Multilingual Information Retrieval with a Monolingual Knowledge Base

Yingying Zhuang[1,*], Aman Gupta[1] and Anurag Beniwal[1]

[1]Amazon.com, Inc., San Francisco, CA, USA

Abstract

Multilingual information retrieval has emerged as powerful tools for expanding knowledge sharing across languages. On the other hand, resources on high quality knowledge base are often scarce and in limited languages, therefore an effective embedding model to transform sentences from different languages into a feature vector space same as the knowledge base language becomes the key ingredient for cross language knowledge sharing, especially to transfer knowledge available in high-resource languages to low-resource ones. In this paper we propose a novel strategy to fine-tune multilingual embedding models with weighted sampling for contrastive learning, enabling multilingual information retrieval with a monolingual knowledge base. We demonstrate that the weighted sampling strategy produces performance gains compared to standard ones by up to 31.03% in MRR and up to 33.98% in Recall@3. Additionally, our proposed methodology is language agnostic and applicable for both multilingual and code switching use cases.

Keywords

Information Retrieval, Text Embedding, Contrastive Learning

1. Introduction

The task of Information Retrieval is to find relevant information from a large collection and a knowledge base often complements this retrieval task and facilitates various downstream tasks as well. With the recent emergence of Large Language Models (LLMs) as a powerful tool to build dialogue systems [1, 2, 3] and conversation AIs, the retrieved information can play a critical role in increasing practicality and reliability of these systems. Application of LLMs has also achieved remarkable advancements in multilingual scenarios and many of the current frontier LLMs, such as Anthropic's Claude 3 [4] and Cohere's Command A [5] are focused to excel in the multilingual settings to support global businesses. On the other hand, because knowledge base construction is labor-intensive and expensive [6], imbalanced data, quality and scarcity issues become the bottleneck for rapid knowledge base development, especially for low-resource languages. Additionally there has been a surge of interest in code-switching recently where communicators switch between two or more languages during linguistic interactions [7, 8]. This communication style is common in bilingual and multilingual communities [9, 10]. For example, in US mixing Spanish and English is a common phenomenon while in Canada mixing French and English has been often observed. However, it remains a challenging task in Information Retrieval as the language IDs are not specified and knowledge base with code-switching context is scarce. Existing work has been mainly focused on automatic multilingual knowledge base construction [11, 12] where knowledge bases of low-resource languages are automatically propagated based on knowledge bases in well-populated high resource languages.

Contributions Different from previous approaches, we propose a novel embedding model development pipeline to align multilingual information retrieval with a **monolingual** knowledge base, therefore removing the bottleneck of multilingual knowledge base construction in information retrieval

GENNEXT@SIGIR'25: The 1st Workshop on Next Generation of IR and Recommender Systems with Language Agents, Generative Models, and Conversational AI, Jul 17, 2025, Padova, Italy

*Corresponding author.

yyzhuang@amazon.com (Y. Zhuang); amangta@amazon.com (A. Gupta); beanurag@amazon.com (A. Beniwal)

https://www.linkedin.com/in/yingyingzhuang/ (Y. Zhuang)

0000-0002-9420-7285 (Y. Zhuang)

systems. Specifically, we design a weighted sampling strategy to select positive and negative pairs for contrastive learning, which guides the model to embed queries with similar semantic meanings into close embedding vector space across different languages. Our approach aims at a more data-efficient multilingual information retrieval system which only requires a high resource language (e.g., English or French) for the knowledge base. Experimental results demonstrate that our approach enables an embedding model to properly handle multilingual queries. The ability to share knowledge in multiple languages is a fundamental component to ensure the development of Conversation AIs systems and to enable both businesses and individuals to reach a broader range of communities, fostering inclusivity, accessibility and collaboration among international teams.

2. Methodology

Embedding models such as E5 [13], GTE [14], Nomic [15], and Arctic-Embed [16] are typically trained with three stages: pretraining via masked language modeling, large scale contrastive pretraining with in-batch negatives, and finally quality focused contrastive fine-tuning [17, 18, 19]. Our work focuses on the contrastive fine-tuning stage and the proposed methodology can be applied on any open-source text embedding models. One key ingredient for effective embedding is supervised fine-tuning with a high quality dataset which contains explicitly labeled positive and negative pairs. Specifically, we focus on the query-to-query multilingual retrieval problem with a monolingual knowledge base.

