YNU-HPCC at SemEval-2025 Task 7: Multilingual and Cross-lingual Fact-checked Claim Retrieval

Yuheng Mao, Jin Wang and Xuejie Zhang
School of Information Science and Engineering
Yunnan University
Kunming, China
maoyuheng@stu.ynu.edu.cn,{wangjin,xjzhang}@ynu.edu.cn

Abstract

This paper presents the system for Task 7, Multilingual and Crosslingual Fact-Checked Claim Retrieval. YNU-HPCC team participated in all subtasks of this task and employed the same unified framework to obtain results. The task includes two subtasks: monolingual and crosslingual. Our approach explores the integration of multiple embedding models to address these subtasks. These embedding models were explicitly fine-tuned for the task, and weighted cosine similarity was utilized for result prediction. Extensive experiments were conducted on development and test datasets. The comparative results show that (I) The integration of multiple embedding models has been demonstrated to significantly enhance retrieval accuracy, particularly in cross-lingual fact-checking retrieval tasks; (II) Translating text may degrade the retrieval performance of cross-lingual embedding models; (III) Using GTE multilingual base model and Jina model for ensemble achieves near-optimal performance, effectively balancing efficiency and computational cost.The code of this paper is available at https://github.com/catoraa/semeval2025-task7.

1 Introduction

Fact-checking is a task designed to evaluate the accuracy of published statements or claims. Manual fact-checking poses significant challenges in the contemporary media ecosystem, which is marked by extensive data volumes and rapid dissemination. This task is often time-consuming and labor-intensive for professional fact-checkers, even within a single language. The complexity increases when claims and fact-checks span multiple languages, making manual completion even more arduous. Previous research has established automated fact-checking retrieval's high feasibility and systematic potential (Guo et al., 2022).

Therefore, SemEval 2025 Task 7 (Peng et al., 2025) focuses on Automated Fact-Checked Claim Retrieval, encompassing Multilingual and Cross-lingual scenarios. This paper designs a retrieval system based on embedding models and semantic similarity for this task. We employed four embedding models, including BGE-M3 model (Chen et al., 2024), GTE base model (Zhang et al., 2024), jina embeddings v3 model (Sturua et al., 2024), and E5 large model (Wang et al., 2024), as foundational models, fine-tuned them using the Hugging Face Trainer, and finally integrated them through a weighted approach to derive the final cosine similarity.

The experimental results of this paper were presented in Task 7 of SemEval 2025. On the original dataset, the system achieved an success@10 of 0.92 in the Monolingual task, ranking $11th$, and 0.77 in the Cross-lingual task, ranking $11th$.

The rest of the paper is structured as follows: Section 2 summarizes recent fact-checking retrieval advancements.Section 3 describes the proposed system and models. In Section 4, the experimental details and parameter selection are elaborated. In Section 5, the comparison and analysis of the experimental results are discussed, and in Section 6, the conclusions are presented.

2 Related Work

Contemporary fact-checking predominantly rely on large LLMs or pre-trained models for retrieval and verification.These approaches typically integrate LLMs with existing evidence for judgment or leverage semantic similarity assessments through pre-trained embedding models.

For instance, Cheung proposed the FactLLaMA system, which enhances the temporal relevance and accuracy of information in fact-checking tasks by employing instruction-based fine-tuning and LoRA methods (hin Cheung and Lam, 2023).Singhal developed a system for fact-checking using LLMs, constructed on the basis of RAG

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 234–240
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

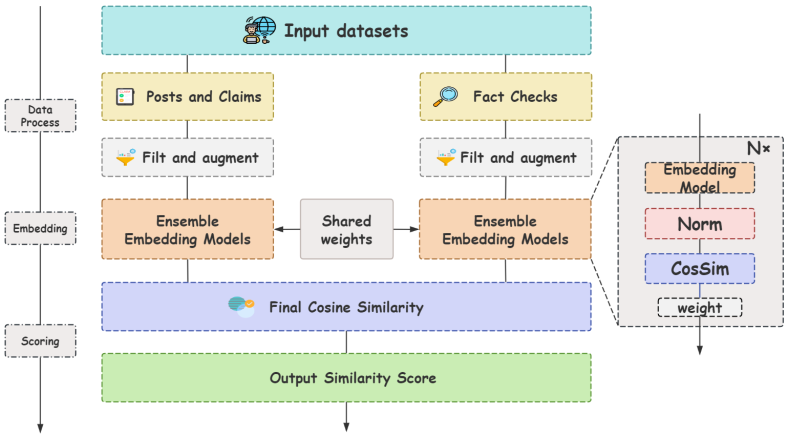

Figure 1: The process of semantic similarity calculation was implemented using the sentence-transformers and embedding models fine-tuned on the training data. Weighted cosine similarity was computed to assess the semantic relevance between posts and fact checks.

