A Large-Scale Parallel Corpus and Multilingual Pretrained Language Model for Machine Translation and Semantic Retrieval for Pāli, Sanskrit, Buddhist Chinese, and Tibetan

Anonymous ACL submission

Abstract

Ancient Buddhist literature features frequent, yet often unannotated, textual parallels spread across diverse languages: Sanskrit, Pāli, Buddhist Chinese, Tibetan, and more. The scale of this material makes manual examination prohibitive. We present the X (name changed for anonymity purposes) framework, which consists of a novel pipeline for multilingual parallel passage mining, X-parallel, a large-scale corpus of 1.74 million parallel sentence pairs between Sanskrit, Chinese, and Tibetan. and the development of the domain-specific pretrained language model `Gemma 2 X`. We present `Gemma 2 X-MT`, a version of this base model fine-tuned on machine translation tasks, reaching state-of-the-art performance for machine translation of these languages into English and outperforming even much larger open-source models. We also present `Gemma 2 X-E`, a semantic embedding model that shows state-of-the-art performance on a novel, detailed semantic embedding benchmark. We make the parallel dataset, model weights, and semantic similarity benchmark openly available to aid both NLP research and philological studies in Buddhist and classical Asian literature.

1 Introduction

Over the course of more than two millennia, the Buddhist tradition has produced a massive body of literature nowadays preserved primarily in Pāli, Sanskrit, Buddhist Chinese, and Tibetan. Within this literature, semantically related or closely matching passages of variable length are frequently encountered. These parallels appear both within literature preserved in the same language and across different languages. Multilingual parallelism is especially prominent due to large-scale translation efforts of Indic Buddhist text material by the Buddhist traditions: first into Chinese (beginning in the 2nd century CE), and later into Tibetan (from the 8th century CE onwards). Translation of this body of literature into modern English is ongoing, but so far only a fraction of the material has been properly translated. To give one example, about 10% of the digitally available Buddhist texts in Chinese have been translated into English so far (Nehrdich et al., 2023).

Textual reuse within Buddhist literature, while occurring frequently, is often not marked explicitly and is therefore a significant research objective for philologists (Freschi, 2014). Exhaustive manual studies remain prohibitively labor-intensive. Consequently, existing research is limited to specific topics within subsections of the literature (for one example, see Hellwig et al. (2023)). Since the Buddhist textual transmission is highly multilingual, parallel data is needed for the training of efficient multilingual retrieval models, which is currently only exhaustively collected between Sanskrit and Tibetan (Nehrdich, 2022).

In order to address these challenges, we propose a framework that utilizes machine translation into English as a pivot step to retrieve alignment candidates of longer passages, which we then further refine with sentence-level alignment. We then combine the generated multilingual parallel data with other monolingual and parallel resources in order to pretrain a domain-specific large language model (LLM) for these languages. We then fine-tune two different versions of this model: one for machine translation and one for multilingual semantic retrieval. Our benchmarks show that this LLM outperforms all other open baselines on machine translation and semantic retrieval for these languages.

This paper makes the following contributions:

- The description of a pipeline for the retrieval of matching parallel passages within literature preserved in different classical Asian languages based on machine translation as a pivot step
- X-parallel, a novel dataset of documents auto-

matically aligned at the sentence level across Sanskrit, Chinese, and Tibetan, comprising 1,742,786 sentence pairs

- Description of the continuous pretraining of `Gemma 2 X`, a domain-specific Gemma 2-based large language model
- Fine-tuning of this model on both machine translation and information retrieval tasks
- Evaluation of the machine translation performance for Sanskrit, Pāli, Tibetan, and Chinese-to-English tasks
- A novel benchmark for multilingual semantic retrieval for ancient Buddhist languages with a dataset covering seven different tasks
- Comprehensive evaluation of different retrieval methods using a unified testing protocol on this evaluation dataset and an ablation study for monolingual retrieval methods

We provide the automatically generated novel dataset of sentence-level aligned ancient Buddhist document pairs, evaluation dataset, testing protocol and links to the trained model weights at `X` Additionally, we provide access to this dataset through a user-friendly, searchable online database designed for philological research in Buddhist and broader classical Asian literature at `X`.

This paper proceeds as follows: Section 2 introduces relevant previous work. Section 3 describes the data mining procedure used to create X-parallel, and presents statistics and a quality assessment of this dataset. Section 4 describes the continuous pretraining of the `Gemma 2 X` base language model, including the composition of its multilingual pretraining data, and the instruction finetuning processes for both machine translation and semantic retrieval tasks. Section 5 presents the evaluation of the machine translation capabilities of `Gemma 2 X-MT`, comparing its performance on translating Sanskrit, Pāli, Tibetan, and Buddhist Chinese into English against other open large language models and existing domain-specific models. Section 6 introduces a novel benchmark for multilingual semantic retrieval and evaluates the performance of `Gemma 2 X-E` against various sparse and dense retrieval methods across four distinct retrieval scenarios. Section 7 conducts an ablation study to assess the monolingual retrieval performance of `Gemma 2 X-E` compared to other models on Sanskrit and Chinese texts. Section 8 summarizes our contributions and discusses potential future work, while Section 9 describes the limitations of this study.