Given a user query q in a target language T, the goal is to retrieve a similar example query q_{EX} from the knowledge base in language E where E is different from T. The knowledge base contains a set of example queries and their corresponding labels

$$KB_E = (q_{\text{EX}_1}, l_1), (q_{\text{EX}_2}, l_2), ..., (q_{\text{EX}_N}, l_N).$$

For example, a key component of task-oriented dialogue systems is to identify the underlying purpose or intent of a user in a conversation [3, 20]. To facilitate information retrieval business can build a monolingual intent knowledge base in English which contains a set of customer query and customer intent pairs based on high quality data source or human annotation. As business expands to other languages, it becomes infeasible to build a separate knowledge base in each language therefore the goal is to share the English knowledge base across all languages. If a user input query is in Spanish (the target language T), the goal is to retrieve a semantically similar English query in the knowledge base, q_{EX_j}, and the corresponding intent label l_j. In order to build a multilingual Information Retrieval system with an English-only (or any other monolingual) knowledge base, it is crucial for the embedding model to map English and non-English queries into a shared representation space to enable non-English retrieval.

Contrastive Training Data In contrastive learning, the model learns by distinguishing between similar and dissimilar queries from positive and negative pairs. An effective embedding model will project similar (positive) pairs closer together and dissimilar (negative) pairs further apart in a embedding vector space. For positive pairs, existing work often uses a noisy data source to identify relevant queries, then applies some heuristic and consistency quality filters to improve data quality [15, 21, 22, 23]. For negative pairs, [24] demonstrated the importance of training on hard negatives, where "hard" refers to the fact that it is not trivial to determine their lower relevance relative to the positive examples, therefore, many existing work has been following this paradigm to identify the hardest negatives for each training example. For example, [16] leveraged a pre-existing text embedding model to identify and score the hardest negatives for each training example while NV Retriever [25] also added a filtering step where any negative with a relevance score exceeding a specified percentage of the known-positive's score is discarded as a potential false negative.

Contrastive Training Data Generation In our methodology, we propose a new strategy to leverage the available labels, $l_1, l_2, ..., l_n$, in the monolingual knowledge base to construct positive and negative

multilingual pairs. Our approach demonstrates the benefits of a weighted sampling strategy of negative mining based on relevance, which in Section 3 we show that it is more effective than focusing purely on selecting the "hardest" negatives possible, or a rank for relevance scores. Algorithm 1 shows the details of our proposed approach step-by-step and Figure 1 is an illustration with examples.

Algorithm 1 Contrastive Training Data Generation

• **Require:**
A knowledge base in language E with N example query and label pairs

$$KB_E = (q_{\text{EX}_1}, l_1), (q_{\text{EX}_2}, l_2), ..., (q_{\text{EX}_N}, l_N).$$

A set of M unlabelled queries q's in a target language T $(q_1, q_2, ..., q_M)$.
• **Data Generation**
◦ **Step 1**: **Initialize** an empty paired dataset $P = \{\}$
◦ **Step 2**: **Split** KB_E into index and training set. We split the N example query and label pairs into two subsets N_1 and N_2, reserve the N_1 pairs for retrieval index, while using the remaining N_2 for systhetic data generation.
◦ **Step 3**: **Generate Positive Pairs**: For each query q_i in N_2 which is in language E, use an LLM to translate it to the target language T into q_i^T. Randomly sample one query from all the queries sharing the same label in the Index set N_1, $q_i^{'}$. Update P with the positive pair $(q_i^T, q_i^{'})$, i.e. $P = P \cup (q_i^T, q_i^{'})$.
◦ **Step 4**: **Generate Negative Pairs based on weighted sampling**: For each query q_i in N_2, use an LLM to translate it to the target language T into q_i^T (or reuse q_i^T from Step 3). For each query q_j in the index set N_1, calculate a similarity score s_{ij} between q_i's label l_i and q_j's label l_j. We generate two types of negative pairs:

- **Random negative**: Randomly sample one query $q_i^{''}$ from all the queries in set N_1 with a different label than l_i.
- **Hard negatives**: Randomly sample k queries from the index set N_1 based on sampling weight s_{ij}. The higher the similarity score s_{ij}, the more likely a query is to be sampled. The intuition is that queries with similar labels are harder to distinguish, making them better candidates for hard negatives.

With $k = 2$, we yield three negative pairs in total: one random negative pair $(q_i^T, q_i^{''})$ and two hard negative pairs $(q_i^T, q_i^{'''})$ and $(q_i^T, q_i^{''''})$. This maintains our desired positive to negative ratio of 1:3 when using contrastive loss. Update P with these negative pairs, i.e., $P = P \cup \{(q_i^T, q_i^{''}), (q_i^T, q_i^{'''}), (q_i^T, q_i^{''''})\}$.
◦ **Step 5**: **Synthetic Data Augmentation**: Compared to the abundance of unlabeled data, high-quality labeled examples in KB_E are more scarce. To address this data scarcity issue, we also use synthetic data creation to construct additional positive and negative pairs. For any query in the target language T, use a LLM to generate a semantically similar query to form a positive pair, and 3 semantically different queries in language E to form 3 negative pairs. Update P with the these pairs.

