BoPC at CheckThat! 2026: Efficient Source Retrieval for Social Media Claims Using Two-Stage Dense and Cross-Attention Networks

Notebook for the CheckThat! Lab at CLEF 2026

Nguyen-Phuc Mac[1,3,*,†], Minh-Phuc Mac[2,3,†], Vi Vo-An Nguyen[1,3] and Dien X. Tran[4]

[1]University of Information Technology - VNUHCM, Han Thuyen, Quarter 6, Linh Xuan Ward, Ho Chi Minh City, Vietnam
[2]University of Science - VNUHCM, 227 Nguyen Van Cu, Cho Quan Ward, Ho Chi Minh City, Vietnam
[3]Vietnam National University, Ho Chi Minh City, Vietnam
[4]Industrial University of Ho Chi Minh City, 12 Nguyen Van Bao, Hanh Thong Ward, Ho Chi Minh City, Vietnam

Abstract

We present the methodology and results of the BoPC team for Task 1 of the CLEF-2026 CheckThat! Lab, which focuses on retrieving the corresponding scientific publication from implicit references in social media posts across English, German, and French. To address this task, we propose a multi-stage multilingual retrieval pipeline inspired by the ViDRILL framework for Vietnamese legal document retrieval, adapted to bridge both the informal-to-formal stylistic gap and the cross-lingual gap between query and document. Specifically, we deliberately omit a sparse BM25 stage and instead we use a multilingual dense encoder multilingual E5-Instruct with a multilingual BGE reranker cross-encoder. Our approach achieves a mean reciprocal rank at 5 (MRR@5) of 63.73% on English, 59.94% on German, 72.67% on French, and 65.45% on average across all languages on the development set, ranking in the top 12 on the development leaderboard. On the official test set, the pipeline secures the 15th position with an MRR@5 of 59.3% on English, 56.2% on German, 62.7% on French, and 59.4% on average. We obtain this performance by running open-source models locally, without external training data, closed-source LLMs, or paid APIs, demonstrating that a carefully designed purely-dense multi-stage pipeline can deliver strong cross-lingual scientific source retrieval.

Keywords

Scientific source retrieval, Multilingual information retrieval, Dense retrieval, Cross-encoder re-ranking, CheckThat! Lab

1. Introduction

The rapid growth of scientific discourse on social platforms has made it routine for users to summarise, paraphrase, or rephrase the findings of peer-reviewed studies in informal posts, typically without providing a citation, a DOI, or a URL [1]. Tracing such implicit mentions back to their underlying publications is essential for evidence-based fact-checking, scholarly attribution, and informed public discourse. However, this task is intrinsically difficult: social-media queries are short, lexically sparse, and stylistically distant from the formal register of scientific abstracts; many posts contain only colloquial paraphrases of the original findings; and the candidate corpus is large enough that exact lexical matching alone rarely suffices [2, 3].

The CLEF-2026 CheckThat! Lab Task 1, *Source Retrieval for Scientific Web Claims* [4, 5], formalises this challenge as a retrieval problem. Given a social-media post that contains a scientific claim and implicitly refers to a scientific paper without providing a URL, DOI, title, or explicit citation, a system must identify and retrieve the mentioned paper from a predefined pool of candidate publications. The task therefore

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
†These authors contributed equally.
22521123@gm.uit.edu.vn (N. Mac); 22280065@student.hcmus.edu.vn (M. Mac); 24521989@gm.uit.edu.vn (V. V. Nguyen); 22650601.dien@student.iuh.edu.vn (D. X. Tran)
0009-0003-9459-0936 (N. Mac); 0009-0005-2304-9311 (M. Mac); 0009-0002-9963-2034 (V. V. Nguyen); 0009-0002-7819-9649 (D. X. Tran)

differs from standard document retrieval: the query is an informal social-media claim rather than a bibliographic description, and the relevant document is the scientific article that originally supports or reports the claim. As a follow-up to the 2025 edition [2, 3, 6], Task 1 in 2026 further extends the setting to English, German, and French posts against a shared, predominantly English-language corpus of scientific articles. The system therefore must bridge two simultaneous gaps: the informal-to-formal stylistic gap between social-media language and scientific writing, and the cross-lingual gap between non-English queries and English-dominant candidate documents, while remaining accurate, efficient, and language-robust.

In this paper, we present a multi-stage multilingual retrieval pipeline for Task 1 that is by the ViDRILL framework [7], originally developed for Vietnamese legal document retrieval at VLSP 2025 DRiLL [8]. Although the original target domain (long, structured Vietnamese statutes) differs substantially from this lab's dataset (short, multilingual social-media queries against English-dominant scientific abstracts) [5, 8], the multi-stage design principles introduced in ViDRILL — specifically dense retrieval, semantic re-ranking, and structured document segmentation — transfer naturally to our setting and motivate the architecture proposed in this paper. Figure 4 illustrates the overall design: each query is processed by a fine-tuned multilingual dense bi-encoder, the retrieved candidate set is re-ranked by a multilingual cross-encoder, and an exploratory listwise generative calibration stage was investigated. Concretely, our method integrates:

1. **Multilingual dense retrieval** based on a fine-tuned multilingual E5 (large) bi-encoder [9] indexed in an in-memory vector store so that English, German, and French posts can be directly mapped against the scientific corpus through a shared continuous embedding space. We deliberately do not include a sparse BM25 [10] stage: in a cross-lingual setting, exact term overlap between non-English queries and English-dominant abstracts is rarely available, and modern multilingual encoders already absorb most lexical signals during large-scale contrastive pre-training [9, 11, 12].
2. **Cross-encoder re-ranking** with a fine-tuned multilingual BGE reranker [11, 13], which directly concatenates the raw social-media post and each candidate abstract to refine the top of the ranking through deep cross-attention and evaluate direct semantic alignment.

