `mistral:7b`, and `gemma2:9b` with `Llama3 70B` decisions. We tested a four-model majority-vote scheme with `Llama3 70B` acting as the final decision model in tied cases. However, this committee did not outperform the stronger single-judge configurations.

E. LLM Judge Prompts

This appendix section documents the main LLM-judge prompts used in the final ablation cycle. All prompts were applied only after dense retrieval and cross-encoder reranking, and in the final selective pipeline only on cases where the retriever top-1 and reranker top-1 disagreed.

E.1. Prompt-Budget Screening

Before fixing the final disagreement prompt, we also screened candidate-list sizes of 3, 5, and 7 documents during prompt design. The later controlled comparison focused on top-5 prompts because this setting provided the best practical balance between candidate coverage and context budget in the downstream pipeline. Table 13 reports the saved top-5 prompt-budget ablations on a balanced development sample using `Llama 3 70B`.

E.2. Baseline Direct-Selection Prompt

```
Instruction: Act as a scientific fact-checker. You are given a claim
and a list of candidate scientific papers.
```

Table 13
LLM-judge prompt ablation on dev set sample using `Llama 3 70B`. Accuracy@1 is Hit@1, parse rate is fraction of examples with a valid parse, and context fail is fraction of examples that exceeded the context limit.

Prompt setting	Accuracy@1	Parse rate	Context fail
top5_full	0.7167	0.9167	0.0833
top5_trunc384	**0.7500**	0.9500	0.0500
letters_no_reason	0.7333	**1.0000**	**0.0000**
letters_with_reason	0.7333	**1.0000**	**0.0000**

```
Choose the single paper that provides the strongest evidence
supporting or verifying the claim.

Claim: <query text>

Candidates:
Paper ID A:
Title: <title>
Abstract: <truncated abstract>

Paper ID B:
...

Output instructions:
Return valid JSON only.
Use this exact format: {"selected_id": "A"}
```

This was the strongest practical prompt. We used letter IDs rather than rank numbers, omitted author names, and did not request justification.

E.3. Pairwise Proxy Prompt

```
Instruction: Act as a scientific fact-checker.
Candidate Paper A is the baseline current best guess.
First decide whether Paper A adequately supports or verifies the claim.
Then compare Papers B-E against Paper A.
If none is clearly better than Paper A, keep A.
If another paper is clearly better, choose the single best alternative.
Retrieval rank is only a weak prior.

Claim: <query text>

Candidates (order randomized):
Paper ID A (retrieval rank: X):
Title: <title>
Abstract: <truncated abstract>
...

Output instructions:
Return valid JSON only.
Use this exact format: {"selected_id": "X"}
```

Despite its intuitive appeal as a cheap pairwise-ranking proxy, this prompt underperformed the baseline direct-selection prompt on the same recoverable disagreement subset.

E.4. RankGPT-Style Permutation Prompt

```
You are RankGPT, an intelligent assistant that can rank passages
based on their relevance to a search query.
I will provide you with N candidate papers, each identified by a
number in square brackets.
Rank the papers based on how well they support or verify the claim.

Claim: <query text>

Candidates:
[1] Title: <title>
Abstract: <truncated abstract>
...

Rank the N candidates above based on their relevance to the claim.
Return only the ranking permutation using the format
[1] > [2] > ... > [N].
Do not explain your answer.
```

This prompt was inspired by permutation-only LLM ranking in the RankGPT family, but in our setting it remained weaker than the simpler direct-selection baseline.

F. Earlier Gold-Cluster Reranker

Before fixing the later comparable reranker protocol used in the main text, we also evaluated a gold-cluster Jina reranker branch. In this setting, HN were sampled from the cluster that contained the gold document. Table 14 reports the corresponding earlier development set MRR@5 values together with the matched nearest-hard and non-gold variants from the same stage of development. The gold-cluster branch was weaker than the strongest competing cluster-based alternatives and was therefore not carried forward into the later fixed-retriever comparison used in Table 6.

Table 14
Earlier Jina reranker ablation including the gold-cluster branch. All numbers are development-set MRR@5.

Model / Negative Strategy	EN	DE	FR
Cluster nearest-hard, best checkpoint	0.6680	**0.6005**	0.7196
Cluster gold, best checkpoint	0.6677	0.5830	0.7182
Cluster non-gold, best checkpoint	**0.6716**	0.5872	**0.7250**