QUST_NLP at SemEval-2025 Task 7: A Three-Stage Retrieval Framework for Monolingual and Crosslingual Fact-Checked Claim Retrieval

Youzheng Liu[1], Jiyan Liu[1], Xiaoman Xu[1], Taihang Wang[1]*, Yimin Wang[2] and Ye Jiang[1]
College of Information Science and Technology[1]
College of Data Science[2]
Qingdao University of Science and Technology, China

Abstract

This paper describes the participation of QUST_NLP in the SemEval-2025 Task 7. We propose a three-stage retrieval framework specifically designed for fact-checked claim retrieval. Initially, we evaluate the performance of several retrieval models and select the one that yields the best results for candidate retrieval. Next, we employ multiple re-ranking models to enhance the candidate results, with each model selecting the Top-10 outcomes. In the final stage, we utilize weighted voting to determine the final retrieval outcomes. Our approach achieved 5th place in the monolingual track and 7th place in the crosslingual track. We release our system code at: https://github.com/warmth27/SemEval2025_Task7.

1 Introduction

SemEval-2025 Shared Task 7 focuses on the retrieval of monolingual and crosslingual fact-checked claims, aiming to tackle the global challenge of misinformation spread (Peng et al., 2025).

We engaged in two tracks of the SemEval-2025 Shared Task 7: monolingual and crosslingual. The monolingual track demands methods capable of retrieving the relationship between social media posts and fact-checked claims within the same linguistic environment. This task presents challenges such as noise arising from the large volume of data and difficulties related to the imbalance of language resources (Xu et al., 2024b). The crosslingual track requires methods that can retrieve fact-checked claims related to social media posts regardless of whether the language of the post matches the language of the related fact-checked claim. The primary challenge in crosslingual retrieval lies in translation inconsistencies, particulaedrly for low-resource languages (Qi et al., 2023; Magueresse et al., 2020). The absence of high-quality translation tools exacerbates the complexity of achieving accurate crosslingual semantic alignment.

To tackle the aforementioned challenges, we propose a three-stage retrieval framework. Initially, we evaluate and employ several pre-trained language models for preliminary retrieval of candidate results (Gao et al., 2024; Huang et al., 2024a), thereby mitigating the noise caused by the large data volume and alleviating the adverse effects of language resource imbalance. Subsequently, a re-ranking model is applied to refine the ranking of the candidate results, enhancing the position of fact-checked claims most relevant to the social media posts. For the crosslingual retrieval task, we utilize machine-translated data for preliminary retrieval, followed by ranking the results using a re-ranking model fine-tuned with English data. Finally, a weighted voting strategy is employed to combine the outputs from multiple re-ranking models, further enhancing the system's accuracy.

Our approach achieved 5th place in the monolingual track and 7th place in the crosslingual track, thereby validating its effectiveness and feasibility in addressing the aforementioned challenges.

2 System Description

Our approach utilizes a three-stage retrieval framework: Retrieval stage, Re-ranking stage, and Weighted Voting stage. This staged design excels at balancing retrieval efficiency and accuracy, making it particularly suitable for handling large-scale datasets. By generating candidate results during the initial retrieval stage, fine-tuning them during the re-ranking phase, and finally aggregating predictions from multiple models in the weighted voting stage, we are able to obtain the final solution. The detailed process is shown in Figure 1.