This two-stage architecture is designed to bridge the semantic gap between informal web claims and formal scientific taxonomy. Building upon the design principles of ViDRILL, we adapt and evaluate specific components for the substantially different setting of short, cross-lingual queries:

- A *structural segmentation strategy* utilizes Docling [14] to partition formal scientific documents into self-contained passages of up to 256 tokens with a 25-token sliding window. While this satisfies the sequence length constraints of the subsequent bi-encoder, its primary architectural purpose is to enable the use of smaller, resource-efficient foundation models. By processing compact chunks rather than full documents, we significantly lower computational and financial overhead—drastically reducing costs during both the model fine-tuning phase and downstream deployment use without sacrificing local semantic cohesion.
- A *task-specific domain adaptation mechanism* bootstraps supervision for both the bi-encoder and the cross-encoder. We utilize pre-mined query-document pairs to fine-tune the bi-encoder via a contrastive Multiple Negatives Ranking loss objective [15], followed by maximum-step fine-tuning on the cross-encoder to produce precise relevance probabilities.
- An *exploratory zero-shot generative re-scoring stage* utilizing a 4-bit quantized Large Language Model (Qwen3.5-4B [16]). Though ultimately excluded from the final submission architecture due to the introduction of semantic noise, we investigated listwise prompting to heuristically fuse binary LLM (LLM - large language model) classifications with cross-attention probabilities.

Importantly, the primary pipeline components rely exclusively on open-source foundation models [9, 11] deployed locally. The final optimized architecture operates without external training data, closed-source LLMs, or paid APIs, making it reproducible and easily transferable to other multilingual retrieval settings [3]. Overall, our main contributions are:

1. A **purely-dense, multilingual two-stage retrieval architecture** for this lab, demonstrating that a fine-tuned mE5-large bi-encoder combined with a BGE reranker cross-encoder provides strong cross-lingual performance (ranking in the top 12 on the development leaderboard) without a sparse BM25 first stage or explicit machine translation.
2. A **cross-domain application of design principles from ViDRILL** [7] - such as structured segmentation and cascading semantic refinement — from long-document Vietnamese legal retrieval to short, cross-lingual scientific source retrieval, providing empirical evidence that these architectural workflows generalize effectively.
3. A **comprehensive experimental analysis**, including ablations on the cross-encoder candidate pool size (k), an evaluation of zero-shot generative LLM listwise calibration, and a critical error analysis highlighting persistent structural linguistic bottlenecks in German claim verification.

By releasing this carefully engineered pipeline, we aim to support ongoing efforts to counter misinformation by attributing informal scientific claims to their original publications, and to offer an open, multilingual blueprint that other teams can extend.

2. Related Work

Identifying the scientific source behind an informal claim sits at the intersection of many research strands: automated fact-checking and scientific claim verification, dense and hybrid retrieval, and multilingual neural retrieval with re-ranking. We organise the literature along these axes, with particular attention to the systems whose design principles most directly motivate our approach.

Scientific Source Retrieval and Fact-Checking. Automated fact-checking pipelines have historically depended on robust evidence retrieval over structured knowledge bases or curated news corpora [17, 18]. More recent work has shifted toward unstructured, domain-specific corpora, particularly scientific literature, where claims must be verified against peer-reviewed studies [19, 20]. The COVID-19 pandemic intensified interest in this setting, since social-media posts frequently invoke biomedical findings without proper attribution [1]. Subtask 4b of CheckThat! 2025 introduced the first large-scale benchmark explicitly framed as tweet-to-study matching against the CORD-19 collection [6]. The participating teams SourceSniffers (1st), AIRwaves (2nd) [2], Deep Retrieval (3rd) [3], and DS@GT (16th) [21], among others converged on a broadly similar recipe: a strong first-stage retrieval followed by a neural re-ranker. This year's lab inherits this framing but adds a cross-lingual dimension by introducing German and French query sets against a shared collection [4], opening new challenges for retrieval models that must align informal multilingual queries with formal, English-dominant scientific abstracts.

Sparse, Dense, and Hybrid Retrieval. Retrieval methods are conventionally grouped into sparse and dense families. While sparse approaches like BM25 [10] are robust in monolingual settings due to exact term matching, they are poorly suited for cross-lingual retrieval where non-English claims share zero lexical overlap with English abstracts [22, 12]. Dense retrievers [23] resolve this by mapping text into a shared semantic vector space, where modern multilingual foundation embeddings — such as multilingual E5 [9], BGE [11], and mGTE [24] — demonstrate powerful zero-shot cross-lingual generalization [25]. Although hybrid frameworks combining sparse and dense signals remain standard in recent scientific [3, 2] and legal [7] tasks, modern multilingual encoders inherently absorb rich lexical signals during contrastive pre-training [9, 11, 12]. Consequently, our system deliberately omits a sparse BM25 stage and avoids the severe computational and financial overhead of multi-model ensembles. Instead, we utilize a single fine-tuned multilingual-e5-large bi-encoder to demonstrate that targeted contrastive domain adaptation can independently maximize candidate recall.

Multi-Stage Pipelines, Re-Ranking, and Architectural Calibration. Two-stage pipelines combining an efficient first-stage retriever with a powerful cross-encoder re-ranker represent the de facto standard for high-precision text retrieval [13], as exemplified by the top-performing CheckThat! 2025 architectures [2, 3]. Drawing inspiration from cascading semantic refinement frameworks like ViDRILL [7], we adapt these multi-stage principles into a highly optimized, resource-efficient pipeline tailored for

short social media claims. Rather than relying on heavy ensembles or complex multi-stage hard negative mining, we couple our single fine-tuned bi-encoder with a fine-tuned cross-encoder (bge-reranker-v2-m3) leveraging deep cross-attention interactions [11]. Crucially, we implement a structural segmentation strategy that partitions long documents into compact 256-token passages to deliberately focus our engineering efforts on smaller foundation models, minimizing both fine-tuning and deployment costs. Finally, we extend this paradigm by evaluating an exploratory third-stage listwise generative calibration step using a locally deployed, 4-bit quantized LLM to empirically chart the performance boundaries of zero-shot generative inference against fine-tuned cross-attention probabilities.

