Monolingual Adapter Networks for Efficient Cross-Lingual Alignment

Pulkit Arya
pulkit.arya.career@gmail.com

Abstract

Multilingual alignment for low-resource languages is a challenge for embedding models. The scarcity of parallel datasets in addition to rich morphological diversity in languages adds to the complexity of training multilingual embedding models. To aid in the development of multilingual models for under-represented languages such as Sanskrit, we introduce GitaDB: a collection of 640 Sanskrit verses translated in 5 Indic languages and English. We benchmarked various state-of-the-art embedding models on our dataset in different bilingual and cross-lingual semantic retrieval tasks of increasing complexity and found a steep degradation in retrieval scores. We found a wide margin in the retrieval performance between English and Sanskrit targets. To bridge this gap, we introduce Monolingual Adapter Networks: a parameter-efficient method to bolster cross-lingual alignment of embedding models without the need for parallel corpora or full finetuning.

1 Introduction

Sanskrit is one of the oldest languages in human history, actively spoken by 25k people in India. The collection of scriptures, written in Vedic and Classic Sanskrit, include Vedas, Bhramanas, Arkanyas, Upnishads, Vedangas, Upvedas, Mahapurans, Upapurans, Darsanas, Smritis, Itihasa, and the Bhagvada Gita. These works have received so little attention that there is no consensus on the total verse count for Brahmanas, Aranyakas, Upanishads, Smritis, Vedangas, Upavedas, and Darsanas. The rest (Vedas, Puranas, Itihasas, and Bhagvada Gita) have an estimated total of 600,000 Sanskrit verses. It is estimated that 30 million documents of Sanskrit exist that are partly digitized (Aralikatte et al., 2021). Being silos of knowledge and wisdom, these are prominent works for cultural and historical studies. However, their accessibility is limited due to a lack of good quality translations and applications to search and analyze these works.

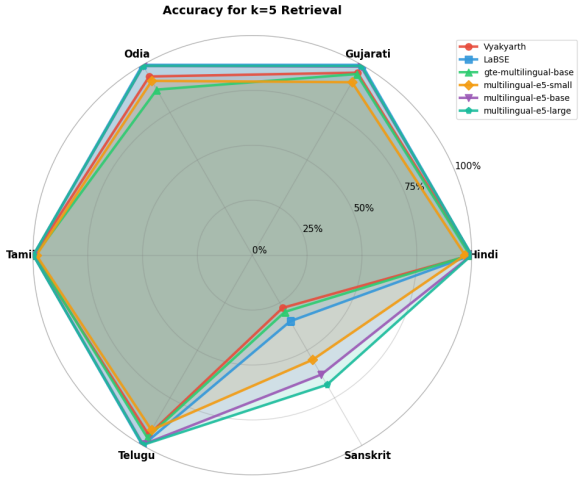

Figure 1: Bilingual retrieval accuracy of embedding models across Indic languages and Sanskrit (k=5) for English translations. SOTA models that excel in Indic-English retrieval, struggle with English-Sanskrit retrieval.

Retrieval-augmented generation (RAG) has emerged as the dominant paradigm for extending large language models' generative question-answering capabilities to new domains (Lewis et al., 2021, Gao et al., 2024, Guo et al., 2025, Han et al., 2025). Their multilingual and cross-lingual performance on question answering tasks have also been evaluated (Liu et al., 2025, Artetxe et al., 2020). The core challenge in creating a RAG system is retrieval of high-quality documents to be passed as part of the context to a LLM for generation. A standard retrieval pipeline uses a variation of semantic retrieval in addition to statistical methods such as BM-25 (Robertson et al., 1994) or graphs (Han et al., 2025, Guo et al., 2025). Semantic retrieval is based on similarity of vector embeddings of a given query and documents in the dataset. The quality of embeddings generated from an embedding model play a crucial role in the retrieval performance of

Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025), pages 360–368

this pipeline.

