Multilingual *vs* Crosslingual Retrieval of Fact-Checked Claims: A Tale of Two Approaches

Alan Ramponi,[1*] **Marco Rovera,**[1*] **Robert Moro,**[2] **Sara Tonelli**[1]
{alramponi,m.rovera,satonelli}@fbk.eu, robert.moro@kinit.sk

[1] Fondazione Bruno Kessler, Trento, Italy
[2] Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia

Abstract

Retrieval of previously fact-checked claims is a well-established task, whose automation can assist professional fact-checkers in the initial steps of information verification. Previous works have mostly tackled the task monolingually, i.e., having both the input and the retrieved claims in the same language. However, especially for languages with a limited availability of fact-checks and in case of global narratives, such as pandemics, wars, or international politics, it is crucial to be able to retrieve claims across languages. In this work, we examine strategies to improve the multilingual *and* crosslingual performance, namely selection of negative examples (in the supervised) and re-ranking (in the unsupervised setting). We evaluate all approaches on a dataset containing posts and claims in 47 languages (283 language combinations). We observe that the best results are obtained by using LLM-based re-ranking, followed by fine-tuning with negative examples sampled using a sentence similarity-based strategy. Most importantly, we show that crosslinguality is a setup with its own unique characteristics compared to the multilingual setup.[1]

1 Introduction

Fighting online mis/disinformation is a challenging task for professional fact-checkers and human moderators alike, given that false information spreads six times faster than true information (Vosoughi et al., 2018). In the fact-checking pipeline, one of the tasks that can remarkably speed up operators' activity and support their work is *previously fact-checked claim retrieval* (PFCR), defined as follows (Pikuliak et al., 2023; Shaar et al., 2020): "Given a text making an input claim (e.g., a social media post) and a set of previously fact-checked claims, the task is to rank the fact-checked claims so that those that are the most relevant with respect to the input claim are ranked as high as possible."

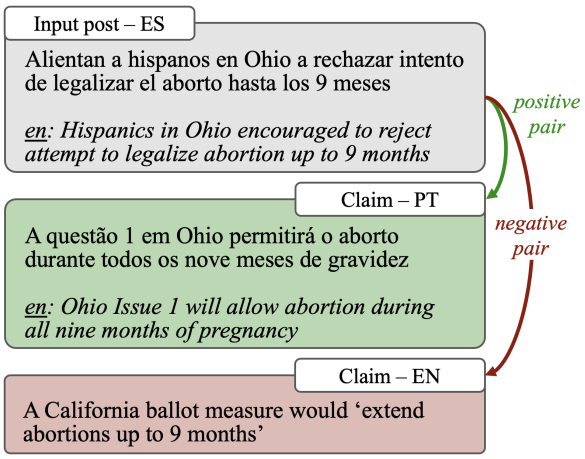

Figure 1: An example of an input post paired with a previously fact-checked claim related to the same event (*positive pair*) and a claim that is similar to the input post but unrelated (*negative pair*), which can be used to fine-tune a retriever. Redacted examples from the MultiClaim dataset (Pikuliak et al., 2023).

The task, whose goal is to avoid fact-checking the same claim again and taking advantage of existing verified knowledge, is meant to address different operators' needs. For example, a fact-checking agency may want to reuse the internal knowledge collected over the years mainly in a *monolingual* fashion, i.e., when input and retrieved claims are in the same language. Another option may concern cases in which a fact-checking agency needs to verify a global narrative (e.g., related to COVID-19) and is interested in retrieving already fact-checked claims in any language, regardless of the language of the input claim. This case requires a *multilingual* approach. Finally, fact-checkers may be aware of the fact that a narrative arising in their country was already present and debunked in other countries (or in other languages of their country), so they need to retrieve fact-checked claims in a language that

[*] These authors contributed equally to this work.
[1] Code and data are publicly available at: https://github.com/kinit-sk/multiclaim-emnlp2025

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 29046–29065
November 4-9, 2025

is different from their input claim. In this case the required approach is *crosslingual*. Figure 1 illustrates this last case, with an input post in Spanish, the associated fact-checked claim in Portuguese and a similar but unrelated claim in English, which can make it difficult to retrieve the correct one.

Most prior works such as Shaar et al. (2020) and Hardalov et al. (2022) tackled the task monolingually focusing on English. Few recent works have proposed to extend the task to multilingual setups (Pikuliak et al., 2023; Panchendrarajan et al., 2025), following the initiatives by international fact-checking agencies aimed at sharing and connecting the different databases of verified information.[2] Only two works so far addressed the task in crosslingual setups, with the input claim being always in a different language than the retrieved fact-checked claims (Pikuliak et al., 2023; Vykopal et al., 2025). Both confirm that crosslingual PFCR is a very challenging task, and show the potential of translating posts and fact-checks into English, especially for low-resource languages.

Despite the promising performance obtained through translation, however, a range of new multilingual embedding models (e.g., mE5; Wang et al., 2024b) and large language models (e.g., Qwen; Yang et al., 2024) has been released, which could greatly benefit crosslingual tasks. Furthermore, prior works have shown a positive impact of fine-tuning these models on PFCR data, specifically when using (multiple) negative examples (Pikuliak et al., 2023; Neumann et al., 2023). However, specific strategies tailored to the task and focused on crosslingual retrieval were left unexplored.

To address this gap, we compare supervised and unsupervised PFCR in multilingual and crosslingual setups, identifying the best strategies and experimental settings for the task without resorting to translation. In particular, we investigate three main aspects: **(RQ1)** how the different text embedding models for retrieval and re-ranking perform on the task, **(RQ2)** what strategies should be used to select negative examples in a supervised framework and how these compare to unsupervised approaches, and **(RQ3)** what the specifics of crosslingual setup are compared to the multilingual one. We carry out our experiments on a dataset derived from the publicly available MultiClaim dataset (Pikuliak et al., 2023). Even if we use a pre-existing dataset, we curate it specifically for multilingual and crosslingual PFCR obtaining 283 post–fact-check language combinations. In particular, we ensure proper data splits of not only posts, but also fact-checks to prevent data contamination, which was not done in prior works and thus can be considered an additional contribution of this work. We publish our subset and data splits to ensure reproducibility.

[2] https://www.poynter.org/ifcn/

2 Related Work

Previously fact-checked claim retrieval, also known as claim matching or claim detection, is a standard task in the fact-checking pipeline, both manual and automated (Panchendrarajan and Zubiaga, 2024; Vykopal et al., 2024), and it can also serve as a signal of content credibility (Srba et al., 2024). It was addressed by a series of CheckThat! Lab shared tasks organized at CLEF in 2020 (Barrón-Cedeño et al., 2020), 2021 (Shaar et al., 2021), and 2022 (Nakov et al., 2022); most recently, it was also part of a SemEval 2025 shared task (Peng et al., 2025).

Due to the relative popularity of the task, there is a range of relevant existing datasets, as summarized in recent surveys by Panchendrarajan and Zubiaga (2024) and Srba et al. (2024). These resources differ in the number of included languages, data volume, means of identification of pairs between input social media posts and fact-checked claims, and in sources of input posts. The highest number of input posts is in CrowdChecked (≈300k; Hardalov et al., 2022) and MuMiN (≈21M; Nielsen and McConville, 2022); however, they either contain a high level of noise in the identified pairs (the former) or do not explicitly provide the pairs (the latter). The most linguistically diverse datasets are MuMiN with 41 languages (Nielsen and McConville, 2022), MultiClaim with 27 languages in posts and 39 languages in fact-checked claims (Pikuliak et al., 2023), and MMTweets with 4 languages in posts and 11 languages in fact-checked claims (Singh et al., 2024). Other relevant multilingual datasets (focusing on distinct but related tasks) include the EUvsDisinfo dataset of disinformation articles matched with trustworthy articles from credible sources (Leite et al., 2024) and MultiClaimNet, a dataset that combines three existing datasets (including MultiClaim) and enriches them with identified claim clusters (Panchendrarajan et al., 2025). Despite this linguistic diversity, crosslingual retrieval remains underexplored for the task, since the identified pairs of fact-checks and input posts

are often in the same language.

Regarding the approaches employed for PFCR, the existing works most typically employ one or more text embedding models to encode input posts and claims for similarity search (Shaar et al., 2021; Pikuliak et al., 2023; Martín et al., 2022). Occasionally, a re-ranker is employed (Shaar et al., 2021), the models are fine-tuned for the task (Pikuliak et al., 2023; Kazemi et al., 2021), or both approaches are combined (Hardalov et al., 2022). Most recently, large language models (LLMs) have been also employed in zero- and few-shot settings for PFCR using a range of prompting strategies (Vykopal et al., 2025; Pisarevskaya and Zubiaga, 2025), highlighting that no single strategy proved as the best overall and also that the performance is lower for low-resource languages. Finally, Neumann et al. (2023) proposed to use multiple negative examples during fine-tuning of the embedding models, improving overall retrieval performance. In our work, we extend this approach by exploring and comparing a wider range of selection strategies and fine-tuned models.

3 Dataset

To perform our experiments, we extract a subset of the MultiClaim v2 dataset.[3] MultiClaim v2 dataset is composed of pairs of *posts* and *fact-checked claims*, which can be in different languages, provided that each post is linked to at least one claim (see Figure 1). It is constructed by extracting claims from fact-checking articles obtained through the Google Fact-Check Explorer API or via custom scrapers and links to posts from the `ClaimReview` schema,[4] provided directly by the fact-checkers in the articles. Additional pairs are created through fact-checking labels on the Meta platforms (i.e., Facebook and Instagram).

To curate a subset of data for multilingual and crosslingual PFCR, we work only with posts' text (i.e., omitting text extracted from OCRed images as it is often noisy) in their original languages. We include only languages originally represented with at least 200 posts and keep only fact-checked claims that have at least one paired post.[5]

Next, we split posts, fact-checked claims, and pairs into training, development, and test sets stratified by language by withholding 10% of the data for development and 10% for testing. We also ensure that no fact-checked claim appears in more than one split (i.e., not only posts are split, but also the search spaces differ across the splits to prevent data contamination and to have a less biased estimate of the true retrieval performance). The data distribution is shown in Table 1. For the full list of supported languages and their distribution across posts and fact-checks, see Table 4 in Appendix A.

