QUERY-SYNERGY: Leveraging English for Improving Retrieval Performance Across Multiple Languages

Anonymous ACL submission

Abstract

We propose QUERY-SYNERGY, a training-free approach to improving retrieval performance using multilingual embeddings. Retrieval systems depend on queries that match the document language, which may not fully exploit the abundant semantic representations available in high-resource languages. Our method utilizes additional queries in English to complement source language queries, and integrates similarity scores from both queries, effectively improving retrieval performance. We evaluate our approach across five languages (Arabic, Chinese, Greek, Thai, and Turkish) using four multilingual embedding models on two datasets. Our experiments show that this approach outperforms conventional source query retrieval methods, achieving superior nDCG scores across various configurations and translation settings. These results confirm that QUERY-SYNERGY is a simple yet effective method for retrieval across multiple languages.

1 Introduction

Text embeddings are currently applied across various natural language processing applications (Karpukhin et al., 2020; Li et al., 2024; Wang et al., 2024c; Su et al., 2024; Wang et al., 2024a). Text embedding models represent documents as high-dimensional vectors, converting semantic relationships into spatial relationships between these vectors. The diverse needs for text embedding models now extend to multilingual embedding requirements. While the available resources and research advancements are most prominent in English, research is actively progressing on developing multilingual embedding models to support various languages (Wang et al., 2024b; Louis et al., 2025).

Conventionally, the performance of embedding models is evaluated in a language-coherent setting where the document and query share the same language (Reimers and Gurevych, 2019; Nogueira and Cho, 2019; Karpukhin et al., 2020). However, this approach remains susceptible to overlooking the unique advantages provided by high-resource languages with strong expressiveness. Language models have been reported to exhibit biases towards stronger representations for a particular language, especially when that language is dominant in pretraining data (Huang et al., 2023b; Yang et al., 2024b; Sharma et al., 2025). For instance, the dominance of English in pretraining data often leads multilingual embedding models to perform better on English (Park and Lee, 2025).

On the other hand, there are cases in which actively leveraging the dominance of high-resource languages such as English has shown promise in improving various multilingual NLP tasks (Seidlhofer, 2005; Liu et al., 2024; Huang et al., 2023a). Through our empirical observations, we actually find that English queries often outperform original-language queries for retrieving non-English documents. Among our experiments using multilingual embedding models, retrieving documents from Thai collections with English queries often yields better results than using queries in Thai. This indicates that although multilingual embedding models have been trained on diverse languages, they still exhibit a noticeable dependency on English representations.

These results indicate that multilingual embeddings trained across diverse languages exhibit substantial cross-lingual abilities, suggesting there remains considerable potential to further leverage these cross-lingual advantages. Motivated by this observation, we propose a QUERY-SYNERGY strategy, which enhances non-English queries by leveraging an English query; retrieval is performed independently for each language query, and the resulting document similarity scores are combined using a weighted mean to produce a unified similarity ranking.

Our extensive empirical evaluations on five ty-

pologically diverse languages including Arabic, Chinese, Greek, Thai, and Turkish from the two datasets demonstrate that our proposed QUERY-SYNERGY approach consistently achieves substantial retrieval performance gains across all examined languages and multilingual embedding models.

2 Methods

We propose QUERY–SYNERGY, a method that enhances retrieval performance in monolingual document collections by leveraging both source-language queries and English queries.

In general, retrieval is performed using a query written in the same language as the documents, referred to as the **Source** query. However, this approach may exhibit limited expressive capacity for languages with relatively insufficient training data, particularly for non-mainstream languages. By leveraging the superior semantic representation capabilities of multilingual embedding models for English, queries written in English, referred to as **Anchor** query, can yield meaningful improvements in several cases. Although anchor queries can achieve strong retrieval performance by leveraging robust semantic representations trained on abundant data, exclusively relying on anchor queries may lead to incomplete capture of document relevance, due to intrinsic cross-lingual semantic gaps.