*Corresponding author

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 834–840

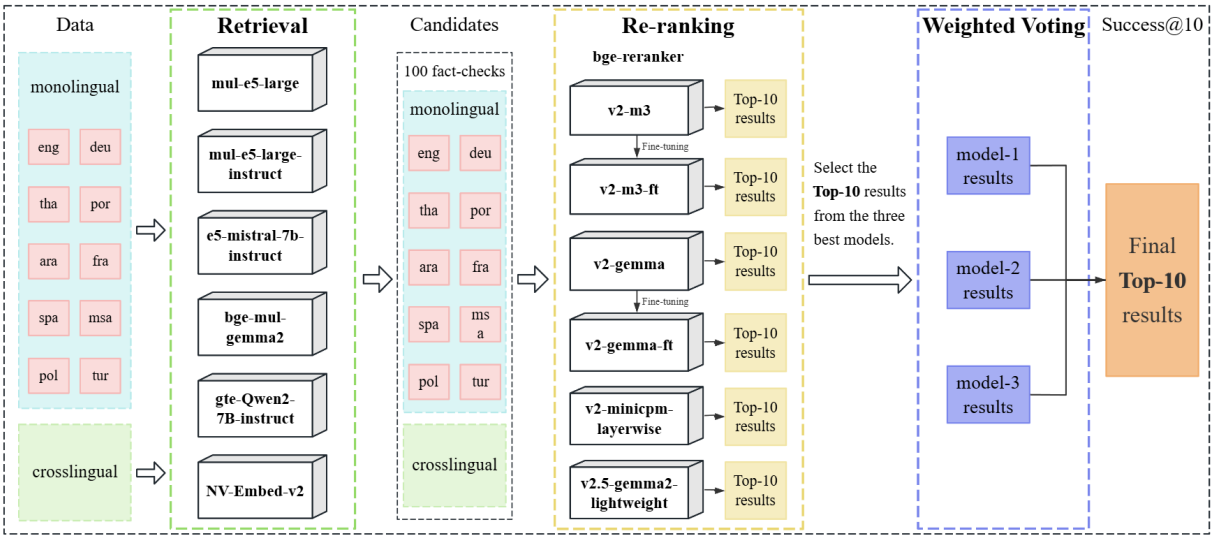

Figure 1: Illustration of the overall workflow in this paper.

2.1 Retrieval Stage

The retrieval model calculates the similarity between the query and the documents, ranking them from most to least relevant based on their similarity. This aids in filtering out a subset of candidate data from a large pool of fact-checked claims, thereby reducing noise. Accordingly, we compare several pre-trained retrieval models and employ the top-performing models in each language to retrieve candidate fact-checked claims.

The key advantage of this strategy lies in significantly reducing the computational complexity of the subsequent re-ranking phase, effectively minimizing the noise caused by the large volume of candidate fact-checked claims by pre-generating the results. The choice of the number of candidate results ensures breadth while avoiding irrelevant information, providing sufficient diversity and selection space for the re-ranking phase, thus ensuring that the final output maintains high relevance.

2.2 Re-ranking Stage

The re-ranking model can make a more refined evaluation of the results of the initial retrieval stage, put the most relevant claims at the front, and further improve the accuracy of retrieval. Therefore, we select a series of re-ranking models and fine-tuning them using the data from the training set, training a set of re-ranking models for each language (Jiang, 2023). Additionally, we combine a re-ranking model with larger parameters to re-rank the 100 candidate results generated in the initial retrieval stage. Through more precise evaluation, we improve the ranking of the most relevant fact-checked claims related to the input social media posts, thereby optimizing the retrieval results.

2.3 Weighted Voting Stage

The weighted voting strategy combines the strengths of different models through weighted fusion, minimizing the potential biases and errors inherent in individual models (Wang et al., 2023b). Therefore, we adopt the weighted voting strategy to integrate the predictions of the re-ranking models. The weight of each re-ranking model is assigned based on its performance on the validation set, with better-performing models receiving higher weights. This ensemble method leverages the strengths of multiple models, synthesizing their predictions to produce more accurate top 10 (Top-10) results. For the crosslingual task, we apply the same strategy.

3 Experimental Setup

3.1 Data

Following the guidelines, we partitioned the original dataset into both monolingual and crosslingual train and development (dev) datasets (Pikuliak et al., 2023). For the monolingual data, we further categorize it by language, producing separate sub-train and sub-dev datasets for eight languages. The data statistics are presented in Appendix A

3.2 Preprocessing

In the data preprocessing phase, we transform the CSV file into JSON format, utilize regular expressions to extract key fields containing the original text and its translation, such as "ocr", "text", "title",

and "claim", and separate them into original text, translated text, and language identifiers. Missing values (NaN) are consistently marked as null, and all text data is encoded in UTF-8 format.