3. Experiments and Results

In this paper, we conduct experiments to evaluate the plausibility of our proposed methods for the business use case of performing Information Retrieval for Hinglish queries using an English knowledge base. Hinglish refers to the conversation style where speakers are mixing Hindi and English in one conversation interaction and often within one single sentence, which is a very common way of communication for the India marketplace.

Our proposed methodology is language agnostic and can be applied to any language with low-resource constraints. Therefore, we simply treat code-switching as a new language. In our experiments we fine-tune an embedding model optimzed to embed English and Hinglish queries into the same vector space for Information Retrieval.

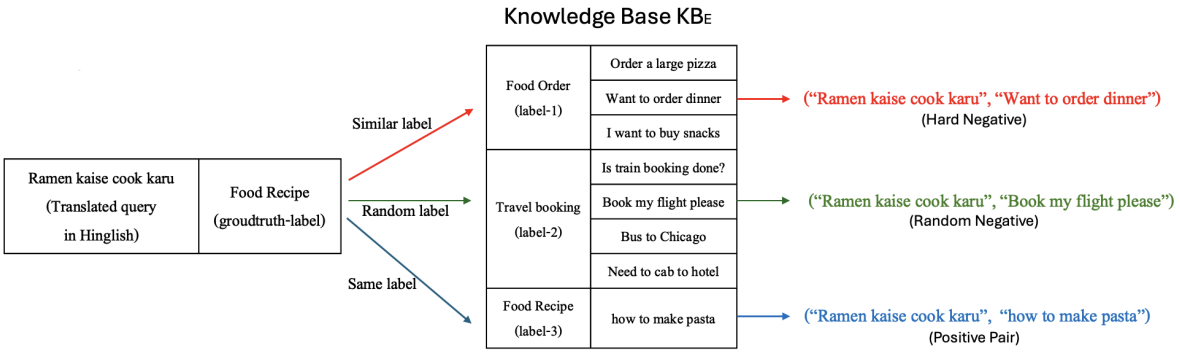


Figure 1: Illustration of our contrastive training data generation algorithm. For each translated query (left), we create a positive pair (blue) with queries sharing the same label, k hard negative pairs (red) from queries with similar labels selected through weighted sampling, and 1 random negative pair (green) from queries with different labels selected randomly from the knowledge base.

3.1. Datasets

Data Anonymization Due to business considerations, we are not permitted to share the results using the original customer queries. As a result, we manually anonymized both the labels and transcripts to ensure no personal information is included. Additionally, specific product and service names were denonymized to prevent the identification of the company from the transcript or label descriptions. Despite these modifications, the conclusions drawn from our experiments remain valid.

We collect about 35,000 dialogue transcripts with customer intent labels in English to construct our KB_E. Following the steps in Algorithm 1 we split these transcripts into two subsets: 3000 to construct the retrieval index and the remaining 32,000 to generate constrastive training data. For Step 5, we collect 10,000 unlabelled dialogue transcripts in Hinglish for data augmentation.

3.2. Experiments

As discussed earlier, we proposed Algorithm 1, which employs a hybrid approach combining both hard and random negative mining, along with synthetic data generation. This mixed strategy helps reduce excessive reliance on translated queries alone, with synthetic data generation performed directly on Hinglish data available natively.

In our experimental setup, we first evaluated several open-source multilingual retrieval models as the foundation for our fine-tuning process:

- `paraphrase-multilingual-mpnet-base-v2`
- `multilingual-e5-base`
- `paraphrase-multilingual-MiniLM-L12-v2`
- `stsb-xlm-r-multilingual`

To validate the effectiveness of our proposed approach and to better understand the contribution of different components, we conducted ablation studies comparing Algorithm 1 against several alternative strategies:

Negative Sampling Variations

- **Random Negative Mining:** This represents the simplest approach where we set $k = 0$ in Step 4 of Algorithm 1, selecting all negatives randomly with equal probability regardless of semantic similarity.
- **Hard Negative Mining:** In this variation, we set $k = 3$ in Step 4, selecting all negatives using weighted sampling based on similarity scores. This targets examples that are more challenging to distinguish.

- **Hardest Negative Mining:** Taking the hard negative approach further, we not only set $k = 3$ but also restrict sampling to only the top-3 most similar labels (excluding exact matches). This focuses exclusively on the most challenging negative examples.

Data Source Variations To evaluate the impact of synthetic data augmentation described in Step 5, we also tested:

- **Labeled Data Only:** This variant trains solely on the high-quality translated examples from the original English dataset, omitting the synthetic data generation Step 5.
- **Synthetic Data Only:** Conversely, this approach relies exclusively on synthetically generated data in Step 5, representing the scenario where no labeled data is available in any language.

These methodical comparisons allowed us to isolate the contribution of each component in our algorithmic design and confirm the superiority of our hybrid approach.

3.3. Implementation Details

We fine-tuned various pretrained multilingual embedding models using the InfoNCE contrastive loss[26] with a maximum sequence length of 512 tokens. Training was performed on 4 NVIDIA A10G GPUs using PyTorch's DistributedDataParallel framework with a batch size of 32. Each model was trained for 15 epochs, with early convergence typically observed around epochs 8-9. For optimization, we used AdamW with a learning rate of 2e-5 and a linear warmup period. Claude Sonnet 3.5[1] was employed for both query translation and generation of synthetic positive and negative pairs for training.

3.4. Evaluation Metrics

For evaluation, we compute Recall@1, Recall@3, Recall@5, and Recall@10 metrics, along with Mean Reciprocal Rank (MRR). These metrics are particularly appropriate for our task since we have a single ground truth label for each query, while our retrieval system returns ranked predictions. Recall@k measures whether the correct label appears within the top-k retrieved results, providing a clear assessment of our model's retrieval performance at various thresholds of interest.

3.5. Results and Discussion

Figure 2 shows the performance comparison of different multilingual embedding models during training. The *multilingual-e5-base* model consistently outperforms other models across all epochs, reaching the highest Recall@1 of approximately 0.54 by the final epoch. The *paraphrase-multilingual-mpnet-base-v2* model performs comparably well, while *stsb-xlm-r-multilingual* and *paraphrase-multilingual-MiniLM-L12-v2* show significantly lower performance. Based on these results, we selected *multilingual-e5-base* as our foundation model for all subsequent experiments. Appendix Section A presents the comparison plots between these models on Recall@3, Recall@5, Recall@10, and MRR, which show a similar pattern as Recall@1.

The evaluation results for different negative sampling and data strategies are presented in Table 1. Our ablation study reveals that combining random and hard negatives yields better performance than approaches using only one type of negative examples. While the Labeled Data Only approach performs comparably to our proposed method, we attribute this to the composition of our test set, which primarily contains samples similar to the labeled data. Nevertheless, Algorithm 1's hybrid approach provides better protection against performance degradation on original queries in the low-resource language. The significantly poorer performance of the purely synthetic data approach indicates the crucial importance of high-quality labeled examples in the training process.

Why does mixed-mining strategy work better? The superior performance of Algorithm 1 can be attributed to its hybrid approach. By combining both random and hard negatives, the model

[1] https://www.anthropic.com/news/claude-3-5-sonnet

optimizes the embedding space globally while simultaneously emphasizing differentiation in local neighborhoods. Pure hard negative mining (0.4613) or hardest negative mining (0.4012) approaches underperform because they overly focus on disambiguating local neighborhoods without maintaining global embedding structure. Additionally, the incorporation of synthetic data in Algorithm 1 helps improve performance on low-resource language queries, reducing overfitting to the translated query distribution while maintaining strong performance on the original dataset.

Table 1
Performance comparison of different negative sampling and data strategies

Method	Top-1	Top-3	Top-10	MRR
Random Negative Mining	0.5308	0.7271	0.8794	0.6520
Hard Negative Mining	0.4613	0.6829	0.8535	0.5982
Hardest Negative Mining	0.4012	0.6678	0.8499	0.5610
Labeled Data Only	**0.5474**	0.7356	0.8803	0.6639
Synthetic Data Only	0.2191	0.4012	0.6444	0.3550
Using Algorithm 1	0.5450	**0.7410**	**0.8842**	**0.6653**