Multilingual and Cross-Lingual Retrieval. A defining challenge of CheckThat! 2026 Task 1 is its cross-lingual setting, requiring non-English claims to be matched against an English-dominant scientific corpus. While dense retrieval historically underperformed sparse baselines in non-English monolingual tasks [22], contrastive pre-training frameworks like mContriever [12] proved that dense models can successfully map non-English queries directly to English target documents without explicit translation supervision. Modern multilingual foundation architectures [9, 11, 24] natively extend this continuous embedding space across over 100 languages. Our pipeline capitalizes on this capability for both initial candidate generation and cross-attention re-ranking via BGE re-ranker [11], handling English, German, and French queries uniformly. By bypassing explicit machine translation, our direct-multilingual pipeline inherently preserves the original surface forms of critical named entities and numerical scientific claims while protecting the system from downstream translation noise propagation.

3. Dataset

This section details the dataset's structural characteristics, including a comprehensive token-length distribution analysis designed to explicitly isolate and quantify cross-lingual tokenization inflation across the English, French, and German claim queries. We also leverage mathematical divergence metrics to evaluate intra-lingual split consistency, verifying that the hidden test partitions maintain structural alignment with our training baselines.

The dataset consists of implicit scientific claims extracted from multilingual web sources and paired with their corresponding formal publications. Explicit identifiers such as URLs or DOIs are deliberately omitted from these pairs to simulate real-world scenarios of implicit scientific discourse. Each query represents an informal, cross-lingual reference to a scientific study, while each target publication within the 10,000-document collection is structured using comprehensive metadata fields, including a unique publication key, title, abstract, venue, and authors. To query this corpus, the claims are partitioned into 19,244 training, 4,993 development, and 8,172 test data points spanning English, French, and German.

3.1. Analytical Setup

To ensure statistical consistency and avoid artifacts introduced by language-specific whitespace or subword segmentation, all claims and documents are tokenized using the multilingual E5-large tokenizer. This specific configuration was selected because both the first-stage dense retriever (mE5-large) and the second-stage cross-encoder (bge-reranker-v2-m3) share the XLM-RoBERTa base architecture and its underlying SentencePiece vocabulary. Consequently, the following analyses accurately reflect the exact input representations processed by the full retrieval pipeline.

To rigorously quantify the observed distribution shifts, we employ the Jensen-Shannon (JS) divergence and the Wasserstein metric (Earth Mover's Distance). While Kullback-Leibler (KL) divergence is often used in information theory, its asymmetric nature ($D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$) and extreme sensitivity to zero-probability bins make it overly brittle for small, sparse datasets. JS divergence resolves this by providing a symmetric, smoothed, and bounded measure of similarity. Furthermore, the Wasserstein metric evaluates physical geometric drift across the x-axis, allowing us to explicitly quantify sequence shift in terms of absolute token counts.

Formally, let P and Q represent the probability distributions of the target and reference sequence

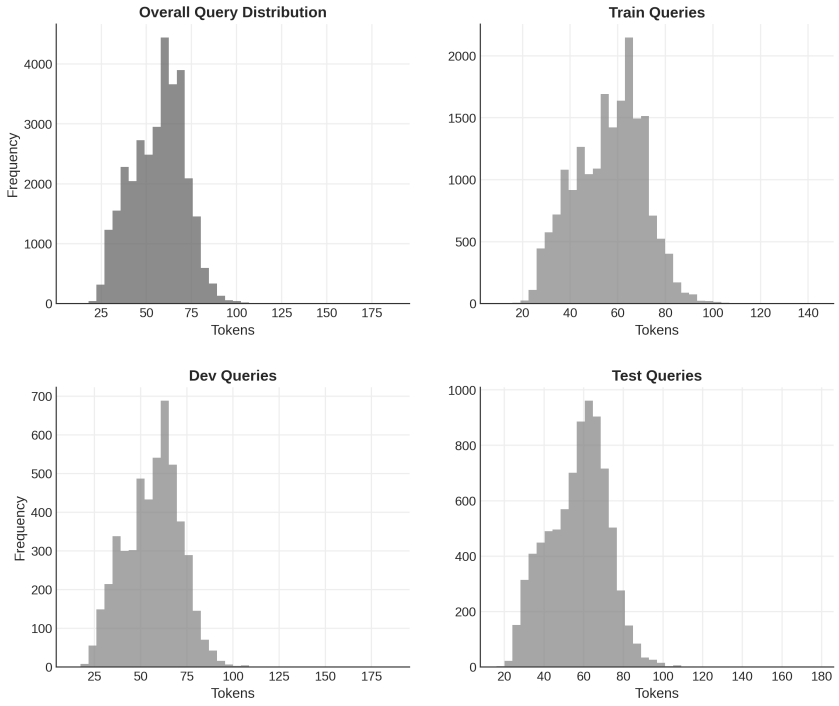

Figure 1: Aggregate cross-language query token-length distributions across the training, development, and test splits.

Table 1: Cross-lingual aggregate distribution shift. The table evaluates the stability of the combined multilingual evaluation splits (Target P) against their respective reference baselines (Reference Q). Distributional similarity is evaluated using JS divergence, while physical sequence-length drift is quantified using the Wasserstein metric.

Split	N (Target)	N (Ref)	JS Div	Wasserstein
Agg. Dev vs Train	4,993	19,244	0.0011	0.3694
Agg. Test vs Train	8,172	19,244	0.0011	0.6528
Agg. Test vs Dev	8,172	4,993	0.0012	0.3853

lengths, respectively. The Jensen-Shannon divergence is defined as:

$$JSD(P \parallel Q) = \frac{1}{2} D_{KL}(P \parallel M) + \frac{1}{2} D_{KL}(Q \parallel M) \tag{1}$$

where $M = \frac{1}{2}(P + Q)$ is the pointwise mixture distribution, and D_{KL} denotes the standard Kullback-Leibler divergence.