A large proportion of the population interested in surveying Sanskrit texts are non-native English speakers and often use their mother tongue, not English for most, as the preferred mode of communication and interaction with applications. The problem for Sanskrit verse retrieval is trivial if both query and translation is available in English or Indic languages (Figure 1). The Sanskrit verse can be retrieved based on semantic similarity of a translation (available in English or Indic language) with the query. The non-trivial cases include:

- Query in English against dataset of Sanskrit documents.
- Query in language X (low-resource indic language) against dataset of Sanskrit documents.
- Retrieving parallel pairs from an unlabeled corpora of Sanskrit/English/Indic languages.

Thus, multilingual embedding models capable of retrieving Sanskrit documents for English and Indic language queries will bolster the efforts in making accessible applications for the analysis of Sanskrit texts.

In our survey we found that existing corpora (Aralikatte et al., 2021, Bakrola and Nasariwala, 2023, Jagadeeshan et al., 2025, Maheshwari et al., 2024, Gala et al., 2023, Ramesh et al., 2022) lacked multilingual parallel translations for Sanskrit verses to benchmark multilingual and cross-lingual retrieval performance of embedding models. To aid in the development of models and applications, we introduce **GitaDB:** a parallel aligned dataset of high quality translations of Sanskrit verses in 5 Indic languages and English to support the development of new models in the field of retrieval, embedding, and question answering. Using our dataset we benchmark the performance of various multilingual embedding models in bilingual and cross-lingual retrieval.

In our analysis, we found a wide gap in the retrieval performance of these models in retrieving Sanskrit documents for English/Indic queries. To bridge this gap we created **Monolingual Adaptation Networks**, as a method to expand coverage of multilingual models to weakly represented languages. Monolingual Adaptation Networks are dense feed-forward neural networks that learn to transform the embeddings for an under-represented language (Sanskrit) to be closer to a pivot language (English) in a parameter and resource efficient manner.

Our main contributions in this paper include:

- **GitaDB** - A parallel aligned corpus of classic Sanskrit in 6 languages: English, Hindi, Gujarati, Odia, Tamil, Telugu
- **Monolingual Adapter Network** - A method to bolster the performance of embedding models for under-represented languages in a resource efficient way.
- **Cross-lingual Alignment** - We showcase the benefits of using a pivot language as training target for contrastive learning in cross-lingual alignment of translations.

2 Related Work

In our survey we found various datasets for Sanskrit translations. Itihasa (Aralikatte et al., 2021) has a collection of 93,000 pairs of Sanskrit and English translations created from two epics: Ramayana and Mahabharata. Sahayaak (Bakrola and Nasariwala, 2023) is a collection of 1.5M pairs of Sanskrit-Hindi translations covering various domains such as daily conversations, Sports, News, History, and ancient Indian literature including the 700 verses from Bhagvada Gita. Anveshana is a dataset of 3400 Sanskrit document-English query pairs used to study the efficacy of translation based retrieval over direct retrieval for cross-lingual retrieval of ancient texts (Jagadeeshan et al., 2025). Samayik (Maheshwari et al., 2024) has a collection of 53,000 Sanskrit-English pairs written in prose form, distinct from the poetic form of verses present in datasets like Itihasa. Other datasets such as IndicTrans2 (Gala et al., 2023) and IndicGenBench (Singh et al., 2024) cover modern Sanskrit, distinct from the Vedic and Classic forms of Sanskrit used in historic literature.

Most of the datasets we surveyed were either bilingual datasets for Sanskrit or were multilingual datasets for low-resource Indic languages *excluding Vedic and Classic Sanskrit*. GitaDB is the first dataset that contains multilingual verse aligned translations of 640 verses in 5 low-resource Indic languages along with English.

Our primary objective is to identify embedding models' ability to retrieve similar verses for a given query, presented in different Indic languages. Roy et al., 2020 introduced the concept of strong cross-lingual alignment and its necessity in a multilin-

gual embedding model's output. Strong cross-lingual alignment is achieved by maximizing inter-cluster distance and minimizing intra-cluster distance for multilingual embeddings of the same information. A model which exhibits low intra-verse distance and high inter-verse distance has strong cross-lingual alignment of translations which produces a high-quality retriever. Thus, we use the concept of strong alignment in our study.

Multilingual alignment methods typically depend on parallel data or bilingual dictionaries, which are scarce for under-represented languages like Sanskrit. More recent multilingual embedding models (e.g., LaBSE (Feng et al., 2020), mE5 (Wang et al., 2024)) aim to create shared representation spaces but still exhibit performance degradation on low-resource languages. Parameter-efficient adaptation methods such as adapter layers (Houlsby et al., 2019) and MAD-X (Pfeiffer et al., 2020) have proven effective for cross-lingual transfer, yet they primarily target task adaptation rather than language alignment. In contrast, *Monolingual Adapter Networks* focus specifically on resource-efficient language-space realignment, enabling embeddings of low-resource languages to be pushed closer to a pivot language without requiring parallel corpora in multiple languages or sacrificing performance on other languages.

3 Dataset

Our dataset is a collection of 640 verses taken from the Bhagvada Gita. The Bhagvada Gita is a subset of 700 verses from the Mahabharata structured as a poetic discourse between Arjuna and Lord Krishna, covering various parts of one's life: duty, knowledge, and devotion. It is also referred to as the summary of the Vedas - the scriptures that form the roots of Sanatan Dharma. The Bhagvada Gita contains a total of 700 verses. After data cleanup, we were left with 640 verses with translations in 6 languages: Hindi, English, Gujarati, Tamil, Telugu, and Odia for a total of 4480 sentences in our dataset.

We sourced our translations from various online sources and align them at the verse level. For each verse, we store the Sanskrit verse along with its translation in each language available: Hindi, English, Gujarati, Tamil, Telugu, and Odia. Each language uses a different script that adds a rich complexity in our dataset.

After initial data collection, we found certain verses were fused together. These verses are translated in pairs/triplets as they provide necessary context for the pair/triplet of verses to be interpreted correctly. We translated each verse of the pair/triplet independently and found the meaning to be skewed without the appropriate context. Thus, we decided to leave the fused verses as a single entity in our dataset. This brought our total verse count from down from 700 to 640.

Our dataset along with all our code for this paper can be found here [1]

4 Methods

4.1 Base Model

We adopt LaBSE (Feng et al., 2020) as the underlying multilingual encoder due to its strong bilingual retrieval performance for Indic queries against a corpus of English translations (Table 1). LaBSE provides sentence-level embeddings for more than 100 languages, but like other multilingual encoders, it performs poorly on low-resource languages such as Sanskrit (Table 2).

4.2 Adapter Network Architecture

On top of the frozen LaBSE encoder, we introduce an *Adapter Network* implemented as a lightweight two-layer feed-forward neural network. This adapter maps Sanskrit embeddings into a space more closely aligned with English embeddings, serving as a post-hoc correction without requiring changes to the base model. By restricting training to the adapter, our approach remains computationally efficient and avoids catastrophic forgetting across other languages. The training and inference setup are showcased in figures 2 and 3 respectively.

4.3 Training Data

We train the Adapter Network on the Itihasa corpus [2] (Aralikatte et al., 2021), which provides paired Sanskrit and English translations. Importantly, only the Sanskrit embeddings are passed through the adapter during training, while the English embeddings from LaBSE remain fixed and serve as alignment targets.

[1] https://github.com/tickloop/gitadb

[2] The Bhagvada Gita is a part of the Mahabharata. To avoid test set leakage, we remove the chapters of Mahabharata that cover the Bhagvada Gita from our training set.

Model	Top-k	Hi (Acc/MRR)	Gu (Acc/MRR)	Od (Acc/MRR)	Ta (Acc/MRR)	Te (Acc/MRR)
Vyakyarth	k=5	99.8 / 98.3	95.9 / 89.5	93.9 / 86.2	99.5 / 98.6	94.4 / 89.1
LaBSE	k=5	100.0 / 99.9	100.0 / 100.0	100.0 / 100.0	99.8 / 99.7	100.0 / 100.0
gte-multilingual-base	k=5	99.2 / 96.6	95.2 / 89.4	87.0 / 75.9	98.9 / 97.2	95.8 / 90.6
multilingual-e5-small	k=5	97.7 / 94.3	90.9 / 79.4	91.6 / 83.7	98.4 / 94.9	91.9 / 83.3
multilingual-e5-base	k=5	100.0 / 98.7	99.1 / 96.1	99.4 / 97.3	99.7 / 99.3	99.5 / 97.7
multilingual-e5-large	k=5	100.0 / 99.5	99.4 / 98.5	99.5 / 98.9	99.8 / 99.5	99.7 / 98.8
Vyakyarth	k=10	99.8 / 98.3	98.0 / 89.8	96.4 / 86.6	99.8 / 98.7	97.3 / 89.5
LaBSE	k=10	100.0 / 99.9	100.0 / 100.0	100.0 / 100.0	99.8 / 99.7	100.0 / 100.0
gte-multilingual-base	k=10	99.5 / 96.7	97.2 / 89.7	92.5 / 76.7	99.4 / 97.2	97.5 / 90.8
multilingual-e5-small	k=10	98.8 / 94.4	95.0 / 80.0	95.5 / 84.2	99.1 / 95.0	95.5 / 83.8
multilingual-e5-base	k=10	100.0 / 98.7	99.5 / 96.1	99.7 / 97.4	99.7 / 99.3	99.8 / 97.8
multilingual-e5-large	k=10	100.0 / 99.5	99.7 / 98.5	100.0 / 99.0	100.0 / 99.5	99.8 / 98.8

Table 1: Top-k retrieval accuracy (Acc) and mean reciprocal rank (MRR) for queries in Indic languages with targets from English translation corpora. Each cell shows Acc / MRR.

4.4 Objective Function

Training is performed with the InfoNCE contrastive loss (van den Oord et al., 2018). For each Sanskrit–English pair, the adapter output for Sanskrit serves as the query, and the corresponding English embedding is treated as the positive key among a set of in-batch negatives. This formulation encourages the adapted Sanskrit embeddings to be "pulled" closer to their English counterparts while being pushed away from non-matching English samples. Use of more advanced loss functions is left as part of future work.

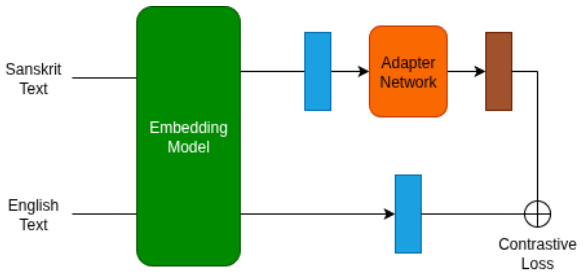

Figure 2: Adapter Network training setup. The embedding model is kept frozen and only the Adapter Network is trained using a contrastive loss. This creates embeddings for Sanskrit that are better aligned with English while the alignment between other languages is unaltered.

5 Experiments

We selected the following models for evaluation on our retrieval benchmarks:

- **Vyakyarth** (Pushkar Singh, 2024) is a 270M sentence embedding model designed for Indic languages, built upon the STSB-XLM-R-Multilingual architecture.
- **GTE-Multilingual** (Zhang et al., 2024) is a 305M parameter General Text Embedding model which is trained on 70+ languages and ranks high on MMTEB (Enevoldsen et al., 2025).
- **LaBSE** (Feng et al., 2020) is a 471M parameter multilingual model that scores well on low-resource languages.
- **Multilingual-e5** (Wang et al., 2024) family of models trained on 100+ languages offer three models: small (118M), base (278M), and large (560M) parameters.

We tested the models in scenarios that resemble real-world application of high-quality embeddings: Bilingual English-Indic Retrieval, Retrieval without translation availability, and Bitext mining in multilingual corpora. Each task requires high bilingual and cross-lingual alignment of embeddings. We use cosine similarity as our distance metric in all our experiments. Since our dataset does not contain queries, we use the translations as a proxy for queries.

5.1 Retrieval from English Corpus

Task: Given a query in Indic language and a corpora of English translations, retrieve the parallel translation of the query. We expect all models

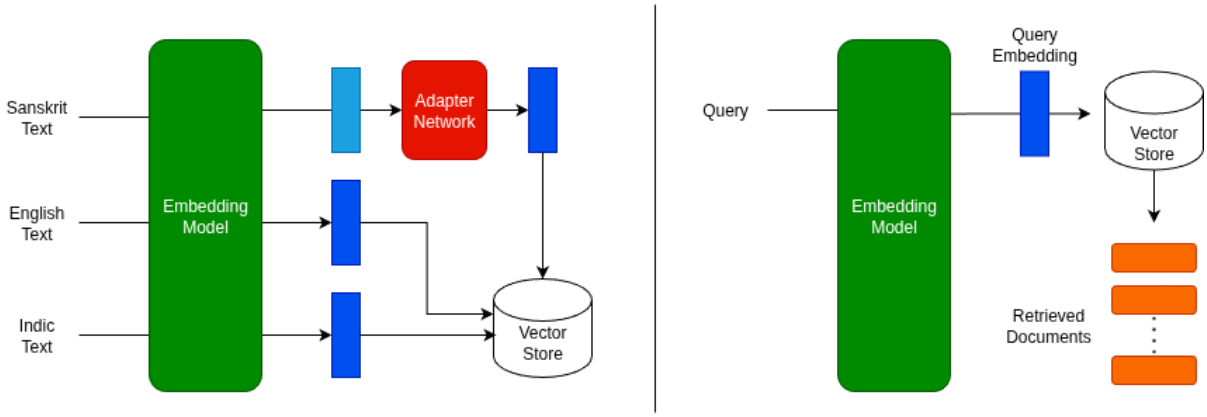

Figure 3: During inference, English and Indic text embeddings are generated via the embedding model while Sanskrit embeddings are generated as a combination of the embedding model and the adapter network. The embeddings are stored in a vector store for retrieval. Queries are embedded using the same embedding model are relevant documents are retrieved from the vector store. Since the adapted embeddings for Sanskrit are better aligned with English (and consequently Indic Languages), the retrieval performance is better for Sanskrit documents against queries from different languages.

to perform well on this task due to the increase in availability of multilingual corpora for all languages tested. This task also serves as a benchmark to identify any language bias in embedding models.

We created embeddings for each English translation in our corpora and stored them in a vector database. Then, for each query we retrieved top-k English translations using Cosine similarity as the distance metric from our vector store. We report the Accuracy@k and Mean Reciprocal Rank (mRR@k) values in Table 1. All models performed near perfectly in this task as we expected. The MRR scores being close to 100 indicate that majority of correct retrievals were the highest scoring result. This showcases a high bilingual alignment between English and Indic language embeddings.

5.2 Retrieval from Sanskrit Corpora

Task: Given a query in language X (En/Indic) and a corpora of Sanskrit verses, retrieve the parallel verse for the query. This task benchmarks the bilingual alignment between English/Indic languages and Sanskrit for each embedding model.

We created embeddings for each Sanskrit verse in our corpora and stored them in a vector database. Then, for each query we retrieved top-k Sanskrit verses using Cosine similarity as the distance metric from our vector store. The results for this study are presented in Table 2.

The multilingual-e5 model family was the dominant model for this task. The base LaBSE model achieves an average score of 29.82. Adding the Adapter Network (+ada) increases performance to 43.3, a 45.2% improvement, highlighting the effectiveness of the adapter. The Adapter Network also aids in alignment between Sanskrit verses and translations in Indic languages which the models was not trained on. For English retrieval, the model exhibited an absolute improvement of 12.6%, whereas for Indic languages it demonstrated a comparatively higher average absolute gain of 13.7%, highlighting the enhancement in cross-lingual alignment. Having a good alignment between pivot and non-pivot languages, as we noted in Task 1, aids in a consistent improvement across all languages (Table 2).