	Multilingual	**Crosslingual**
# posts	**55,421**	**7,975**
training set	44,553	6,343
development set	5,185	782
test set	5,683	850
# fact-checks	**52,911**	**7,869**
training set	41,060	6,118
development set	5,706	850
test set	6,145	901
# pairs	**63,913**	**9,066**
training set	51,658	7,261
development set	5,880	876
test set	6,375	929

Table 1: Distribution of social media posts, fact-checked claims, and their pairs across train, development, and test splits for the multilingual and crosslingual setups.

With 47 languages in total (30 languages represented in the posts, 46 languages in the fact-checked claims) and 283 language combinations, the experiments reported in this paper are carried out, to the best of our knowledge, with the most linguistically diverse dataset for PFCR to date.

4 Methodology

In line with previous work, we cast PFCR as a ranking task. Specifically, given a post p and a set of fact-checked claims $c_1, ..., c_n \in C$ that includes the gold claim c_p for the given post, the goal is to rank the fact-checked claims so that c_p ranks as high as possible. We design unsupervised and supervised approaches, and for both of them we preliminarily compute and index, for each post and fact-checked claim in the dataset, the corresponding text embedding representation.

In the **unsupervised setting**, in the first stage, we use similarity-based dense retrieval to rank the available fact checks. In a further step, we apply re-ranking to the retrieved fact-checks in order to improve accuracy. To this end, we evaluate two re-ranking techniques: cross-encoder re-

[3] MultiClaim v2 is an updated version of the original MultiClaim dataset (Pikuliak et al., 2023) and is available at: https://doi.org/10.5281/zenodo.15413169

[4] https://www.claimreviewproject.com/

[5] This leads to a cut-off of 180 posts after additional filtering to ensure non-overlapping fact-checks in the data splits.

ranking and LLM-based re-ranking (for a comparison, see Déjean et al., 2024). While cross-encoder re-rankers (Nogueira and Cho, 2020) work by comparing query-document pairs and produce a relatedness or similarity score as output, LLM-based re-rankers (Muennighoff, 2022; Sun et al., 2023) are *instructed* to *generate* a ranking of a set of documents given a query, thus harnessing the reasoning capabilities of the underlying model.

In the **supervised setting**, both positive and negative examples are required to fine-tune text embedding models. Although positive examples can be easily obtained from the pairs of posts and fact-checks in the dataset, there is not a single and well-established way to select negative examples for training. Previous work has mainly focused on random sampling (Pikuliak et al., 2023), i.e., creating a negative example by pairing a given post with a fact-checked claim randomly picked from those not associated with the input post. Albeit straightforward, such a strategy leads to rapid saturation of the training set because negative and positive examples can be easily discriminated after a few training iterations.[6] We therefore design two approaches to mitigate training set saturation during fine-tuning through the sampling of challenging negative pairs.

We experiment by sampling topically relevant (`topic`) and semantically similar (`similarity`) negative pairs, using the `random` strategy as a baseline for comparison. An example of negative pair based on `similarity` is reported in Figure 1, with the fact-checked claim sharing terms such as 'abortion(s)' and '9 months' with the input post. This example is very challenging for the model, since it should be able to discriminate between measures on abortion proposed in California and in Ohio. On the contrary, pairs sampled randomly may be fully unrelated, and the classifier may learn to discriminate them simply because they are about different topics. Furthermore, we investigate the impact of using a varying number k of negative examples on performance across our negative sampling strategies and the random sampling baseline.

For `topic`, we compute text embeddings for both posts and fact-checked claims, cluster them through topic modeling,[7] and then create k negative pairs by selecting and associating to the post a number k of fact-checks at random from within the same cluster to which the post belongs.[8] For `similarity`, we compute the cosine similarity between each post and all the fact-checked claims and create k negative pairs by associating each post to the top-k most similar fact-checks. In creating all negative pairs, we ensure that each sampled fact-checked claim is not already associated with the post as a positive pair. To avoid costly computation during fine-tuning and ensure reproducibility, we create negative examples across strategies offline and serialize them to be used at training time. Details and hyper-parameters are in Section 5.2.

5 Experimental Setup

We experiment with unsupervised retrieval with and without re-ranking (Section 5.1) and supervised retrieval using three negative sampling strategies and a varying amount of negatives (Section 5.2). We evaluate these approaches on the test set in two main data settings: *i)* a *multilingual* setting, without distinction between monolingual and crosslingual pairs, and *ii*) a *crosslingual* setting, including only post–fact-check pairs in different languages. We rely on two widely used text embedding models of varying size to compare the two approaches, namely `multilingual-e5-large` and `paraphrase-multilingual-mpnet-base-v2`, and further compare the performance of 14 additional models in the unsupervised setting.[9] All are evaluated using *Pair Success at 10* (S@10) and *Mean Reciprocal Rank at 10* (MRR@10). The former measures, for each post, whether the paired claim appears in the top-10 retrieved results, while the latter also considers its position (i.e., rank).

5.1 Unsupervised Text Embedding Models

We select 16 among the most recent, top performing multilingual embedding models, resulting in a diverse set of architectures, pre-training techniques, number of parameters (278M–7B) and data, including the use of synthetic data. Models have been selected based on their performance for the 'Retrieval' task in the Multilingual Text

[6] This is also likely to happen if we consider the set of (filtered) claims without any paired post as negative examples.

[7] We use BERTopic (Grootendorst, 2022) for topic modeling and `multilingual-e5-large` for text embeddings.

[8] In the case there are $\leq k$ fact-checked claims in the cluster, we draw them from a cluster of uncategorized fact-checks (i.e., without assigned topic) as a fallback. This is similar in spirit to random selection.

[9] We select two widely-used models for the supervised setting due to computational constraints. `paraphrase-multilingual-mpnet-base-v2` was selected as it was the best multilingual model in Pikuliak et al. (2023).

Embedding Benchmark[10] (MTEB; Enevoldsen et al., 2025). Experiments[11] are run on both the *test* set (in order to ensure fair comparison with the supervised approach) and the *full* set of claims in the dataset. This enables us to estimate the scalability capabilities of the models. In fact, while in the test set the set of claims C consists of over 6,100 claims (Table 1), in the full set the model is required to rank over 52,000 claims for each input post. To this end, we test traditional models like `paraphrase-multilingual-mpnet-base-v2` and `multilingual-MiniLM-L12-v2` (Reimers and Gurevych, 2019), encoder-based models like `bge-m3` (Chen et al., 2024), `jina-embeddings-v3` (Sturua et al., 2024), `multilingual-e5-large` (Wang et al., 2024b), `snowflake-arctic-embed-l-v2.0` (Yu et al., 2024), and `gte-multilingual-base` (Zhang et al., 2024), as well as LLM-based models like `bge-multilingual-gemma2` (Chen et al., 2024), `E5-mistral-7b` (Wang et al., 2024a), `Linq-Embed-Mistral` (Choi et al., 2024), `KaLM-embedding-multilingual-mini-v1` (Hu et al., 2025), and `gte-Qwen2-7B-instruct` (Li et al., 2023b). Where available, we include instructed versions. While prioritizing open source and research models, we also test OpenAI's `text-embedding-3-large` (OpenAI, 2024) as a benchmark. For re-rankers, we use `bge-reranker-v2-m3`[12] (Li et al., 2023a) (cross-encoder) and `RankGPT` (Sun et al., 2023) (LLM-based). We test the RankGPT re-ranker on the three best performing embedding models in retrieval (on the test set).[13] The only hyper-parameter for re-ranking, i.e., the top-n claims to re-rank, has been explored in a preliminary phase, experimenting with $n \in [20, 30, 50, 100]$. We found 20 and 30 to be the best values. To not introduce additional burden to our already composite experimental setup, throughout the paper we report re-ranking results with top-$n = 30$.

[10] https://huggingface.co/spaces/mteb/leaderboard (last visited: 05/13/2025).

[11] All experiments have been conducted on a single Nvidia A40 GPU, equipped with 48 GB RAM.

[12] https://huggingface.co/BAAI/bge-reranker-v2-m3

[13] This was motivated by computational and funding reasons, as each test set experiment costed approximately $50. The three models have been chosen because of their high performance in pure retrieval and to enable comparison with the supervised approach (`multilingual-e5-large`).

5.2 Supervised Text Embedding Models

We employ `multilingual-e5-large` and `paraphrase-multilingual-mpnet-base-v2` as our pretrained text embedding models to fine-tune. We select them since they are widely used models, represent varied parameter sizes (278M–560M), and can be fine-tuned on an average GPU[14] without the need for a costly computing infrastructure. We use multiple negatives ranking loss and perform hyper-parameter tuning for selecting the best value for the learning rate (among 1e-9, 5e-9, 1e-8, 5e-8, 1e-7), the batch size (4, 8, 16), and warm-up steps (800, 1,600) based on S@10 multilingual performance on the development set for both models using all negative sampling approaches. We select 1e-8 as the learning rate, 8 as batch size, and 1,600 as warm-up steps. The full list of hyper-parameters is reported in Appendix C.

To select the model configurations to be used for test set evaluation, we further investigate which number k of negative examples provides the best overall performance across models and the three negative sampling strategies. For stability, we run each model configuration three times, each with a different seed, and report average results. We experiment with $k \in [1, 2, 3, 4, 5, 10]$ and find that using $k = 10$ negative examples provides the best overall performance for all strategies on the development set. As shown in Figure 2 for the fine-tuned `multilingual-e5-large` model, `similarity` consistently outperforms both the `topic` and `random` strategies according to the S@10 multilingual score, with the best results obtained when using 10 negatives. These results are consistent with those obtained when fine-tuning `paraphrase-multilingual-mpnet-base-v2` (cf. Figure 15 in Appendix C) and motivate our selection of $k = 10$ for test set evaluation.

6 Results and Discussion

In this section, we present the results on the test set of unsupervised (Section 6.1) and supervised (Section 6.2) approaches, along with a detailed discussion. We then compare the approaches and discuss limitations and future directions (Section 6.3).

6.1 Unsupervised Results

Retrieval In the *multilingual* setting (Figure 3 (a), purple bars) `text-embedding-3-large` leads the ranking, with `bge-multilingual-gemma2`

[14] We rely on a single Tesla V100-SXM2-32GB GPU.