Motivated by the complementary properties between these two query types, our proposed QUERY–SYNERGY leverages both the Source and Anchor queries into a unified retrieval result. Specifically, given a source query q_{src}, an anchor query q_{anc}, and a collection of documents D_{src} written in the source language, we first embed them using a multilingual embedding model E. We then compute the cosine similarity between each query's embedding and the embedding of every document in D_{src}, yielding two similarity vectors: one for the source query and one for the anchor query.

$$\text{sim}(q, D) = [\cos(E(q), E(d_i))]_{i=1}^{|D|}, D = d_1, d_2, \ldots, d_{|D|}$$

$$\text{sim}_{\text{src}} = \text{sim}(q_{\text{src}}, D_{\text{src}}), \quad \text{sim}_{\text{anc}} = \text{sim}(q_{\text{anc}}, D_{\text{src}})$$

To comprehensively combine the advantages of the two similarity vectors, we compute a weighted mean between them to obtain a single final similarity vector. The final ranked list of documents is generated by sorting the documents according to the values of this final similarity vector.

$$S_{\text{ranked}} = sorted\,(\lambda \cdot \text{sim}_{\text{src}} + (1 - \lambda) \cdot \text{sim}_{\text{anc}})$$

By combining similarity scores through the weighting factor λ[1], the approach leverages complementary strengths and compensates for inherent limitations of each query type. As a result, our method achieves consistent improvements in retrieval performance across multiple languages by leveraging the robust semantic representations provided by high-resource languages.

3 Experimental Setup

Models We utilize four publicly available multilingual embedding models capable of encoding high-quality semantic representations across diverse languages: bge-m3 (Chen et al., 2024), gte-multilingual-base (Zhang et al., 2024), jina-embeddings-v3 (Sturua et al., 2024), and gte-Qwen2-7B-instruct (Li et al., 2023). Despite differences in training data, architectures, and supported languages, these models uniformly exhibit robust retrieval performance.

Dataset To systematically evaluate the effectiveness of the proposed method in multiple languages, we utilize two multilingual datasets, XQuAD (Artetxe et al., 2019) and BELEBELE (Bandarkar et al., 2024), both included in MMTEB (Enevoldsen et al., 2025),which provide consistent and comparable query-document pairs across multiple languages. These benchmark collections offer fully parallel queries and corresponding documents spanning numerous languages, including English, enabling rigorous evaluations and reliable performance comparisons.

Evaluation Details To ensure a comprehensive quantitative assessment of retrieval performance across different languages, we evaluate five languages: Arabic, Chinese, Greek, Thai, and Turkish. Using these languages, we evaluate retrieval performance across three approaches (Source query, Anchor query, and QUERY-SYNERGY) on document collections consisting exclusively of documents written in each respective language. In our experiments, we use English as the Anchor.

We employ nDCG@K as a metric for evaluating performance, presenting results for $K = \{1, 3\}$. This metric allows us to evaluate how effectively the model ranks relevant documents toward the top of the retrieval list. The experiments are conducted using the MTEB (Muennighoff et al., 2022)[2], en-

[1]Where λ determines the contribution of the Source query, and $(1 - \lambda)$ that of the Anchor query.

