RACAI at SemEval-2025 Task 7: Efficient adaptation of Large Language Models for Multilingual and Crosslingual Fact-Checked Claim Retrieval

Radu Chivereanu* **and Dan Tufis**
Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy
rchivereanu@gmail.com, tufis@racai.ro

Abstract

This paper presents our approach to SemEval 2025 Shared Task 7: Multilingual and Crosslingual Fact-Checked Claim Retrieval. We investigate how large language models (LLMs) designed for general-purpose retrieval can be adapted for fact-checked claim retrieval across multiple languages. This includes cases where the original claim and the fact-checked claim are in different languages. The experiments involve fine-tuning with a contrastive objective, resulting in notable gains in accuracy over the baseline. We evaluate cost-effective techniques such as LoRA, QLoRA and Prompt Tuning. Additionally, we demonstrate the benefits of Matryoshka embeddings in minimizing the memory footprint of stored embeddings, reducing the system requirements for a fact-checking engine. The final solution, using a LoRA adapter, achieved 4th place for the monolingual track (0.937 S@10) and 3rd place for crosslingual (0.825 S@10).

1 Introduction

Verifying claims across various languages is becoming increasingly difficult for fact-checkers as the amount of Internet content keeps growing. Additionally, cross-lingual fact-checking not only requires a deep understanding of nuanced linguistic differences and diverse syntactical structures, but also demands bridging significant resource gaps in less-represented languages (Huang et al., 2022).

Recent advances in large language models (LLMs) have shown promise in tackling these challenges. Besides generative capabilities, LLMs can provide high-quality textual embeddings (Wang et al., 2024a) that can be leveraged for textual retrieval (Chen et al., 2024b).

As part of SemEval 2025 Shared Task 7: Multilingual and Crosslingual Fact-Checked Claim Retrieval (Peng et al., 2025), we explore different methods to adapt general-purpose retrieval LLMs to the downstream task of fact-checked claim retrieval.

We experiment with cost-effective fine-tuning techniques, including Prompt Tuning (Lester et al., 2021), Low-Rank Adaptation (LoRA) (Hu et al., 2021), and QLoRa (Dettmers et al., 2024), focusing on improving the models performance with minimal trade-offs between accuracy and resources.

Moreover, the amount of memory required to store fact checks in large databases grows along with the number of posts and claims on social media (Lauer, 2024). To address this issue, we propose using Matryoshka learning (Kusupati et al., 2024) to compress fact representations while maintaining their retrieval utility and speeding up processing.

Our contributions are threefold:

1. We adapt a general purpose retrieval LLM for multilingual fact-checked claim retrieval by using a contrastive objective between social media posts and claims, resulting in improved performance.

2. We evaluate LoRA, QLoRA, and Prompt Tuning strategies.

3. We show that Matryoshka representation learning can help significantly reduce the memory footprint of fact-checked claims storage with minimal impact on accuracy.

The implementation is available at https://github.com/racai-ro/FactCheckRetrieval.

2 Related Work

Previous work for fact-checked claim retrieval explored systems based on classical IR-models such as BM25 and semantic similarity searches using BERT like models. Shaar et al. (2020) showed

*PhD Candidate

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 551–557

that a hybrid approach results in increased performance. They proposed a two-step retrieval pipeline, comprised of a BM25 initial retrieval and re-ranking step that takes advantage of both scores from BM25 and a fine-tuned version of sentence-BERT (Reimers and Gurevych, 2019).

Subsequent studies (Chernyavskiy et al., 2021; Mansour et al., 2022) have adopted similar approaches, incorporating semantic similarity as a standard component of the retrieval pipeline.

Notably, these systems were developed mainly for English, with some also addressing Arabic, as The CLEF 2021 CheckThat! (Shaar et al., 2021) challenge introduced a separate track for the language. This is a short-coming, given the universal aspect of the task.

To address the language limitation and enable multilingual retrieval of previously fact-checked claims, Pikuliak et al. (2023) proposed the MultiClaim dataset. The best results were obtained using a fine-tuned GTR-T5-Large model (Ni et al., 2021). They observed that most embedding models tested—even those explicitly designed for multiple languages—performed better when applied to the English-translated version of the dataset. In contrast, our work focuses **solely on the original multilingual data**.