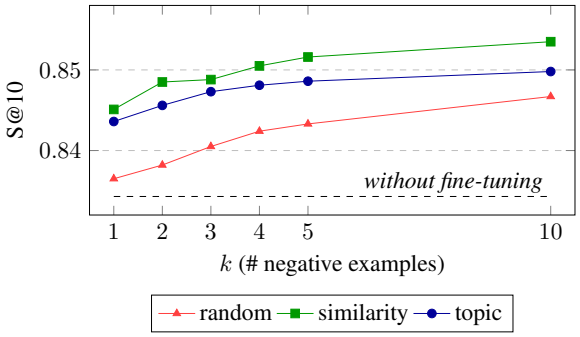

Figure 2: **Multilingual** S@10 performance across negative sampling strategies and number of negative examples k for the fine-tuned `multilingual-e5-large` model on the **development** set. Reported results are averaged over three runs using different seeds. The dashed line indicates results when no fine-tuning is conducted.

as the best open source model and `multilingual-e5-large` as the best lightweight model. This trend is consistent across the test and the full data settings, as well as across both metrics (for all details on multilingual evaluation, see Table 5 and Figure 5 in Appendix B). The *crosslingual* setting (Figure 3 (b), purple bars) which proved more challenging, provides a different picture, with `bge-multilingual-gemma2` scoring the best result on the test set and `gte-multilingual-base` ranking first on the full set of claims. The two models outperform `text-embedding-3-large` by a margin of 3.27 and 3.58 MRR@10 points respectively. Also in this case, results are consistent across the two adopted metrics (for all details on crosslingual evaluation, see Table 6 and Figure 6 in Appendix B). In order to better understand models' performance, we further filtered the results by removing pairs containing a post or a claim in English, which is the dominant language in the dataset. In this setting, open-source models like `snowflake-arctic-embed-l-v2.0` and `bge-multilingual-gemma2` significantly outperform `text-embedding-3-large`. Results are reported in Figure 3 (c), purple bars.

Re-ranking For cross-encoder re-ranking, results show different effects across models and linguistic settings. In the *multilingual* setting (see Figure 3 (a), orange bars), the effect of re-ranking is nuanced: in terms of MRR@10, and compared to pure retrieval, cross-encoder re-ranking does not yield *better* rankings, except for low-performing models (for the complete results on multilingual re-ranking, please refer to Table 5 and Figures 7, 8, 9 and 10 in Appendix B). This trend is also reflected by the S@10 metric, which shows that re-ranking boosts the performance of weaker and moderately increases performance in mid-performing models, but it reduces the performance of high-performing models. In fact, in the multilingual setup, cross-encoder re-ranking has an average MRR@10 gain of 0.60 and 0.02 points on the test and the full set respectively, while S@10 shows an average increase of 1.37 and 1.77 points, respectively. The effects of re-ranking, by contrast, emerge much more clearly in the *crosslingual* setting (Figure 3 (b), orange and cyan bars). In fact, in this case cross-encoder re-ranking proves effective in boosting performance across all models, both in terms of MRR@10 and S@10, thus yielding better quality rankings. Indeed, the average gain is 8.11 and 7.04 points for MRR@10 (see details in Tables 12 and 14 in Appendix B) and 5.32 and 4.93 points for S@10 (Tables 11 and 13 in Appendix B). A comparable performance increase can be observed in the crosslingual setup when English data are excluded (see Figure 3 (c), orange bars), with gains of 8.08 (test) and 5.91 (full) MRR@10 points. LLM-based re-ranking, on the other hand, demonstrates superior performance in multilingual, crosslingual, and crosslingual without English setups. It results in an average increase of 4.54, 13.80, and 6.93 MRR@10 points respectively, besides yielding the best overall results (Figure 3, cyan bars).

To better understand the results, we also conducted a correlation analysis between *a)* the embedding dimension and *b)* the number of parameters of each embedding model on the one hand, and the observed performance on the other. The analysis, conducted using Pearson's r coefficient, shows that there is no correlation between these variables and the observed performance (Table 8 in Appendix B).

Overall, our unsupervised experiments highlight the following relevant trends: *a)* with the exception of `bge-multilingual-gemma2`, which shows high performance in both contexts, models that perform best in the multilingual setting do not always perform equally well in the crosslingual setting, and vice versa. This, once again, reflects the uniqueness of the crosslingual context. Moreover, *b)* smaller, encoder-only embedding models (like `multilingual-e5-large` and `gte-multilingual-base`) often challenge or even outperform larger decoder-only models (like `bge-multilingual-gemma2` or `text-embedding-3-large`) in the task of claim

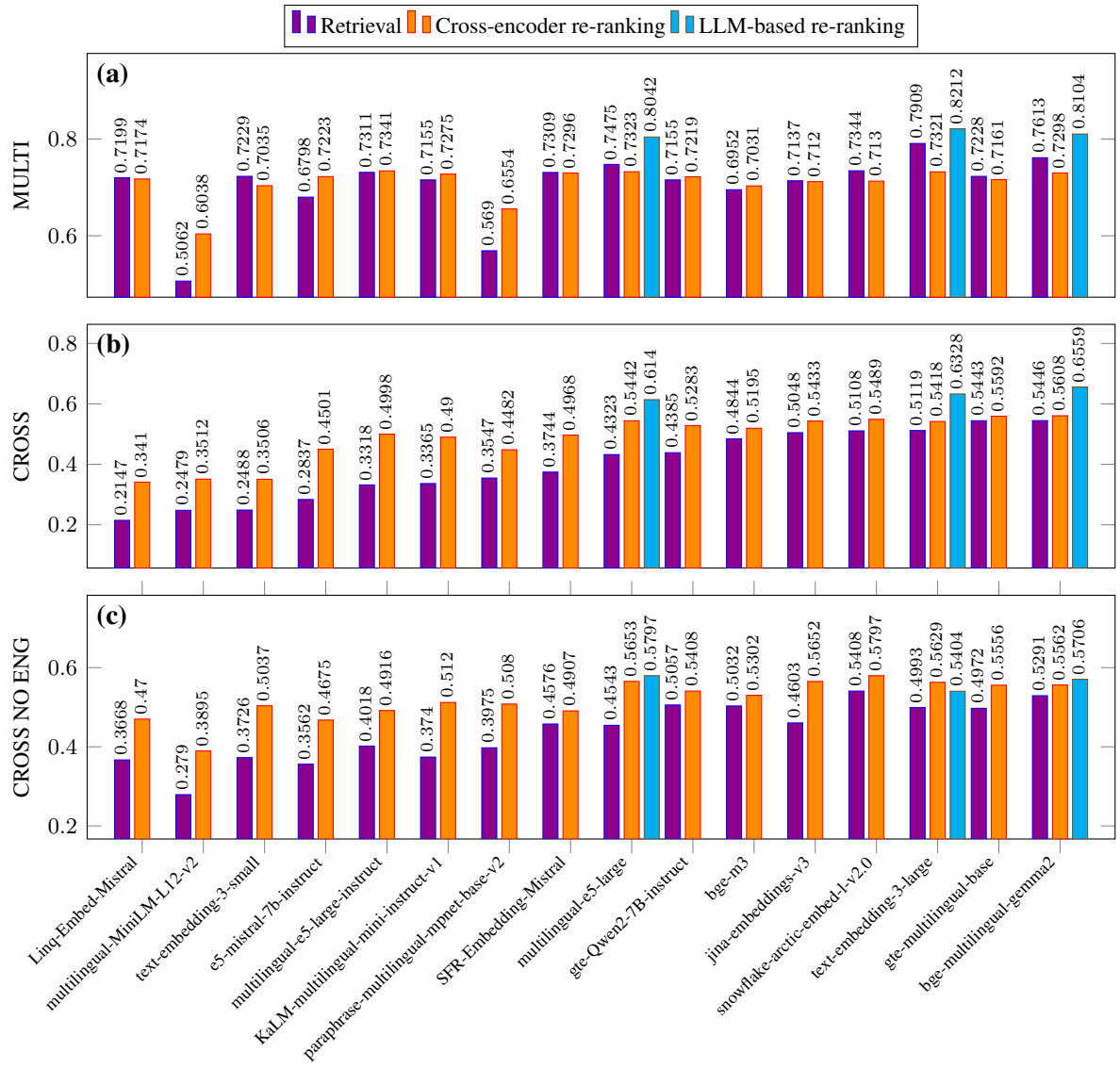

Figure 3: **Test** set performance (MRR@10) for retrieval and re-ranking across models in **multilingual** (a), **crosslingual** (b) and **crosslingual no eng** (i.e., without English) (c) settings. Results refer to 6,375 pairs (a), 929 pairs (b), and 118 pairs (c), and are sorted by crosslingual *retrieval* performance.

retrieval, in particular in the crosslingual setup. When combined with the above-mentioned correlation analysis, this indicates that model size, embedding dimensionality, and model architecture do not impact significantly performance, suggesting that other factors, such as language coverage, data variety, and the used pre-training method may have greater influence on results. Finally, *c)* re-ranking proves effective in many cases, although its contribution is more evident in the crosslingual setup. Albeit tested on a limited number of models, LLM-based re-ranking yields major performance improvements with respect to cross-encoder re-ranking, but at a higher computational cost.

6.2 Supervised Results

For supervised experiments, we observe that fine-tuning the models using `similarity` as a negative sampling strategy consistently improves the MRR@10 performance over the `topic` approach as well as the `random` selection baseline in both *multilingual* and *crosslingual* settings (Table 2). This is further confirmed by S@10 scores (Table 9 in Appendix D). Among the two models, `multilingual-e5-large` provides the best overall results across all strategies, showing an improvement of 1.45 and 5.31 MRR@10 points and 0.94 and 4.96 S@10 points over `random` when using the `similarity` strategy in multilingual and crosslingual setups, respectively. Although also

the topic strategy outperforms the commonly employed random selection baseline (Pikuliak et al., 2023), it still lags behind the similarity approach (by 0.41 and 0.54 MRR@10 points and 0.25 and 0.38 S@10 points in multilingual and crosslingual setups). The similarity sampling strategy therefore appears to be a viable approach for selecting hard negative examples for fine-tuning due to stable performance improvements over the other methods. Furthermore, it does not rely on the costly computation of topic clusters of the topic strategy to draw negative examples from (Section 4).