The 1-Wasserstein distance (W_1), which measures the minimum work required to transform distribution P into Q, is computed for these one-dimensional empirical distributions as the integral of the absolute difference between their respective cumulative distribution functions (CDFs), F_P and F_Q:

$$W_1(P, Q) = \int_{-\infty}^{\infty} |F_P(x) - F_Q(x)| \, dx \tag{2}$$

3.2. Distribution Analysis

Figure 1 and Table 1 initially present the aggregated sequence-length metrics across all languages. While the aggregated Development and Test splits appear exceptionally stable relative to the Training baseline—and demonstrate near-perfect alignment with each other (Wasserstein 0.3853)—it is critical to note that this stability is a manifestation of Simpson's Paradox. The dominant English sample size

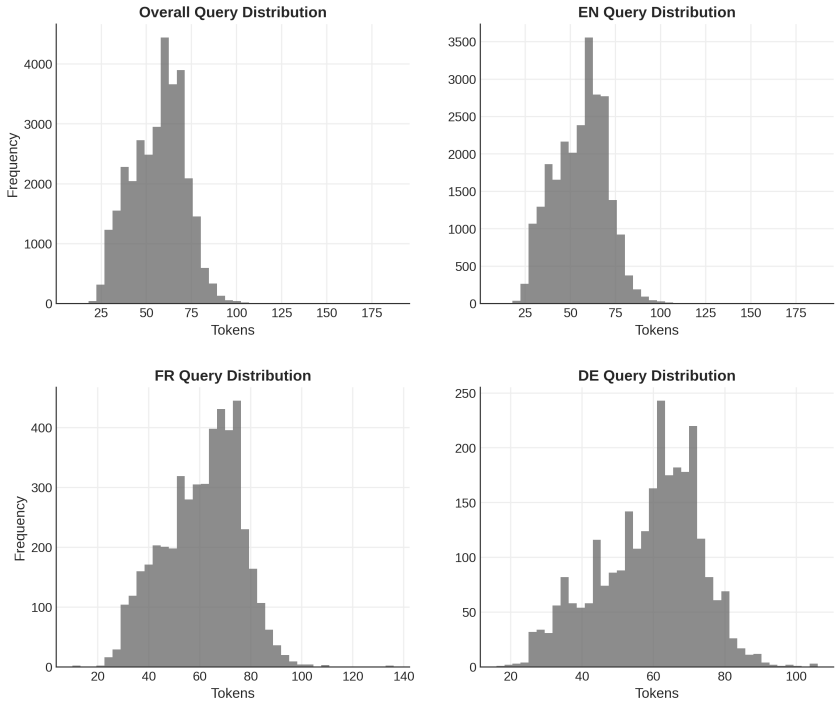

Figure 2: Target language query distributions compared to the overall and English baseline distributions.

Table 2: Target versus baseline language distribution shift. The table isolates cross-lingual tokenization inflation by comparing the target split (Target P) against the reference split (Reference Q).

Comparison	N (Target)	N (Ref)	JS Div	Wasserstein
FR vs EN	4,729	24,958	0.0216	5.2939
DE vs EN	2,722	24,958	0.0097	3.3286
FR vs DE	4,729	2,722	0.0106	2.0467

mathematically masks the underlying tokenization inflation and sparsity present in the French and German sub-corpora.

To isolate this masking effect, Figure 2 and Table 2 evaluate the cross-lingual tokenization inflation explicitly. German and French claims exhibit distinct morphological profiles compared with English, partly due to compounding and noun-formation patterns that fracture more heavily under SentencePiece subword tokenization. The Wasserstein distances confirm that French and German queries require approximately 5.3 and 3.3 additional tokens on average, respectively, to represent semantic claims equivalent to the English baseline. Furthermore, the direct comparison between the two European languages isolates their specific morphological differences, revealing that French claims undergo even heavier subword fracturing, requiring roughly 2.0 additional tokens on average compared to German.

Having established the cross-lingual discrepancies, Table 3 quantifies the intra-lingual distribution shift to verify split integrity. When evaluated strictly within their own language boundaries, the distributions exhibit strong alignment. The English subsets show near-zero drift. While the French and German test sets show slightly higher Wasserstein distances primarily due to sample sparsity ($N < 1,500$), they remain well within the acceptable tolerances of standard 512-token context windows, indicating that the test sets do not introduce severe out-of-distribution (OOD) length shifts.

Finally, Figure 3 illustrates the token distribution of the evidence document collection. These target lengths, combined with the query inflation observed in the European languages, motivate the evaluation of multiple dynamic chunking configurations, such as max_len = 256 and max_len = 512, to balance local context preservation against the sequence-length constraints imposed by the embedding and

Table 3: Intra-lingual distribution shift across the official development and test splits. The table quantifies sequence-length drift between the target evaluation splits (Target P) and their respective training baselines (Reference Q).

Split	N (Target)	N (Ref)	JS Div	Wasserstein
EN Dev	3,905	14,977	0.0013	0.2884
EN Test	6,076	14,977	0.0009	0.4800
FR Dev	702	2,807	0.0068	1.0416
FR Test	1,220	2,807	0.0068	0.5875
DE Dev	386	1,460	0.0121	1.0472
DE Test	876	1,460	0.0104	1.8566

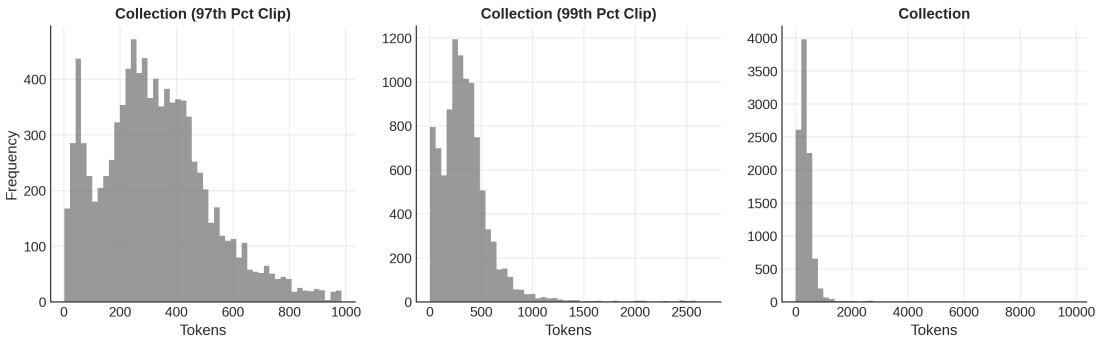

Figure 3: Document collection token-length distributions, including the 97th and 99th percentile truncation clips.

cross-encoder models.

4. Methodology

This section details the empirical framework and operational pipeline developed for the cross-lingual source retrieval task. The methodology is designed to structurally evaluate the efficacy of a purely dense retrieval architecture. The following subsections systematically delineate the high-level system architecture, the preparation of the scientific corpus and query data, and the specific mechanics governing the multi-stage retrieval cascade.

4.1. System Architecture

The proposed system for the source retrieval task (illustrated in Figure 4) employs a purely dense, two-stage retrieve-and-rerank architecture. This pipeline explicitly omits sparse lexical components (e.g., BM25) to test the hypothesis that a highly optimized, task-specific dense retrieval framework can independently bridge the semantic gap between informal, cross-lingual web claims and formal scientific taxonomy. The architecture processes queries through a deterministic sequence:

First-Stage Dense Retrieval. Informal web claims are ingested and vectorized using a fine-tuned multilingual bi-encoder. These query embeddings are evaluated against a pre-computed, in-memory index of chunked scientific documents using exact vector similarity operations. This stage is designed to maximize recall, rapidly isolating a broad candidate pool of the top-k relevant passages.

Second-Stage Re-ranking. The retrieved candidate pool is subsequently processed by a fine-tuned cross-encoder. By evaluating the direct semantic alignment of the concatenated claim and candidate document pairs, this stage refines the ordering to maximize precision for the final MRR@5 evaluation.

Additionally, an exploratory pipeline variant was investigated to evaluate a potential third-stage generative re-scoring stage utilizing a pre-trained Large Language Model (LLM). Due to scope constraints,

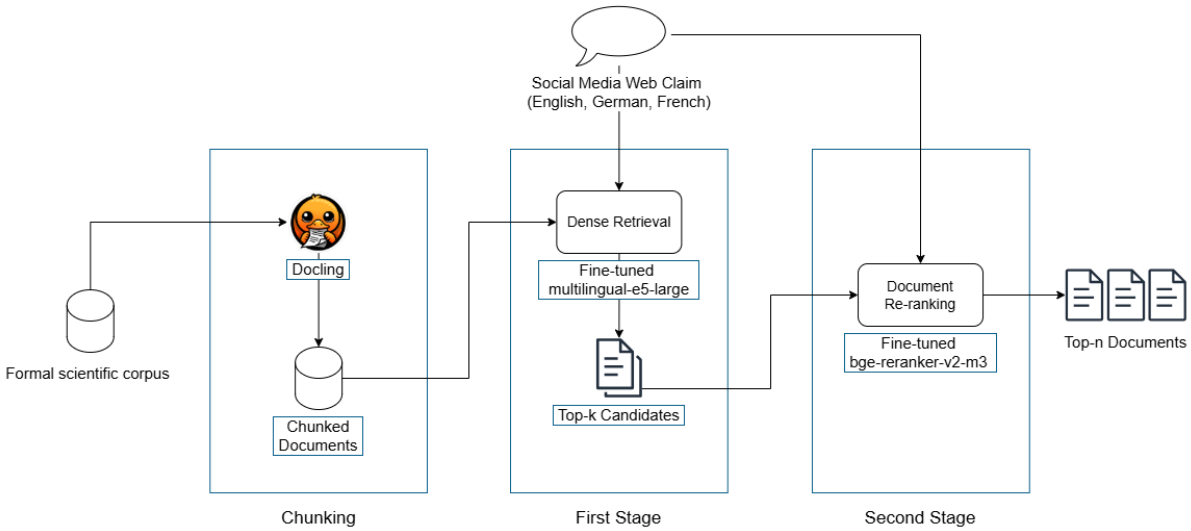

Figure 4: The proposed two-stage dense retrieval architecture. The formal scientific corpus is structurally segmented via docling to construct the in-memory index (Chunking). Raw, cross-lingual web claims bypass heuristic preprocessing and are directly vectorized by a fine-tuned multilingual E5-large bi-encoder for initial candidate retrieval against the chunked index (First Stage). The original claims and the retrieved top-k candidate passages are subsequently concatenated and evaluated by a fine-tuned bge-reranker-v2-m3 cross-encoder to produce the final relevance ordering (Second Stage).

the model was deployed for direct inference without task-specific fine-tuning. This zero-shot application introduced semantic noise and resulted in a degradation of primary retrieval metrics compared to the highly optimized two-stage pipeline, ultimately leading to its exclusion from the final submission architecture.

4.2. Data Processing

The data processing pipeline was strictly bifurcated into corpus preparation and claim ingestion. To process the formal scientific corpus, documents were parsed and structurally segmented utilizing the docling library. To maintain semantic cohesion while adhering to the sequence length constraints of the subsequent bi-encoder, the parsed documents were partitioned into atomic passages with a maximum length of 256 tokens and a sliding window overlap of 25 tokens.

Conversely, query preprocessing for the informal web claims was deliberately minimized. The system ingested raw claim text without applying heuristic cleaning, lexical normalization, or explicit entity extraction. Furthermore, to address the cross-lingual dimension of the task, the architecture relied entirely on the native continuous embedding space of the multilingual bi-encoder to map German and French claims directly to the English corpus. Explicit machine translation of non-English queries prior to vectorization was excluded from the current pipeline due to temporal constraints and remains an avenue for future comparative work.

4.3. Multi-Stage Retrieval Pipeline

The inference phase operates as a sequential cascade, progressively refining the candidate document pool from broad semantic recall to strict relevance ordering. The pipeline consists of a foundational dense retriever for initial candidate generation, a second-stage cross-encoder for fine-grained semantic ranking, and an experimental third-stage step utilizing zero-shot generative inference for final score calibration.

Initial Candidate Generation. The retrieval phase is initiated by projecting the encoded cross-lingual query into the pre-computed in-memory index utilizing the multilingual E5-large foundation model. This specific architecture was selected not only for its robust cross-lingual alignment

space—which natively supports mapping French and German queries to English documents—but also because it serves as the official competition baseline, providing a strictly controlled foundation to empirically isolate the impact of our contrastive fine-tuning. The system executes an exhaustive vector similarity computation, utilizing cosine similarity to retrieve a broad candidate pool of the most semantically adjacent scientific passages. This stage prioritizes high recall while maintaining low computational latency.

Cross-Attention Scoring. To refine the initial candidate pool, a secondary re-ranking stage is implemented utilizing a state-of-the-art cross-encoder (bge-reranker-v2-m3). The BGE-m3 architecture was deliberately chosen for this stage due to its optimized capacity for multilingual dense retrieval and its proven efficacy in modeling highly asymmetrical text pairs (e.g., short informal claims versus dense scientific chunks). The cross-encoder directly concatenates the raw query and the retrieved candidate passage, modeling their deep cross-attention to output a precise relevance probability.

Generative Listwise Calibration (Exploratory). As an exploratory pipeline extension, a final generative re-evaluation stage was implemented using a 4-bit quantized Large Language Model (Qwen3.5-4B). To minimize inference bottlenecks, this stage was restricted to the top 10 candidates outputted by the secondary cross-encoder. The LLM was deployed in a zero-shot configuration. The system utilized a listwise prompting strategy, presenting the model with the raw query and the 10 candidate chunks (randomly shuffled to mitigate positional bias). The LLM was strictly instructed to act as a binary relevance classifier, outputting an array of indices corresponding to relevant passages. A heuristic scoring fusion was then applied: passages identified by the LLM received an absolute $+1.0$ score boost, which was additively combined with the prior cross-encoder probability. The fused scores dictated the absolute ordering for the final deduplicated top-5 documents. As noted earlier, this zero-shot configuration introduced out-of-distribution noise during testing and was ultimately excluded from the primary submission architecture.

5. Results

This section presents the empirical evaluation of the proposed multi-stage dense retrieval architecture. We first outline the experimental setup and baseline benchmarks, followed by the quantitative performance on both the development and official blind test splits. Subsequent subsections provide a component analysis isolating the impact of the cross-encoder candidate pool size, a critical error analysis addressing cross-lingual discrepancies, and an evaluation of the zero-shot LLM re-scoring stage.

5.1. Experimental Setup

Our experimental pipeline is built upon the multilingual E5-large (mE5) foundation model. Transitioning from high-level architecture to empirical execution, the exact hyperparameter configurations and optimization regimes utilized to train and deploy each pipeline stage are detailed below.

Bi-Encoder Fine-Tuning. To adapt the base embeddings to the specific semantic requirements of scientific claim verification, the mE5 model was fine-tuned using a contrastive MultipleNegativesRankingLoss objective. The training pipeline utilized a dataset of pre-mined query-document pairs partitioned into a 95/5 random split. Fine-tuning was executed over 20 epochs with a global batch size of 64 and a dynamically scaled learning rate of 1.6×10^{-5} utilizing BF16 mixed precision.

Cross-Encoder Adaptation. The second-stage cross-encoder was fine-tuned using the FlagEmbedding framework. The optimization parameters included a train group size of 21, restricted to a maximum query length of 128 tokens and a passage length of 256 tokens. The fine-tuning was constrained to 200 maximum steps with a peak learning rate of 1×10^{-5}, a per-device batch size of 4, and a weight decay coefficient of 0.01.

LLM Inference Configuration. For the exploratory third-stage generative re-scoring, the Large Language Model was deployed using 4-bit quantization to optimize GPU memory utilization (by using Unsloth [26]). Inference was executed in batches of 16, necessitating explicit left-side padding

Table 4
Performance evaluation (MRR@5) on the official development and test splits. The table compares the official zero-shot baseline against various cross-encoder candidate pool sizes (k) and the third-stage LLM. All reported metrics reflect official submissions to the competition leaderboard. All of the evaluations at stage 2 are models fine-tuned on train set, and evaluated using zero-shot technique.

Split	Model Configuration	Overall	English	French	German
Dev	CLEF Official Baseline (mE5 Zero-Shot)	0.4400	0.5000	0.4600	0.3700
Dev	Stage 2: BGE Reranker ($k = 20$)	0.6301	**0.6420**	0.6759	0.5725
Dev	Stage 2: BGE Reranker ($k = 50$) *	**0.6545**	0.6373	**0.7267**	**0.5994**
Dev	Stage 2: BGE Reranker ($k = 100$)	0.6406	0.6210	0.7175	0.5833
Dev	Stage 3: LLM (Inference)	0.6466	0.6254	0.7181	0.5962
Test	Stage 2: BGE Reranker ($k = 50$) *	0.5940	0.5930	0.6270	0.5620

* Official BoPC leaderboard submission.

(padding_side="left") to accommodate the model's auto-regressive generation. To ensure the model adhered strictly to outputting structural indices rather than conversational text, the generation parameters were constrained to deterministic greedy decoding (do_sample=False) and capped at a maximum of 15 new tokens. The system parsed these integer outputs via regular expressions, mapping them back to the randomized top-10 listwise prompts to apply the final heuristic score fusion.

5.2. Retrieval Performance and Error Analysis

The performance of the proposed architecture was evaluated against the official competition benchmarks, with our verified leaderboard validation metrics summarized in Table 4. Crucially, because our upstream document processing partitions long-form texts into smaller overlapping passages, multiple retrieved chunks can share identical source identifiers. To isolate unique document targets and evaluate true retrieval depth, all experimental runs—including our baseline reproduction—apply post-retrieval deduplication to retain only unique publication keys within the top-k candidate pool. Beyond these official submission metrics, understanding the operational dynamics of the retrieval pipeline requires a granular component analysis exploring cross-encoder pool sensitivities, cross-lingual bottlenecks, and late-stage generative re-scoring.

Standalone First-Stage Evaluation. To establish a comparative baseline for the re-ranking stage, we first evaluated the impact of contrastive fine-tuning on the primary mE5 retriever strictly on the development set. The official competition baseline, utilizing an off-the-shelf, zero-shot implementation of mE5, achieved an overall development MRR@5 of 0.4400 (English: 0.5000, French: 0.4600, German: 0.3700). Independent of official leaderboard submissions, our internal evaluation of the fine-tuned first-stage architecture established an overall MRR@5 of 0.6340 (English: 0.6343, French: 0.6794, German: 0.5494) on the official development set. This substantial performance delta confirms that task-specific domain adaptation is strictly necessary to bridge the semantic gap before any secondary re-ranking occurs.

Candidate Pool Expansion and Re-ranking Precision. Transitioning to the complete two-stage pipeline, the empirical results demonstrate a high sensitivity to the retrieval candidate pool size (k) passed to the secondary cross-encoder. The optimal configuration was observed at $k = 50$, which elevated the overall MRR@5 to 0.6545, substantially outperforming the standalone first-stage retrieval baseline. Varying the pool size reveals a strict recall-precision trade-off. Restricting the candidate pool to $k = 20$ bottlenecks initial recall, preventing the cross-encoder from evaluating the true positive document and resulting in a degraded overall MRR@5 of 0.6301. Conversely, over-expanding the candidate pool to $k = 100$ introduces an excess of hard negatives that disrupts precision. Interestingly, cross-lingual sensitivity varies within this dynamic: while $k = 50$ maximizes French and German performance, the English split actually achieves its peak precision at the tighter $k = 20$ threshold.

Cross-Lingual Discrepancies and Generalization. Analysis reveals a persistent bottleneck in

German claim verification. Across all configurations, the German MRR@5 consistently trailed English and French performance by a substantial margin. This discrepancy is likely attributable to structural linguistic factors. German's heavy reliance on noun compounding interacts poorly with standard subword tokenization algorithms, resulting in fragmented embeddings that degrade the bi-encoder's ability to maintain semantic alignment. Furthermore, the evaluation on the official blind test split revealed a noticeable generalization gap. The overall MRR@5 decreased from 0.6545 on the development split to 0.5940 on the test split for the optimal $k = 50$ configuration, suggesting that the cross-encoder slightly overfitted to the semantic distribution of the development set.

Generative LLM Exploratory Evaluation. To evaluate the efficacy of late-stage generative listwise ranking, a third-stage analysis was conducted utilizing a zero-shot Large Language Model. The LLM was tasked with extracting relevant passages from the top-10 cross-encoder candidates, applying an absolute $+1.0$ heuristic score boost to identified targets. The empirical results demonstrate that this zero-shot generative inference universally degraded retrieval performance. Compared to the foundational Stage 2 ($k = 50$) baseline, the Stage 3 LLM integration reduced the overall MRR@5 from 0.6545 to 0.6466. This degradation confirms that applying LLMs in a zero-shot capacity for strict, precision-based ranking tasks introduces uncalibrated noise, and that continuous cross-attention probabilities remain superior unless the generative model undergoes task-specific fine-tuning.

6. Conclusion

In this working note, we presented a purely dense, two-stage retrieval architecture for the CLEF 2026 CheckThat! Lab Task 1, explicitly omitting sparse lexical components to evaluate the independent capabilities of fine-tuned dense embeddings. Our evaluation demonstrated that the combination of an mE5 bi-encoder and a BGE cross-encoder provides a robust mechanism for cross-lingual scientific claim verification. Furthermore, our empirical distribution analysis quantified critical cross-lingual sequence-length drift, providing a mathematical diagnosis for the performance discrepancies observed across the English, French, and German partitions.

While the system established a robust baseline, future work will focus on optimizing the pipeline sequentially from initial ingestion to final ranking. Beginning with data preprocessing and query formulation, we intend to refine the training corpus by filtering noisy metadata. To further bridge the semantic gap prior to vectorization, we will investigate LLM-guided lexical style transfer alongside explicit query translation pipelines, particularly to mitigate the tokenization inefficiencies currently degrading the German partition. Transitioning to the first-stage dense retrieval, systematic hyperparameter optimization will be explored, specifically tuning the batch size and calibrating the temperature for the bi-encoder's contrastive loss.

At the second-stage re-ranking level, optimization will expand beyond basic hyperparameter calibration—such as tuning the learning rate and batch size for stable convergence—to include a comparative architectural evaluation of alternative, domain-specific cross-encoder models. Finally, at the third-stage terminal ranking, future iterations will replace the noisy zero-shot generative inference with a task-specific, fine-tuned Large Language Model adapted explicitly for the structural nuances of scientific claim verification.

Acknowledgments

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

Declaration on Generative AI

During the preparation of this work, we used Gemini, Claude, and Prism in order to draft content, improve writing style, grammar and spelling check, and manage citations. After using these tools, we

reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] F. Alam, S. Shaar, F. Dalvi, H. Sajjad, A. Nikolov, H. Mubarak, G. Da San Martino, A. Abdelali, N. Durrani, K. Darwish, A. Al-Homaid, W. Zaghouani, T. Caselli, G. Danoe, F. Stolk, B. Bruntink, P. Nakov, Fighting the COVID-19 Infodemic: Modeling the Perspective of Journalists, Fact-Checkers, Social Media Platforms, Policy Makers, and the Society, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 611–649. URL: https://aclanthology.org/2021.findings-emnlp.56/. doi:10.18653/v1/2021.findings-emnlp.56.

[2] 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: [27], 2025, pp. 826–844.

[3] 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: [27], 2025, pp. 1141–1155.

[4] 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.

[5] 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.

[6] S. Hafid, Y. S. Kartal, S. Schellhammer, K. Boland, D. Dimitrov, S. Bringay, K. Todorov, S. Dietze, Overview of the CLEF-2025 CheckThat! Lab Task 4 on Scientific Web Discourse, in: [27], 2025, pp. 722–732.

[7] D. X.Tran, T. D.Truong, K. C.Nguyen, ViDRILL: A Multi-Stage Retrieval Framework for Vietnamese Legal Document Search, in: L. C. Mai, N. T. M. Huyen, N. T. T. Trang (Eds.), Proceedings of the 11th International Workshop on Vietnamese Language and Speech Processing, Association for Computational Linguistics, Hanoi, Vietnam, 2025, pp. 120–126. URL: https://aclanthology.org/2025.vlsp-1.17/.

[8] T.-H.-Y. Vuong, T.-M. Nguyen, H.-T. Nguyen, T.-K. Dao, H.-T. Nguyen, H.-Q. Le, DRILL Shared Task 2025: The Challenge of Deep Retrieval in the Expansive Legal Landscape, in: L. C. Mai, N. T. M. Huyen, N. T. T. Trang (Eds.), Proceedings of the 11th International Workshop on Vietnamese Language and Speech Processing, Association for Computational Linguistics, Hanoi, Vietnam, 2025, pp. 113–119. URL: https://aclanthology.org/2025.vlsp-1.16/.

[9] 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.

[10] S. Robertson, H. Zaragoza, The Probabilistic Relevance Framework: BM25 and Beyond, Foundations and Trends® in Information Retrieval 4 (2009) 1–174. URL: https://www.emerald.com/ftinr/article/4/1-2/1/1326508/The-Probabilistic-Relevance-Framework-BM25-and. doi:10.1561/1500000019.

[11] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 2318–2335. URL: https://aclanthology.org/2024.findings-acl.137/. doi:10.18653/v1/2024.findings-acl.137.

[12] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, E. Grave, Unsupervised

Dense Information Retrieval with Contrastive Learning, Trans. Mach. Learn. Res. 2022 (2022). URL: https://openreview.net/forum?id=jKN1pXi7b0.
[13] R. Nogueira, K. Cho, Passage Re-ranking with BERT, 2020. URL: https://arxiv.org/abs/1901.04085. arXiv:1901.04085.
[14] D. S. Team, Docling Technical Report, Technical Report, 2024. URL: https://arxiv.org/abs/2408.09869. doi:10.48550/arXiv.2408.09869. arXiv:2408.09869.
[15] M. Henderson, R. Al-Rfou, B. Strope, Y. hsuan Sung, L. Lukacs, R. Guo, S. Kumar, B. Miklos, R. Kurzweil, Efficient natural language response suggestion for smart reply, 2017. URL: https://arxiv.org/abs/1705.00652. arXiv:1705.00652.
[16] Qwen Team, Qwen3.5: Towards native multimodal agents, 2026. URL: https://qwen.ai/blog?id=qwen3.5.
[17] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a Large-scale Dataset for Fact Extraction and VERification, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 809–819. URL: https://aclanthology.org/N18-1074/. doi:10.18653/v1/N18-1074.
[18] A. Vlachos, S. Riedel, Fact Checking: Task definition and dataset construction, in: C. Danescu-Niculescu-Mizil, J. Eisenstein, K. McKeown, N. A. Smith (Eds.), Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, Association for Computational Linguistics, Baltimore, MD, USA, 2014, pp. 18–22. URL: https://aclanthology.org/W14-2508/. doi:10.3115/v1/W14-2508.
[19] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, H. Hajishirzi, Fact or Fiction: Verifying Scientific Claims, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7534–7550. URL: https://aclanthology.org/2020.emnlp-main.609/. doi:10.18653/v1/2020.emnlp-main.609.
[20] D. Wadden, K. Lo, B. Kuehl, A. Cohan, I. Beltagy, L. L. Wang, H. Hajishirzi, SciFact-Open: Towards open-domain scientific claim verification, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 4719–4734. URL: https://aclanthology.org/2022.findings-emnlp.347/. doi:10.18653/v1/2022.findings-emnlp.347.
[21] J. Schofield, S. Tian, H. T. T. Truong, M. Heil, DS@GT at CheckThat! 2025: Exploring Retrieval and Reranking Pipelines for Scientific Claim Source Retrieval on Social Media Discourse, in: [27], 2025, pp. 1193–1202.
[22] X. Zhang, X. Ma, P. Shi, J. Lin, Mr. TyDi: A Multi-lingual Benchmark for Dense Retrieval, in: D. Ataman, A. Birch, A. Conneau, O. Firat, S. Ruder, G. G. Sahin (Eds.), Proceedings of the 1st Workshop on Multilingual Representation Learning, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 127–137. URL: https://aclanthology.org/2021.mrl-1.12/. doi:10.18653/v1/2021.mrl-1.12.
[23] 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: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6769–6781. URL: https://aclanthology.org/2020.emnlp-main.550/. doi:10.18653/v1/2020.emnlp-main.550.
[24] X. Zhang, Y. Zhang, D. Long, W. Xie, Z. Dai, J. Tang, H. Lin, B. Yang, P. Xie, F. Huang, M. Zhang, W. Li, M. Zhang, mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval, in: F. Dernoncourt, D. Preoţiuc-Pietro, A. Shimorina (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, Association for Computational Linguistics, Miami, Florida, US, 2024, pp. 1393–1412. URL: https://aclanthology.org/2024.emnlp-industry.103/. doi:10.18653/v1/2024.emnlp-industry.103.
[25] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, BEIR: A Heterogeneous Benchmark

for Zero-shot Evaluation of Information Retrieval Models, in: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021. URL: https://openreview.net/forum?id=wCu6T5xFjeJ.
[26] D. Han, M. Han, Unsloth team, Unsloth, 2023. URL: https://github.com/unslothai/unsloth.
[27] G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 2025.