(Retrieval-Augmented Generation) and ICL (In-Context Learning) (Singal et al., 2024).Li explored self-instruction techniques to improve LLMs' capability in evaluating semantic similarity and relevance (Li et al., 2024).Khaliq introduced the RAGAR system, which combines LLMs with RAG to enhance verification accuracy (Khaliq et al., 2024).

In contrast to conventional methods, Liu investigated the use of LSTM integrated with the Attention Mechanism for the classification and judgment of factual information, demonstrating promising results. (Liu et al., 2019a).

A critical challenge in these approaches lies in efficiently retrieving relevant information, as they heavily depend on direct factual evidence or contextual corroboration. Samarinas designed the Quin+ passage retrieval module, which employs embedding models and semantic similarity metrics for evidence retrieval (Samarinas et al., 2021). Similarly, Nanekhan proposed corpus compression and index compression techniques to improve retrieval efficiency through vector quantization (Nanekhan et al., 2025).

Recent advances in cross-lingual pre-trained models, such as XLM-RoBERTa (Liu et al., 2019b) and mBERT (Pires et al., 2019), have significantly advanced multilingual fact-checking research. Sawiński explored the use of fine-tuned multilingual BERT models for fact-checking retrieval tasks (Sawiński et al., 2024). Liu investigated multilingual sentence embedding representations using sentence-transformers and XLM-RoBERTa architectures, demonstrating the potential of these models in the evaluation of cross-lingual semantic similarity (Liu et al., 2022).

3 System Overview

3.1 Sentence-Transformers

Sentence-Transformers (Reimers and Gurevych, 2019) (Reimers and Gurevych, 2020) is an architecture developed based on pre-trained Transformer models, which generates fixed-dimensional sentence embeddings by adding a pooling layer. Since sentence embeddings are precomputed and stored, this approach is well-suited for efficient retrieval in large-scale scenarios.

Unlike cross-encoders, Sentence-Transformers employ a dual-encoder structure, where the embedding model with shared weights independently encodes the input sentences, mapping each fact-check and posts sentence x_i to a vector v_i. Using a cosine similarity function F, the distance between v_i and all other vectors can be calculated, thereby measuring the semantic similarity between sentences . Finally, we select the top 10 vectors v_j with the closest distances, and the corresponding fact-checks are identified as similar instances. The following functions can formulate this process:

$$v_i = S(x_i) \tag{1}$$

$$F(v_i, v_j) = \|v_i - v_j\|^2 \quad \text{or} \quad \frac{v_i \cdot v_j}{\|v_i\|^2 \|v_j\|^2} \tag{2}$$

$$v_i^{\text{closest}} = \arg \min_{j \in \{1,\ldots,N\} \cap j \neq i} F(S(x_i), S(x_j)) \tag{3}$$

The Sentence-Transformers architecture aims to bring semantically similar sentences closer together in the embedding space while pushing semantically dissimilar sentences further apart. Therefore, this architecture can be trained using methods based on contrastive learning and triplet loss to optimize the capability of calculating semantic similarity for sentence embeddings.

3.2 Embedding Models

Four high-performing embedding models were integrated into the system, including BGE-M3, gte-multilingual-base, jina-embeddings-v3, and multilingual-e5-large-instruct.

BGE-M3 (Chen et al., 2024), developed by BAAI, is a multilingual embedding model capable of handling input data at varying granularities. Its standout feature is self-knowledge distillation, which integrates relevance scores from diverse retrieval functions as teacher signals to enhance training quality. Additionally, the model employs an optimized batching strategy, enabling large batch sizes and high training throughput to improve embedding distinctiveness. The proposed system's baseline was constructed based on this embedding model, which yielded favorable results.