[2]https://github.com/embeddings-benchmark/mteb

Method	XQuAD										BELEBELE									
	Arabic		Chinese		Greek		Thai		Turkish		Arabic		Chinese		Greek		Thai		Turkish	
	@1	@3	@1	@3	@1	@3	@1	@3	@1	@3	@1	@3	@1	@3	@1	@3	@1	@3	@1	@3
bge-m3																				
Source	89.0	93.7	91.5	95.2	91.1	95.2	90.3	95.0	91.9	95.5	84.1	88.9	88.3	92.8	**86.8**	91.4	84.9	90.3	87.8	92.0
Anchor	85.4	91.1	86.8	92.1	88.2	93.4	88.1	92.9	89.7	93.8	80.3	85.5	82.6	88.6	80.4	86.3	79.6	85.9	85.2	90.0
Query-Synergy	**89.4**	**94.1**	**91.9**	**95.4**	**91.4**	**95.4**	**91.4**	**95.4**	**92.0**	**95.6**	**84.6**	**89.2**	**88.7**	**93.0**	86.6	**91.4**	**85.3**	**90.5**	**88.6**	**92.2**
gte-multilingual-base																				
Source	82.0	89.1	87.2	92.1	84.8	90.9	82.5	89.4	85.2	90.9	78.6	85.2	87.2	91.2	81.4	86.7	78.7	85.4	83.0	88.1
Anchor	83.2	90.3	84.4	90.0	84.5	91.6	83.3	90.3	85.3	91.5	73.7	81.0	82.1	88.2	78.1	84.4	74.2	81.1	78.3	85.8
Query-Synergy	**87.7**	**93.0**	**88.2**	**92.9**	**87.1**	**92.9**	**87.8**	**93.2**	**87.6**	**92.7**	**79.1**	**85.8**	**87.8**	**91.6**	**81.7**	**87.2**	**79.9**	**86.3**	**83.3**	**88.7**
jina-embeddings-v3																				
Source	85.7	91.4	87.1	92.8	88.1	92.9	85.8	92.4	88.5	93.3	83.4	88.0	85.2	89.9	85.9	90.4	83.1	88.5	87.2	91.2
Anchor	86.4	92.1	85.2	91.6	88.2	93.4	86.6	92.4	88.2	93.5	78.4	84.7	81.9	87.5	80.7	86.2	78.3	84.0	83.3	88.0
Query-Synergy	**88.7**	**93.4**	**87.8**	**93.7**	**89.8**	**94.0**	**88.7**	**94.0**	**89.4**	**94.0**	**83.8**	**88.1**	**86.0**	**90.5**	**86.0**	**90.4**	**83.1**	**88.6**	**87.4**	**91.5**
gte-Qwen2-7B-instruct																				
Source	89.7	94.3	93.7	96.9	85.7	91.2	88.7	93.9	88.9	93.5	89.6	93.5	93.6	95.9	**86.2**	89.5	86.7	90.8	88.4	92.4
Anchor	**92.4**	95.5	92.8	96.6	88.5	**93.7**	91.8	95.8	91.2	95.4	85.8	90.9	90.4	94.1	81.4	86.5	86.0	90.0	85.0	89.8
Query-Synergy	91.8	**95.7**	**94.6**	**97.4**	**88.5**	93.3	**92.0**	**95.9**	**91.4**	**95.6**	**89.8**	**93.5**	**93.7**	**96.0**	85.8	**89.6**	**87.1**	**91.0**	**88.7**	**92.6**

Table 1: Evaluation of multilingual embedding models on the XQuAD and BELEBELE benchmarks, reporting nDCG@1 and @3 scores for Source, Anchor, and QUERY-SYNERGY. The highest scores are highlighted in bold.

suring consistency and reproducibility with standardized evaluation metrics and procedures.

4 Results

4.1 Main Comparisions

Table 1 presents retrieval performance results comparing our proposed QUERY-SYNERGY approach with retrieval using source and anchor queries. The source query employs queries in the same language as the document, a conventional retrieval method. It demonstrates robust performance across multiple languages, achieving particularly high performance in Chinese, a language with abundant resources, while showing relatively lower performance in Thai, a language with limited resources. The anchor query enhances retrieval performance using English queries. In some cases, such as Arabic and Thai in XQuAD, anchor queries outperform source queries in nDCG@1 and nDCG@3, suggesting that English representations can alleviate limitations in retrieval for low-resource languages. However, anchor queries alone do not consistently surpass source queries across all languages and models. Performance variations exist due to factors including data resources, linguistic similarity, and other language-specific characteristics. These results highlight the strengths and limitations intrinsic to retrieval approaches based exclusively on either source or anchor queries.