Looking at the crosslingual MRR@10 performance across strategies and the number of negative examples (see Figure 4), we observe a large performance gain between the proposed strategies and the random baseline and the unsupervised setting over all k values (results using S@10 show the same trend, see Figure 16 in Appendix C). The topic strategy seems more effective when few negative examples are used for fine-tuning (i.e., ≤ 5), whereas similarity confirms to be the most effective approach in the PFCR task when sampling more negative examples for each pair. Overall, compared to using multilingual-e5-large in an unsupervised fashion, fine-tuning it with just 10 negative examples selected using the similarity strategy leads to a crosslingual MRR@10 score of 0.4947 (+6.24 points) and a crosslingual S@10 score of 0.7076 (+7.52 points). This indicates that our approach is effective even in the more challenging crosslingual setup. When excluding the highly-represented English language in the data from crosslingual evaluation (i.e., "crosslingual (no eng)" in Table 2), we observe the same trend, with topic and similarity achieving large performance gains compared to random.

6.3 Comparing Unsupervised and Supervised Approaches

Overall, by comparing the *test* results for the unsupervised and supervised approaches obtained by the best shared model (i.e., multilingual-e5-large, see Table 2 and Table 9 in Appendix D), we observe that in both the multilingual and crosslingual setups the best results in terms of MRR@10 score are obtained by using LLM-based re-ranking (0.8042 and 0.6140, respectively). This is confirmed also by S@10 performance, with 0.8324 and 0.7283 S@10 for the multilingual and crosslingual setups, respectively. The second best approach is

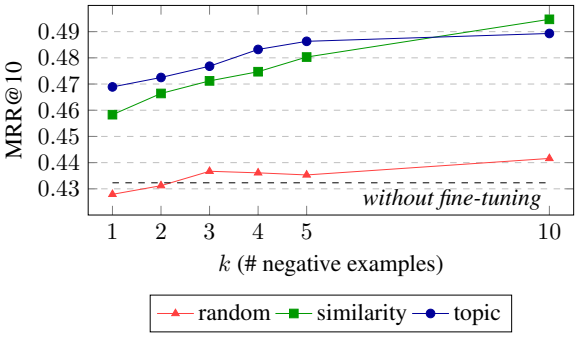

Figure 4: **Crosslingual** MRR@10 performance across negative sampling strategies and number of negative examples k for the fine-tuned multilingual-e5-large model on the **test** set. Reported results are averages over three runs using different seeds. The dashed line indicates results when no fine-tuning is conducted.

cross-encoder re-ranking, which proves effective in producing better rankings than supervised approaches, especially in the crosslingual setup (0.5442 MRR@10). However, we observe that fine-tuning with negative examples sampled using the similarity strategy leads to better MRR@10 and S@10 performance than cross-encoder re-ranking in the multilingual setup (0.7768 MRR@10 and 0.8228 S@10, respectively). All other approaches follow these two, with plain unsupervised retrieval showing the worst (or the second worst) results in both scenarios, namely obtaining MRR@10 scores of 0.7475 in the multilingual setup and 0.4323 in the crosslingual setup, and S@10 scores of 0.7971 in the multilingual setup and 0.6324 in the crosslingual setup. As regards paraphrase-multilingual-mpnet-base-v2, results confirm that fine-tuning with similarity-sampled negatives is the best strategy in the supervised scenario (0.5537 and 0.3500 MRR@10 and 0.6320 and 0.5157 S@10 in multilingual and crosslingual settings, respectively), but retrieval in an unsupervised fashion provides better performance (0.5690 and 0.3547 MRR@10 and 0.6416 and 0.5188 S@10), ranking after cross-encoder re-ranking (0.6485 and 0.4482 MRR@10 and 0.6796 and 0.5779 S@10). We speculate that this could be due to the small parameter size of this model, which could limit the learning of nuanced patterns from negative and positive pairs.

Moreover, in both unsupervised and supervised approaches, we observe a notable difference in the contribution of re-ranking and negative sampling, with a much more significant benefit in the crosslingual context. This shared pattern highlights the

Model	Strategy	Multilingual	Crosslingual	Crosslingual (no eng)
multilingual-e5-large	re-rank	0.7323	0.5442	0.5653
	retrieve	0.7475	0.4323	0.4543
	random	$0.7623_{\pm 0.0002}$	$0.4416_{\pm 0.0007}$	$0.4810_{\pm 0.0017}$
	topic	$0.7727_{\pm 0.0001}$	$0.4893_{\pm 0.0004}$	$0.5349_{\pm 0.0000}$
	similarity	$0.7768_{\pm 0.0001}$	$0.4947_{\pm 0.0001}$	$0.5269_{\pm 0.0004}$
	llm re-rank	**0.8042**	**0.6140**	**0.5797**
paraphrase-multilingual-mpnet-base-v2	random	$0.5202_{\pm 0.0003}$	$0.3177_{\pm 0.0001}$	$0.3518_{\pm 0.0000}$
	topic	$0.5412_{\pm 0.0002}$	$0.3367_{\pm 0.0001}$	$0.3729_{\pm 0.0004}$
	similarity	$0.5537_{\pm 0.0002}$	$0.3500_{\pm 0.0003}$	$0.3961_{\pm 0.0002}$
	retrieve	0.5690	0.3547	0.3975
	re-rank	**0.6485**	**0.4482**	**0.5080**

Table 2: **Multilingual**, **crosslingual**, and **crosslingual (no eng)** MRR@10 performance across negative sampling strategies (`random`, `topic`, `similarity`) for fine-tuned models on the **test** set compared to unsupervised strategies (`retrieve`, cross-encoder `re-rank`, `llm re-rank`). Results are ordered by increasing multilingual score; for the supervised setup, we report averages with standard deviation over three runs using different seeds.

Language	Strategy	Monolingual	Crosslingual
English	retrieve	0.7173	0.4024
	random	$0.7265_{\pm 0.0003}$	$0.4627_{\pm 0.0015}$
	topic	$0.7325_{\pm 0.0002}$	$0.5024_{\pm 0.0005}$
	similarity	$0.7378_{\pm 0.0006}$	$0.5159_{\pm 0.0005}$
	re-rank	0.6958	0.5503
	llm re-rank	**0.7661**	**0.5919**
Hindi	random	$0.8214_{\pm 0.0010}$	$0.4531_{\pm 0.0010}$
	retrieve	0.8005	0.4817
	topic	$0.8328_{\pm 0.0001}$	$0.4974_{\pm 0.0001}$
	similarity	$0.8423_{\pm 0.0006}$	$0.5060_{\pm 0.0001}$
	re-rank	0.8036	0.5817
	llm re-rank	**0.8507**	**0.6335**

Table 3: **Monolingual** and **crosslingual** MRR@10 performance across negative sampling strategies (`random`, `topic`, `similarity`) for fine-tuned models on the **test** set compared to unsupervised strategies (`retrieve`, cross-encoder `re-rank`, `llm re-rank`) for the most represented languages in terms of post–fact-check pairs (i.e., English and Hindi) using the `multilingual-e5-large` model. Results are ordered by increasing crosslingual score; for the supervised setup, we report averages with standard deviation over three runs using different seeds.

specificity of the crosslingual context and will be the subject of further investigation.

Overall, our experiments show that fine-tuning embeddings models with negative sampling leads to significant performance improvements, more evident for the crosslingual setup, but requires enough training data for sampling and implies an additional computational effort in the fine-tuning step. Unsupervised PFCR performance, instead, is more dependent on the re-ranking method: while cross-encoder re-ranking underperforms the supervised approach, LLM-based re-ranking yields the best overall performance, but comes at a high computational cost. On the other hand, the unsupervised approach does not require training data and scales well on larger amounts of data.

Additionally, we investigate the monolingual performance for English and Hindi, the two most represented post–fact-check language pairs in the dataset (Table 3). Also in this case, empirical observations confirm the already observed trends, namely that supervised approaches (in particular with `similarity`-based negative sampling) produce better results in a mono- or multilingual setup, whereas unsupervised, re-ranking based strategies prove more effective in the crosslingual setup.

7 Conclusion

We carried out an extensive evaluation of unsupervised and supervised PFCR focusing on multilingual and crosslingual settings. We showed that results in the two settings are remarkably different, with crosslingual retrieval being much more challenging. Unsupervised learning with LLM-based re-ranking yields the best results, even outperforming the best supervised approach. Overall, our study highlights the importance of a thorough evaluation of embedding models and the impact of re-ranking and negative sampling on retrieval performance. We believe that this kind of work could guide the development of PFCR systems tailored to fact-checkers' needs (multilingual *vs* crosslingual) and to the available computational resources.

Limitations

Despite the extensive set of experiments and comparisons, this work still has some limitations. A general issue affecting datasets like MultiClaim, which are created by merging different fact-

checkers' databases, is the possibility that some pairs of posts and fact-checked claims have not been annotated as positive pairs, even if they should. This is mainly due to the fact that different fact-checking agencies tend to better curate their (monolingual) database, while an extensive effort to annotate positive pairs also crosslingually is often lacking. This may affect the performance of the models and the set of sampled negatives.

Another possible issue is about the different representation of languages in the dataset. As shown in Table 4, English posts are paired with fact-checked claims in almost all languages in our dataset, while some others, such as Dutch or Romanian, have only few crosslingual pairs. Our results and findings could change with a different distribution of languages (and cross-lingual pairs) in the data.

Finally, as detailed in Appendix A, we rely on automatic means to detect languages in posts and fact-checked claims. Although we combine several approaches to increase detection robustness and have also corrected outlier cases (e.g., when Latin was recognized as a language for posts and fact-checks), there might still be some noise in the assigned languages in the data.

Ethics Statement

In this paper, we work with the MultiClaim dataset that has been published for research purposes only. We reuse it in line with its terms and conditions; e.g., we do not re-share the subset of data we used, but publish only the IDs together with information about additional metadata (i.e., identified languages), our defined data splits, a list of identified negative examples, and code to load the dataset and run the models and their fine-tuning.

As a part of the paper, we fine-tune text embedding models for the PFCR task. They are intended to be used as an assistance to human fact-checkers or moderators and not to be used in an automated way to fact-check input claims, i.e., to ascertain their veracity based on the retrieved claims.

Acknowledgments

This work was partially funded by the European Media and Information Fund (grant number 291191). The sole responsibility for any content supported by the European Media and Information Fund lies with the author(s) and it may not necessarily reflect the positions of the EMIF and the Fund Partners, the Calouste Gulbenkian Foundation and the European University Institute. It was also partially supported by the European Union under the Horizon Europe project AI-CODE, GA No. 101135437, the Slovak Research and Development Agency under the project Modermed, GA No. APVV-22-0414, and the PNRR project FAIR – Future AI Research (PE00000013), under the NRRP MUR program funded by NextGeneration EU.

The authors also wish to acknowledge the TAILOR project funded by the European Union under the EU Horizon 2020, GA No. 952215, which supported the research mobility that started the collaboration on this paper under the TAILOR Connectivity fund.

References

Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, Shaden Shaar, and Zien Sheikh Ali. 2020. Overview of CheckThat! 2020: Automatic identification and verification of claims in social media. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020*, volume 12260 of *Lecture Notes in Computer Science*, pages 215–236. Springer.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Chanyeol Choi, Junseong Kim, Seolhwa Lee, Jihoon Kwon, Sangmo Gu, Yejin Kim, Minkyung Cho, and Jy-Yong Sohn. 2024. Linq-Embed-Mistral technical report. *arXiv preprint arXiv:2412.03223*.

Hervé Déjean, Stéphane Clinchant, and Thibault Formal. 2024. A thorough comparison of cross-encoders and LLMs for reranking SPLADE. *arXiv preprint arXiv:2403.10407*.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömerand Çagatan, et al. 2025. MMTEB: Massive multilingual text embedding benchmark. In *The Thirteenth International Conference on Learning Representations*.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Momchil Hardalov, Anton Chernyavskiy, Ivan Koychev, Dmitry Ilvovsky, and Preslav Nakov. 2022. CrowdChecked: Detecting previously fact-checked claims in social media. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 266–285, Online only. Association for Computational Linguistics.

Xinshuo Hu, Zifei Shan, Xinping Zhao, Zetian Sun, Zhenyu Liu, Dongfang Li, Shaolin Ye, Xinyuan Wei, Qian Chen, Baotian Hu, Haofen Wang, Jun Yu, and Min Zhang. 2025. KaLM-Embedding: Superior training data brings a stronger embedding model. *arXiv preprint arXiv:2501.01028*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Ashkan Kazemi, Kiran Garimella, Devin Gaffney, and Scott Hale. 2021. Claim matching beyond English to scale global fact-checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4504–4517, Online. Association for Computational Linguistics.

João A. Leite, Olesya Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. 2024. EUvsDisinfo: A dataset for multilingual detection of prokremlin disinformation in news articles. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, pages 5380–5384, New York, NY, USA. Association for Computing Machinery.

Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. 2023a. Making large language models a better foundation for dense retrieval. *arXiv preprint arXiv:2312.15503*.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023b. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Alejandro Martín, Javier Huertas-Tato, Álvaro Huertas-García, Guillermo Villar-Rodríguez, and David Camacho. 2022. FacTeR-Check: Semi-automated fact-checking through semantic similarity and natural language inference. *Knowledge-Based Systems*, 251:109265.

Niklas Muennighoff. 2022. SGPT: GPT sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*.

Preslav Nakov, Giovanni Da San Martino, Firoj Alam, Shaden Shaar, Hamdy Mubarak, and Nikolay Babulkov. 2022. Overview of the CLEF-2022 CheckThat! lab task 2 on detecting previously fact-checked claims. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, Bologna, Italy. CEUR-WS.org.

Anna Neumann, Dorothea Kolossa, and Robert M Nickel. 2023. Deep learning-based claim matching with multiple negatives training. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 134–139, Online. Association for Computational Linguistics.

Dan S. Nielsen and Ryan McConville. 2022. MuMiN: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3141–3153, New York, NY, USA. Association for Computing Machinery.

Rodrigo Nogueira and Kyunghyun Cho. 2020. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

OpenAI. 2024. text-embedding-3-large. `https://platform.openai.com/docs/guides/embeddings`. Accessed: 2025/01/05.

Rrubaa Panchendrarajan, Rubén Míguez, and Arkaitz Zubiaga. 2025. MultiClaimNet: A massively multilingual dataset of fact-checked claim clusters. *arXiv preprint arXiv:2503.22280*.

Rrubaa Panchendrarajan and Arkaitz Zubiaga. 2024. Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *Natural Language Processing Journal*, 7:100066.

Qiwei Peng, Robert Moro, Michal Gregor, Ivan Srba, Simon Ostermann, Marian Simko, Juraj Podrouzek, Matúš Mesarčík, Jaroslav Kopčan, and Anders Søgaard. 2025. SemEval-2025 task 7: Multilingual and crosslingual fact-checked claim retrieval. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2498–2511, Vienna, Austria. Association for Computational Linguistics.

Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smoleň, Martin Melišek, Ivan Vykopal, Jakub Simko, Juraj Podroužek, and Maria Bielikova. 2023. Multilingual previously fact-checked claim retrieval. In *Proceedings of the 2023*

Conference on Empirical Methods in Natural Language Processing, pages 16477–16500, Singapore. Association for Computational Linguistics.

Dina Pisarevskaya and Arkaitz Zubiaga. 2025. Zero-shot and few-shot learning with instruction-following LLMs for claim matching in automated fact-checking. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9721–9736, Abu Dhabi, UAE. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online. Association for Computational Linguistics.

Shaden Shaar, Fatima Haouari, Watheq Mansour, Maram Hasanain, Nikolay Babulkov, Firoj Alam, and Preslav Nakov. 2021. Overview of the CLEF-2021 CheckThat! lab task 2 on detecting previously fact-checked claims in tweets and political debates. In *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum*, Bucharest, Romania. CEUR-WS.org.

Iknoor Singh, Carolina Scarton, Xingyi Song, and Kalina Bontcheva. 2024. Breaking language barriers with MMTweets: Advancing cross-lingual debunked narrative retrieval for fact-checking. *arXiv preprint arXiv:2308.05680*.

Ivan Srba, Olesya Razuvayevskaya, João A. Leite, Robert Moro, Ipek Baris Schlicht, Sara Tonelli, Francisco Moreno García, Santiago Barrio Lottmann, Denis Teyssou, Valentin Porcellini, Carolina Scarton, Kalina Bontcheva, and Maria Bielikova. 2024. A survey on automatic credibility assessment of textual credibility signals in the era of large language models. *arXiv preprint arXiv:2410.21360*.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task LoRA. *arXiv preprint arXiv:2409.10173*.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. 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*, pages 14918–14937, Singapore. Association for Computational Linguistics.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Ivan Vykopal, Matúš Pikuliak, Simon Ostermann, Tatiana Anikina, Michal Gregor, and Marián Šimko. 2025. Large language models for multilingual previously fact-checked claim detection. *arXiv preprint arXiv:2503.02737*.

Ivan Vykopal, Matúš Pikuliak, Simon Ostermann, and Marián Šimko. 2024. Generative large language models in automated fact-checking: A survey. *arXiv preprint arXiv:2407.02351*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual E5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Puxuan Yu, Luke Merrick, Gaurav Nuti, and Daniel Campos. 2024. Arctic-Embed 2.0: Multilingual retrieval without compromise. *arXiv preprint arXiv:2412.04506*.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.

Appendix

A Dataset

Pre-processing The MultiClaim v2 dataset consists of posts, fact-checked claims, and their pairs. We preprocess the dataset to curate a subset of data for multilingual and crosslingual PFCR. More specifically, we work with posts' (anonymized) texts (`post_body` field) and fact-checked claims (`claim` field), both in their original languages. We thus omit texts extracted with OCR from the images associated with some of the posts because they are too noisy, as well as the titles of the fact-checks, as these are often missing or duplicate the information already present in the claims. We work with pairs, whose `relationship` is identified either as `claim_review` or `backlink`, since only these relationships were present in the original version of MultiClaim and they represent the ground truth mappings provided by the fact-checkers.

MultiClaim v2 uses Google Translate to identify the languages of posts and fact-checks. However, this being a third party black-box API, we implement our own pipeline to identify languages using open-source tools to have more control and increase robustness of the predictions. Specifically, we use a combination of four language detectors: *i)* `fastText`,[15] which supports 176 languages (Joulin et al., 2016, 2017), *ii)* `gCLD3` (Compact Language Detector v3),[16] which supports more than 100 languages, *iii)* `langdetect`,[17] which supports 55 languages, and *iv)* `polyglot`,[18] which supports 196 languages. We combine the outputs of these detectors as follows: we first filter out detected languages that appear only once. Then, we average the normalized detection scores for the remaining ones and filter out those whose average score is <0.5. Finally, we take the language with the highest average score as the post/claim detected language.

We also filter out posts whose languages did not appear in at least 200 posts (leading to a cut-off of 180 posts after additional filtering to ensure non-overlapping fact-checks in the data splits; see Section 3). Yet, we took all fact-checked claims associated with these remaining posts irrespective of their language – as a result, there are more languages in claims than in posts. We manually checked the languages of claims that appeared <10 times (e.g., Esperanto, Latin, Welsh, Corsican). Since these were all misclassifications, we manually corrected the language identified in these cases.

Covered languages Our subset of the dataset covers 47 languages in total: 30 languages in posts and 46 languages in fact-checked claims. All posts' languages appear in claims except for Urdu, which is represented only in posts. We obtain 283 language combinations in total (see Figure 4). The full list of covered languages is the following:

Afrikaans (`af`), Arabic (`ar`), Assamese (`as`), Azerbaijani (`az`), Bulgarian (`bg`), Bengali (`bn`), Bosnian (`bs`), Catalan (`ca`), Czech (`cs`), Danish (`da`), German (`de`), Modern Greek (`el`), English (`en`), Spanish (`es`), Persian (`fa`), Finnish (`fi`), French (`fr`), Hindi (`hi`), Croatian (`hr`), Hungarian (`hu`), Indonesian (`id`), Italian (`it`), Kazakh (`kk`), Korean (`ko`), Macedonian (`mk`), Malayalam (`ml`), Malay (`ms`), Burmese (`my`), Nepali (`ne`), Dutch (`nl`), Norwegian (`no`), Punjabi (`pa`), Polish (`pl`), Portuguese (`pt`), Romanian (`ro`), Russian (`ru`), Sinhala (`si`), Slovak (`sk`), Slovenian (`sl`), Serbian (`sr`), Telugu (`te`), Thai (`th`), Tagalog (`tl`), Turkish (`tr`), Ukranian (`uk`), Urdu (`ur`), and Chinese (`zh`).