GTE (Zhang et al., 2024), introduced by Alibaba, is a multilingual embedding model that utilizes Rotary Position Embedding (RoPE) as its text encoder, effectively capturing semantic information in long texts. Furthermore, the model was trained and fine-tuned using contrastive learning, achieving performance comparable to BGE-M3.

Jina (Sturua et al., 2024), built on the XLM-RoBERTa architecture, was optimized for multilingual long-text and multi-task scenarios. To enhance the efficiency of long-text encoding, the model also incorporates Rotary Position Embedding (RoPE). Additionally, it supports the integration of LoRA adapters to generate task-specific embeddings, significantly reducing fine-tuning costs.

Multilingual-E5 (Wang et al., 2024), an enhanced version of multilingual-e5-large, is distinguished by its support for instruction tuning, allowing task-specific adaptation through guided instructions. This method was employed in our system to provide appropriate instruction guidance during the fine-tuning process.

3.3 Models Ensemble

We integrated the four models by computing a weighted cosine similarity (Henderson et al., 2017). Specifically, we first calculated the cosine similarity arrays individually for each model. These arrays were then summed using respective weights to obtain the final cosine similarity array, which could subsequently be used for semantic similarity calculations. The following function can represent this process:

$$F_{\text{final}} = \sum_{i=1}^{n} w_i \cdot F_i \tag{4}$$

where w_i represents the weights assigned to the cosine similarity arrays from the embedding models, respectively. Given that the performance of the four models was comparable, assigning them equal weights was considered reasonable. We set $w_1 = w_2 = w_3 = w_4$ to ensure an equal contribution from each model.

4 Experiment Details

4.1 Datasets

The datasets used for monolingual and cross-lingual tasks are consistent, with the distinction between tasks made through the [task.json] file. The dataset comprises three subsets: fact-checks, posts, and pairs. Fact-checking data includes verification results for claims made on social media or the internet, with each record containing information and content about a verified claim. The post data contains information about posts on social media platforms (such as Facebook and Twitter) and their authenticity assessments. The pairs file annotates the IDs of highly relevant fact-check-post pairs, which can be used for subsequent model training and fine-tuning.

To facilitate model training and fine-tuning, we filtered and augmented the datasets. We extracted the *claim* and *title* columns from the fact-checks dataset as key information, removed unnecessary symbols, and concatenated them to create a new fact-checks file. For the posts data, we extracted the *ocr* and *verdict* columns, performed similar filtering and concatenation, and formed a new posts file. We directly replaced the sentence IDs with the sentences for the pairs file, ultimately obtaining a highly usable training set.

4.2 Model Selection

Based on the ranking results from the MTEB leaderboard (Muennighoff et al., 2023), we selected four high-performing embedding models to integrate into our system to address monolingual and cross-lingual tasks. Specifically, our system incorporates the BGE-M3, gte-multilingual-base, jina-embeddings-v3, and multilingual-e5-large-instruct embedding models. We fine-tuned these models

Finetune	Task.json	Translate	Max_len	Epoch	Batch_size	Mono_S@10	Cross_S@10
no	no	no	512	1	8	0.62	0.57
finetune	no	translate	512	3	8	0.78	0.77
finetune	no	translate	384	3	8	0.77	0.76
finetune	no	no	512	3	8	0.89	0.83
finetune	no	no	512	4	8	0.89	0.83
finetune	task.json	no	512	4	8	**0.91**	**0.83**

Note:The [task.json] file released by SemEval2025-Task7 restricts the retrieval scope of different monolingual tasks

Table 1: The results of experimental parameters during the development phase

Model	Mono	Cross
bge-m3	0.91	0.83
gte-multilingual-base	0.92	0.83
jina-embeddings-v3	0.91	0.81
multilingual-e5-large	0.91	0.84
bge + gte	0.92	0.86
bge + gte + jina	**0.93**	**0.88**
bge + gte + jina + e5	**0.93**	**0.88**

Table 2: The success@10 of models in the dev phase

Model	Mono	Cross
bge-m3	0.892	0.698
gte-multilingual-base	0.903	0.736
jina-embeddings-v3	0.917	0.748
multilingual-e5-large	0.897	0.702
bge + e5	0.907	0.730
gte + jina	**0.922**	0.769
bge + gte + jina + e5	**0.922**	**0.770**

Table 3: The success@10 of models in the test phase

using bf16 precision and leveraged the sentence-transformers framework to map text into vector representations.

4.3 Loss Function Selection

Given that our dataset contains only positive samples, we employed MultipleNegativesRankingLoss (Henderson et al., 2017) as the loss function for fine-tuning. This loss function allows for input in the format of anchor-positive pairs and automatically samples negative examples within the batch. Through this approach, we can effectively utilize contrastive learning and triplet loss methods for training. The preprocessed posts data, which consists of a series of positive sample pairs, is highly compatible with this loss function, enabling us to achieve strong performance.

4.4 Hyper-Parameter Selection

We utilized AdamW (Kingma and Ba, 2015) as the optimizer. During the training process, we set the warmup rate to 0.1 and the learning rate to 2e-5.

Given that we employed MultipleNegatives-RankingLoss as the loss function, we configured the *batch_sampler* as *BatchSamplers. NO_DUPLICATES*, for the reason that MultipleNegativesRankingLoss benefits from having no duplicate samples within a batch.

Considering our hardware's performance and memory limitations, we set reasonable parameters for the bge-m3, gte-multilingual-base, and multilingual-e5-large-instruct models, including a train batch size of 8 and train epochs of 4. For jina-embeddings-v3, due to its excellent performance and efficiency, we increased its batch size to 32 and set train epochs to 10.

5 Main Result and Analysis

5.1 Results on Dev Dataset

During the initial development stage, experiments were conducted using the BGE-M3 model. As shown in Table 1, contrastive fine-tuning significantly enhances the adaptability of the embedding model to the task, with 3-4 rounds of fine-tuning yielding substantial improvements compared to a single round. Given that most of the dataset's text lengths are around 300 tokens, the value of max_len was further explored. Setting max_len to 512 slightly improved success@10 over 384.

A noteworthy observation is that translating the original text into English led to a decrease in success@10. This suggests that translation may disrupt the inherent linguistic characteristics, potentially hindering the performance of multilingual embedding models.

After the fine-tuning parameters were tested, experiments on the model ensemble were conducted based on the development datasets. It was observed that the prediction success@10 gradually improved

Model	Pol	Eng	Msa	Por	Deu	Ara	Spa	Fra	Tha	Tur
bge-m3	0.848	0.818	0.978	0.828	0.880	0.938	0.892	0.926	0.967	0.846
gte-multilingual-base	0.848	0.834	0.978	0.860	0.894	0.948	0.930	0.932	0.951	0.854
jina-embeddings-v3	0.864	0.850	0.978	0.862	0.910	0.952	0.932	0.940	**0.978**	**0.908**
multilingual-e5-large	0.860	0.822	0.989	0.824	0.888	0.924	0.908	0.946	0.962	0.846
bge + e5	0.870	0.830	0.989	0.840	0.906	0.940	0.922	0.950	0.962	0.862
gte + jina	0.876	**0.858**	**0.989**	**0.876**	**0.914**	**0.958**	0.934	0.946	0.973	0.898
bge + gte + jina + e5	**0.878**	0.852	**0.989**	0.874	0.904	**0.958**	**0.940**	**0.954**	0.973	0.896

Table 4: The results of each model in the test phase monolingual task

as the number of ensemble models increased. In the monolingual task, a slight improvement in success@10 was achieved by integrating three and four models, increasing from 0.9-0.92 for single models to 0.93. A more significant improvement was observed in the cross-lingual task, with success@10 rising from 0.81 to 0.83 for single models to 0.88 when three and four models were integrated. These results indicate that the ensemble of multiple embedding models significantly enhances the capability of cross-lingual fact-checking retrieval.

5.2 Results on Test Dataset

Given the larger scale and higher reference value of the test phase, more comprehensive testing was conducted. In the monolingual task, the highest success@10 achieved by a single model was limited to 0.917, whereas the ensemble of multiple models increased the success@10 to 0.922. Notably, the success@10 achieved by the Jina+GTE combination was marginally higher than that of the four-model ensemble, indicating that the inclusion of BGE and E5, which exhibited slightly inferior performance, caused a minor reduction in success@10. Overall, the enhancement from the model ensemble in the monolingual task was not particularly substantial. Examination of multiple language subtasks within the monolingual task demonstrated that the four models displayed varying strengths across different languages, and their ensemble provided a complementary effect, resulting in improved success@10.