2 Previous Research

Recent scholarship has increasingly employed computational methods to detect textual parallels within ancient Buddhist literature. Efforts include identifying reuse in Sanskrit corpora (Hellwig, 2013), Tibetan texts using string similarity (Klein et al., 2014), and Buddhist Chinese literature via word embeddings and alignment algorithms (Nehrdich, 2020).

Cross-lingual investigations have explored aligning Tibetan and Buddhist Chinese sentences with static embeddings (Felbur et al., 2022) and matching Sanskrit and Tibetan documents using deep neural transformer-based sentence representations (Nehrdich, 2022). Despite these advances, there is a lack of large-scale, sentence-aligned multilingual resources spanning Sanskrit, Pāli, Buddhist Chinese, and Tibetan.

While deep neural sentence representations are state-of-the-art for semantic similarity (Reimers and Gurevych, 2019) and bitext mining (Schwenk, 2018; Artetxe and Schwenk, 2019a), with powerful multilingual models available for high-resource languages (Artetxe and Schwenk, 2019b; Feng et al., 2022), the application of this approach to this specific multilingual ancient setting is not yet explored. Similarly, powerful dense retrieval models (Karpukhin et al., 2020), crucial for tasks like RAG (Lewis et al., 2020), often lack specialization for these historically significant, low-resource languages. Our work addresses these gaps by developing a comprehensive parallel corpus and domain-specific models.

3 X-parallel dataset

Ancient Buddhist literature features extensive multilingual parallels, primarily from Indic texts translated into Chinese and Tibetan. However, these are often uncataloged and may exist only as fragments (e.g., chapters or paragraphs within larger works), making identification challenging. We define our task as detecting common sub-passages of semantically equivalent parallel sentences of at least paragraph-length. The highly repetitive nature of Buddhist texts means standard sentence-level

mining approaches (Schwenk, 2018; Artetxe and Schwenk, 2019a) are prone to generating excessive noise, thus requiring a more constrained mining pipeline.

3.1 Data Mining Procedure

Our pipeline leverages the tendency of parallel sentences to appear in continuous chains. It proceeds in three main stages:

Machine Translation All documents are translated into English using a MADLAD-400 model (Kudugunta et al., 2023) fine-tuned on domain-specific data (Nehrdich et al., 2023), including 2 million Tibetan-English pairs (monlam.ai) and a forthcoming Sanskrit-English dataset.

Candidate Clusters On these English translations, we generate overlapping sliding windows (concatenated adjacent sentences to a minimum length for higher retrieval precision). These windows are embedded using BGE M3 (Chen et al., 2024). Corpus-wide kNN search on cosine similarity identifies initial candidate pairs (x_i, y_i) (source/target text positions). Spatial hashing then efficiently groups these pairs into clusters $C_k = C_1, C_2, \ldots$, each representing a contiguous region of likely parallelism.

Sentence Alignment The identified candidate regions are refined to precise sentence-level alignments using BERTALIGN (Liu and Zhu, 2022) with a domain-finetuned `LaBSE` model (Feng et al., 2022). Crucially, this alignment operates on the original language sentences, not the translations, helping to remove noisy matches at cluster peripheries. A final filtering step applies a moving average threshold on the sentence pairs' average cosine similarity to ensure quality.

This three-step process consisting of machine translation, coarse region identification, and fine-grained sentence alignment, significantly reduces noise from the repetitive nature of Buddhist literature, yielding high-quality parallels suitable for both machine learning and philological research.

3.2 Generated Dataset

This method yielded 1,742,786 parallel sentence pairs across Sanskrit<>Tibetan, Chinese<>Tibetan, and Sanskrit<>Chinese (see Figure 1 for an overview). This significantly expands existing resources. For instance, our 596,812 Sanskrit<>Tibetan pairs represent an 89% increase over the SansTib dataset (Nehrdich, 2022). For Sanskrit<>Chinese and Chinese<>Tibetan, this work provides the first large-scale dedicated digital parallel datasets.

To assess quality, we manually evaluated 100 randomly sampled pairs, categorizing them as 'Correct' (perfect match), 'Partially Correct' (majority overlap with minor mismatch), or 'Wrong' (<50% correct correspondence). As shown in Table 1, 73% achieved perfect alignment, indicating a strong baseline for unsupervised retrieval. While 11% were 'Wrong', some may still correctly identify document pairs with the help of such matches, and deterministic filters (e.g., segment length ratios) can remove such misalignments effectively for ML applications.