Our proposed QUERY-SYNERGY enhances retrieval by performing a weighted mean of similarity results from both source and anchor queries, consistently achieving the highest nDCG scores across most languages and models. By integrating the complementary abilities of both source queries and high-resource language representations (from anchor queries), QUERY-SYNERGY effectively combines their strengths, compensating for individual limitations. This approach captures relevant information neither query type could retrieve alone, consistently delivering improved overall retrieval performance. Furthermore, our experiments confirm that QUERY-SYNERGY achieves these gains without additional training data or fine-tuning.

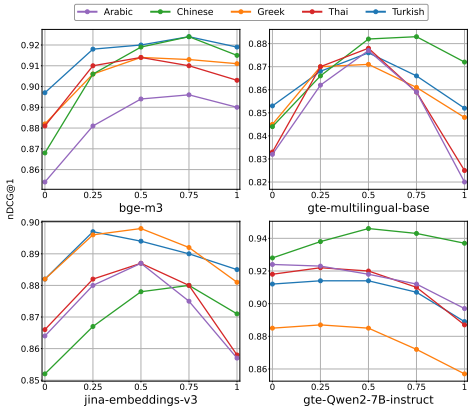

Figure 1: nDCG@1 performance on XQuAD for varying λ (0: English, 1: Source Language).

4.2 Effect of Balancing Lambda

Figure 1 presents the changes in nDCG@1 for each model with varying weighted average λ values. We find that setting λ between 0.25 and 0.75 generally yields higher scores than source query retrieval with only the source language ($\lambda = 1$) or only English ($\lambda = 0$). This indicates that a balanced combination of information from both languages significantly enhances retrieval performance. Additionally, consistent improvements are observed across all models, with some achieving optimal results at intermediate λ values. These findings

Anchor \ DB	Arabic	Chinese	Greek	Thai	Turkish
Arabic	0.00	-0.66	-0.75	-0.52	-0.15
Chinese	-0.01	0.00	-0.11	0.96	0.00
Greek	1.21	-0.55	0.00	0.89	0.11
Thai	-0.30	-0.44	-0.99	0.00	-0.44
Turkish	-0.39	-0.44	0.00	1.33	0.00
English	0.45	0.44	0.33	1.22	0.11
Espanol	0.46	-0.17	-0.11	0.69	0.67
Vietnamese	-0.48	-0.72	-0.75	0.41	0.39
Hindi	-0.11	-1.37	-0.57	-0.06	-0.24

Table 2: Comparison of performance change rates ($\Delta\%$) across anchor languages relative to source queries, using bge-m3 ($\lambda = 0.5$) on the XQuAD

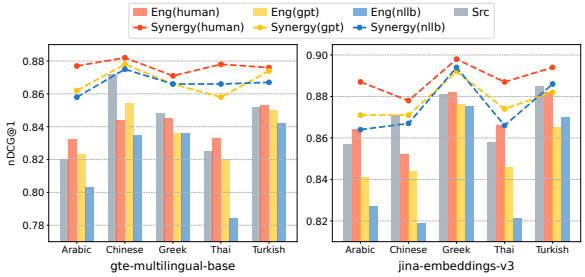

Figure 2: Comparison of performance by translation models on XQuAD. The bars and lines represent Source query and QUERY-SYNERGY ($\lambda = 0.5$).

support that QUERY-SYNERGY is more effective than relying on a single language. Moreover, adjusting the λ values for each language can enhance performance robustness.

4.3 Anchor Language Comparison

Table 2 presents the changes (%) in nDCG@1 performance compared to using the source query when different anchor languages are integrated in QUERY-SYNERGY. Anchor choices significantly affect retrieval results: high-resource anchors such as English and Spanish generally enhance performance across language pairs, exemplified by English improving Thai retrieval by approximately 1.22%. Conversely, lower-resource anchors like Arabic for Thai (0.52%) or languages such as Thai, Turkish, Vietnamese, and Hindi often result in minor improvements or performance declines.

These results show that while high-resource languages clearly lead to notable retrieval improvements. Additionally, careful selecting anchor languages based on structural and lexical properties with the source language is important for achieving consistent improvements in QUERY-SYNERGY.