In the cross-lingual task, the success@10 of Jina and GTE was significantly higher than that of BGE and E5. Compared to single embedding models, the ensemble of multiple models showed a more pronounced improvement, with success@10 increasing from 0.698-0.748 for single models to a maximum of 0.773. Integrating only two embedding models resulted in a significant success@10 boost. Two combinations were compared: the integration of BGE and E5, which had single-model success@10 around 0.7, improved to 0.73, while the integration of GTE and Jina, with single-model success@10 around 0.74, achieved an success@10 close to 0.77.

The ensemble experiments conducted in both the development and test phases concluded that the ensemble of multiple models substantially improves cross-lingual fact-checking retrieval more than monolingual fact-checking retrieval. Additionally, the integration of Jina and GTE achieved success@10 close to that of the four-model ensemble, offering a balanced performance and computational cost solution.

6 Conclusion

This paper delineates the contributions of the YNU-HPCC team in the *Multilingual and Cross-lingual Fact-Checked Claim Retrieval (SemEval 2025 Task 7)*. We participated in all subtasks and developed a corresponding system through the fine-tuning and ensemble of embedding models. By leveraging preprocessed data containing only essential information, we constructed a fine-tuning dataset, individually fine-tuned four embedding models, and integrated them using a weighted approach. Semantic relevance was subsequently evaluated through the computation of cosine similarity.

Our methodology demonstrated significant efficacy, achieving robust performance across all subtasks. Specifically, we attained an success@10 of 0.92 in the monolingual task and 0.77 in the cross-lingual task, exceeding the baseline performance and securing the $11th$ position on the leaderboard. In future work, we intend to explore the impact of dynamic weighting and alternative ensemble methods on the results, aiming to devise more efficient and accurate retrieval methods utilizing embedding models to enhance the scalability and effectiveness of large-scale fact-checking retrieval.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61966038 and 62266051. We would like to thank the anonymous reviewers for their constructive comments.

References

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.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient Natural Language Response Suggestion for Smart Reply. *arXiv e-prints*, arXiv:1705.00652.

Tsun hin Cheung and Kin Man Lam. 2023. Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking. *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 846–853.

Mohammed Abdul Khaliq, Paul Yu-Chun Chang, Mingyang Ma, Bernhard Pflugfelder, and Filip Miletić. 2024. RAGAR, your falsehood radar: RAG-augmented reasoning for political fact-checking using multimodal large language models. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 280–296, Miami, Florida, USA. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Weijie Li, Jin Wang, and Xuejie Zhang. 2024. YNU-HPCC at SemEval-2024 task 1: Self-instruction learning with black-box optimization for semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 792–799, Mexico City, Mexico. Association for Computational Linguistics.

Kuanghong Liu, Jin Wang, and Xuejie Zhang. 2022. YNU-HPCC at SemEval-2022 task 2: Representing multilingual idiomaticity based on contrastive learning. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 211–216, Seattle, United States. Association for Computational Linguistics.

Peng Liu, Jin Wang, and Xuejie Zhang. 2019a. YNU-HPCC at SemEval-2019 task 8: Using a LSTM-attention model for fact-checking in community forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1180–1184, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Kevin Nanekhan, V Venktesh, Erik Martin, Henrik Vatndal, Vinay Setty, and Avishek Anand. 2025. Flashcheck: Exploration of efficient evidence retrieval for fast fact-checking. *ArXiv preprint arXiv:2502.05803*.

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.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. 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.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Chris Samarinas, Wynne Hsu, and Mong Li Lee. 2021. Improving evidence retrieval for automated explainable fact-checking. In *Proceedings of the 2021 Con-*

ference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations, pages 84–91, Online. Association for Computational Linguistics.

Marcin Sawiński, Krzysztof Węcel, and Ewelina Księżniak. 2024. Openfact at checkthat! 2024: Cross-lingual transfer learning for check-worthiness detection. In *Conference and Labs of the Evaluation Forum*.

Ronit Singal, Pransh Patwa, Parth Patwa, Aman Chadha, and Amitava Das. 2024. Evidence-backed fact checking using RAG and few-shot in-context learning with LLMs. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 91–98, Miami, Florida, USA. Association for Computational Linguistics.

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

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

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.