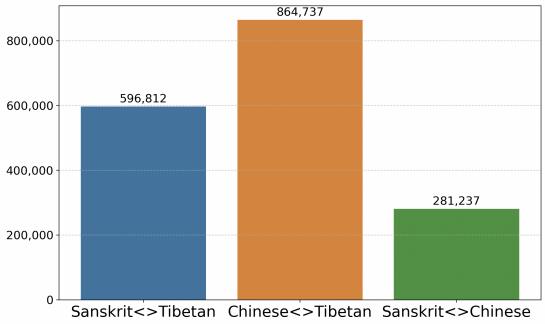

Figure 1: Number of datapoints per language pair in the sentence-level aligned parallel dataset for Ancient Buddhist Languages.

Category	Percentage (%)
Perfect	73
Partly correct	16
Wrong	11

Table 1: Manual examination of error rates in the mined dataset. Perfect means that both source and target segment match up without any errors. Partly correct means that more than 50% of source and target segment match up. Wrong means that less than 50% of source and target segment match.

4 Gemma 2 X Base LLM

We pretrain our own language model on a large dataset of relevant Buddhist data which serves as the basis for the machine translation and semantic retrieval model. We follow the training recipe

of TOWER (Alves et al., 2024): We take a strong baseline model, in this case Gemma 2 (Gemma Team, 2024), and continuously pretrain it on a large domain-specific multilingual corpus that consists of both monolingual data in all four languages as well as high-quality parallel data. We decided to build on top of Gemma 2 since in our evaluation, it has shown the most consistent baseline performance on machine translation into English compared to other open LLMs such as Llama 3, Mistral, or the Qwen family (see Section 5). Our continuous pretraining dataset consists of a total number of 4.4 billion tokens.

Monolingual data 40% of the data is domain-specific English data consisting of academic works on Buddhist Studies and related disciplines, as well as translations of Buddhist texts into English acquired from various academic sources, which we processed with Google Cloud Vision OCR. We applied simple rule-based cleaning to remove lines that consist largely of non-alphabetic characters. We deduplicate the dataset on the document level. 20% of the data is Sanskrit and Pāli data collected from various digital resources of Sanskrit and Pāli texts. We do not use any non-corrected OCR Indic material directly. 15% is Buddhist Chinese from the CBETA collection[1] and 5% is Tibetan sourced via the Asian Classics Input Project (ACIP).

Parallel data 20% of the data consists of multilingual parallel sentence pairs. The basis of the dataset are the mined sentence pairs described in Section 3.2. We further collected 1M sentence pairs between Sanskrit and English (publication under preparation). We added 2M sentences between Tibetan and English sourced via our collaborative effort with monlam.ai. The Kumarajiva project[2] contributed 41,000 gold-quality Tibetan↔Buddhist Chinese sentence pairs. We also collected 31,000 gold-quality Sanskrit↔Buddhist Chinese sentences. We further use 149,418 Pāli-to-English sentence pairs.

Continuous Pretraining We train the Gemma 2 model (without instruction fine-tuning) with a size of 9B parameters for two epochs on this dataset. We use an effective batch size of 2M tokens per gradient step. We set the maximum length at 1024 tokens. We used the DeepSpeed library[3] with ZeRO Stage 3 for half-precision training in fp16. The pretraining took four weeks on 8× A100. We refer to this continuously fine-tuned version of Gemma 2 as `Gemma 2 X` in this paper.

Machine Translation Instruction Fine-tuning We use the Claude 3.5 Sonnet API (September 2024) in order to mine 10,000 multi-direction translation examples and 30,979 document-level Sanskrit/Pāli/Tibetan/Chinese-to-English examples. We fine-tuned the model for four epochs on this dataset. In our examination, mining instruction data from high-performing LLMs leads to preferable results over using gold-quality human created sentence pairs, which lead to frequent repetitive hallucinations when used for instruction fine-tuning, which occur much less frequently when using LLM-generated instruction data.

Semantic Retrieval Fine-tuning We fine-tune `Gemma 2 X` for semantic retrieval using contrastive loss with task-specific prompts (Li et al., 2024). Original data for these tasks is scarce, so we augmented it using the Gemini 2.0 Flash API (March 2025). The fine-tuning dataset is detailed in Table 2. In this table, the first two blocks describe retrieval tasks where the goal is to find a matching translation sentence in the target language given a source sentence. Regarding the various tasks, **Keywords (eng)** describes the retrieval of a passage in any of the four languages based on a number of English keywords. **Keywords** describes the same task but with keywords in the same language, i.e., Sanskrit keywords for the retrieval of a Sanskrit sentence. **Questions (eng)** describes the retrieval of a passage in any of the four languages based on a question in English. **Summary (eng)** is the same task, but based on an English summary rather than a question. **Sentence** describes the retrieval of a larger section based on a single, short sentence in the same language.