B Unsupervised Approach

In this section, we report the complete, per-model experimental results for the unsupervised setup. Tables 5 and 6 report the results for the baseline retrieval and cross-encoder re-ranking, in both the *test* set and the *full* set of claims. Figures 5 and 6 report the same S@10 results in graphical format, also integrating LLM-based re-ranking for the three models involved. Figures 7 to 14 instead focus on the S@10 and MRR@10 difference in performance (delta) between base retrieval and cross-encoder re-ranking on the *test* and *full* set, respectively. Moreover, Table 8 reports the full results for the correlation analysis embedding dimension/model dimension *vs.* model performance.

C Supervised Approach

We report hyper-parameter values in Table 7. The search space was: learning rate: {1e-9, 5e-9, 1e-8, 5e-8, 1e-7}, batch size: {4, 8, 16}, warm-up steps: {800, 1,600}, # negatives (k): {1, 2, 3, 4, 5, 10}. Further details are provided in Section 5.2.

In Figure 15, we present multilingual S@10 results across negative sampling strategies and number of negative examples for `mpnet` on the *devel-*

[15] https://fasttext.cc/docs/en/language-identification.html

[16] https://github.com/google/cld3

[17] https://pypi.org/project/langdetect/

[18] https://polyglot.readthedocs.io/en/latest/

Post → Claim ↓	ar	bg	bn	cs	de	el	en	es	fr	hi	hr	hu	id	it	ko	mk	ml	my	nl	pl	pt	ro	ru	si	sk	th	tl	tr	ur	zh
af	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	3	0	0	0	0	0	0
ar	3,360	0	0	0	0	0	14	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
as	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
az	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	7	0	0
bg	0	189	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0
bn	3	0	786	0	0	0	151	0	2	125	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0
bs	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ca	0	0	0	0	0	0	5	75	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cs	1	0	0	246	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	11	0	0	0	0	0
da	2	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0
de	1	2	0	1	2,381	1	94	5	3	0	2	0	1	3	0	0	1	0	0	1	1	1	20	0	0	0	0	1	1	0
el	3	3	0	2	0	430	4	4	1	0	0	0	0	0	0	2	0	0	0	1	1	3	2	0	0	0	0	0	0	0
en	88	7	233	9	32	12	19,341	118	78	3,673	3	9	220	23	138	6	76	157	19	35	23	9	74	366	4	198	310	7	451	463
es	6	2	0	0	3	0	69	7,470	3	0	0	0	0	7	0	1	0	0	0	5	33	3	2	0	1	0	1	2	0	0
fa	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
fi	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
fr	20	0	0	0	1	0	66	6	2,699	0	0	0	0	5	0	0	0	0	3	2	6	0	3	0	0	0	0	0	4	0
hi	3	0	10	0	0	2	471	3	0	2,470	0	0	1	0	0	0	4	0	0	1	1	0	0	0	0	0	0	0	6	0
hr	4	0	0	1	1	0	5	0	0	0	301	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0
hu	0	1	0	0	0	0	2	0	0	0	0	178	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
id	1	0	1	0	1	0	23	2	1	2	0	0	939	0	1	0	0	0	0	0	2	0	0	1	0	0	0	0	0	1
it	2	2	0	0	5	0	81	7	4	1	0	1	0	1,154	0	0	0	0	2	4	0	3	9	0	0	1	0	0	0	0
kk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ko	0	0	0	0	0	0	4	0	0	0	0	0	0	0	410	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mk	0	1	0	0	0	0	1	0	0	0	3	0	0	0	0	1,137	0	0	0	0	0	0	29	0	0	0	0	0	0	0
ml	0	0	0	0	0	0	60	0	0	11	0	0	0	0	0	0	482	0	0	0	0	0	0	0	0	4	0	0	1	0
ms	0	0	0	0	0	0	6	0	0	0	0	0	94	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5
my	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	191	0	0	0	0	0	0	0	0	0	0	0	0
ne	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
nl	0	0	0	0	0	0	29	1	0	0	0	0	0	0	0	1	0	0	923	0	0	1	0	0	0	0	0	0	0	0
no	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
pa	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
pl	2	1	0	1	1	0	10	7	1	0	0	0	0	3	0	0	0	0	0	1,309	1	0	0	0	0	0	0	0	2	0
pt	4	1	0	0	1	0	49	18	5	0	0	0	0	7	0	0	0	0	0	0	3,628	3	2	0	1	0	0	0	1	0
ro	0	0	0	0	0	0	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	288	0	0	0	0	0	0	0	0
ru	0	4	0	1	2	0	18	3	3	0	0	0	0	1	0	3	0	0	0	1	0	0	454	0	0	0	0	0	0	0
si	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	67	0	0	0	0	0	0
sk	0	1	0	11	0	0	9	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	321	0	0	0	0	0
sl	0	0	0	0	0	0	1	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sr	0	0	0	0	0	0	0	1	0	0	31	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
te	3	0	1	0	0	0	14	1	0	21	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
th	0	0	0	0	2	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	449	0	0	0	2
tl	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0
tr	3	0	0	1	2	0	91	4	3	0	0	0	0	1	0	0	0	0	0	0	2	0	1	0	0	0	0	3,179	0	0
uk	0	0	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0
zh	0	0	0	0	0	0	23	1	0	1	0	0	0	0	0	0	0	0	1	2	0	0	1	0	0	0	0	0	0	55

Table 4: Combinations of the languages in pairs of posts (columns) and fact-checked claims (rows) across all data splits. Languages are represented via their ISO 639-1 two-letter codes. Full language names are in Appendix A.

opment set. Moreover, in Figure 16 we report the crosslingual S@10 performance on the *test* set for `multilingual-e5-large`, according to negative sampling strategies and number of negatives.

D Additional Results

Table 9 shows S@10 *test* scores for all approaches.

Finally, since prior works showed the benefits of translation of posts and fact-checks to English over the use of original data with multilingual models (see e.g., Pikuliak et al., 2023), we also provide results of the unsupervised `multilingual-e5-large` model with the data translated to English for comparison and completeness despite our focus on multi- and crosslinguality. The achieved score of 0.6996 MRR@10 and 0.7647 S@10 (on the test set in the multilingual setting) show that it is outperformed when using the data with its original languages in contrast to the results by Pikuliak et al. (2023). This demonstrates the recent progress in the multilingual text embedding models. We hypothesize that the original language can help to retrieve the correct fact-check, especially in the case when both post and fact-check are in the same language. On the other hand, the results on the test set in the crosslingual setting – 0.6017 MRR@10 (the second best when compared to the MRR@10 results in Table 2) and 0.7514 S@10 (the best when compared to the S@10 results in Table 9) – show that translation can still help when the original language of the post and the fact-check differ.

Model	Retrieve				Re-rank			
	Test		Full		Test (*top-n = 30*)		Full (*top-n = 30*)	
	S@10	MRR@10	S@10	MRR@10	S@10	MRR@10	S@10	MRR@10
multilingual-MiniLM-L12-v2	0.5762	0.5062	0.4550	0.3829	0.6950	0.6554	0.5037	0.4561
paraphrase-multilingual-mpnet-base-v2	0.6416	0.5690	0.5180	0.4283	0.6796	0.6485	0.5714	0.5068
e5-mistral-7b-instruct	0.7516	0.6798	0.6257	0.5509	0.7811	0.7223	0.6631	0.5811
bge-m3	0.7525	0.6952	0.6586	0.5599	0.7692	0.7031	0.6752	0.5686
Linq-Embed-Mistral	0.7630	0.7199	0.6662	0.6027	0.7731	0.7174	0.6836	0.5952
text-embedding-3-small	0.7687	0.7229	0.6797	0.6015	0.7731	0.7035	0.6819	0.5770
KaLM-multilingual-instruct-v1	0.7756	0.7155	0.6661	0.5902	0.7938	0.7275	0.6907	0.5952
gte-Qwen2-7B-instruct	0.7769	0.7155	0.6791	0.5781	0.7922	0.7219	0.6937	0.5852
jina-embeddings-v3	0.7863	0.7137	0.6792	0.5737	0.7900	0.7120	0.6974	0.5766
snowflake-arctic-embed-l-v2.0	0.7870	0.7344	0.7103	0.6124	0.7902	0.7130	0.7068	0.5807
gte-multilingual-base	0.7881	0.7228	0.6953	0.5812	0.7939	0.7161	0.7068	0.5808
multilingual-e5-large-instruct	0.7912	0.7311	0.6758	0.6035	0.8055	**0.7341**	0.7017	0.6003
SFR-Embedding-Mistral	0.7951	0.7309	0.6830	0.6030	0.8039	0.7296	0.7057	0.5976
multilingual-e5-large	0.7971	0.7475	0.7072	0.6280	0.8075	0.7323	0.7212	**0.6034**
bge-m-gemma2	0.8168	0.7613	0.7265	0.6376	0.8111	0.7298	0.7230	0.5970
text-embedding-3-large	**0.8462**	**0.7909**	**0.7570**	**0.6661**	**0.8241**	0.7321	**0.7399**	0.6013

Table 5: Performance of retrieval and retrieval+re-ranking (bge-reranker-v2-m3) across models in the *test* set and the *full* set of claims in the **multilingual** setting.

Model	Retrieve				Re-rank			
	Test		Full		Test (*top-n = 30*)		Full (*top-n = 30*)	
	S@10	MRR@10	S@10	MRR@10	S@10	MRR@10	S@10	MRR@10
Linq-Embed-Mistral	0.3292	0.2147	0.1865	0.1138	0.4152	0.3410	0.2410	0.1994
multilingual-MiniLM-L12-v2	0.3737	0.2479	0.2225	0.1429	0.4497	0.3512	0.2617	0.1980
text-embedding-3-small	0.3899	0.2488	0.2463	0.1434	0.4658	0.3506	0.2855	0.1930
e5-mistral-7b-instruct	0.4751	0.2837	0.2647	0.1513	0.5840	0.4501	0.3354	0.2608
paraphrase-multilingual-mpnet-base-v2	0.5188	0.3547	0.3315	0.2048	0.5779	0.4482	0.4137	0.3002
KaLM-multilingual-mini-instruct-v1	0.5272	0.3365	0.3162	0.1936	0.6178	0.4900	0.3883	0.2903
multilingual-e5-large-instruct	0.5395	0.3318	0.3047	0.1764	0.6332	0.4998	0.4029	0.2962
SFR-Embedding-Mistral	0.5687	0.3744	0.3630	0.2094	0.6562	0.4968	0.4413	0.3177
gte-Qwen2-7B-instruct	0.6301	0.4385	0.4536	0.2720	0.6792	0.5283	0.5081	0.3556
multilingual-e5-large	0.6324	0.4323	0.4597	0.2703	0.7069	0.5442	0.5234	0.3570
bge-m3	0.6754	0.4844	0.5127	0.3104	0.6961	0.5195	0.5403	0.3558
text-embedding-3-large	0.7199	0.5119	0.5487	0.3276	0.7299	0.5418	0.5748	0.3669
snowflake-arctic-embed-l-v2.0	0.7260	0.5108	0.5687	0.3352	0.7337	0.5489	0.5840	0.3714
jina-embeddings-v3	0.7283	0.5048	0.5142	0.3157	0.7360	0.5433	0.5610	0.3688
gte-multilingual-base	0.7360	0.5443	**0.5948**	**0.3634**	0.7422	0.5592	**0.5979**	0.3818
bge-multilingual-gemma2	**0.7621**	**0.5446**	0.5787	0.3478	**0.7598**	**0.5608**	0.5956	**0.3910**

Table 6: Performance of retrieval and retrieval+re-ranking (bge-reranker-v2-m3) across models in the *test* set and the *full* set of claims in the **crosslingual** setting.