Previous approaches have primarily relied on similarity search, which in turn depends on high-quality textual embeddings. Wang et al. (2024a) demonstrated that the effectiveness of such embeddings can be significantly improved by leveraging large language models (LLMs) for both data augmentation and embedding generation.

In line with this, BGE-Multilingual-Gemma2 (Xiao et al., 2023; Chen et al., 2024a) is a multilingual text embedding model built on the Gemma2 LLM architecture (Riviere et al., 2024). Trained on a wide variety of multilingual tasks, it has achieved state-of-the-art performance on the MIRACL benchmark (Zhang et al., 2023), which is specifically designed for multilingual retrieval.

This paper evaluates the effectiveness of adapting BGE-Multilingual-Gemma2 for claim retrieval in fact-checking, focusing on parameter-efficient tuning methods.

3 Dataset

The dataset proposed for the SemEval task builds upon and extends the original MultiClaim dataset (Pikuliak et al., 2023). It was assembled from a variety of social media platforms, with post-fact pairs formed according to the fact-check alerts issued by these platforms.

Quantitatively, the development dataset comprises 28,092 social media posts in 27 languages, 205,751 fact-checks in 39 languages authored by professional fact-checkers, and 31,305 connections between these two categories. Among these, 4,212 post-fact pairs are crosslingual.

Qualitatively, about 13% of the posts included multimedia attachments (image/video) and did not accurately convey the claims in text. In some of these cases, text was extracted from images using Optical Character Recognition (OCR), but other errors may have been introduced as a consequence. The dataset remains biased toward major languages and the Indo-European language family, despite its diversity. Furthermore, the crosslingual pairs' applicability to other language pairs is limited because they primarily consist of posts in East or South Asian languages coupled with English fact-checks.

4 Proposed Methodology

In this section, we introduce our overall strategy for our retrieval system for fact-checks: Contrastive Fine-Tuning with Low-Rank Adaptation and Matryoshka Embeddings. We separately use Prompt Tuning to evaluate the contribution of the instruction in the prompt to the baseline's performance on the task.

Specifically, we investigate:

1. **LoRa Contrastive Fine-Tuning with Large In-Batch Negatives:** We utilize Multiple Negative Ranking Loss (MNRL) to separate positive and negative pairs in the latent space, and leverage GradCache to overcome hardware memory limitations.

2. **Prompt Tuning:** We add trainable prompt embeddings and tune them for the fact-checked claim retrieval task, keeping the model's weights frozen.

3. **Matryoshka Embeddings:** To reduce the memory footprint of high-dimensional embeddings, we explore the matryoshka training approach.

In the following subsections, we provide details for each of these experiments.

4.1 Contrastive Fine-Tuning

The MNRL function is particularly well-suited for datasets that contain only positive pairs. Given that MultiClaim consists of pairs of social media posts and corresponding facts, we find the objective to be an appropriate choice.

For clarity, we briefly describe the MNRL function. Let $\mathbf{p}$ denote the post's textual embedding and $\mathbf{f}$ represent the fact-checked claim's textual embedding. To construct positive pairs, we utilize the post-fact pairs provided in the dataset. During training, the model aims to maximize the similarity between $\mathbf{p}$ and its corresponding positive $\mathbf{f}_+$, while minimizing the similarity to all other $\mathbf{f}_-$ in the batch, which act as negative examples. Consequently, the loss function can be expressed as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(\mathbf{p}i, \mathbf{f}i+)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{p}i, \mathbf{f}_j)/\tau)}, \quad (1)$$

where:

- $\text{sim}(\mathbf{p}, \mathbf{f})$: cosine similarity between $\mathbf{p}$ and $\mathbf{f}$.
- τ: temperature scaling parameter.
- N: total number of fact checks in the batch.

