DKE-Research at SemEval-2025 Task 7: A Unified Multilingual Framework for Cross-Lingual and Monolingual Retrieval with Efficient Language-specific Adaptation

Yuqi Wang[1,2], **Kangshi Wang**[3]
[1]Xi'an Jiaotong Liverpool University
[2]University of Liverpool
[3]Midea Group (Shanghai) Co., Ltd
yuqi.wang17@student.xjtlu.edu.cn, wangks24@midea.com

Abstract

The global spread of misinformation has become a critical challenge, making multilingual and cross-lingual fact-checking increasingly essential for ensuring the credibility of information across diverse languages. This paper presents a unified framework for fact-checked claim retrieval, integrating contrastive learning with an in-batch multiple negative ranking loss and a conflict-aware batch sampler to enhance query-document alignment across languages. Additionally, we introduce language-specific adapters for efficient fine-tuning, enabling adaptation to previously unseen languages. Our results demonstrate significant improvements in retrieval performance in both monolingual and cross-lingual settings, underscoring the importance of developing scalable, multilingual systems to combat misinformation and ensure the reliability of information on a global scale.

1 Introduction

With the rapid dissemination of information in the digital age, the global spread of misinformation has become a significant challenge. For instance, a recent study (Vosoughi et al., 2018) found that false news spreads around six times faster than true news on social media, highlighting the urgency of addressing this issue. Moreover, false posts and disinformation on popular social media platforms often transcend boundaries of linguistic and cultural, reaching diverse audiences before they can be effectively countered (Wang et al., 2024b). This dynamic underscores the critical importance of multilingual and cross-lingual natural language processing (NLP) techniques (Chen et al., 2024b; Peng et al., 2023; Wang et al., 2024c), which enable fact-checkers to break down language barriers, access and verify information across languages for the rapid identification of relevant content. Therefore, such techniques are essential, as they not only enhance the efficiency and scalability of fact-checking efforts but also bridge gaps between disparate sources of information.

In this paper, we propose a multilingual and cross-lingual information retrieval (CLIR) system designed to enhance fact-checked claim retrieval in a multilingual context. We present a unified framework for both cross-lingual and monolingual retrieval tasks, demonstrating the effectiveness of our system through detailed experiments. Our method combines a contrastive learning approach with a conflict-aware batch sampler to improve the alignment of query-document pairs in different languages. Additionally, we introduce language-specific adapters for efficient fine-tuning, which significantly improves the performance of the system on unseen languages.

2 Background

Multilingual and CLIR have evolved significantly, transitioning from lexical matching to semantic-aware neural architectures. Early approaches relied on statistic-based lexical methods (Robertson et al., 1995), which performed exact term matching but struggled with cross-lingual lexical gaps, such as polysemy or morphological variations across languages (Oard and Diekema, 1998).

To address these limitations, translation-based CLIR methods emerged, leveraging machine translation (MT) to bridge languages. These methods either translate queries into the document language (query translation) or documents into the query language (document translation) (Sokolov et al., 2013; Järvelin et al., 2008). However, such approaches were prone to error propagation from imperfect MT systems, particularly for morphologically rich or under-resourced languages (Nie, 2010).

The advent of pre-trained multilingual language representation models, such as multilingual BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), marked a paradigm shift for dense retrieval.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 155–159

These models enabled embedding queries and documents in different languages into a shared semantic space. For example, the XLM-R-based models, such as multilingual E5 (Wang et al., 2024a) and BGE-M3 (Chen et al., 2024a), leverage contrastive pre-training on large multilingual corpora to align cross-lingual representations, capturing meaningful semantic relationships among multiple languages and better generalize across linguistic and cultural contexts.

3 System Overview

Our multilingual and CLIR system is designed to address the challenges of cross-lingual and monolingual retrieval for different languages in a unified framework. The system leverages a Multilingual E5 (Wang et al., 2024a) (M-E5) model, pretrained on extensive corpora including more than 100 languages, as the backbone for the language-agnostic representations and further enhances it with language-specific adapters for parameter-efficient adaptation for each language indicated in this task. The system is fine-tuned using contrastive learning with an in-batch multiple negative ranking loss, enhanced with the conflict-aware batch sampling constraint, ensuring better alignment across languages. Below, we describe the key components of the system in detail.