5.3 Bitext Retrieval

Task: Given a Sanskrit verse and a multilingual corpora of English and Indic language translations, retrieve all the parallel translations for the verse. This task benchmarks the cross-lingual alignment and retrieval ability of embedding models.

For each verse in our Sanskrit corpora, we retrieve top-k results from a multilingual corpora of English and Indic translations. We report the Accuracy@k for k=6 in Table 3. We use k=6 as there are 6 parallel translations for each verse in our dataset. Since there are multiple correct candidates for retrieval, we also report Mean Average Precision (mAP) (Roy et al., 2020) values in Table 4 along with average accuracy (count of correct retrievals / total correct translations in dataset).

The base LaBSE model achieves an average score of 11.9 which is a massive drop compared to the bilingual setting with only one target lan-

guage. The presence of multiple correct translations was a challenge for every embedding model. The sharp drop in performance indicates that the embedding space contains clusters with low interverse distance resulting in weak alignment between Sanskrit and English/Indic languages. Incorporating the Adapter Network (+ada) raises the average to 23.9, highlighting that the adapter helps align low-resource language embeddings even without full fine-tuning. It provided a +25.6% absolute improvement in retrieval accuracy for English translations, while also providing a consistent improvement in cross-lingual alignment for non-pivot languages: +3.2% in Hindi, +9.5% in Gujarati, +13.6% in Odia, +10.3% in Tamil, and +10.4% in Telugu.

We also note a wide variation in the performance of multilingual-e5 models across languages. Their performance for English retrieval dropped significantly in the presence of multiple translations from Indic languages. The e5-large model's top-k retrieval accuracy for Gujarati was 61.6 whereas for English it was only 11.1. There are similar language biases in e5-base and e5-small embedding models. To investigate this bias, we trained an Adapter Network for e5-base model.

While the performance of m-e5-base (+ada) on English, Odia, and Tamil increased by an absolute average of 23.8%, it dropped for Hindi, Gujarati and Telugu by an absolute average of 10.2%. The top-6 average retrieval accuracy for m-e5-base was 30.7, which was boosted to 37.5 with the help of Adapter Networks. The average retrieval performance of this combination of multilingual-e5-base with Adapter networks (37.5) is comparable to multilingual-e5-large (38.1). It is clear that multilingual-e5 family of models' Indic language embeddings do no cluster around English as a pivot language and an interesting future work will be to investigate the choice of pivot language for different embedding models.

6 Results

Overall, our experiments reveal a clear stratification in model performance across tasks and languages. While nearly all multilingual embedding models exhibited decent performance in bilingual retrieval from parallel corpora, their effectiveness dropped substantially when moving to tasks that required cross-script and cross-lingual alignment with Sanskrit. The multilingual-e5 family consistently ranked at the top for bilingual scenarios, particularly the large variant, which demonstrated strong resilience to performance degradation.

In the Indic/English-to-Sanskrit retrieval task (Table 2), the models encountered a significant challenge. The shift from modern language corpora to a under-represented, morphologically rich language introduced substantial difficulty in semantic alignment. Even top-performing models exhibited a marked decline in retrieval accuracy, indicating that bi-lingual alignment learned from contemporary corpora does not directly transfer to Sanskrit. The relative resilience of the multilingual-e5 family suggests that broader multilingual coverage and larger model capacity help preserve alignment in low-resource or structurally distant target languages, but performance gaps remain large enough to affect real-world applicability in downstream RAG systems. Adapter Networks consistently improved retrieval accuracy for English and Indic

Model	En	Hi	Gu	Od	Ta	Te
Vyakyarth	27.7	29.2	24.2	22.7	22.8	19.7
gte-m-base	29.8	41.4	25.8	24.2	27.5	27.3
m-e5-small	55.0	63.4	56.1	53.6	50.8	56.4
m-e5-base	62.8	70.8	66.9	66.7	59.7	67.5
m-e5-large	68.1	75.8	77.5	74.1	68.9	77.5
LaBSE	34.7	26.7	30.2	29.2	25.8	32.3
LaBSE (+ada)	47.3	40.5	42.7	42.8	42.0	44.5

Table 2: Top-5 retrieval accuracy for queries in English/Indic language with targets from Sanskrit verse corpora. (+ada) uses adapted embeddings for retrieval targets. The Adapter Network not only increased performance on English, but also across non-pivot languages that were not included in training.

Model	En	Hi	Gu	Od	Ta	Te
Vyakyarth	12.0	15.0	10.2	6.9	6.1	9.7
gte-m-base	19.8	17.5	8.4	7.5	9.8	11.7
m-e5-small	2.3	14.4	44.2	15.5	5.0	19.4
m-e5-large	11.1	38.3	61.6	48.9	21.1	48.0
LaBSE	10.0	12.3	13.0	10.0	8.8	17.3
LaBSE (+ada)	35.6	15.0	22.5	23.6	19.1	27.7
m-e5-base	8.4	41.7	49.2	27.8	19.1	38.0
m-e5-base (+ada)	59.7	40.3	28.1	33.1	33.8	29.8

Table 3: Top-6 parallel alignment accuracy for each language. There is a stark decline in performance for all models as compared to retrieval from English corpora in Task 1 and for bilingual retrieval with Sanskrit targets in Task 2. The multilingual-e5 family also showcases a heavy language bias.

Model	acc@6	acc@10	mAP@6	mAP@10
Vyakyarth	9.97	13.26	0.18	0.19
gte-m-base	12.47	16.67	0.26	0.25
m-e5-small	16.80	20.44	0.39	0.37
m-e5-large	38.15	46.35	0.63	0.60
LaBSE	11.90	16.09	0.18	0.18
LaBSE (+ada)	23.91	31.30	0.35	0.34
m-e5-base	30.70	37.16	0.52	0.50
m-e5-base (+ada)	37.47	46.28	0.54	0.52

Table 4: Average accuracy and mAP values for parallel translation retrieval. mAP values closer to 1 are better. All models struggled in retrieving parallel translations in the presence of multiple targets from different languages. A stark contrast from the bilingual performance highlights the complexity of this task.

languages, even with the lack of parallel corpora or full-finetuning of embedding models.

The bitext retrieval task (table 3-4), which required retrieving all valid translations of a Sanskrit verse from a multilingual pool, proved the most difficult. The presence of multiple correct answers across diverse scripts and languages compounded alignment complexity, amplifying the effects of language bias and imperfect semantic clustering. Here, accuracy dropped sharply for most models, and mAP values were substantially below 1 indicating the lack of correct answers in majority of retrievals. The multilingual-e5 models again emerged as the most robust, though their performance in English retrieval degraded noticeably in this multi-target setting, suggesting that even strong multilingual alignment is strained when faced with semantically overlapping candidate sets. This result underscores the need for embedding strategies explicitly optimized for multi-answer, multilingual retrieval scenarios in low-resource languages. While the use of Adapter Networks showed improvement in cross-lingual alignment for all non-pivot languages for LaBSE, the lack of strong cross-lingual alignment between English and Indic language translations resulted in a split performance for the E5 family of models.

7 Conclusion

In this work, we introduced GitaDB, a parallel-aligned multilingual dataset of 640 Bhagavad Gita verses in Sanskrit with translations in five Indic languages and English. We benchmarked a range of multilingual embedding models on retrieval tasks of increasing complexity, revealing the strengths and limitations of current embedding models for cross-lingual and cross-script retrieval in a classical language setting. While state-of-the-art models such as the multilingual-e5 family demonstrated strong performance in parallel multilingual retrieval, their performance dropped substantially in bilingual Sanskrit alignment and multilingual bitext retrieval scenarios. These results underscore the unique challenges of handling morphologically rich, low-resource languages with diverse scripts, even for models trained on extensive multilingual corpora. Our method of creating resource efficient Adapter Networks proved effective in extending the capabilities of embedding models to an under-represented languages without full finetuning or parallel multilingual corpora.

8 Future Work

Our findings suggest several promising directions for future work. There is a clear need for embedding models explicitly trained on classical language corpora and capable of handling cross-script alignment without relying solely on translations. This work has uncovered the use of Adapter Networks as a strategy to improved cross-lingual retrieval performance with a simple architecture. Adapter Networks can be further studied with varying architectures, loss functions, and pivot languages based on the choice of underlying embedding model. Using hard in-batch negatives has also shown promising results in contrastive training. We leave the exploration of using hard in-batch negatives for a future study.

The multi-answer retrieval setting presents an open challenge; techniques that better cluster semantically equivalent translations while maintaining separation between distinct verses could yield significant zero-shot improvements. For RAG systems in particular, such advances could enable more accurate context retrieval across languages, improving both coverage and relevance for end users who query in non-English languages. By closing the alignment gap between Sanskrit and modern Indic languages, future systems will be better equipped to serve as multilingual gateways to the cultural and philosophical heritage embedded in these texts.

References

Rahul Aralikatte, Miryam de Lhoneux, Anoop Kunchukuttan, and Anders Søgaard. 2021. Itihasa: A large-scale corpus for Sanskrit to English translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 191–197, Online. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Vishvajitsinh Bakrola and Jitendra Nasariwala. 2023. Sahaayak 2023 – the multi domain bilingual parallel corpus of sanskrit to hindi for machine translation. *Preprint*, arXiv:2307.00021.

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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. In *Proceedings of ACL*.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Preprint*, arXiv:2305.16307.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2025. Lightrag: Simple and fast retrieval-augmented generation. *Preprint*, arXiv:2410.05779.

Haoyu Han, Yu Wang, Harry Shomer, Kai Guo, Jiayuan Ding, Yongjia Lei, Mahantesh Halappanavar, Ryan A. Rossi, Subhabrata Mukherjee, Xianfeng Tang, Qi He, Zhigang Hua, Bo Long, Tong Zhao, Neil Shah, Amin Javari, Yinglong Xia, and Jiliang Tang. 2025. Retrieval-augmented generation with graphs (graphrag). *Preprint*, arXiv:2501.00309.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. *Preprint*, arXiv:1902.00751.

Manoj Balaji Jagadeeshan, Prince Raj, and Pawan Goyal. 2025. Anveshana: A new benchmark dataset for cross-lingual information retrieval on English queries and Sanskrit documents. In *Computational Sanskrit and Digital Humanities - World Sanskrit Conference 2025*, pages 161–180, Kathmandu, Nepal. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Wei Liu, Sony Trenous, Leonardo F. R. Ribeiro, Bill Byrne, and Felix Hieber. 2025. Xrag: Cross-lingual retrieval-augmented generation. *Preprint*, arXiv:2505.10089.

Ayush Maheshwari, Ashim Gupta, Amrith Krishna, Atul Kumar Singh, Ganesh Ramakrishnan, Anil Kumar Gourishetty, and Jitin Singla. 2024. Samayik: A benchmark and dataset for English-Sanskrit translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14298–14304, Torino, Italia. ELRA and ICCL.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Rajkiran Panuganti Pushkar Singh, Sandeep Kumar Pandey. 2024. Vyakyarth: A multilingual sentence embedding model for indic languages. GitHub.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Stephen E. Robertson, Steve Walker, Susan Jones, M. Hancock-Beaulieu, and M. Gatford. 1994. Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, pages 109–121.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. LAReQA: Language-agnostic answer retrieval from a multilingual pool. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5919–5930, Online. Association for Computational Linguistics.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073, Bangkok, Thailand. Association for Computational Linguistics.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

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

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.