Contrastive learning with in-batch negatives achieves the best results with large batch sizes (Chen et al., 2022), but this approach demands significant memory resources. To address this challenge, we leverage GradCache (Gao et al., 2021). GradCache caches the encoder's output embeddings along with their corresponding gradients, thereby decoupling gradient computation for the encoder from that of the loss function. As a result, the memory footprint is significantly reduced, and large batch sizes can be simulated by processing smaller micro-batches sequentially.

Moreover, based on its training methodology, BGE-Multilingual-Gemma2 allows an **instruction** to be prepended to the query. This instruction provides a description of the task and its context. The authors originally provided the following template:

"`Given a web search query, retrieve relevant passages that answer the query.`"

In our work, we defined the prompt according to the context of the task as follows:

"`Given a social media post as a query, retrieve fact checks that verify or debunk the post.`"

The instruction prompt is delimited from the query using the special tokens `<instruct>` and `<query>`, resulting in the final model input format:

`<instruct>[PROMPT]<query>[QUERY]`.

To build the training data for our contrastive learning pipeline, we use the following inputs:

- **Social media post:** The original post text concatenated with any text extracted via OCR. For QLoRA and LoRA settings, we also prepend the instruction prompt to this combined text.
- **Fact-checked claim:** The title concatenated with the claim text.

4.2 Prompt Tuning

During development, we noticed that eliminating the instruction from the prompt for the base model had the following effects: (a) -2% in S@10 for monolingual evaluation, compared to (b) +1% for crosslingual evaluation (Table 2).

For our experiments, we defined the instruction part solely on the task's context, but in order to push the **baseline's performance** to its upper limit, we used prompt tuning to find the optimal value.

Prompt tuning (Lester et al., 2021) is a technique in which small, trainable prompt embeddings are prepended to the input of a pre-trained language model, enabling the model to adapt to specific tasks without updating its main parameters. This approach optimizes only a small set of prompt parameters while keeping the rest of the model frozen. In our case, the trainable embeddings are concatenated exclusively to the queries, positioned between the special `<instruct>` and `<query>` tokens, as illustrated in Figure 1. We use the previously defined prompt as the starting point for these embeddings.

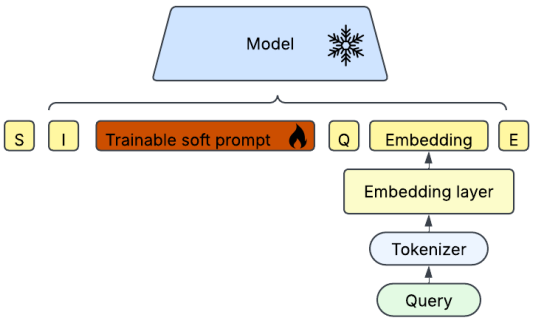

Figure 1: Prompt Tuning approach for query/anchor. S, I, Q, E are special tokens (<start>, <instruct>, <query>, <end>).

4.3 Matryoshka Embeddings

To address the memory implications associated with storing high-dimensional embeddings, we explore Matryoshka embedding learning.

This method allows for hierarchical training of embeddings at multiple resolutions, enabling the system to store compact representations when required while retaining the ability to utilize higher-dimensional embeddings when memory permits.

Specifically, we compute intermediate losses at predefined embedding dimensions: $[128, 256, 512, 768, 1024, 2048, 3584]$. The losses are summed together with equal weights, becoming the final training objective:

$$\mathcal{L}_{Matryoshka} = \sum_{d \in \mathcal{D}} \mathcal{L}_d$$

where $\mathcal{D} = \{128, 256, 512, 768, 1024, 2048, 3584\}$ denotes the set of predefined truncation dimensions applied to the post and fact embeddings, and each $\mathcal{L}_d$ is computed as specified in Equation 1.

This hierarchical loss computation encourages each subspace of the embedding to maintain a meaningful semantic structure, allowing for progressive reduction of dimensionality, limiting the loss in retrieval performance.

5 Experiments and results

In this section, we present details on the training environment, followed by an overview of our experiments and the results obtained.

5.1 Training environment

For training, we randomly split our available data (post-fact pairs) into 90% train and 10% validation.

The environment consists of 3 H100 GPUs for optimized training speed. For optimized memory, we train in bfloat16, and we use Flash Attention 2 (Dao, 2023).

Due to computation limitations, we truncate the model's input to **512 tokens**. Analyzing the character length of the posts in the dataset, we find that 95 % of the posts have fewer than 1899 characters. Averaging 4 characters per token, our truncation does not cut substantial data from the posts. The analysis is shown in Figure 2.

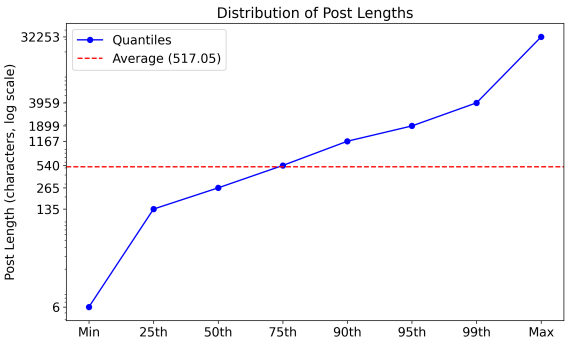

Figure 2: Character length of social media posts.

During training, we ensure **that no duplicates** (post or fact) are present in the batch per device. The hyperparameters selected for training are displayed in table 1.

Learning Rate	Scheduler	Per device batch size
0.0002	cosine	1024

Table 1: Training Hyperparameters for Fine-tuning

5.2 Finetuning

The results of our experiments for the development stage of the shared task are shown in Table 2, and those for the testing stage in Table 3. To gain a deeper understanding of our failure cases, we manually categorized errors on the development set (for our best solution) into four groups:

- **Missing Context** (independent from the verdict label): The context lacked sufficient information for non-expert humans to align the golden fact check with the post.
- **Similar**: One or more predicted fact checks fully covered the golden fact check.
- **Similar but Insufficient**: Predicted fact checks partially covered the golden fact check.
- **Missed**: Information from the golden fact check was entirely absent from the predictions.

Classification was based **on English translations**, which may introduce artifacts.

In the monolingual setting, most errors were labeled as Missing Context (51.4%) and Similar (27.1%), with fewer cases labeled as Similar but Insufficient (11.6%) and Missed (10%). Missed cases often involved longer posts with **multiple**

Technique	Model	Mono (S@10)	Cross (S@10)	Trained Params	Rank
Base	NV-Embed-v2 (Lee et al., 2024)	0.76	0.64	0	NA
Base	gte-Qwen2-7B-instruct (Li et al., 2023)	0.79	0.55	0	NA
Base	multilingual-e5-large (Wang et al., 2024b)	0.83	0.66	0	NA
Base	bge-multilingual-gemma2	0.85	0.71	0	NA
Prompt engineering	bge-multilingual-gemma2	0.87	0.70	0	NA
Prompt tuning	bge-multilingual-gemma2	0.91	0.84	75,264	NA
QLoRA	bge-multilingual-gemma2	0.95	0.91	216,072,192	64
LoRA	bge-multilingual-gemma2	0.955	**0.927**	54,018,048	16
LoRA (large)	bge-multilingual-gemma2	**0.958**	**0.927**	216,072,192	64

Table 2: Results on Shared Task Development Stage

Technique	Model	Mono (S@10)	Cross (S@10)	Trained Params	Rank
LoRA	bge-multilingual-gemma2	**0.937**	**0.82**	54,018,048	16

Table 3: Results on Shared Task Testing Stage

claims, where the model focused on non-target claims. Similar distributions were observed in the crosslingual setting (Missing Context: 42.2%, Similar: 24.5 %, Similar but Insufficient: 22.2 %, Missed: 11.1 %).

Multimodal cases (image or video information) lacking enough descriptive text were labeled Missing Context. Prediction errors were sometimes caused by similar events occurring at different times. Including posting dates could help narrow the search for relevant fact checks.

5.3 Matryoshka training

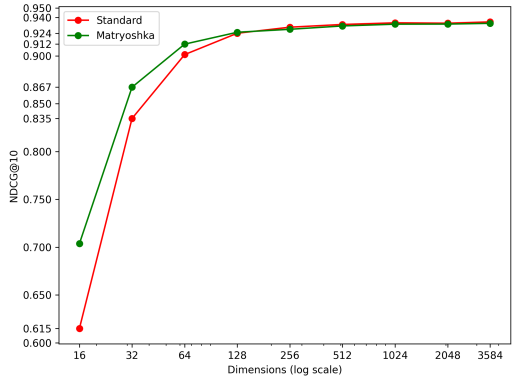

Figure 3: Performance with respect to truncated dimension.

To assess the efficiency of Matryoshka training with respect to the trade-off between embedding size and accuracy, we compared two models:

- **Standard**: bge-multilingual-gemma2 model finetuned with LoRA and MNRL loss (4.1).
- **Matryoshka**: The same strategy as Standard, also incorporating the Matryoshka training objective 4.3.

For each of the two models, we truncated the computed embeddings to different sizes and ran our evaluation. The comparative results are shown in Figure 3. We noticed that by using Matryoshka representation learning, we can reduce the embedding size and therefore memory by **98.2%** (from 3584 to 64) with **only a 2% loss** in NDCG@10.

6 Conclusions and Limitations

In this work, we successfully adapted a general-purpose retrieval LLM for multilingual and cross-lingual fact-checking through contrastive fine-tuning and parameter-efficient techniques, increasing its performance on the task. Also, using Matryoshka learning, we can significantly lower memory requirements while still achieving competitive accuracy.

However, a number of limitations still exist:

- **Bias in the dataset:** Generalizability to non-Indo-European and low-resource languages is restricted by the Indo-European bias.
- **Single-Step Retrieval:** Relying on a single-step retrieval approach limits the system's capability in handling complex, multi-hop fact-checking scenarios.

Future work should improve information extraction from images/videos, expand the dataset, and explore multi-step retrieval methods.

References

Changyou Chen, Jianyi Zhang, Yi Xu, Liqun Chen, Jiali Duan, Yiran Chen, Son Tran, Belinda Zeng, and Trishul Chilimbi. 2022. Why do we need large batchsizes in contrastive learning? a gradient-bias perspective. In *Advances in Neural Information Processing Systems*, volume 35, pages 33860–33875. Curran Associates, Inc.

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. *Preprint*, arXiv:2402.03216.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024b. 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.

Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. 2021. Aschern at checkthat! 2021: lambda-calculus of fact-checked claims. *Faggioli et al.[12]*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *Preprint*, arXiv:2307.08691.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. 2021. Scaling deep contrastive learning batch size under memory limited setup. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 316–321, Online. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Kung-Hsiang Huang, ChengXiang Zhai, and Heng Ji. 2022. CONCRETE: Improving cross-lingual fact-checking with cross-lingual retrieval. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1024–1035, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2024. Matryoshka representation learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Laurens Lauer. 2024. Understanding the global rise of fact-checking. In *Similar Practice, Different Rationales: Political Fact-Checking around the World*, pages 47–76. Springer.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *Preprint*, arXiv:2104.08691.

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

Watheq Mansour, Tamer Elsayed, and Abdulaziz Al-Ali. 2022. Did i see it before? detecting previously-checked claims over twitter. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*, page 367–381, Berlin, Heidelberg. Springer-Verlag.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2021. Large dual encoders are generalizable retrievers. *Preprint*, arXiv:2112.07899.

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.

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.

Gemma Team Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L'eonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram'e, Johan Ferret, Peter Liu, Pouya Dehghani Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stańczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Boxi Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Christoper A. Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozi'nska, D. Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Pluci'nska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost R. van Amersfoort, Josh Gordon, Josh Lipschultz, Joshua Newlan, Junsong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, L. Sifre, L. Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Gorner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, S. Mc Carthy, Sarah Perrin, S'ebastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomás Kociský, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Brian Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeffrey Dean, Demis Hassabis, Koray Kavukcuoglu, Cl'ement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118.

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, Giovanni Da San Martino, Tamer Elsayed, and Preslav Nakov. 2021. Overview of the clef-2021 checkthat! lab task 2 on detecting previously fact-checked claims in tweets and political debates.

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*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.