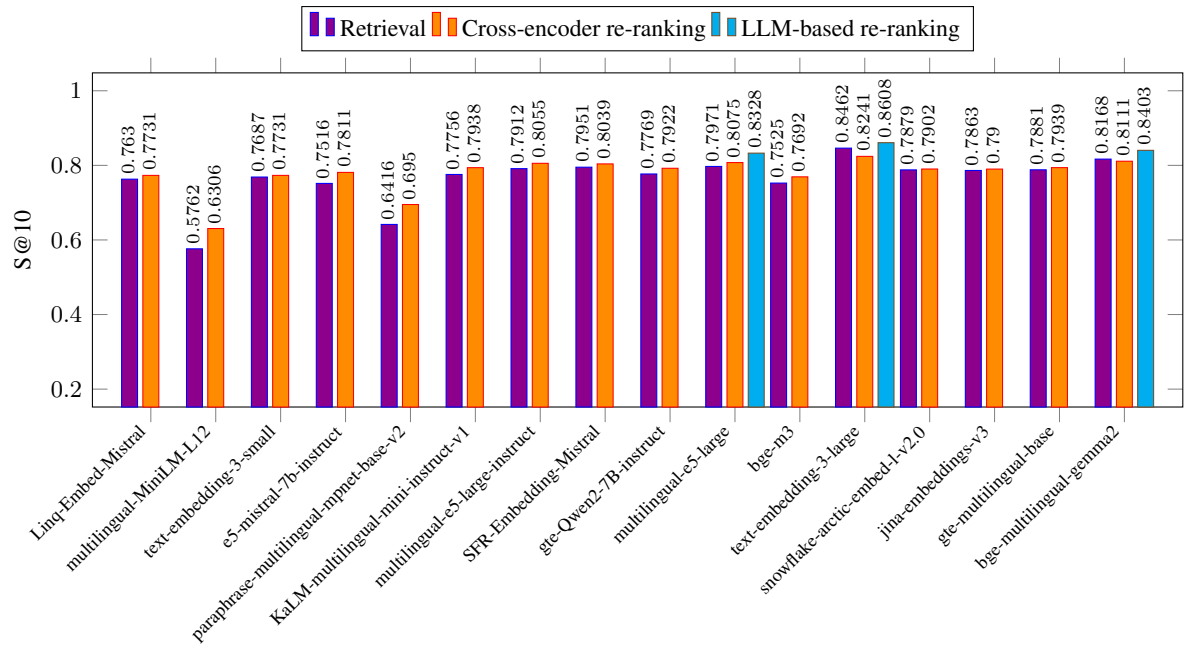

Figure 5: S@10 performance of retrieval and re-ranking across models in the **test** set in the **multilingual** setting.

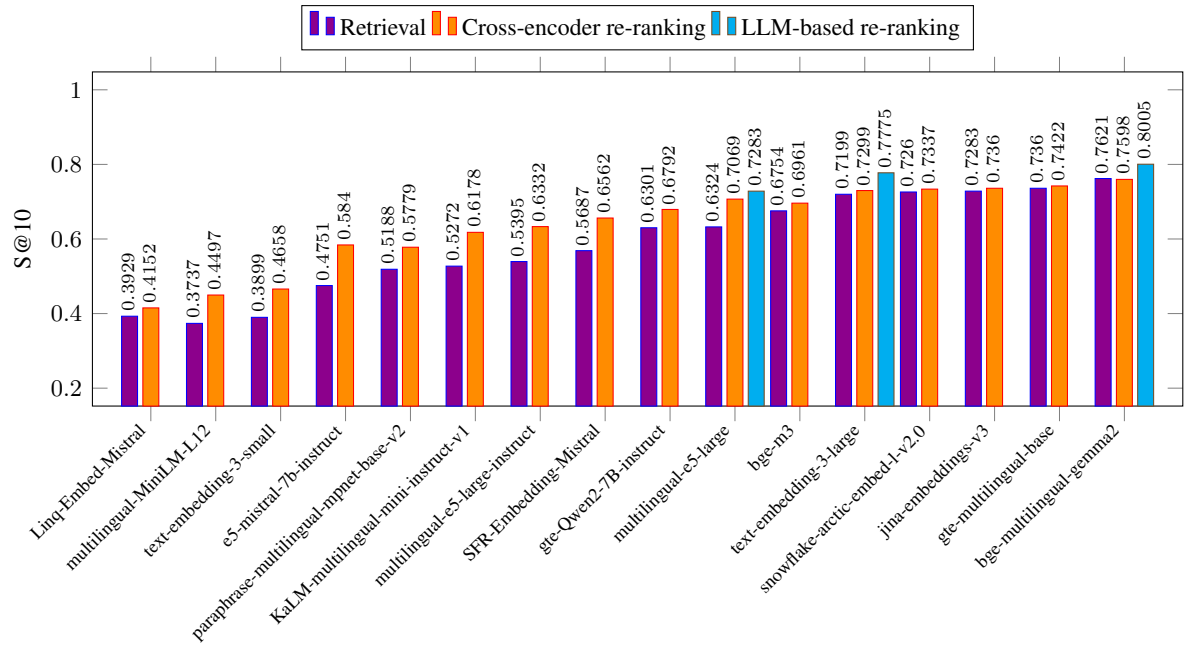

Figure 6: S@10 performance of retrieval and re-ranking across models in the **test** set in the **crosslingual** setting.

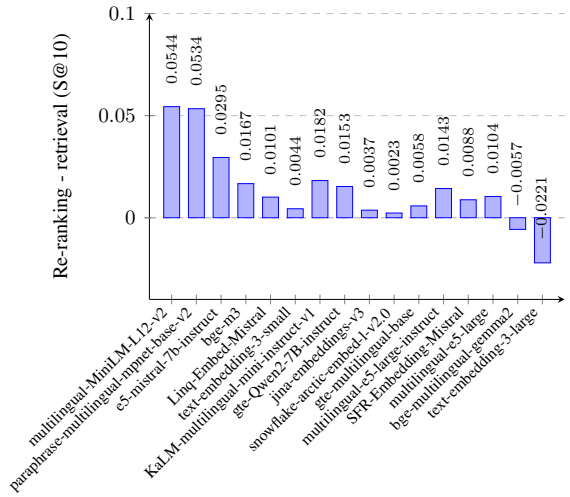

Figure 7: Difference in **S@10** **test** set performance for re-ranking - retrieval in the **multilingual** setting, with `bge-reranker-v2-m3`. 6,375 pairs, 5,683 posts, and 6,145 claims. Top-$n = 30$.

Figure 8: Difference in **MRR@10** **test** set performance for re-ranking - retrieval in the **multilingual** setting, with `bge-reranker-v2-m3`. 6,375 pairs, 5,683 posts, and 6,145 claims. Top-$n = 30$.

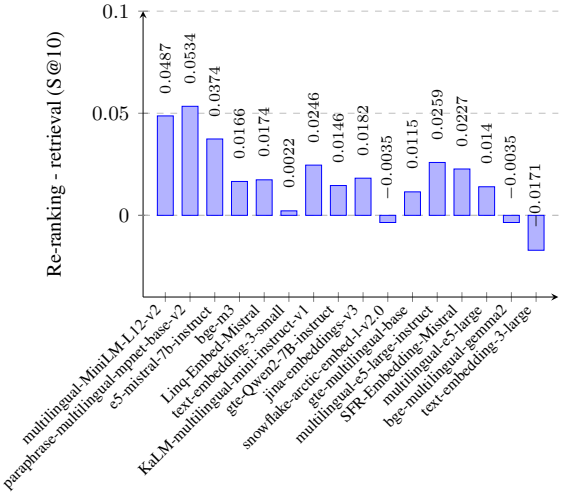

Figure 9: Difference in **S@10** **full** set performance for re-ranking - retrieval in the **multilingual** setting, with `bge-reranker-v2-m3`. 63,913 pairs, 55,421 posts, and 52,911 claims. Top-$n = 30$.

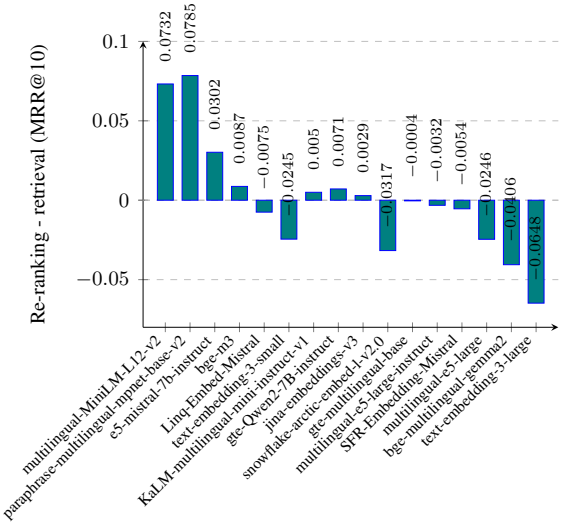

Figure 10: Difference in **MRR@10** **full** set performance for re-ranking - retrieval in the **multilingual** setting, with `bge-reranker-v2-m3`. 63,913 pairs, 55,421 posts, and 52,911 claims. Top-$n = 30$.