3.3 Evaluation Metrics

The task employs Success@10 (S@10) as the primary evaluation metric. Specifically, when multiple correct fact-checked claims are present, the task is deemed successful if any one of the correct results appears on the Top-10 list, allowing the social media post to receive a score.

4 Experiments and Results

4.1 Retrieval

Monolingual During the development phase, we conduct systematic experiments to confirm the effectiveness of the multilingual feature combination strategy. As shown in Table 1, the combination of the original text (O) and the machine translated text (T) improves the S@10 score of the model, which means that the combination of the original text and the translated text (O, T) can improve the performance of the model on monolingual retrieval tasks (Muennighoff et al., 2022). The introduction of the translated text can help the model better understand the content of the original text, thereby improving the retrieval accuracy, especially when dealing with complex or ambiguous expressions.

Building upon the combined input of the original text and the translation, we further incorporate the "verdicts" field (V). The experimental results reveal that for several languages (Spanish, Portuguese, Malay, and Thai), the scores decrease by about 2% after including the "verdicts" field. However, the scores for English, French, and Arabic improve after incorporating the "verdicts" field, with Arabic showing an increase of approximately 5%. This suggests that in specific languages, the "verdicts" field can offer valuable supplementary information, aiding in the enhancement of retrieval accuracy.

Simultaneously, we compare the performance of several retrieval models, including mul-e5-large (Wang et al., 2024), mul-e5-large-instruct, e5-mistral-7b-instruct (Wang et al., 2023a), bge-mul-gemma2 (Xiao et al., 2023), NV-Embed-v2 (Lee et al., 2024), and gte-Qwen2-7B-instruct[1] (Li et al., 2023). As shown in Table 1, the experimental results demonstrate that e5-mistral-7b-instruct achieves the best performance in eng (85.98%), spa (91.21%), deu (80.72%), por (88.41%), fra (93.61%) and tha (97.61%), while bge-mul-gemma2 performs better in msa (91.42%) and mul-e5-large-instruct achieves the highest score in ara (89.74%). This indicates that different retrieval models exhibit specific advantages or limitations depending on the language.

Crosslingual In the crosslingual task, we compare the performance of multiple retrieval models. The results shown in Table 2 demonstrate that the effect of using pure translation input is better than mixed input (combining original and translated text) and pure original text input. Among them, e5-mistral-7b-instruct performs the best, with S@10 reaching 72.64%. In crosslingual retrieval, the semantic expression of the translation is more consistent and accurate, thereby improving retrieval performance. On the contrary, combining the original and translated text or using pure original text input may introduce noise due to language differences, resulting in reduced retrieval performance. This further highlights the key role of translation consistency in crosslingual semantic matching.

4.2 Re-ranking

We utilize the models that performe well during the retrieval phase to extract 100 fact-checked claims as candidate data from the sub-dev sets of each language. Subsequently, we employ re-ranking models from the BAAI/bge-reranker[2] series to reorder the candidate data and derive the Top-10 as the re-ranking results.

Monolingual The experimental results in Table 3 demonstrate that v2.5-gemma2-lightweight outperforms the other models, achieving an S@10 score of 92.74%. In comparison, v2-minicpm-layerwise and v2-m3 (Chen et al., 2024) exhibit weaker performance. This outcome can be attributed to the disparities in model parameters. v2.5-gemma2-lightweight benefits from a larger parameter size and enhanced learning capability, enabling it to capture semantic information more efficiently. Conversely, v2-minicpm-layerwise and v2-m3 have relatively smaller parameter sizes, which hinder their ability to handle complex retrieval tasks, likely contributing to their suboptimal performance.

We experiment with applying the rerank model directly to reorder Arabic (ara) data, resulting in an S@10 score of 85.89% . In comparison to us-

[1] https://huggingface.co/models

[2] https://huggingface.co/BAAI

Model	Plan	Avg	eng	spa	deu	por	fra	ara	msa	tha
mul-e5-large	O	82.18	78.03	81.30	79.51	80.46	84.04	80.76	78.09	95.23
	T	82.11	78.87	85.85	72.28	78.80	85.10	80.76	80.00	95.23
	O, T	84.52	74.89	86.99	79.51	83.77	85.63	85.89	86.66	92.85
mul-e5-large-instruct	O	83.14	79.07	86.99	74.69	76.49	86.17	80.76	85.71	95.23
	T	83.53	78.03	87.31	68.67	80.79	85.63	85.89	86.66	95.23
	O, T	84.16	79.49	86.50	79.51	77.81	88.82	85.89	80.00	95.23
	O, T, V	83.22	77.19	85.20	66.26	81.78	85.63	**89.74**	84.76	95.23
e5-mistral-7b-instruct	O	80.96	81.79	84.06	68.67	78.80	86.70	74.35	78.09	95.23
	T	85.20	82.00	86.99	75.90	85.43	88.29	82.05	85.71	95.23
	O, T	87.79	84.72	**91.21**	**80.72**	**88.41**	92.55	79.48	87.61	**97.61**
	O, T, V	**87.90**	**85.98**	89.91	**80.72**	87.41	**93.61**	84.61	85.71	95.23
bge-mul-gemma2	O, T	87.71	81.38	91.05	79.51	83.11	90.42	87.17	**91.42**	**97.61**
gte-Qwen2-7B-instruct	O, T	84.43	81.38	88.78	75.90	80.46	86.17	79.48	85.71	**97.61**
NV-Embed-v2	O, T	85.94	82.42	87.47	78.31	81.45	91.48	85.89	87.61	92.85

Table 1: The Success@10 (S@10) scores (%) for the monolingual track, where O uses original text, T uses translation, O,T combines both, and O,T,V includes the verdict field. **Bold** highlights the best score.

Model	Plan	Avg
mul-e5-large	O	52.89
	T	58.51
	O, T	46.92
mul-e5-large-instruct	O	57.78
	T	62.86
	O, T	58.87
e5-mistral-7b-instruct	O, T	**72.64**
bge-mul-gemma2	O, T	71.37

Table 2: The Success@10 (S@10) scores (%) of the crosslingual results. **Bold** indicates the best S@10.

ing the retrieval model alone, the rerank model does not demonstrate a clear advantage and introduces additional computational costs. This implies that relying exclusively on the rerank model does not fully utilize the system's overall performance, and combining the retrieval model with the rerank model is clearly the more effective strategy.

Additionally, we conduct fine-tuning on the complete training set for the v2-m3 and v2-gemma models. The experimental results reveal that the fine-tuned models demonstrate a significant improvement, with gains of 19.74% and 2.17% over the original models, respectively. Remarkably, the fine-tuned v2-gemma outperforms the larger parameter model v2.5-gemma2-lightweight, which provides strong evidence of the effectiveness of fine-tuning the rerank model.

Crosslingual In crosslingual, we fine-tuning the v2-m3 and v2-gemma models using the translation data in the crosslingual training set and compare them with v2-gemma and v2.5-gemma2-lightweight (Cui et al., 2025). Considering that the performance is better when only English translation data is used in crosslingual tasks, we also add a comparison with two models (v2-m3 and v2-gemma) fine-tuned on the English monolingual training set. The experimental results show that the v2.5-gemma2-lightweight model performs best with an S@10 of 80.25%, while the model fine-tuned on crosslingual data performs worse than the model fine-tuned on English monolingual data. We speculate that this difference may be attributed to the quality of the translations in the crosslingual training set, which may affect the language representation learned by the model, resulting in poor crosslingual matching performance.

4.3 Weighted Voting

Finally, we employ a weighted voting ensemble strategy to integrate the results of the monolingual and cross-lingual re-ranking models in Table 3 and Table 4. Experimental results demonstrate that the S@10 scores after integration are equal to or exceed those of the individual re-ranking models in all languages. The average of monolingual S@10 is 1.41% higher than that of the highest re-ranking model, and the crosslingual S@10 is 3.8% higher than that of the highest re-ranking model.

4.4 Evaluation on the Test Set

Test Data Augmentation For the newly introduced Polish (pol) and Turkish (tur) in the test set, we translate the original text from training set data in other languages into pol for data augmenta-

Model	Avg	eng	spa	deu	por	fra	ara	msa	tha
v2-m3	72.83	72.17	72.52	65.06	68.87	79.78	79.48	59.04	85.71
v2-m3-ft	92.57	89.33	93.82	89.15	**92.71**	**94.68**	**92.30**	93.33	95.23
v2-gemma	91.56	89.12	93.17	89.15	88.74	91.48	**92.30**	88.57	**100.0**
v2-gemma-ft	**93.73**	**89.74**	**94.95**	**93.97**	**92.71**	93.61	91.02	**96.19**	97.61
v2.5-gemma2-lightweight	92.74	89.53	93.82	90.36	92.05	93.08	89.74	93.33	**100.0**
v2-minicpm-layerwise	87.25	86.82	92.19	80.72	89.07	90.95	85.89	86.66	85.71
Voting	95.14	91.21	95.44	93.97	94.03	95.74	93.58	97.14	100.0

Table 3: Results of the re-ranking model for the monolingual track. The -ft indicates the model fine-tuned on this model. **Bold** indicates the S@10 score (%) of the best re-ranking model. **Voting** is the result of the third-stage weighted voting.

Model	Avg
v2-m3-ft(cross)	76.44
v2-m3-ft(eng)	79.16
v2-gemma	75.72
v2-gemma-ft(cross)	78.62
v2-gemma-ft(eng)	79.34
v2.5-gemma2-lightweight	**80.25**
Voting	84.05

Table 4: The results of the re-ranking model in the crosslingual track. The (cross) indicates the model fine-tuned using the crosslingual training set data, and the (eng) indicates the model fine-tuned utilizing the English monolingual training set data. **Bold** indicates the S@10 score (%) of the best re-ranking model. **Voting** is the result of the third-stage weighted voting.

tion and to fine-tune the re-ranking model (Huang et al., 2024b; Xu et al., 2024a). The results reveal that the S@10 score of the re-ranking model v2-m3, fine-tuned using the data augmentation method on Polish, is 83.2%, while the v2-gemma model achieved a score of 85.4%. Both scores are lower than the 88.6% score obtained by directly using the re-ranking model with a larger parameter size.

Furthermore, for Portuguese (por), which exhibited relatively low scores during the test phase, we aimed to augment the data by translating training data from other languages into Portuguese (Hangya et al., 2022). The experiment showed that models fine-tuned with the augmented training data achieved an S@10 score of 87.2% for Portuguese, representing a 1.4% decrease compared to the previous models and falling short of the anticipated improvement.

This suggests that while translated data can enhance the dataset's diversity, the translation process may introduce semantic distortions and information loss. The meaning and context of the original text may not be entirely preserved, leading to issues like translation inconsistency and data mismatch, which prevent the model from benefiting from the augmented training data.

Official Test Results In the final test phase, based on the official evaluation metric S@10, our approach achieved the highest score of 93.64% in the monolingual track, securing the 5th place (5/28). In the crosslingual track, our best score was 79.25%, ranking 7th (7/29), further validating the effectiveness of our method. The scores of each language are shown in Table 5.

Mono_Avg	eng	fra	deu	por	spa
93.65	89.40	95.00	90.20	89.00	94.80
tha	msa	ara	tur	pol	Cross_Avg
99.45	100.0	97.0	93.00	88.60	79.25

Table 5: Final S@10 score (%) on the official test set

5 Conclusion and Limitation

This paper introduces a monolingual and crosslingual fact-checked claim retrieval method utilizing a three-stage retrieval framework. By integrating retrieval models, re-ranking models, and weighted voting, we effectively address challenges such as data noise and imbalanced language resources. Our findings suggest that employing a mixed input strategy markedly enhances retrieval performance, while fine-tuning further optimizes re-ranking efficacy. Our method achieved 5th place in the monolingual track and 7th place in the crosslingual track.

We acknowledge that our method has limitations in terms of translation consistency and quality. Future work will focus on enhancing translation quality and refine model fine-tuning strategies to overcome these challenges.

Acknowledgments

This work is funded by the Natural Science Foundation of Shandong Province under grant ZR2023QF151 and the Natural Science Foundation of China under grant 12303103.

References

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

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, et al. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. *arXiv preprint arXiv:2502.02481*.

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. 2024. Multilingual pretraining and instruction tuning improve cross-lingual knowledge alignment, but only shallowly. *arXiv preprint arXiv:2404.04659*.

Viktor Hangya, Hossain Shaikh Saadi, and Alexander Fraser. 2022. Improving low-resource languages in pre-trained multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11993–12006.

Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, et al. 2024a. A survey on large language models with multilingualism: Recent advances and new frontiers. *arXiv preprint arXiv:2405.10936*.

Yue Huang, Chenrui Fan, Yuan Li, Siyuan Wu, Tianyi Zhou, Xiangliang Zhang, and Lichao Sun. 2024b. 1+ 1> 2: Can large language models serve as cross-lingual knowledge aggregators? *arXiv preprint arXiv:2406.14721*.

Ye Jiang. 2023. Team qust at semeval-2023 task 3: A comprehensive study of monolingual and multilingual approaches for detecting online news genre, framing and persuasion techniques. *arXiv preprint arXiv:2304.04190*.

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*.

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*.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

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 Smolen, Martin Melisek, Ivan Vykopal, Jakub Simko, Juraj Podrouzek, and Maria Bielikova. 2023. Multilingual previously fact-checked claim retrieval. *arXiv preprint arXiv:2305.07991*.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. *arXiv preprint arXiv:2310.10378*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023a. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

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

Taihang Wang, Jianxiang Tian, Xiangrun Li, Xiaoman Xu, and Ye Jiang. 2023b. Ensemble pre-trained multimodal models for image-text retrieval in the newsimages mediaeval.

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

Xiaoman Xu, Xiangrun Li, Taihang Wang, Jianxiang Tian, and Ye Jiang. 2024a. Team qust at semeval-2024 task 8: A comprehensive study of monolingual and multilingual approaches for detecting ai-generated text. *arXiv preprint arXiv:2402.11934*.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin XU, Yuqi Ye, and Hanwen Gu. 2024b. A survey on multilingual large language models: Corpora, alignment, and bias. *arXiv preprint arXiv:2404.00929*.

A Appendix

Mono	Post			Fact_check	
	train	dev	test	train, dev	test
eng	4,351	478	500	85,734	145,287
spa	5,628	615	500	14,082	25,440
deu	667	83	500	4,996	7,485
por	2,571	302	500	21,569	32,598
fra	1,596	188	500	4,355	6,316
ara	676	78	500	14,201	21,153
msa	1,062	105	93	8,424	686
tha	465	42	183	382	583
pol	-	-	500	-	8,796
tur	-	-	500	-	12,536
Cross	4,972	552	4,000	153,743	272,447

Table A1: Statistics on monolingual and crosslingual tracks. These languages are: English (eng), Spanish (spa), German (deu), Portuguese (por), French (fra), Arabic (ara), Malay (msa), Thai (tha), Polish (pol) and Turkish (tur).