3.1 Contrastive Learning with In-batch Multiple Negative Ranking Loss

To effectively utilise the cross-lingual data, we employ a contrastive learning framework that uses an in-batch multiple negative ranking loss (Henderson et al., 2017). Let the batch size be N. For the i-th positive pair in the batch, denote the query post as q_i and the corresponding fact-checked claim as d_i. The similarity between a query post and a fact-checked claim is measured using a function such as cosine similarity, denoted as $\text{sim}(q_i, d_j)$. The overall loss $\mathcal{L}$ is computed by averaging the loss over all positive pairs in the mini-batch:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left(\log\frac{\exp\left(\frac{\text{sim}(q_i,d_i)}{\tau}\right)}{\sum_{j=1,i\neq j}^{N}\exp\left(\frac{\text{sim}(q_i,d_j)}{\tau}\right)}\right),$$

where τ is a temperature parameter that controls the smoothness of the distribution.

This formulation drives the model to maximize the similarity between positive query-document pairs while treating all other documents in the same batch as hard negatives. By leveraging these in-batch negatives, the M-E5 model learns highly discriminative representations that effectively capture cross-lingual semantic correspondences for the fact-checking task.

3.2 Conflict-aware Batch Sampler

In practice, applying the in-batch multiple negative ranking loss for the fact-checking task often encounters scenarios where a single query may be associated with multiple relevant fact-checked claims, and similarly, a fact-checked claim may be relevant to multiple queries. Without careful batching, this can lead to conflicts where the same query or document appears multiple times within the same batch. Such conflicts can result in a situation where an instance inadvertently acts as both a positive and a negative example, thereby contaminating the loss signal during the training.

To mitigate this issue, we design a conflict-aware batch sampler. The sampler ensures that, within any given batch $B = \{(q_i, d_i)\}_{i=1}^{N}$, each query q_i and each fact-checked claim d_i appears only once. Formally, for any two distinct pairs (q_i, d_i) and (q_j, d_j) in the batch (with $i \neq j$), we enforce

$$\forall i, j \in \{1, \ldots, N\},\ i \neq j \implies (q_i \neq q_j \ \wedge\ d_i \neq d_j)$$

To avoid the situation where q_i and d_j appear in the same batch if they belong to positive pairs in other batches, we can add the following constraint:

$$\forall i, j \in \{1, \ldots, N\},\ (q_i, d_j) \in P \implies i = j$$

In this way, we ensure that whenever a query q_i and fact-checked claim d_j form a positive pair in some batches, they cannot appear as separate negative pairs in different batches. Here, P represents the set of all positive query-document pairs.

Adding these constraints in the batch sampler guarantees that every instance in the batch is unique, thereby preventing any query or document from inadvertently acting as a negative example for itself or for another positive pair. By leveraging such a conflict-aware strategy, we can confirm that the in-batch negatives are truly negative.

3.3 Language-specific Adapter

After establishing the foundation with the cross-lingual training data for the fact-checking task, the system is further refined with monolingual data

Split	Mono.										Cross.
	eng	spa	deu	por	fra	ara	msa	tha	pol	tur	all
train	4,351	5,628	667	2,571	1,596	676	1,062	465	-	-	4,972
dev	478	615	83	302	188	78	105	42	-	-	552
test	500	500	500	500	500	500	93	183	500	500	4,000

Table 1: We present the number of queries for each language in the monolingual setting, and the total number of queries in the cross-lingual setting.

for each language through the incorporation of language-specific adapters. Specifically, for each language l, a low-rank adaptation (LoRA) (Hu et al., 2022) is introduced to efficiently fine-tune the model without modifying the base parameters. Let $\mathbf{W} \in \mathbb{R}^{m \times k}$ denote a pre-trained weight matrix in the transformer layers. The adapter injects a low-rank update $\Delta\mathbf{W}_l$ into $\mathbf{W}$, parameterized as:

$$\Delta\mathbf{W}_l = \mathbf{B}_l\mathbf{A}_l,$$

where $\mathbf{B}_l \in \mathbb{R}^{m \times r}$ and $\mathbf{A}_l \in \mathbb{R}^{r \times k}$ are trainable low-rank matrices with rank $r \ll \min(m, k)$. During forward propagation, the adapted output for language l becomes:

$$(\mathbf{W} + \Delta\mathbf{W}_l)\mathbf{x} = \mathbf{W}\mathbf{x} + \mathbf{B}_l\mathbf{A}_l\mathbf{x}.$$

Here, $\mathbf{W}$ remains frozen, while only $\mathbf{B}_l$ and $\mathbf{A}_l$ are updated during training.

The language-specific adapter for language l is optimized using the same in-batch loss as in Section 3.1 and Section 3.2. Training leverages both monolingual positive pairs (q_i^l, d_i^l) and cross-lingual positive pairs, with high-resource English as the pivot, i.e. either (q_i^l, d_i^{en}) or (q_i^{en}, d_i^l), where q_i^{en} and d_i^{en} are English translations of q_i^l and d_i^l, respectively. In this way, the model retains language-specific features while benefiting from the semantic consistency provided by the pivot, thus enhancing the representation for each language.

For unseen languages, we propose to merge language-specific adapters from morphologically similar seen languages. Let $\mathcal{S}$ denote the set of languages in the training set, and let u represent an unseen language (e.g., Turkish (tur) or Polish (pol)). For each u, we define a subset $\mathcal{S}_u \subset \mathcal{S}$ comprising seen languages morphologically similar to u. The adapter for the unseen language is then constructed by averaging the adapters of the languages in $\mathcal{S}_u$:

$$\Delta\mathbf{W}_u = \frac{1}{|\mathcal{S}_u|} \sum_{l \in \mathcal{S}_u} \Delta\mathbf{W}_l.$$

For instance, considering that Polish is an Indo-European language, we select:

$$\mathcal{S}_{\text{pol}} = \{\text{eng}, \text{spa}, \text{deu}, \text{por}, \text{fra}\}$$

Similarly, for Turkish, due to the absence of direct Turkic counterparts in the training set, we merge adapters from languages with non-Latin scripts and typological diversity to mitigate biases from Indo-European languages. In this case, we define:

$$\mathcal{S}_{\text{tur}} = \{\text{ara}, \text{msa}, \text{tha}\}$$

4 Experiments

4.1 Dataset

In this work, we evaluated our proposed system using the SemEval 2025 Task-7 dataset (Peng et al., 2025), which consists of two sub-tasks: cross-lingual and monolingual fact-checked claim retrieval. The dataset is the modified version of the MultiClaim dataset developed by Pikuliak et al. (2023), including three key files for training and validation: a database of fact-checked claims, posts extracted from social media platforms, and mappings between posts and corresponding claims. Additionally, for each post or fact-checked claim, the English translation of each non-English content is also provided via Google API. The statistics (#query) of the dataset are shown in Table 1.

4.2 Setup

We used Success@10 as our evaluation metric, where retrieval is considered successful if all relevant fact-checked claims are found within the top 10 retrieved results. This system was implemented using Python 3.10 and Pytorch 2.1.1. The M-E5 model was downloaded from the huggingface repository[1]. During the training, the batch size was set to 24, We used the AdamW as the optimizer and the learning rate was set to 2×10^{-5}.

[1]https://huggingface.co/intfloat/multilingual-e5-large

Models	Mono.									Cross.
	eng	spa	deu	por	fra	ara	msa	tha	avg	avg
M-E5 (w/o constraint)	80.5	87.8	83.1	86.4	87.8	83.3	93.3	100	87.8	82.4
M-E5	81.4	89.9	83.1	86.4	88.8	83.3	93.3	100	88.3	83.5
LADA-M-E5	85.4	94.0	88.0	89.1	91.5	83.3	93.3	100	**90.6**	**86.1**
M-E5-Instruct	84.1	91.5	81.9	86.4	89.4	83.3	90.4	97.6	88.1	80.6
LADA-M-E5-Instruct	87.0	94.5	81.9	90.1	89.4	83.3	91.4	97.6	89.5	81.0

Table 2: Results on development set measured in Success@10

Models	Mono.											Cross.
	eng	spa	deu	por	fra	ara	msa	tha	pol*	tur*	avg	avg
M-E5	80.4	89.6	85.2	80.2	91.4	92.2	97.8	97.6	83.6	81.8	87.6	70.6
LADA-M-E5	82.0	91.6	86.8	83.4	92.4	93.6	100	97.6	85.6	87.4	89.8	71.3

Table 3: Results on test set measured in Success@10, * indicates the unseen language in the training and development sets.

4.3 Results

We present the results on the development set measured in Success@10 in Table 2. We show the result of M-E5 trained with the provided cross-lingual data. M-E5 w/o constraint means we did not apply the proposed conflict-aware constraints on the batch sampler, and "LADA-M-E5" stands for the proposed system language-specific adapter in this work. Apart from the original M-E5, we also implemented the instruction-tuned embedding model, namely, M-E5-Instruct. The instruction for the query is "*Given a web search query, retrieve relevant passages that answer the query*", which is consistent in the pre-training phase.

We observed that incorporating the conflict-aware batch sampler improved performance on several languages in the monolingual set. For instance, English, Spanish, and French saw improvements of 0.9%, 2.1%, and 1.0%, respectively, highlighting the importance of selecting truly negative samples for the in-batch loss. Additionally, the M-E5 model outperformed M-E5-Instruct in most languages within the monolingual setting, with the exception of a few Latin-based languages. Furthermore, M-E5 showed significant performance gains in the cross-lingual setting. This advantage may be attributed to the fact that the instruction is primarily in English, which could benefit Latin languages more than others. It can be seen that both M-E5 and M-E5-Instruct gain improvements from our proposed LADA in both monolingual and cross-lingual settings, enhancing performance on the fact-checking task.

In the testing, we present the performance of M-E5 and LADA-M-E5 in Table 3. For the unseen languages, Polish and Turkish, we merged the language-specific adapters as described in Section 3.3. It is evident that the proposed language-specific adapters significantly improve the monolingual information retrieval performance for both Polish and Turkish. However, in the cross-lingual setting, we observe a large performance discrepancy between the testing and development sets for both M-E5 and M-E5-Instruct. This discrepancy may stem from the fact that the unseen languages might not be well-aligned with the other seen languages in the training set.

5 Conclusion

In this paper, we introduced a unified multilingual framework for cross-lingual and monolingual fact-checked claim retrieval, leveraging contrastive learning, conflict-aware batch sampling, and language-specific adapters. Our approach effectively improves retrieval performance by aligning multilingual representations while maintaining language-specific features. The integration of low-rank adapters allows efficient adaptation to individual languages, with a strategy for handling unseen languages based on morphologically similar counterparts. Experimental results on the SemEval 2025 Task 7 dataset demonstrate the effectiveness of our

method, achieving strong performance across multiple languages. Future work will explore extending our approach to more low-resource languages and further optimizing retrieval efficiency in real-world applications.

References

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Tong Chen, Procheta Sen, Zimu Wang, Zhengyong Jiang, and Jionglong Su. 2024b. Knowledge base-enhanced multilingual relation extraction with large language models.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, and Guillaume Wenzek. 2020. Francisco guzmá n, edouard grave, myle ott, luke zettlemoyer, and veselin stoyanov. 2020. unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers). In *Association for Computational Linguistics*, pages 4171–4186.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Antti Järvelin, Tuomas Talvensaari, and Anni Järvelin. 2008. Data driven methods for improving mono-and cross-lingual ir performance in noisy environments. In *Proceedings of the second workshop on analytics for noisy unstructured text data*, pages 75–82.

Jian-Yun Nie. 2010. *Cross-language information retrieval*, volume 8. San Rafael, CA: Morgan & Claypool Publishers.

Douglas W Oard and Anne R Diekema. 1998. Cross-language information retrieval. *Annual Review of Information Science and Technology (ARIST)*, 33:223–56.

Hao Peng, Xiaozhi Wang, Feng Yao, Zimu Wang, Chuzhao Zhu, Kaisheng Zeng, Lei Hou, and Juanzi Li. 2023. Omnievent: A comprehensive, fair, and easy-to-use toolkit for event understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 508–517.

Qiwei Peng, Robert Moro, Michal Gregor, Ivan Srba, Simon Ostermann, Marian Simko, Juraj Podroužek, 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, Vienna, Austria.

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.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Artem Sokolov, Laura Jehl, Felix Hieber, and Stefan Riezler. 2013. Boosting cross-language retrieval by learning bilingual phrase associations from relevance rankings. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1688–1699.

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

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Xinyu Wang, Wenbo Zhang, and Sarah Rajtmajer. 2024b. Monolingual and multilingual misinformation detection for low-resource languages: A comprehensive survey. *arXiv preprint arXiv:2410.18390*.

Yuqi Wang, Zimu Wang, Nijia Han, Wei Wang, Qi Chen, Haiyang Zhang, Yushan Pan, and Anh Nguyen. 2024c. Knowledge distillation from monolingual to multilingual models for intelligent and interpretable multilingual emotion detection. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 470–475.