5 Machine Translation Evaluation

We use manually selected sentence pairs for evaluation of machine translation quality into English for each language: 2,662 sentence pairs of Buddhist Chinese↔English taken from Nehrdich et al. (2025). For Sanskrit, we use a total of 5,552 sentence pairs selected from a number of domains, including Buddhist Sūtras as well as

[1] https://github.com/cbeta-org/xml-p5

[2] https://khyentsefoundation.org/kf-projects/kumarajiva-project/

[3] https://www.deepspeed.ai

Task	Source	Target	Type	Samples
English→X	English	Chinese	Orig	50,000
	English	Pali	Orig	50,000
	English	Sanskrit	Orig	50,000
	English	Tibetan	Orig	50,000
Multilingual	Pali	Chinese	Orig	4,809
	Pali	Pali	Orig	247
	Pali	Sanskrit	Orig	4,613
	Pali	Tibetan	Orig	11,184
	Sanskrit	Chinese	Orig	50,000
	Sanskrit	Tibetan	Orig	50,000
	Tibetan	Chinese	Orig	50,000
Various	Keywords (eng)		Synth	47,223
	Keywords		Synth	47,997
	Questions (eng)		Synth	43,321
	Summary (eng)		Synth	38,882
	Sentence		Synth	51,382

Table 2: Instruction finetuning dataset for semantic retrieval. Type "Orig" describes original data, while "Synth" describes synthetic data.

other domains such as Vedic ritual and poetry (publication under preparation). For Tibetan, we use 4,053 sentence pairs randomly sampled from the entirety of the sentence pair data. Since the Pāli canon is heavily dominated by the Sutta collection with its repetitive language, we sampled 1,900 sentence pairs of mostly non-canonical material (*Jātakagāthāvaṇṇanā*, *Navapadamañjarī*, and *Cariyāpiṭaka*) for which domain-wise very little intersection with our existing training corpus exists. All evaluation data points have been removed from the pretraining/fine-tuning stages of `Gemma 2 X`. While `Gemma 2 X-MT` was fine-tuned on translation into a number of different languages, we limit our evaluation here to translation into English.

We compare the following models: Mistral 7B v0.3 IT (Jiang et al., 2023), Llama 3.1 8B Instruct (Grattafiori et al., 2024), Qwen2.5 7B (Qwen, 2025), Gemma 2 9B IT (Gemma Team, 2024), and Gemma 3 in the 12B IT and 27B IT variants (Gemma Team, 2025).

We present the results of the machine translation evaluation of Gemma 2 X against other open models in Figure 2. We use GEMBA with Gemini 2.0 Flash as evaluation metric due to its strong performance on ancient Asian languages (Nehrdich et al., 2025). The results show that `Gemma 2 X-MT` outperforms all other open LLMs by a significant margin. All models do best on Buddhist Chinese, which indicates that transfer learning from a closely related high-resource language pair, Modern Chinese and English in this case, strongly benefits this idiom. Mistral 7B IT is comparatively strong on Sanskrit, but struggles with Tibetan and Pāli. Llama 3.1 8B Instruct performs rather well on Tibetan, but falls slightly behind on the other languages. Qwen2.5 7B performs comparatively well on Chinese and Pāli, but shows extremely weak performance on Tibetan. Gemma 2 9B IT is very consistent across all four languages. Gemma 3 12B IT further outperforms Gemma 2 9B IT, showing consistent performance increases over time in the Gemma family. Gemma 3 27B IT further outperforms the 12B version on all languages, indicating that larger parameter count does lead to better MT performance. `Gemma 2 X-MT` outperforms Gemma 3 27B IT by a significant margin for all languages. While the performance of Sanskrit, Tibetan, and Buddhist Chinese converge on a similar plateau after fine-tuning, Pāli falls behind. The likely reason for this is that our evaluation data heavily relies on commentarial material, for which very little has been translated into English at all, and it is therefore heavily underrepresented in the training data.

Model	chrF	BLEURT	GEMBA
MITRA NMT ZH-EN	32.14	0.551	67.41
Gemma 2 X-MT	**36.59**	**0.579**	**82.78**

Table 3: Chinese-English translation performance on the MITRA ZH-eval test set. M

We also evaluate the performance of `Gemma 2 X-MT` against the only other domain-specific open-source model `MITRA NMT ZH-EN`[4] for Buddhist Chinese (Nehrdich et al., 2023). `MITRA NMT ZH-EN` is a model based on Facebook AI's 2021 WMT submission (Tran et al., 2021) with further domain-specific fine-tuning. We present the results in Table 3. In this setting, we also evaluate on chrF (Popović, 2017) and BLEURT (Sellam et al., 2020) scores in addition to GEMBA. `Gemma 2 X-MT` outperforms `MITRA NMT ZH-EN` on all metrics, establishing a new state-of-the-art for open models on Buddhist Chinese-to-English machine translation among open models.

6 Retrieval System Evaluation

Our evaluation framework assesses cross-lingual and monolingual passage retrieval across four distinct scenarios:

[4] https://huggingface.co/buddhist-nlp/mitra-mnt-zh-en

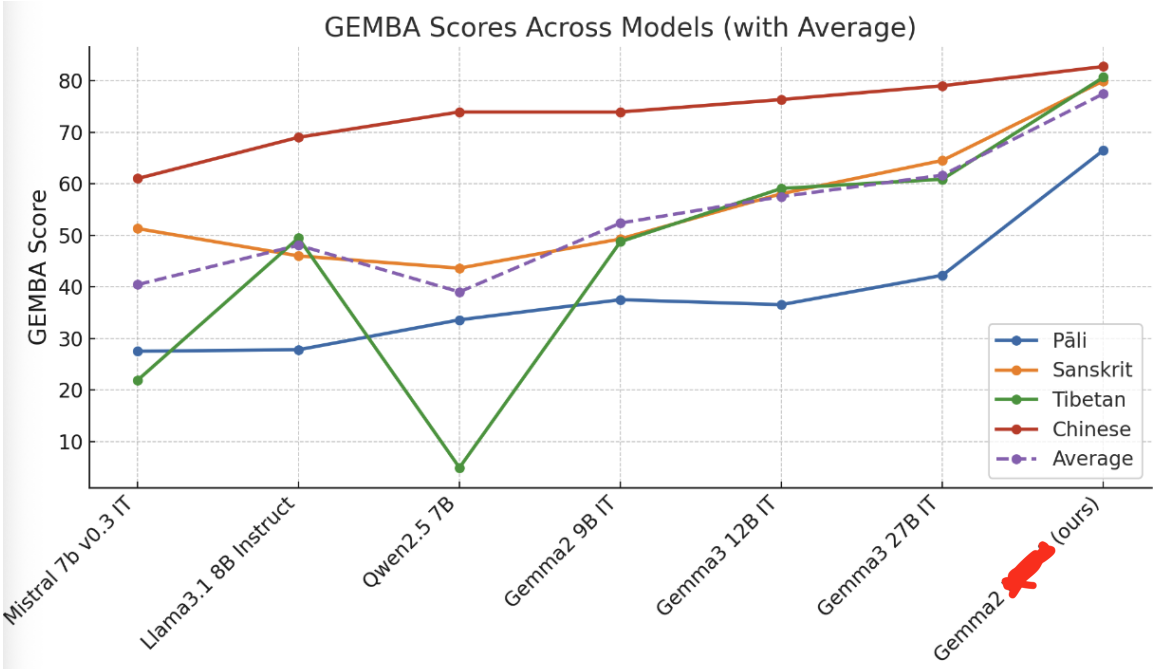

Figure 2: Machine translation performance of different open base models compared to our finetuned model. Performance is measure in GEMBA score, which we implemented with Gemini 2.0 Flash as judge.

Modern English -> Classical Retrieval Retrieving matching Sanskrit, Tibetan, Chinese, and Pāli segments using contemporary English queries. Data consists of manually verified English translations paired with original classical sentences, analogous to the BUCC mining task (Pierre Zweigenbaum and Rapp, 2017, 2018).

Cross-lingual Parallel Retrieval Evaluating precision for retrieving parallel Classical Buddhist text segments across Sanskrit-Tibetan, Sanskrit-Chinese, and Chinese-Tibetan pairs. This manually verified data is also analogous to BUCC.

Verse → Commentary Retrieval Specialized tasks involving retrieving commentary passages from root texts. We test four cases: Sanskrit root → Sanskrit commentary, Chinese root → Chinese commentary, Tibetan root → Tibetan commentary, and the cross-lingual Sanskrit root → Tibetan commentary. These present challenges due to variable passage relatedness, length, and content imbalances.

Cross-lingual QA Retrieval Retrieving answers in Classical languages (Pāli, Sanskrit, Tibetan, Chinese) to English questions.

Evaluation Protocol In order to simulate a real-world retrieval scenario where the right data point needs to be retrieved from a large search space of potentially millions of sentences, we randomly sample a total of 400,412 sentences from the Pāli, Sanskrit, Tibetan, and Chinese target corpora, giving each corpus 25% weight. These sentences are used as negatives in the retrieval setup. We decided on this number of hard negatives to strike a balance between being close to the real-world retrieval application and at the same time making the evaluation, where multiple tasks need to be evaluated in multiple different setups, feasible in reasonable time.

We compare both sparse and dense information retrieval approaches. As sparse method we evaluate `BM25` on the English pivot machine translation. We implement `BM25` via the rank-bm25 Python library utilizing the `BM25Okapi` algorithm. We evaluate the `BGE-M3` embeddings on the English pivot in two configurations: `BGE-M3 base`, where we apply the model as it is, and `BGE-M3 ft`, where we fine-tune it on 100k samples of English-to-English pairs achieved by machine-translating source and target of random sentence pairs of the X-parallel dataset described in Section 3 into English. We evaluate LaBSE directly on the input languages without English pivot, after fine-tuning it on domain-specific parallel sentence data taken from X-parallel. In the same way we evaluate `Gemma 2 X-E` directly on the input languages without English pivot.

Results We present the evaluation results in Table 4. Gemma 2 X-E outperforms all other models by a significant margin on all tasks, the only exception being the P@1 accuracy on the Sanskrit verse to Tibetan commentary retrieval task, where it is outperformed by BGE-M3 base. BGE-M3 ft outperforms BGE-M3 base consistently, demonstrating that the domain-specific fine-tuning of this model yields considerable performance improvements even when pivoting through a different language. LaBSE, which was fine-tuned only on parallel sentence pair data, performs best on the modern English→classical and cross-lingual parallel retrieval tasks, while lagging behind BGE-M3 and Gemma 2 X-E. The gap is more pronounced for the more semantically distant tasks like verse→commentary and cross-lingual QA retrieval. BM25 with English as pivot is outperformed by LaBSE on the first two tasks, but on the commentary and QA retrieval tasks their performance is very comparable. The results show that machine translation as a pivot step in combination with a versatile embedding model like BGE-M3 is an improvement over native multilingual embedding systems such as LaBSE that are primarily trained on parallel data. While BM25 on the machine translation yields results somewhat comparable to LaBSE, it is outclassed by BGE-M3 in all cases, which makes it the preferable model for this data, especially after fine-tuning on domain-specific English data. Gemma 2 X-E significantly outclassing all other models demonstrates that the combination of monolingual and parallel pretraining data, comparatively high parameter count, and task-appropriate instruction tuning creates a model that adapts well to all given tasks.

7 Ablation Study

In order to understand how Gemma 2 X-E performs in comparison to other models in strictly monolingual retrieval settings, i.e., locating a Sanskrit commentary based on a Sanskrit verse without pivoting through English, we conduct an ablation study where we use only monolingual data of about 100k sentences per language as search space. We compare Gemma 2 X-E against three different approaches: BM25, FastText, and BGE-M3. For BM25 and FastText, we apply word segmentation and lemmatization in the case of Sanskrit with the model presented in Nehrdich et al. (2024). For Chinese, we split after each character, effectively treating individual characters as words. For BGE-M3 and Gemma 2 X-E, we use the raw, unprocessed original sentences as input.

The results show that even in this setting, Gemma 2 X-E significantly outperforms all other approaches, the only exception being the Sanskrit text MSABh, where the P@1 retrieval precision of BGE-M3 is higher. In this monolingual setting, BM25 outperforms FastText for Buddhist Chinese, and it matches the performance of BGE-M3 closely. This shows that individual Chinese characters are very efficient signal for sparse retrieval methods. For Sanskrit, on the other hand, FastText shows significant performance advantages over BM25. We assume that even after lemmatization, remaining word segmentation ambiguities and the rich derivational morphology of Sanskrit, where new words can be derived via suffixes or prefixes from existing ones, and nouns can be derived from verbs and vice versa with similar but not completely identical lemmas, yield better performance for the subword-aware FastText algorithm compared to solely word-based BM25 retrieval. In the case of Sanskrit, FastText also outperforms BGE-M3 in 4 out of 6 texts. All in all, the results of our ablation study show that Gemma 2 X-E also performs excellently in a monolingual setting in comparison with other monolingual retrieval techniques. Since it can operate directly on the input string without additional preparation steps such as word segmentation and lemmatization, which are not trivial in the case of Sanskrit, it also allows for a less complex and more unified retrieval pipeline in the monolingual setting as well.

8 Conclusion

We presented X, a complete pipeline that (1) mines noisy multilingual corpora for sentence-level parallels via an MT-pivot + filter/sentence align scheme, (2) compiles the 1.74M-pair X-parallel corpus (89% perfect/mostly-correct alignments in a manual audit), and (3) continuously pretrains and task-fine-tunes a 9B-parameter base LLM, Gemma-2-X.

The machine translation fine-tuned model Gemma-2-X-MT sets a new open-model state-of-the-art on Sanskrit, Pāli, Tibetan and Buddhist Chinese-to-English translation, outperforming the much larger Gemma-3-27B by at least +15

Task Type	Source	Target	BM25	LaBSE	BGE M3 (MT)		Gemma 2 MITRA-E
			(MT)	ft	base	ft	ft
Modern-English-> Classical Retrieval	English	Sanskrit	33·45·50	48·64·70	74·82·85	84·90·92	**95·98·99**
	English	Tibetan	38·50·55	73·85·88	68·79·82	76·86·89	**95·98·99**
	English	Chinese	23·36·40	53·68·73	58·70·74	69·79·83	**90·95·96**
	English	Pāli	28·43·48	30·52·59	53·70·75	60·76·81	**86·97·98**
Cross-lingual Parallel Retrieval	Sanskrit	Tibetan	42·57·62	54·69·74	69·82·85	77·88·91	**93·98·98**
	Sanskrit	Chinese	14·23·28	19·33·39	29·45·51	40·59·65	**79·94·96**
	Chinese	Tibetan	17·27·31	32·50·57	36·54·58	46·63·68	**72·85·87**
Sanskrit Verse-> Sanskrit Commentary Retrieval	BGh	BGh	29·43·48	31·40·45	53·61·63	60·70·74	**89·95·96**
	MSABh	MSABh	05·18·22	07·21·26	11·28·34	11·36·44	**14·64·70**
	DrāhŚS	DrāhŚS	11·16·18	06·10·12	11·16·19	14·21·24	**37·50·53**
	LāṭŚS	LāṭŚS	24·35·40	19·27·31	31·42·46	34·44·49	**54·64·67**
	ŚāṅkhŚS	ŚāṅkhŚS	14·22·25	10·15·18	18·26·29	20·27·30	**50·65·69**
	JaimŚS	JaimŚS	11·18·21	07·11·12	13·21·23	12·21·23	**35·52·58**
Chinese Verse-> Chinese Commentary Retrieval	T1552	T1552	16·26·32	17·27·29	36·52·59	44·60·68	**81·92·94**
	T1600	T1600	12·21·25	15·21·22	26·43·47	37·55·63	**85·90·92**
	T1604	T1604	14·24·29	29·43·50	38·56·64	44·69·74	**88·96·97**
Tibetan Verse-> Tibetan Commentary Retrieval	Text 1	Text 1	11·23·24	19·36·37	13·26·3	14·24·3	**31·90·93**
	Text 2	Text 2	10·21·25	14·21·25	19·41·48	24·49·57	**32·76·83**
	Text 3	Text 3	04·06·08	06·09·11	05·09·12	05·09·12	**15·29·36**
Sanskrit Verse-> Tibetan Commentary Retrieval	Text 1	Text 1	11·23·24	02·05·07	**15**·33·37	03·07·14	14·**43·46**
	Text 2	Text 2	10·21·25	02·03·03	03·06·08	20·41·48	**30·68·77**
	Text 3	Text 3	04·06·08	00·00·03	03·12·16	03·06·09	**10·20·26**
Cross-lingual QA Retrieval	English Q	Pāli A	02·04·05	01·03·05	13·23·27	21·34·41	**36·55·62**
	English Q	Sanskrit A	03·06·08	03·07·09	28·39·45	46·60·64	**56·72·76**
	English Q	Tibetan A	02·04·06	05·11·14	15·26·31	28·43·49	**49·64·69**
	English Q	Chinese A	02·03·05	06·10·13	12·21·26	25·39·44	**57·72·78**

Table 4: Evaluation of different retrieval strategies for various retrieval tasks. All results are reported in P@1·P@5·P@10 accuracy. BM25, FastText, and BGE use machine translation into English as pivotal to enable the crosslingual mapping, while LaBSE and Gemma 2 Mitra Embed use the native language data directly.

Text	BM25	FastText	BGE M3	Gemma 2 MITRA-E
BGh	28·46·54	63·70·75	69·78·80	**90·95·96**
MSABh	16·29·36	21·51·57	**29**·53·58	14·**64·70**
DrāhŚS	05·12·16	26·34·38	10·16·18	**37·50·54**
LāṭŚS	25·41·46	65·76·78	47·60·65	**77·89·92**
ŚāṅkhŚS	11·21·25	39·51·54	24·36·39	**50·65·69**
JaimŚS	10·20·54	28·38·43	20·25·30	**35·51·57**
T1552	61·79·84	11·29·35	68·83·86	**81·92·94**
T1600	59·73·75	25·43·52	59·73·75	**85·90·92**
T1604	71·88·91	20·41·49	69·84·88	**89·96·97**

Table 5: Evaluation of retrieval strategies for monolingual specialized retrieval tasks without pivoting through English machine translation. Results are reported in P@1·P@5·P@10 accuracy scores.

GEMBA on average and the best previous domain-specific model for Buddhist Chinese by +15 GEMBA. Its retrieval sibling, `Gemma-2-X-E`, beats `LaBSE` and `BGE-M3` on our seven-task evaluation.

Beyond NLP tasks, these resources can be of substantial help for philologists: locating a Sanskrit–Chinese parallel or a relevant commentary passage now takes seconds via a search system powered by `Gemma-2-X-E`. Because the mining recipe is language-agnostic, it can be ported to other historical traditions given a sufficiently performing MT and sentence alignment model.

All code, models, evaluation scripts, and the semantic retrieval evaluation dataset are released under open licenses, providing a reproducible testbed for future research. In future work, we hope to expand the `Gemma-2-X` model to include relevant resources in other languages of the Buddhist tradition such as Tocharian or classical Japanese and modern research languages such as Japanese, French, or modern Chinese as well. Another vector of future work is the distillation of the semantic embedding model `Gemma-2-X-E`, since its high parameter count and high dimensionality of the vectors currently make it challenging to apply on large corpora.

9 Limitations

The semantic similarity benchmark currently does not involve any cross-lingual retrieval tasks between Pāli and other ancient Buddhist languages, which is a noteworthy limitation. Adding these data points requires significant manual data annotation. While the current evaluation benchmark is distributed openly, the data used for the machine translation benchmark cannot be made accessible

since we do not hold the rights to these works.

While our paper focuses on Sanskrit, Pāli, Tibetan, and Chinese, other languages of the Buddhist traditions such as classical Japanese, Tocharian, Mongolian, as well as modern material in Japanese, Korean, modern Chinese, French, and more are not yet taken into account.

References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Rafal Felbur, Marieke Meelen, and Paul Vierthaler. 2022. Crosslinguistic semantic textual similarity of buddhist chinese and classical tibetan. *Journal of Open Humanities Data*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Elisa Freschi. 2014. The reuse of texts in indian philosophy: Introduction. *Journal of Indian Philosophy*, 42(2-3):311–331. 10 citations, 120 reads on ResearchGate.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size.

Gemma Team. 2025. Gemma 3 technical report.

Grattafiori et al. 2024. The llama 3 herd of models.

Oliver Hellwig. 2013. Googling the Rishi - Graph Based Analysis of Parallel Passages in Sanskrit Literature. In *Recent Researches in Sanskrit Computational Linguistics: Fifth International Symposium IIT Mumbai, India, January 2013 Proceedings*.

Oliver Hellwig, Sven Sellmer, and Kyoko Amano. 2023. The Vedic corpus as a graph. an updated version of bloomfields Vedic concordance. In *Proceedings of the Computational Sanskrit & Digital Humanities: Selected papers presented at the 18th World Sanskrit Conference*, pages 188–200, Canberra, Australia (Online mode). Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Benjamin Eliot Klein, Nachum Dershowitz, Wolf Lior, Orna Almogi, and Dorji Wangchuk. 2014. Finding inexact quotations within a Tibetan Buddhist corpus. In *Digital Humanities 2014 Conference Abstracts*, pages 486–488.

Sneha Kudugunta, Isaac Rayburn Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Chaofan Li, Minghao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Yingxia Shao, Defu Lian, and Zheng Liu. 2024. Making text embedders few-shot learners. *ArXiv*, abs/2409.15700.

Lei Liu and Min Zhu. 2022. Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities*.

S. Nehrdich, O. Hellwig, and K. Keutzer. 2024. One model is all you need: Byt5-sanskrit, a unified model for sanskrit nlp tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (Findings)*. Association for Computational Linguistics.

Sebastian Nehrdich. 2020. A method for the calculation of parallel passages for buddhist chinese sources based on million-scale nearest neighbor search. *Journal of the Japanese Association for Digital Humanities*, 5:132–153.

Sebastian Nehrdich. 2022. SansTib, a Sanskrit - Tibetan parallel corpus and bilingual sentence embedding model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6728–6734, Marseille, France. European Language Resources Association.

Sebastian Nehrdich, Marcus Bingenheimer, Justin Brody, and Kurt Keutzer. 2023. MITRA-zh: An efficient, open machine translation solution for buddhist Chinese. In *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pages 266–277, Tokyo, Japan. Association for Computational Linguistics.

Sebastian Nehrdich, Avery Chen, Marcus Bingenheimer, Lu Huang, Rouying Tang, Xiang Wei, Leijie Zhu, and Kurt Keutzer. 2025. MITRA-zh-eval: Using a buddhist Chinese language evaluation dataset to assess machine translation and evaluation metrics. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 129–137, Albuquerque, USA. Association for Computational Linguistics.

Serge Sharoff Pierre Zweigenbaum and Reinhard Rapp. 2017. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada. Association for Computational Linguistics.

Serge Sharoff Pierre Zweigenbaum and Reinhard Rapp. 2018. Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Qwen. 2025. Qwen2.5 technical report.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *ACL*.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook AI's WMT21 news translation task submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 205–215, Online. Association for Computational Linguistics.