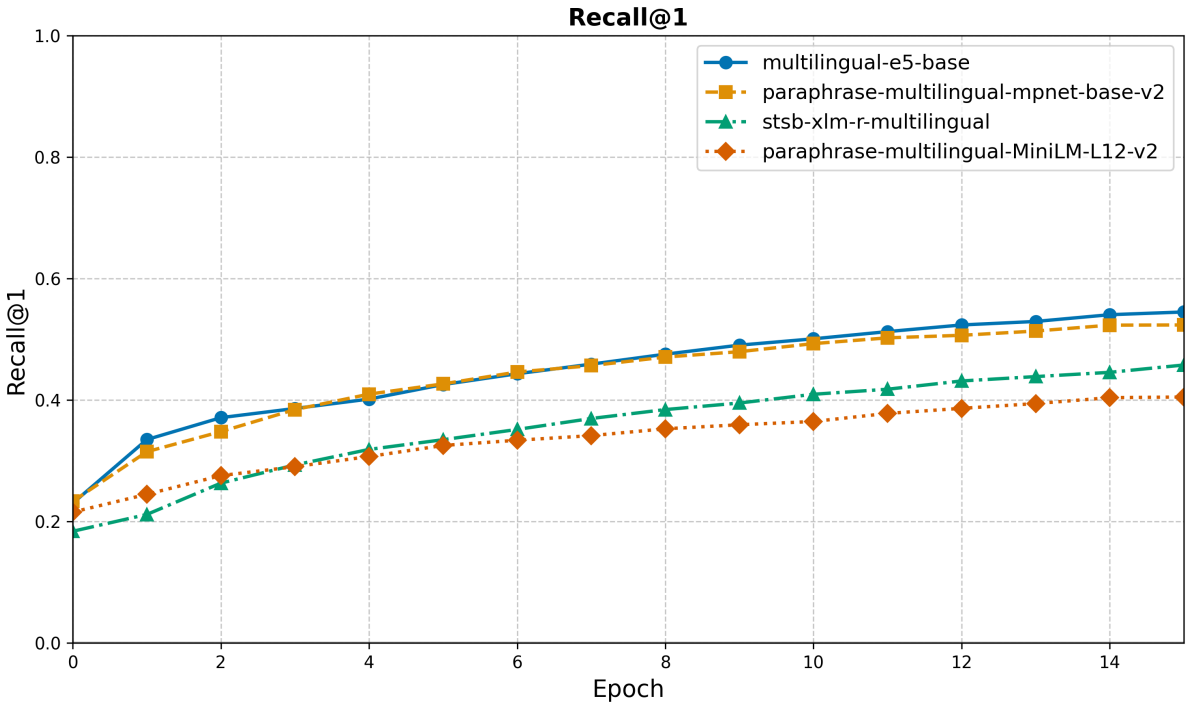


Figure 2: Performance of Recall@1 at different points during contrastive fine-tuning.

4. Conclusion and Future Work

Our proposed embedding model development pipeline to align multilingual information retrieval with a monolingual knowledge base removes the bottleneck of multilingual knowledge base construction in information retrieval systems. The ability to share knowledge across languages facilitates faster and easier global expansion of conversation AI systems, specifically to lower resource languages or code-switching use cases. Our comprehensive experiment results in a code-switching use case demonstrate the effectiveness and robustness of our proposed weighted sampling strategy for contrastive learning. Our framework is language agnostic therefore applicable for any target language and we hope our

paper can help push forward the research in the multilingual information retrieval communities and facilitate faster global expansion for businesses.

In the future, we seek to continue our experimentation to leverage monolingual knowledge base for multilingual dialogue systems. Additionally, we hope to explore more robust embedding models with compression approaches such as binarization or quantization to further improve performance.

Declaration on Generative AI

(by using the activity taxonomy in ceur-ws.org/genai-tax.html):
During the preparation of this work, the author(s) used Claude Sonnet 3.5 in order to: Text Translation. Specifically, the authors used Claude Sonnet 3.5 for query translation and synthetic training data generation. Please refer to Section 3.3 for implementation details.

References

[1] Y. Zhuang, J. Song, N. Sadagopan, A. Beniwal, Self-supervised pre-training and semi-supervised learning for extractive dialog summarization, in: Companion Proceedings of the ACM Web Conference 2023, WWW '23 Companion, Association for Computing Machinery, New York, NY, USA, 2023, p. 1069–1076. URL: https://doi.org/10.1145/3543873.3587680. doi:10.1145/3543873.3587680.

[2] Y. Zhuang, Y. Lu, S. Wang, Weakly supervised extractive summarization with attention, in: H. Li, G.-A. Levow, Z. Yu, C. Gupta, B. Sisman, S. Cai, D. Vandyke, N. Dethlefs, Y. Wu, J. J. Li (Eds.), Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, Association for Computational Linguistics, Singapore and Online, 2021, pp. 520–529. URL: https://aclanthology.org/2021.sigdial-1.54/. doi:10.18653/v1/2021.sigdial-1.54.

[3] M. C. Z. Z. Y. P. S.-T. L. Q. M. R. N. V. B. A. Zhang, Ziji; Yang, Reic: Rag-enhanced intent classification at scale, arXiv preprint arXiv:6498308 (2025).

[4] Introducing the next generation of claude, https://www.anthropic.com/news/claude-3-family, ???? Accessed: 2025-04-20.

[5] T. Cohere, :, Aakanksha, A. Ahmadian, M. Ahmed, J. Alammar, M. Alizadeh, Y. Alnumay, S. Althammer, A. Arkhangorodsky, V. Aryabumi, D. Aumiller, R. Avalos, Z. Aviv, S. Bae, S. Baji, A. Barbet, M. Bartolo, B. Bebensee, N. Beladia, W. Beller-Morales, A. Bérard, A. Berneshawi, A. Bialas, P. Blunsom, M. Bobkin, A. Bongale, S. Braun, M. Brunet, S. Cahyawijaya, D. Cairuz, J. A. Campos, C. Cao, K. Cao, R. Castagné, J. Cendrero, L. C. Currie, Y. Chandak, D. Chang, G. Chatziveroglou, H. Chen, C. Cheng, A. Chevalier, J. T. Chiu, E. Cho, E. Choi, E. Choi, T. Chung, V. Cirik, A. Cismaru, P. Clavier, H. Conklin, L. Crawhall-Stein, D. Crouse, A. F. Cruz-Salinas, B. Cyrus, D. D'souza, H. Dalla-Torre, J. Dang, W. Darling, O. D. Domingues, S. Dash, A. Debugne, T. Dehaze, S. Desai, J. Devassy, R. Dholakia, K. Duffy, A. Edalati, A. Eldeib, A. Elkady, S. Elsharkawy, I. Ergün, B. Ermis, M. Fadaee, B. Fan, L. Fayoux, Y. Flet-Berliac, N. Frosst, M. Gallé, W. Galuba, U. Garg, M. Geist, M. G. Azar, E. Gilsenan-McMahon, S. Goldfarb-Tarrant, T. Goldsack, A. Gomez, V. M. Gonzaga, N. Govindarajan, M. Govindassamy, N. Grinsztajn, N. Gritsch, P. Gu, S. Guo, K. Haefeli, R. Hajjar, T. Hawes, J. He, S. Hofstätter, S. Hong, S. Hooker, T. Hosking, S. Howe, E. Hu, R. Huang, H. Jain, R. Jain, N. Jakobi, M. Jenkins, J. Jordan, D. Joshi, J. Jung, T. Kalyanpur, S. R. Kamalakara, J. Kedrzycki, G. Keskin, E. Kim, J. Kim, W.-Y. Ko, T. Kocmi, M. Kozakov, W. Kryściński, A. K. Jain, K. K. Teru, S. Land, M. Lasby, O. Lasche, J. Lee, P. Lewis, J. Li, J. Li, H. Lin, A. Locatelli, K. Luong, R. Ma, L. Mach, M. Machado, J. Magbitang, B. M. Lopez, A. Mann, K. Marchisio, O. Markham, A. Matton, A. McKinney, D. McLoughlin, J. Mokry, A. Morisot, A. Moulder, H. Moynehan, M. Mozes, V. Muppalla, L. Murakhovska, H. Nagarajan, A. Nandula, H. Nasir, S. Nehra, J. Netto-Rosen, D. Ohashi, J. Owers-Bardsley, J. Ozuzu, D. Padilla, G. Park, S. Passaglia, J. Pekmez, L. Penstone, A. Piktus, C. Ploeg, A. Poulton, Y. Qi, S. Raghvendra, M. Ramos, E. Ranjan, P. Richemond, C. Robert-Michon, A. Rodriguez, S. Roy, S. Ruder, L. Ruis, L. Rust, A. Sachan, A. Salamanca, K. K. Saravanakumar,

I. Satyakam, A. S. Sebag, P. Sen, S. Sepehri, P. Seshadri, Y. Shen, T. Sherborne, S. S. Shi, S. Shivaprasad, V. Shmyhlo, A. Shrinivason, I. Shteinbuk, A. Shukayev, M. Simard, E. Snyder, A. Spataru, V. Spooner, T. Starostina, F. Strub, Y. Su, J. Sun, D. Talupuru, E. Tarassov, E. Tommasone, J. Tracey, B. Trend, E. Tumer, A. Üstün, B. Venkitesh, D. Venuto, P. Verga, M. Voisin, A. Wang, D. Wang, S. Wang, E. Wen, N. White, J. Willman, M. Winkels, C. Xia, J. Xie, M. Xu, B. Yang, T. Yi-Chern, I. Zhang, Z. Zhao, Z. Zhao, Command a: An enterprise-ready large language model, 2025. URL: https://arxiv.org/abs/2504.00698. arXiv:2504.00698.

[6] H. Paulheim, How much is a triple? estimating the cost of knowledge graph creation (2018).

[7] A. S. Doğruöz, S. Sitaram, B. E. Bullock, A. J. Toribio, A survey of code-switching: Linguistic and social perspectives for language technologies, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 1654–1666. URL: https://aclanthology.org/2021.acl-long.131/. doi:10.18653/v1/2021.acl-long.131.

[8] G. I. Winata, S. Cahyawijaya, Z. Liu, Z. Lin, A. Madotto, P. Fung, Are multilingual models effective in code-switching?, in: T. Solorio, S. Chen, A. W. Black, M. Diab, S. Sitaram, V. Soto, E. Yilmaz, A. Srinivasan (Eds.), Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching, Association for Computational Linguistics, Online, 2021, pp. 142–153. URL: https://aclanthology.org/2021.calcs-1.20/. doi:10.18653/v1/2021.calcs-1.20.

[9] Z. Zhang, L. Zhou, C. Wang, S. Chen, Y. Wu, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, F. Wei, Speak foreign languages with your own voice: Cross-lingual neural codec language modeling, 2023. URL: https://arxiv.org/abs/2303.03926. arXiv:2303.03926.

[10] K.-P. Huang, C.-K. Yang, Y.-K. Fu, E. Dunbar, H. yi Lee, Zero resource code-switched speech benchmark using speech utterance pairs for multiple spoken languages, 2024. URL: https://arxiv.org/abs/2310.03018. arXiv:2310.03018.

[11] S. Ji, S. Pan, E. Cambria, P. Marttinen, P. S. Yu, A survey on knowledge graphs: Representation, acquisition, and applications, IEEE Transactions on Neural Networks and Learning Systems 33 (2022) 494–514. URL: http://dx.doi.org/10.1109/TNNLS.2021.3070843. doi:10.1109/tnnls.2021.3070843.

[12] W. Zhou, F. Liu, I. Vulić, N. Collier, M. Chen, Prix-lm: Pretraining for multilingual knowledge base construction, arXiv preprint arXiv:2110.08443 (2021).

[13] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, F. Wei, Text embeddings by weakly-supervised contrastive pre-training, ArXiv abs/2212.03533 (2022). URL: https://api.semanticscholar.org/CorpusID:254366618.

[14] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, M. Zhang, Towards general text embeddings with multi-stage contrastive learning, ArXiv abs/2308.03281 (2023). URL: https://api.semanticscholar.org/CorpusID:260682258.

[15] Z. Nussbaum, J. X. Morris, B. Duderstadt, A. Mulyar, Nomic embed: Training a reproducible long context text embedder, 2025. URL: https://arxiv.org/abs/2402.01613. arXiv:2402.01613.

[16] L. Merrick, D. Xu, G. Nuti, D. Campos, Arctic-embed: Scalable, efficient, and accurate text embedding models, 2024. URL: https://arxiv.org/abs/2405.05374. arXiv:2405.05374.

[17] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual e5 text embeddings: A technical report, 2024. URL: https://arxiv.org/abs/2402.05672. arXiv:2402.05672.

[18] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. URL: https://arxiv.org/abs/2402.03216. arXiv:2402.03216.

[19] X. Zhang, Y. Zhang, D. Long, W. Xie, Z. Dai, J. Tang, H. Lin, B. Yang, P. Xie, F. Huang, M. Zhang, W. Li, M. Zhang, mgte: Generalized long-context text representation and reranking models for multilingual text retrieval, 2024. URL: https://arxiv.org/abs/2407.19669. arXiv:2407.19669.

[20] L. Degand, P. Muller, Introduction to the special issue on dialogue and dialogue systems, Traitement Automatique des Langues 61 (2020) 7–15. URL: https://aclanthology.org/2020.tal-3.1/.

[21] M. Günther, L. Milliken, J. Geuter, G. Mastrapas, B. Wang, H. Xiao, Jina embeddings: A novel set

of high-performance sentence embedding models, 2023. URL: https://arxiv.org/abs/2307.11224. arXiv:2307.11224.

[22] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Improving text embeddings with large language models, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 11897–11916. URL: https://aclanthology.org/2024.acl-long.642/. doi:10.18653/v1/2024.acl-long.642.

[23] Z. Dai, V. Y. Zhao, J. Ma, Y. Luan, J. Ni, J. Lu, A. Bakalov, K. Guu, K. B. Hall, M.-W. Chang, Promptagator: Few-shot dense retrieval from 8 examples, 2022. URL: https://arxiv.org/abs/2209.11755. arXiv:2209.11755.

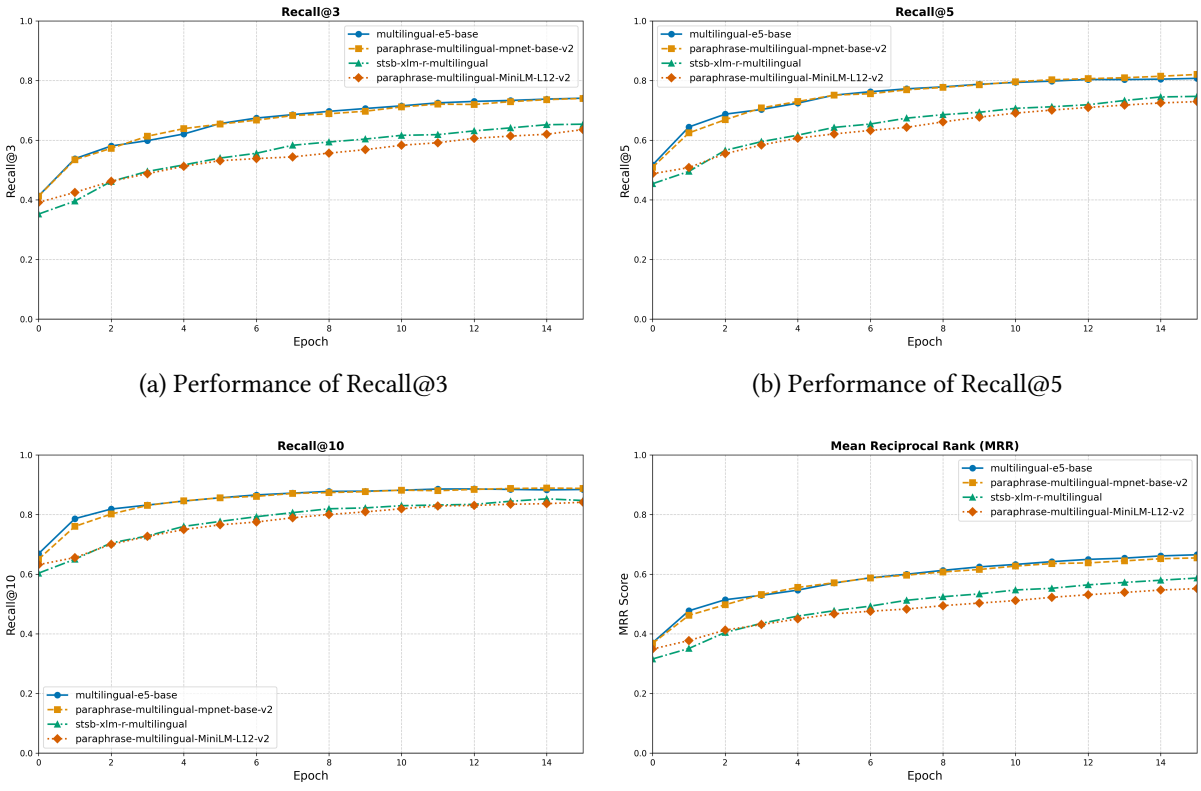
[24] Y. Qu, Y. Ding, J. Liu, K. Liu, R. Ren, W. X. Zhao, D. Dong, H. Wu, H. Wang, Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering, 2021. URL: https://arxiv.org/abs/2010.08191. arXiv:2010.08191.

[25] G. de Souza P. Moreira, R. Osmulski, M. Xu, R. Ak, B. Schifferer, E. Oldridge, Nv-retriever: Improving text embedding models with effective hard-negative mining, 2025. URL: https://arxiv.org/abs/2407.15831. arXiv:2407.15831.

[26] A. van den Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, ArXiv abs/1807.03748 (2018). URL: https://api.semanticscholar.org/CorpusID:49670925.

A. Plots for Model Comparison

The following plots represent some other comparisons between the models we compared between.



(a) Performance of Recall@3

(b) Performance of Recall@5

(c) Performance of Recall@10

(d) Mean Reciprocal Rank (MRR)

Figure 3: A comprehensive comparison of performance metrics (Recall@3, Recall@5, Recall@10, and MRR) during contrastive fine-tuning across different multilingual embedding models.