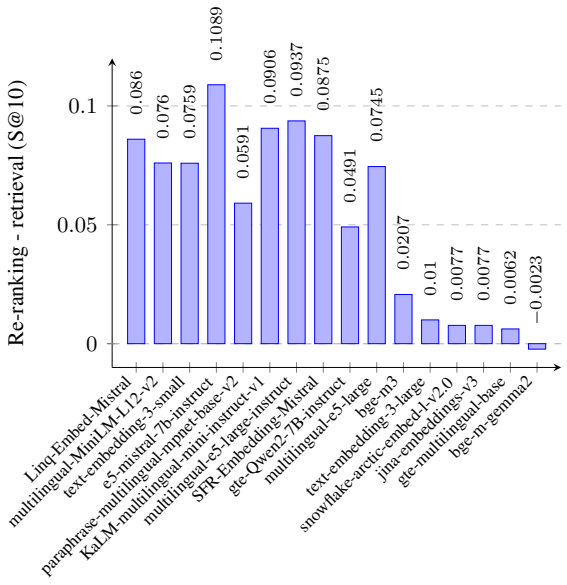

Figure 11: Difference in **S@10** **test** set performance for re-ranking - retrieval in the **crosslingual** setting, with `bge-reranker-v2-m3`. 929 pairs, 850 posts, and 901 claims. Top-$n = 30$.

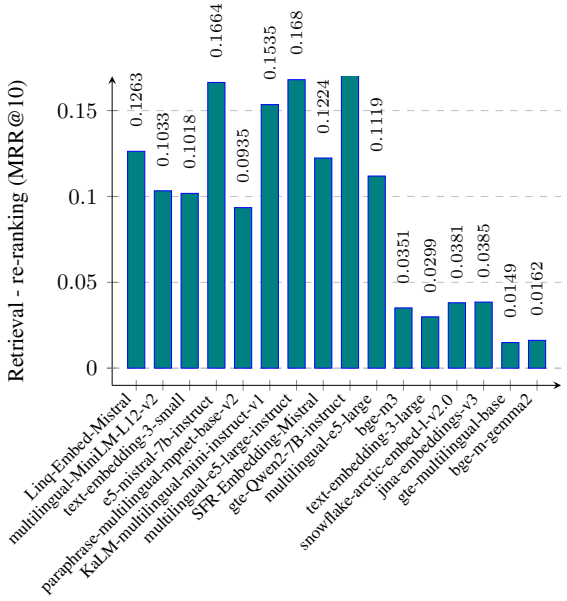

Figure 12: Difference in **MRR@10** **test** set performance for re-ranking - retrieval in the **crosslingual** setting, with `bge-reranker-v2-m3`. 929 pairs, 850 posts, and 901 claims. Top-$n = 30$.

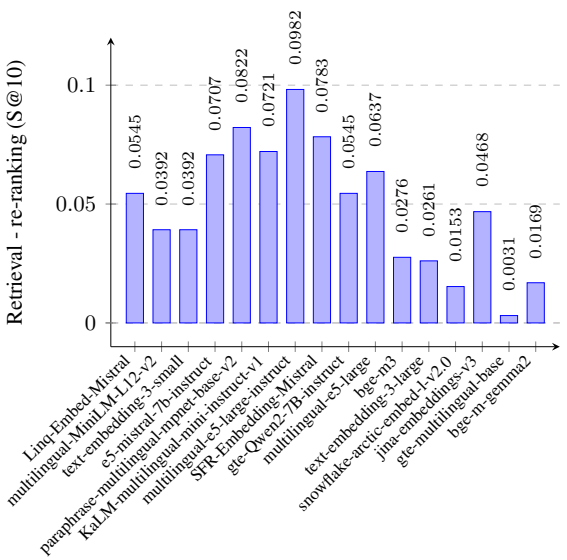

Figure 13: Difference in **S@10** **full** set performance for re-ranking - retrieval in the **crosslingual** setting, with `bge-reranker-v2-m3`. 9,066 pairs, 7,975 posts, and 7,869 claims. Top-$n = 30$.

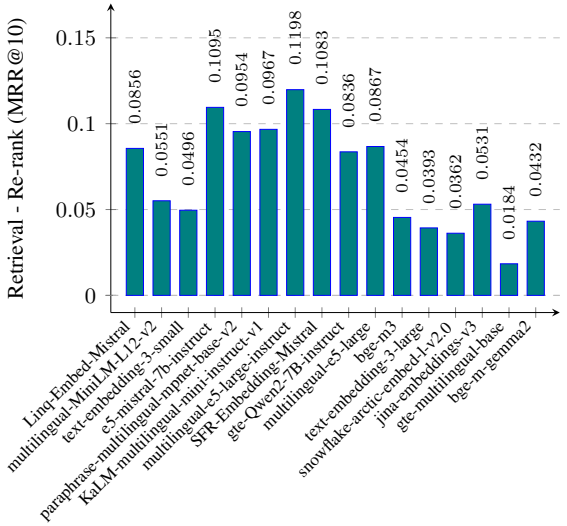

Figure 14: Difference in **MRR@10** **full** set performance for re-ranking - retrieval in the **crosslingual** setting, with `bge-reranker-v2-m3`. 9,066 pairs, 7,975 posts, and 7,869 claims. Top-$n = 30$.

Hyper-parameter	Value
Optimizer	AdamW
Epochs	3
Batch size	8
Learning rate	1e-8
Warm-up steps	1,600
Label smoothing	0.1
Similarity function	cosine
Similarity scale	20
Weight decay	8e-5
Decay factor	0.38
Cut fraction	0.3
Clip value	1
# negatives (k)	10
Sampling strategy	`similarity`, `topic`, `random`

Table 7: Final hyper-parameter values used for the supervised models in all our experiments.

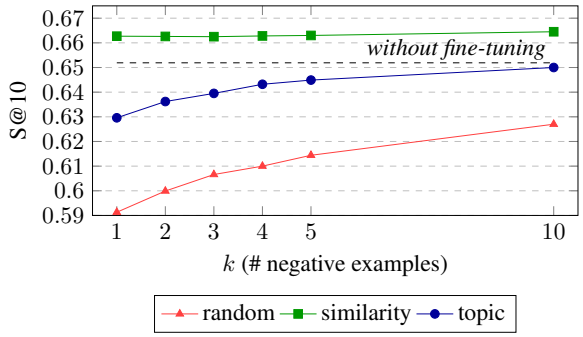

Figure 15: **Multilingual** S@10 performance across negative sampling strategies and number of negative examples k for the fine-tuned `paraphrase-multilingual-mpnet-base-v2` model on the **development** set. Reported results are averages over three runs using different seeds. The dashed line indicates results when no fine-tuning is conducted.

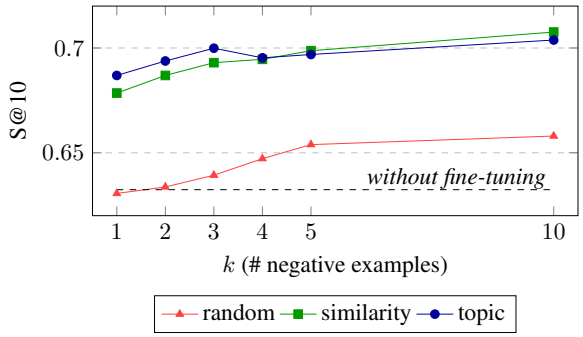

Figure 16: **Crosslingual** S@10 performance across negative sampling strategies and number of negative examples k for the fine-tuned `multilingual-e5-large` model on the **test** set. Reported results are averages over three runs using different seeds. The dashed line indicates results when no fine-tuning is conducted.

(A) Multilingual - Retrieval

	test		full	
	r	p-value	r	p-value
Emb. dimension	0.306	0.286	0.327	0.252
Model size (params)	0.330	0.249	0.348	0.222

(B) Multilingual - Re-rank

	test		full	
	r	p-value	r	p-value
Emb. dimension	0.374	0.188	0.376	0.185
Model size (params)	0.372	0.191	0.370	0.193

(C) Crosslingual - Retrieval

	test		full	
	r	p-value	r	p-value
Emb. dimension	-0.244	0.401	-0.270	0.352
Model size (params)	-0.120	0.684	-0.145	0.620

(D) Crosslingual - Re-rank

	test		full	
	r	p-value	r	p-value
Emb. dimension	-0.197	0.500	-0.185	0.525
Model size (params)	-0.100	0.735	-0.073	0.803

Table 8: Correlation analysis between embedding dimension / model size (in terms of number of parameters) and performance (MRR@10) in the unsupervised setting. We report Pearson r and the respective p-value. The scores have been computed on the 16 embedding models described in Section 5.1.

Model	Strategy	Multilingual	Crosslingual	Crosslingual (no eng)
multilingual-e5-large	retrieve	0.7971	0.6324	0.6622
	re-rank	0.8075	0.7069	**0.7780**
	random	$0.8134_{\pm 0.0002}$	$0.6580_{\pm 0.0012}$	$0.6853_{\pm 0.0000}$
	topic	$0.8203_{\pm 0.0002}$	$0.7038_{\pm 0.0000}$	$0.7224_{\pm 0.0000}$
	similarity	$0.8228_{\pm 0.0001}$	$0.7076_{\pm 0.0000}$	$0.7270_{\pm 0.0000}$
	llm re-rank	**0.8324**	**0.7283**	0.7409
paraphrase-multilingual-mpnet-base-v2	random	$0.5955_{\pm 0.0002}$	$0.4758_{\pm 0.0008}$	$0.5788_{\pm 0.0000}$
	topic	$0.6206_{\pm 0.0003}$	$0.5014_{\pm 0.0012}$	$0.5819_{\pm 0.0027}$
	similarity	$0.6320_{\pm 0.0002}$	$0.5157_{\pm 0.0000}$	$0.5880_{\pm 0.0000}$
	retrieve	0.6416	0.5188	0.5834
	re-rank	**0.6796**	**0.5779**	**0.6900**

Table 9: **Multilingual**, **crosslingual**, and **crosslingual (no eng)** S@10 performance across negative sampling strategies (random, topic, similarity) for fine-tuned models on the **test** set compared to unsupervised strategies (retrieve, cross-encoder re-rank, llm re-rank). Results are ordered by increasing multilingual score; for the supervised setup, we report averages with standard deviation over three runs using different seeds.