4.4 Robustness to Translation Variation

We explore whether significant performance improvements can be achieved through query translation and the QUERY-SYNERGY, even in cases where human translation is unavailable. Figure 2 compares retrieval performance of queries translated by human experts provided within the original dataset (`human`), GPT-4o[3] (OpenAI, 2024), and NLLB-200[4] (Team et al., 2022), respectively.

Our findings show clear performance variations across different translation models. For example, for Thai retrieval using the `gte-multilingual-base` model, human translation achieves an nDCG@1 of 0.833, outperforming `gpt`(0.820) and `nllb`(0.784). However, incorporating these translated queries via our proposed QUERY-SYNERGY enhances retrieval results, yielding scores of 0.887(human), 0.874(`gpt`), and 0.866(`nllb`). Similar improvements are consistently observed across different embedding models such as `jina-embeddings-v3`. Simultaneously, regardless of the translation model used, QUERY-SYNERGY consistently achieves better performance compared to the source query only. This demonstrates the practicality and effectiveness of our method for real-world applications, offering robust retrieval performance improvements without dependency on specific translation models.

5 Conclusion

In this paper, we propose QUERY-SYNERGY, a method that leverages the representational ability of English in single multilingual embeddings to improve retrieval performance. To be specific, we compute similarity scores between the Source and Anchor queries and all documents, then combine them using a weighted average to produce enhanced retrieval results. Through various experiments involving anchor weight lambda tuning, varying anchor languages, and employing multiple machine translation models, our proposed method is validated to consistently improve retrieval performance across diverse multilingual embedding models and datasets. Overall, QUERY-SYNERGY effectively exploits existing multilingual embeddings without requiring additional training or data augmentation, offering a simple yet practical approach across multiple languages.

[3]The prompt used can be found in Appendix B.

[4]https://huggingface.co/facebook/nllb-200-3.3B

Limitations

Our study is limited to evaluating QUERY-SYNERGY with five languages: Arabic, Chinese, Greek, Thai, and Turkish. Extending the evaluation to a broader set of languages in future research would improve our understanding of the method's general applicability. Additionally, we use GPT-4o and NLLB to translate source queries in our translation variation analysis, which may be subject to translation quality issues, such as failing to fully capture subtle cultural nuances or context-specific meanings. Such limitations can potentially affect retrieval performance. Moreover, our experiments are conducted on only two parallel datasets that cover all selected languages, potentially limiting the generalizability of our findings. Future analyses involving diverse datasets, domains, and language combinations would further clarify the practical applicability and robustness of our approach.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, and 67 others. 2025. Mmteb: Massive multilingual text embedding benchmark. *Preprint*, arXiv:2502.13595.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023a. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394.

Zhiqi Huang, Puxuan Yu, and James Allan. 2023b. Improving cross-lingual information retrieval on low-resource languages via optimal transport distillation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, page 1048–1056. ACM.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and 1 others. 2022. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249.

Shuo Li, Sangdon Park, Insup Lee, and Osbert Bastani. 2024. TRAQ: Trustworthy retrieval augmented question answering via conformal prediction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3799–3821, Mexico City, Mexico. Association for Computational Linguistics.

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*.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024. Is translation all you need? a study on solving multilingual tasks with large language models. *arXiv preprint arXiv:2403.10258*.

Antoine Louis, Vageesh Kumar Saxena, Gijs van Dijck, and Gerasimos Spanakis. 2025. ColBERT-XM: A modular multi-vector representation model for zero-shot multilingual information retrieval. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4370–4383, Abu Dhabi, UAE. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Jeonghyun Park and Hwanhee Lee. 2025. Investigating language preference of multilingual rag systems. *Preprint*, arXiv:2502.11175.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Preprint*, arXiv:1606.05250.

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.

Barbara Seidlhofer. 2005. English as a lingua franca. *ELT journal*, 59(4):339–341.

Nikhil Sharma, Kenton Murray, and Ziang Xiao. 2025. Faux polyglot: A study on information disparity in multilingual large language models. *Preprint*, arXiv:2407.05502.

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. *Preprint*, arXiv:2409.10173.

Zhenpeng Su, Xing W, Wei Zhou, Guangyuan Ma, and Songlin Hu. 2024. Dial-MAE: ConTextual masked auto-encoder for retrieval-based dialogue systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 820–830, Mexico City, Mexico. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Hongru Wang, Wenyu Huang, Yang Deng, Rui Wang, Zezhong Wang, Yufei Wang, Fei Mi, Jeff Z. Pan, and Kam-Fai Wong. 2024a. Unims-rag: A unified multi-source retrieval-augmented generation for personalized dialogue systems. *Preprint*, arXiv:2401.13256.

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

Yuhao Wang, Ruiyang Ren, Junyi Li, Xin Zhao, Jing Liu, and Ji-Rong Wen. 2024c. REAR: A relevance-aware retrieval-augmented framework for open-domain question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5613–5626, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024a. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Jinrui Yang, Fan Jiang, and Timothy Baldwin. 2024b. Language bias in multilingual information retrieval: The nature of the beast and mitigation methods. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 280–292, Miami, Florida, USA. Association for Computational Linguistics.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, and 1 others. 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.

A Experiments Details

A.1 Models

In our experiments, we employed four multilingual embedding models that vary substantially in their architectures, training data, supported languages, and embedding dimensions. Specifically, the bge-m3 (Chen et al., 2024) model utilizes the XLM-RoBERTa architecture (Conneau et al., 2019), supporting over 100 languages and delivering embeddings of 1024 dimensions. The gte-multilingual-base (Zhang et al., 2024) employs an encoder-only Transformer architecture trained on approximately 70 languages; it creates 768-dimensional embeddings designed with a focus on computational efficiency. Jina-embeddings-v3 (Sturua et al., 2024), also based on the XLM-RoBERTa architecture, was fine-tuned on datasets covering over 30 languages. Notably, it integrates Rotary Position Embeddings along with Matryoshka Embeddings (Kusupati et al., 2022), enabling dynamic embedding dimensions adjustable from 32 to 1024. Lastly, the gte-Qwen2-7B-instruct (Li et al., 2023) model is built upon the decoder-only transformer architecture qwen2-7b (Yang et al., 2024a), trained on multilingual data from diverse domains. It handles sequences up to 32,000 tokens and produces

embeddings of dimension 3584. Despite clear differences regarding supported languages, embedding dimensions, model architectures, and training corpus specifications, these models uniformly demonstrate strong capabilities in generating reliable multilingual embeddings.

A.2 Datasets

In our experiments, we adapted datasets originally designed for question answering (QA) and machine reading comprehension (MRC) tasks into query-document retrieval benchmarks. Specifically, XQuAD (Artetxe et al., 2019) originates from the validation set of the SQuAD v1.1 dataset (Rajpurkar et al., 2016), and has been manually translated into ten different languages, such as Spanish, German, and Greek. The resultant dataset includes fully parallel contexts along with corresponding question-answer pairs across all ten languages, significantly reducing linguistic discrepancies and ensuring uniformity in comparative analyses.

The BELEBELE dataset (Bandarkar et al., 2024) was initially created for multilingual machine reading comprehension and contains data translated into a wide array of 122 different languages. Due to its fully parallel structure, BELEBELE offers a precise mechanism for evaluating variations in linguistic expressions between languages, which is particularly useful for examining performance differences across multiple translated versions.

A.3 Hardware

We utilized one NVIDIA A6000 GPU with 48GB memory capacity and AMD EPYC 7513 32-core Processor CPUs in this experiments.

B Prompts

This section details the prompts used for translations in our experiments, ensuring consistency and replicability across different models. These prompts were employed by the translation models to obtain English translations of source queries, which then served as anchor queries for the QUERY-SYNERGY approach.

Translation Prompt

Translate the following query from `{Source Language}` to English.

please return only the translated question.

Input Sentence: `{QUERY}`

Output Sentence: