Mitrasaṃgraha: A Comprehensive Classical Sanskrit Machine Translation Dataset

Anonymous ACL submission

Abstract

While machine translation is regarded as a "solved problem" for many high-resource languages, close analysis quickly reveals that this is not the case for content that shows challenges such as poetical language, philosophical concepts, multi-layered metaphorical expressions, and more. Sanskrit literature is a prime example of this, as it combines a large number of such challenges in addition to inherent linguistic features like sandhi, compounding, and heavy morphology, which further complicate NLP downstream tasks. It spans multiple millennia of text production time as well as a large breadth of different domains, ranging from ritual formulas via epic narratives, philosophical treatises, poetical verses up to scientific material. As of now, there is a strong lack of publicly available resources that cover these different domains and temporal layers of Sanskrit.

We therefore introduce Mitrasaṃgraha, a high-quality Sanskrit-to-English machine translation dataset consisting of 391,548 bitext pairs, more than four times larger than the largest previously available Sanskrit dataset Itihāsa (Aralikatte et al., 2021). It covers a time period of more than three millennia and a broad range of historical Sanskrit domains. In contrast to web-crawled datasets, the temporal and domain annotation of this dataset enables fine-grained study of domain and time period effects on MT performance. We also release a validation set consisting of 5,587 and a test set consisting of 5,552 post-corrected bitext pairs. We conduct experiments benchmarking commercial and open models on this dataset and fine-tune NLLB and Gemma models on the dataset, showing significant improvements, while still recognizing significant challenges in the translation of complex compounds, philosophical concepts, and multi-layered metaphors. We also analyze how in-context learning on this dataset impacts the performance of commercial models.

1 Introduction

Sanskrit is one of humanity's longest-attested literary traditions, encompassing ritual hymns, epics, scientific treatises, philosophical commentaries, poetry and more, composed over a span of three millennia. Yet, despite its enduring impact on South Asian culture, Sanskrit remains a severely low-resource language for contemporary NLP. In machine translation (MT), existing parallel corpora are confined to narrow domains and periods, contain relatively few sentence pairs, and therefore fail to represent the topical breadth and temporal depth of Sanskrit literature. Consequently, Sanskrit-to-English MT, particularly for open, non-commercial systems, lags far behind the progress observed for high-resource modern languages, hindering both digital humanities research and public access to Sanskrit sources.

To address this gap we present Mitrasaṃgraha, a large-scale, carefully curated Sanskrit–English parallel corpus that spans from the early Vedic period until late-medieval scholastic works. Beyond assembling the data, we compare two sentence-alignment strategies tailored to noisy historical texts, create manually inspected development and test splits, validate automatic MT metrics against expert judgments, and benchmark both commercial LLMs and fine-tuned open-source models. We further provide a qualitative error analysis that highlights the specific linguistic phenomena such as compounding, free word order, meter, and polysemy, that make Sanskrit translation uniquely challenging.

This paper makes the following contributions:

- **Mitrasaṃgraha**, the largest public Sanskrit→English MT corpus to date with 391,548 aligned sentence pairs plus 5,587-sentence development and 5,552-sentence test splits covering six literary domains and three millennia of Sanskrit.

- A detailed **sentence-alignment study** on noisy historical texts; among four systems, we find BERTALIGN delivers the highest performance for Sanskrit–English.

- Manually post-corrected **development/test sets** and **document-level metadata** (title, author, period, genre) released to support fine-grained evaluation and domain-transfer research.

- An **empirical comparison of four automatic MT metrics**, where neural BLEURT and LLM-based GEMBA/GEMBA* show the strongest correlation with expert judgments, outperforming BLEU and chrF.

- **Comprehensive MT baselines**: zero-shot commercial LLMs, fine-tuned open-source models, and retrieval-augmented generation setups, illustrating how the new dataset improves Sanskrit machine translation.

We make the dataset and fine-tuned model checkpoint available at XXX. This paper is organized as follows. We first provide background on the Sanskrit language in Section 1.1 and review related work in Section 2. Section 3 evaluates different sentence alignment approaches, while Section 4 describes our data collection workflow. We present the Mitrasaṃgraha dataset in Section 5. We then evaluate machine translation metrics for Sanskrit in Section 6. We show benchmark experiments of different commercial and open-source models in Section 7, followed by an analysis of translation challenges in Section 8 and an ablation study on data cleaning and prompt augmentation in Section 9. Section 10 summarizes the paper. Section 11 describes its limitations.

1.1 The Sanskrit Language

Sanskrit is a classical language of the Indo-Aryan branch of the Indo-European languages. It was widely used as a lingua franca by religious, scientific and literary communities of ancient India. Since the second half of the first millennium BCE, Sanskrit was used for the production of Buddhist literature. Early Buddhist works in Sanskrit show a strong influence of Middle Indian dialects, which has decreased over time (Edgerton, 1985). Sanskrit relies heavily on morphology to indicate grammatical relations and has a relatively free word order. It follows the nominative-accusative alignment, has a complex verbal system and a rich nominal declension. Nominal compounds are frequently found. Another special characteristic and challenge in the computational processing of Sanskrit is the phenomenon of sandhi, by which the contact phonemes of neighboring word tokens are changed and merged, and which creates unseparated strings spanning multiple tokens (Hellwig and Nehrdich, 2018).

2 Related Work

For Classical Sanskrit, Aralikatte et al. (2021) introduced a dataset consisting of 93,000 pairs of Sanskrit verses with their English translations. This dataset consists of two central epic Sanskrit texts, the Mahābhārata and the Rāmāyaṇa, composed between the 4th century BCE and the 4th century CE. Both texts have been translated by the same person, M.N. Dutt in the late 19th century. Regarding contemporary Sanskrit, Team et al. (2022) provide 244k automatically mined sentence pairs. Gala et al. (2023) expand this dataset by adding 27.7k sentences from the Wiki and 5.4k sentences from the Daily domains (day-to-day spoken language). Another Sanskrit dataset was published in Maheshwari et al. (2024). This dataset consists of about 53k sentence pairs, which are written in contemporary prose and therefore do not cover the classical domain. When it comes to other language directions, Nehrdich (2022) provides a Sanskrit-Classical Tibetan sentence-aligned dataset that covers 317k sentence pairs.

3 Sentence Alignment For Sanskrit and English

Aligning Sanskrit-English sentence pairs presents unique challenges compared to high-resource language pairs. Sanskrit's sandhi (phonological changes at word boundaries), frequent compounding, and free word order weaken lexical alignment signals. Additionally, divergent punctuation conventions between Sanskrit editions and English translations create frequent one-to-many/many-to-one mappings, while English texts often contain extraneous material (footnotes, headers) with no Sanskrit counterpart. We evaluated two state-of-the-art alignment algorithms on manually aligned gold-standard data with synthetic noise injection to mimic real-world conditions: VECALIGN (Thompson and Koehn, 2019), and BERTALIGN (Liu and Zhu, 2022). In line with findings on San-

Metric	Itihasa	NLLB	BPCC	Sāmayik	Mitrasaṃgraha
Released	2021	2022	2022	2023	2026
Domains	1	1	1	5	6
Sanskrit composition time	Epic	Modern	Modern	Modern	Vedic, Epic, Classical, Medieval
Number of Translators	1	unknown	unknown	unknown	40+
Number of Datapoints	93,000	244,367	277,467	53,000	391,548
Document-level	✗	✗	✗	✗	✓
Contains Prose	✗	✗	✗	✗	✓
Contains Metrical	✓	✗	✗	✗	✓

Table 1: Comparison of Mitrasamgraha with existing published Sanskrit-English datasets.

skrit–Tibetan alignment by Nehrdich (2022), we focus our benchmark on embedding-based aligners and omit purely length- and lexicon-based methods, which performed substantially worse in our setting. Following Lamraoui and Langlais (2013), we report sentence-level F-measure (F_S), which measures whether correct sentences are paired regardless of many-to-many mappings. BERTALIGN demonstrated superior performance across all evaluated domains, achieving F_S scores ranging from 70.15 (Tattvasaṃgraha) to 97.12 (Manusmṛti), significantly outperforming the other algorithms. Based on these results, we adopted BERTALIGN for creating Mitrasaṃgraha. While certain domains still show room for improvement, the alignment quality is sufficient for downstream machine learning applications. Detailed alignment results and methodology are provided in Appendix A.

4 Data Collection Workflow

We follow a structured multi-step workflow for data collection and preprocessing/postprocessing. First, we collaborate with domain experts of Sanskrit literature to identify Sanskrit texts that have corresponding English translations. We then source these text pairs from publicly available resources.
For data extraction, we employ a systematic preprocessing pipeline. For texts available on public websites, we use automated scraping systems to extract high-quality content from HTML sources. For public domain PDFs, we apply an OCR pipeline using Google Cloud Vision, followed by manual cleanup to remove irrelevant sections such as introductions, indices, and notes. Once the text is preprocessed, we segment the Sanskrit source as well as the English translation into individual sentences using a rule-based approach with manual review, since depending on the input format, different rules need to be applied for successful sentence segmentation. These rules need to be selected and adjusted for each text individually. After the sentence segmentation, we utilize `BertAlign` to generate aligned bitext pairs. Once the automatic alignment is completed, we manually review the output and dismiss files with insufficient alignment quality. For those that we keep, we do a coarse manual inspection where we remove sections that are obviously misaligned.

5 Dataset Description

We provide an overview of the domain composition of the dataset in Figure 2 and compare Mitrasaṃgraha with previously published datasets in Table 1. A detailed per-document breakdown of the dataset is given in Appendix B.
To demonstrate the domain distribution, we assigned high-level categories to each text. These categories are intentionally broad; for instance, "treatises" encompasses Dharmaśāstra literature, medical texts, and commentaries on religious scriptures, among others. The temporal classification of Sanskrit literature presents significant challenges, as philological research typically establishes only approximate ranges, often spanning several centuries. Given these constraints, we divide our dataset into four broad periods: Vedic (before 500 BCE), Epic (500 BCE–200 CE), Classical (200–1000 CE), and Medieval (1000–1800 CE), while acknowledging the inherent imprecision of such categorization.
The Sanskrit sources in our dataset span nearly three millennia, from the second millennium BCE to the early eighteenth century CE. This breadth is complemented by the diversity of translators represented in the corpus, with at least forty identified contributors. Some works, such as the *Mahāsubhāṣitasaṃgraha*, are anthologies compris-

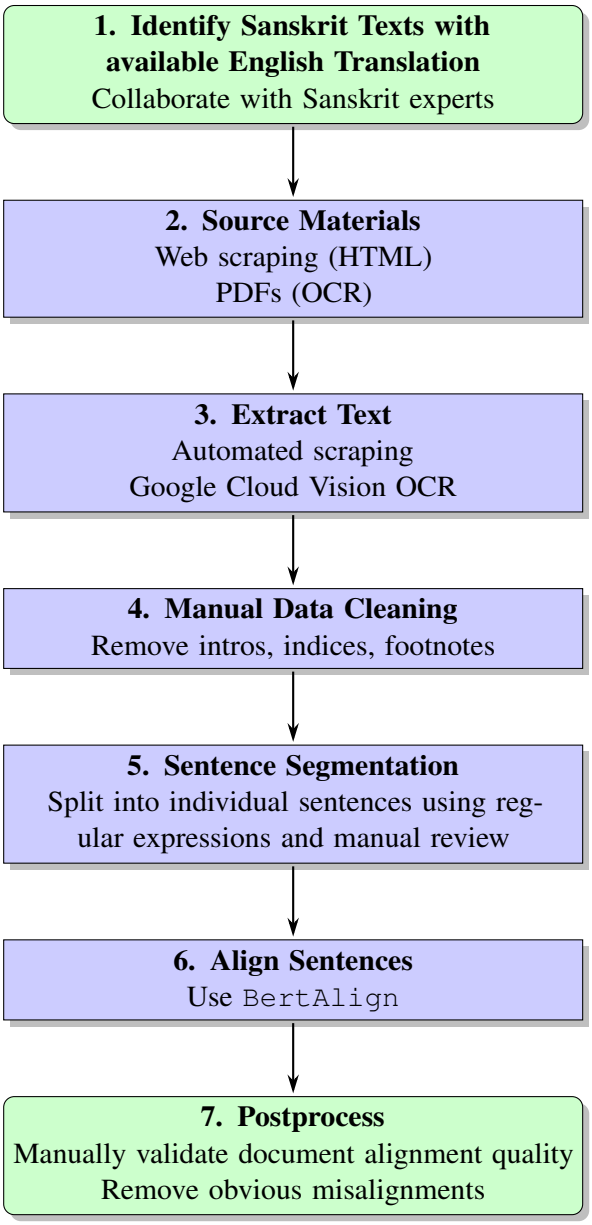

Figure 1: Flowchart of our Sanskrit-English bitext collection and alignment procedure.

ing contributions from numerous additional translators. The combination of temporal range and domain variety means that our dataset captures the linguistic and stylistic evolution of Sanskrit literature across its major historical periods, from the earliest Vedic hymns to late medieval philosophical treatises.

5.1 Compliance of Mitrasaṃgraha with ARR Guidelines for New Dataset Release

We release Mitrasaṃgraha in two versions to accommodate different licensing requirements: (1) a complete dataset under CC BY-NC-ND 4.0 containing all 391,548 sentence pairs, and (2) a research-friendly subset under CC BY-SA 4.0 containing

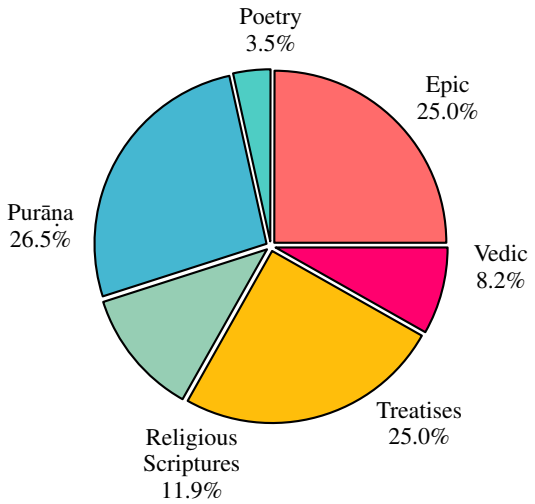

Figure 2: Distribution of Sanskrit Literature Categories

372,791 sentence pairs (excluding 18,757 pairs with BY-NC-ND restrictions). Detailed licensing information for each text is provided in Appendix B. All data collected or scraped from various sources (e.g., websites, PDF books) has been manually verified to be publicly accessible without copyright restrictions. The resources and pre-existing datasets incorporated into our collection are used in a manner consistent with their originally intended purposes (where specified) and are aligned with our declared intended use. Our dataset does not raise any anonymity concerns, as it contains no personally identifiable information or offensive content.

6 Machine Translation Metrics Evaluation for Sanskrit

In order to assess the effectiveness of different automatic metrics for Sanskrit-to-English machine translation evaluation, we asked two expert annotators to independently assess machine-generated translations for 100 sentences. For each of the 100 sentences, we generated three translations with GPT-4o-mini, Claude 3.5 Haiku, and Claude 3.5 Sonnet. We then randomized the order of the translations, without disclosing the identity of the models to the annotators. We asked the annotators to determine for each sentence which of the three translations is the best, and which is the worst. All annotators either hold PhDs or are doctoral candidates specializing in Sanskrit literature. We conducted a one-hour annotation training session where the annotators together evaluated the machine translations of ten sentences, in order to ensure common

Measure	Best	Worst	Avg
Pearson's ρ	0.549	0.554	0.552
Spearman's ρ	0.544	0.551	0.548

Table 2: Inter-annotator correlation between Annotator A and Annotator B on "best" vs. "worst" translation rankings, with the average of both.

evaluation standards. We present the results for the inter-annotator agreement on determining the best and worst translations in Table 2.

Evaluating Sanskrit-English machine translation presents unique challenges that impact the suitability of automatic metrics. Standard n-gram-based metrics such as BLEU and chrF are limited by specific features of Sanskrit. There is the phenomenon of sandhi, where phonemes at word boundaries change and merge, creating ambiguity in word segmentation. Second, Sanskrit has a relatively free word order compared to English, leading to variations in sentence structure that metrics struggle to capture. Additionally, Sanskrit literature often employs complex compounding, figurative language, and compressed verses, making direct surface-level comparison difficult. Finally, many core semantic concepts in philosophical or religious texts lack direct English equivalents, leading humans to adopt different valid approaches. All these factors lead to significant stylistic and structural variation even among high-quality human reference translations.

To quantify the extent of this diversity in translation and its potential impact on evaluation, we conducted a focused analysis. We manually collected seven different high-quality human translations of 100 verses of the *Bhagavad-gītā*. We then measured corpus-level BLEU, chrF, and reference-based BLEURT scores for all 42 possible directed pairwise comparisons among these seven human translations. The average pairwise BLEU score was 25.2, and the average pairwise chrF score was 44.0. This highlights the considerable surface-level divergence even among expert human renditions, showing that multiple perfectly fine English translations can substantially disagree when it comes to surface characteristics such as wording and structure.

We investigate how different metrics correlate with expert human judgment in light of these challenges. In Figure 3 we show how the automatic metrics BLEU, chrF, BLEURT, and GEMBA perform against the human annotations. BLEU (Papineni et al., 2002a) and chrF (Popović, 2015a) capture word n-gram precision/recall and character-level overlap and are calculated on document level. BLEURT (Sellam et al., 2020) uses a BERT model fine-tuned on human judgments of translation quality. We use the "BLEURT-20" checkpoint and average the score across all samples. GEMBA (Kocmi and Federmann, 2023b) uses LLMs to rate machine translation quality. We implement GEMBA using the Gemini 2.0 Flash API (April 2025). In our evaluation, GEMBA denotes the reference-based implementation, where we used the Sanskrit source sentence, the English reference translation, and the machine translation candidate for scoring. GEMBA* denotes the reference-free implementation which only uses the Sanskrit source sentence and the machine translation candidate for evaluation.

Our results show weak correlations for the string-based metrics BLEU and chrF, which is in line with previous research for high-resource languages (Kocmi et al., 2024), which established that these metrics are not suitable to compare machine translation quality across different model types. BLEURT, GEMBA*, and GEMBA all show significantly stronger performance. While the reference-based GEMBA slightly outperforms BLEURT and GEMBA*, it is noteworthy that the reference-free GEMBA* performs very competitively compared to the two reference-based neural models. This outcome is encouraging for the evaluation of domains where reference evaluation data is scarce. In summary, our results show that the neural model based score BLEURT and the LLM-based GEMBA*/GEMBA scores outperform string similarity-based metrics by a clear margin, making them the preferred choice for assessing Sanskrit-to-English machine translation quality across different model types.

7 Benchmark Experiments

In order to quantify the impact of Mitrasaṃgraha and to give an overview of the performance landscape of current MT systems for Sanskrit-to-English, we evaluate three families of translation systems.

Commercial, off-the-shelf LLM/MT APIs (zero-shot): Google Translate, GPT-4o-mini / GPT-4o (OpenAI), Claude 3.5 Haiku (1022) / Claude 3.5 Sonnet (0620) (Anthropic), and Gemini 1.5 Pro / Gemini 2.0 Flash (Google).

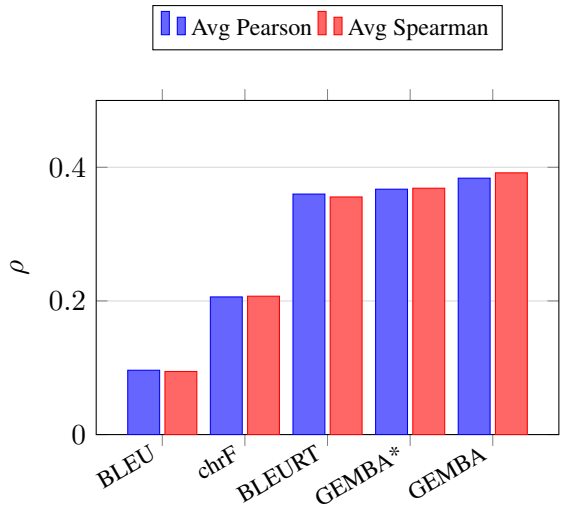

Figure 3: Average Pearson (blue) and Spearman (red) correlations of automatic metrics vs. human judgments. GEMBA* is the reference-free implementation of GEMBA (Kocmi and Federmann, 2023b), while GEMBA is reference-based.

Open-source multilingual MT models, both in their original "zero-shot" form and after supervised fine-tuning on our training split: NLLB-200 600M & 3.3B (Team et al., 2022), and MADLAD-400-3B-MT (Kudugunta et al., 2023). We also test the general-purpose instruction LLMs Llama-3-Instruct (8B) (Grattafiori et al., 2024) and Gemma-2-Instruct (9B) (Team, 2024) by prompting them as translators and by fine-tuning Gemma 2 9B on our corpus.

Retrieval-augmented generation (RAG): the top-performing commercial LLMs (Claude Sonnet, Gemini 1.5 Pro, Gemini 2.0 Flash) are paired with a sentence-level fine-tuned LaBSE retriever that feeds the most similar Sanskrit–English pairs from the training set into the prompt, following the recipe of Ram et al. (2023).

All non-commercial models are decoded with greedy search. Performance is reported on the held-out test split using BLEU (Papineni et al., 2002b), chrF (Popović, 2015b), and GEMBA (Kocmi and Federmann, 2023a); the latter is the metric that showed the strongest correlation with expert judgments in Section 6. We show the results in Table 3. The first group are the commercial baselines, the second group the commercial baselines retrieval-augmented with the Mitrasaṃgraha dataset, the third group the open model baselines, and the fourth group the open models after full-parameter fine-tuning on our dataset.

8 Analysis

We present the results of our comparative benchmark in Table 3.[1] Regarding the results for the **commercial baselines**, it is evident that the machine translation provider Google Translate underperforms compared to all chat systems provided by OpenAI, Claude, and Google. The best performance is achieved by Claude 3.7 Sonnet in BLEURT and Gemini 2.0 Flash in GEMBA. All commercial models that have in-context learning abilities (which excludes Google Translate) benefit significantly from retrieval augmentation, with the effect most strongly pronounced for Claude 3.7 Sonnet.

When it comes to the open model baseline comparison, the decoder-only LLMs Llama-3.1-8B-Instruct and Gemma-2 9B significantly outperform the encoder-decoder models NLLB, MADLAD-400, and IndicTrans2. The encoder-decoder-only models without any fine-tuning struggle to achieve meaningful performance at all and frequently generate nonsensical output. As shown with NLLB, the higher parameter count is not leading to better baseline performance. The decoder-only LLMs are able to provide much better results, but they fall short compared to even the weakest commercial chat systems GPT-4o-mini and Claude 3.7 Haiku, while being somewhat comparable to Google Translate in chrF and BLEURT scores.

The open models all benefit significantly from full-parameter fine-tuning as shown in the last section of Table 3. NLLB with 3.3B size outperforms NLLB 600M in all metrics. Gemma-2 9B fine-tuned significantly outperforms the encoder-decoder models here.

9 Ablation Study

To study the influence of noisy samples on model performance, we conduct an ablation study with

[1]The results in this table were computed using an expanded training corpus that includes an additional augmentation set, increasing the size to a total of 516,000 sentence pairs for training. For licensing reasons, this augmentation material cannot be redistributed, and is therefore excluded from the publicly released version of the dataset. We report results on the expanded corpus to quantify the effect of larger-scale training, while expecting the qualitative trends and relative comparisons to remain stable under the redistributable release.

	Model	chrF	BLEURT	GEMBA
Commercial Baselines	Google Translate	28.28	47.42	56.09
	GPT4o-mini	29.44	50.16	66.39
	GPT4o	32.63	53.53	82.37
	Claude 3.5 Haiku	31.34	52.45	79.10
	Claude 3.7 Sonnet	33.42	**55.28**	86.24
	Gemini 1.5 Pro	31.62	51.20	79.61
	Gmini 2.0 Flash	**34.24**	53.87	**87.04**
Commercial + RAG	Claude 3.7 Sonnet	**38.83**	**57.54**	**89.32**
	Gemini 1.5 Pro	37.07	55.46	84.70
	Gemini 2.0 Flash	38.80	56.69	88.61
Open Model Baselines	NLLB 600M	7.65	33.84	8.08
	NLLB 3.3B	1.43	10.07	0.82
	MADlad400-3b-mt	12.89	33.04	5.19
	IndicTrans2 1B	16.76	39.35	12.23
	Llama-3.1-8B-Instruct	24.35	43.53	33.70
	Gemma2 9b it	**27.58**	47.97	50.64
	Gemma3 12b it	25.78	47.92	55.68
	Gemma3 27b it	27.5	**49.13**	**64.55**
Open Model Finetuned	NLLB 600M	35.59	52.88	71.07
	NLLB 3.3B	36.06	53.89	75.12
	Gemma2 9b	**38.36**	**55.10**	**83.34**

Table 3: Translation performance across different models. Values shown are chrF, BLEURT and GEMBA scores on the test set. For each section, best scores are in **bold**.

different data cleaning methods. Given the alignment results shown in Table 5, we can expect noise to be present in our dataset. We therefore evaluate two approaches for removing noisy datapoints: (1) a deterministic rule-based method combining fixed criteria such as string length ratios with MT evaluation scores using multiple translation models as references; and (2) an LLM-based cleaning method utilizing a specialized prompt pattern.

For the deterministic cleaning approach, we begin with 391,548 datapoints and first filter out samples where the input-output string length ratio falls outside a predetermined range. We then train a Gemma-2-based machine translation model on the remaining dataset to serve as our initial reference. Each sample is evaluated against the MT output of this reference model using a combination of BLEU, chrF, and BERTScore metrics, identifying approximately 20% of datapoints as potentially faulty.

For these flagged sentences, we generate additional reference translations using both Google Translate and Claude 3.5 Sonnet. We then compute evaluation scores (BLEU, chrF, and BERTScore) between the flagged datapoints and these MT/LLM-generated references. Samples with scores below a certain threshold for all three metrics across all three models (our fine-tuned Gemma-2, Google Translate, and Claude 3.5 Sonnet) are removed. This process yields a cleaned dataset of 368,415 datapoints. While using a model trained on the same data to detect faulty samples introduces potential bias, we mitigate this by incorporating external models as additional voters in the pruning process.

For the LLM-based approach, we process the complete set of 391,548 datapoints using Gemini 1.5 Pro, employing a specialized prompt pattern to clean and post-correct the alignments.

We present the results in Table 4. The results show that our deterministic filtering leads to performance increases of +5.01 GEMBA and +0.52 chrF on the fine-tuned NLLB model, while BLEURT shows no improvement. For the RAG setup, the cleaned dataset does not perform significantly better on any metric. Adding grammatical analysis to the Gemini 1.5 Pro as a prompt augmentation technique yields

slight performance increases. However, when combining grammatical analysis with the RAG setup, this significantly decreases performance compared to using the RAG setup alone. The dataset cleaned with Gemini 1.5 prompting visibly underperforms compared to both the uncleaned base dataset and the deterministically cleaned dataset across all metrics.

Model	chrF	BLEURT	GEMBA
NLLB 600M ft (unclean)	35.59	**52.88**	71.07
NLLB 600M ft (clean)	**36.11**	52.85	**76.08**
G1.5 Pro Base	31.62	51.20	79.61
G1.5-Pro + Grammar	32.42	51.52	81.07
G1.5-Pro + RAG	**37.07**	55.46	**84.70**
G1.5-Pro + RAG + Gram	34.57	54.01	84.66
G1.5-Pro + RAG (clean)	36.94	**55.52**	84.49
G1.5-Pro + RAG (G-clean)	35.81	54.83	84.13

Table 4: Results of the ablation study on different data cleaning strategies and augmentation via grammatical annotation.

10 Conclusion

In this paper, we introduced the creation and release of the novel, comprehensive Mitrasaṃgraha dataset for Sanskrit-to-English machine translation, consisting of 391,548 bitext pairs. This dataset is now the largest and most comprehensive publicly available Sanskrit-English parallel corpus, being more than four times larger than Itihāsa and covering multiple domains across Vedic, Epic, Classical, and Medieval time periods. Alongside this training corpus, we contribute manually verified validation and test sets, each comprising approximately 5,500 bitext pairs. Our work included a careful examination of automatic alignment methods, establishing BertAlign in combination with a fine-tuned LaBSE embedding model as the most reliable method for aligning Sanskrit-English sentences.

Our MT metrics evaluation compared different scores, confirming that neural BLEURT and the LLM-based GEMBA score correspond much more closely with human judgment compared to string-similarity-based metrics like BLEU and chrF. We conducted a comprehensive benchmark of various open and commercial models, showing that fine-tuning open models such as NLLB and Gemma 2 on Mitrasaṃgraha leads to consistent performance increases. Furthermore, we demonstrated that in-context learning, by retrieving relevant samples from this dataset, clearly enhances the performance of commercial models.

Despite these advances, we recognize that the complex features of Sanskrit such as sandhi, compounding, multi-layered metaphors, free word order, and more present a uniquely challenging case for machine translation. Current MT solutions still exhibit significant weaknesses in adequately capturing these features. The Mitrasaṃgraha dataset, therefore, offers a foundational resource for low-resource Sanskrit NLP, can aid digital humanities research on Sanskrit literature, and improve public access to Sanskrit cultural heritage.

11 Limitations

Several limitations should be acknowledged. First, the creation of Mitrasaṃgraha faced challenges due to frequent mismatches in sentence segmentation between Sanskrit sources and their English translations, leading to complex one-to-many or many-to-one alignments. While our chosen aligner, BertAlign, is designed for such scenarios, these alignments inherently carry a higher risk of error, potentially introducing noise. Additionally, as large portions of the dataset were sourced from OCR-processed texts, persistent OCR mistakes are likely present despite preprocessing.

Second, while Mitrasaṃgraha provides extensive parallel data, the dataset itself does not inherently solve deep semantic challenges in Sanskrit MT, such as the interpretation of multi-layered metaphors or complex philosophical concepts. Current models, even when trained on this dataset, might still struggle to adequately capture these nuanced linguistic phenomena.

Finally, although Mitrasaṃgraha significantly expands domain and temporal coverage, certain specialized domains within Sanskrit literature, such as technical treatises on astronomy or mathematics, remain entirely missing or are severely underrepresented. Addressing these gaps will require dedicated future data collection and curation efforts.

References

Rahul Aralikatte, Miryam de Lhoneux, Anoop Kunchukuttan, and Anders Søgaard. 2021. Itihasa: A large-scale corpus for sanskrit to english translation. *ArXiv*, abs/2106.03269.

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 169–176, Berkeley, California, USA. Association for Computational Linguistics.

Franklin Edgerton. 1985. *Buddhist Hybrid Sanskrit Grammar and Dictionary. Volume I: Grammar*, volume 1. Rinsen Book Co., Kyoto, Japan. Reprint of the 1953 Yale University Press edition (William Dwight Whitney linguistic series).

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. *Transactions on Machine Learning Research*.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

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

Oliver Hellwig and Sebastian Nehrdich. 2018. Sanskrit Word Segmentation Using Character-level Recurrent and Convolutional Neural Networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2754–2763, Brussels, Belgium. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.

Sneha Kudugunta, Isaac 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 *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks*.

Fethi Lamraoui and Philippe Langlais. 2013. Yet another fast, robust and open source sentence aligner. time to reconsider sentence alignment? In *Proceedings of Machine Translation Summit XIV: Papers*.

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*, 38(2):621–634.

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.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015a. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2015b. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

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

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved Sentence Alignment in Linear Time and Space. 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 1342–1348, Hong Kong, China. Association for Computational Linguistics.

A Detailed Sentence Alignment Evaluation

A.1 Evaluation Methodology

Sentence alignment is considered largely solved for high-resource language pairs with matching punctuation and clean input, where even simple length-based algorithms perform reliably (Gale and Church, 1993; Brown et al., 1991). Sanskrit<>English, however, poses far greater challenges. Sanskrit editions and their English translations frequently diverge in sentence boundaries, omit or reorder passages, and the English side is often cluttered with footnotes, headers, and page numbers that survive OCR but have no Sanskrit counterpart. Divergent punctuation conventions produce many one-to-many or many-to-one mappings, and Sanskrit's pervasive sandhi and prolific compounding weaken the surface lexical cues that many alignment approaches implicitly benefit from.

Following Lamraoui and Langlais (2013), we report two complementary scores: the alignment-level F-measure (F_A), which counts a link as correct only when the entire bisegment exactly matches the gold reference, and the sentence-level F-measure (F_S), which checks whether the correct sentences are paired irrespective of how they are grouped into bisegments. Using both metrics captures performance under a strict criterion that penalizes many-to-many links (F_A) and a more permissive view that tolerates them (F_S). Because many-to-many alignments can harm F_A while leaving F_S relatively unaffected, both measures are necessary to characterize alignment quality.

We begin with manually aligned gold pairs and synthetically corrupt them to mimic realistic noise. Specifically, we (i) randomly split some source or target segments on punctuation, creating one-to-many / many-to-one cases, and (ii) inject random extraneous English sentences (footnotes, headers, etc.). This stresses the aligners' ability to match uneven sentence counts while discarding irrelevant material.

	VecAlign		BertAlign	
Dataset	F_A	F_S	F_A	F_S
Manusmṛti	60.64	82.67	**82.88**	**97.12**
Tattvasaṃgraha	15.11	63.55	**30.00**	**70.15**
Buddhacarita	63.07	79.55	**85.50**	74.62
Itihasa Test	33.54	78.66	**40.44**	**83.39**
Bhāgavata-Purāṇa	64.41	74.79	**72.63**	**85.05**

Table 5: Comparison of VecAlign and BertAlign scores across datasets.

We present the results of the alignment benchmark in Table 5. Across datasets, the embedding-based aligners VECALIGN and BERTALIGN perform strongly under the sentence-level metric F_S, indicating that they typically identify the correct sentence pairings even when boundaries are noisy. BERTALIGN generally outperforms VECALIGN in both F_A and F_S, suggesting it is the more reliable choice for Sanskrit–English alignment in this setting.

Even BERTALIGN attains relatively modest F_A scores, which is expected given frequent many-to-many correspondences and boundary divergence between editions and translations. For downstream use, the F_S scores are more indicative; they range from 70.15 to 97.12, suggesting accuracy sufficient for many machine learning applications. At the same time, the lower-performing domains (e.g., *Tattvasaṃgraha* and *Buddhacarita*) highlight that alignment quality remains uneven and warrants further attention in future work.

B Dataset Catalog

Title	Date	Translator	Period	Category	License	Pairs
Abhidharmakośabhāṣya	350-450	Pruden	Classical	treatises	cc0	10,827
Abhidharmakośakārikā	350-450	Pruden	Classical	treatises	cc0	597
Agni-Purāṇa	700-1000	Ganghadharan	Classical	purana	BY-NC-ND 3.0	14,207
Āpastamba-Dharmasūtra	600BCE-300BCE	Bühler	Vedic	vedic	cc0	307
Arthaśāstra	300BCE-100BCE	Shama Sastry	Epic	treatises	cc0	5,346
Arthaviniścayasūtra	100-800	Ānandajoti	Epic	rel_script	cc0	5,052
Atharvaveda-Saṃhitā	1200BCE-1000BCE	Griffiths	Vedic	vedic	cc0	2,822
Bhāmatī	850-900	Sastri Raja	Classical	treatises	cc0	1,350
Bhāvanākrama	770-780	Sharma	Classical	treatises	cc0	873
Chāndogya-Upaniṣad (Śaṅkarabhāṣya)	700-900	Jha	Classical	treatises	cc0	5,109
Dhammapada	100-300	Ānandajoti	Epic	rel_script	cc0	842
Garga Saṃhitā	1000-1300	Goswami	Medieval	rel_script	cc0	4,436
Gautama-Dharmasūtra	600BCE-400BCE	Bühler	Vedic	vedic	cc0	518
Harivaṃśa	200-500	Dutt	Epic	epic	cc0	7,468
Jātakamālā	400-500	Speyer	Classical	treatises	cc0	2,683
Jīvanmuktiviveka	1350-1400	Swamy	Medieval	rel_script	cc0	1,756
Kathāsaritsāgara	1000-1100	Tawney	Medieval	poetry	cc0	12,515
Kāvyādarśa	500-700	S K Belvalkar	Classical	poetry	cc0	417
Kṛṣṇa-Yajurveda	1200BCE-1000BCE	Keith	Vedic	vedic	cc0	7,501
Kumārasambhava	350-450	Kale	Classical	poetry	cc0	424
Madhyāntavibhāgabhāṣya	350-450	Anacker	Classical	treatises	cc0	630
Madhyāntavibhāgaṭīkā	550-600	Stanley	Classical	treatises	cc0	2,601
Madhyāntavibhāgaṭīkā 1	550-600	Shcherbatskoy	Classical	treatises	cc0	888
Mahābhārata	400BCE-400CE	Ganguly	Epic	epic	cc0	90,789
Manubhāṣya	800-900	Jha	Classical	treatises	cc0	19,919
Manusmṛti	200BCE-200CE	Bühler	Epic	rel_script	cc0	1,384
Meghadūta	350-450	Kale	Classical	poetry	cc0	103
Mīmāṃsāsūtra Tantravarttika	650-750	Jha	Classical	treatises	cc0	6,171
Nyāyamañjari	900-1000	Bhattacharyya	Classical	treatises	cc0	7,763
Pañcaskandhaprakaraṇa	350-450	Anacker	Classical	treatises	cc0	151
Raghuvaṃśa	350-450	Kale	Classical	poetry	cc0	488
Rigveda	1800BCE-1200BCE	Oldenberg	Vedic	vedic	cc0	11,645
Sāmaveda-Saṃhitā	1200-1000BCE	Griffths	Vedic	vedic	cc0	1,343
Sāṃkhyakārikā	300-400	Horace Hayman Wilson	Classical	treatises	cc0	606
Sarva-darśana-saṃgraha	1370-1380	Cowell	Medieval	treatises	cc0	1,755
Śatapatha-Brāhmaṇa	800-699BCE	Eggeling	Vedic	vedic	cc0	6,193
Śiva-Purāṇa	750-1200	Shastri	Classical	purana	cc0	19,136
Skanda-Purāṇa	700-1200	Tagare	Classical	purana	cc0	67,176
Śukla-Yajurveda	1100BCE-900BCE	Griffiths	Vedic	vedic	cc0	1,776
Suśrutasaṃhitā	200-500	Bhishagratna	Classical	treatises	cc0	10,233
Tantrāloka	800-1100	Singh	Classical	rel_script	BY-NC-ND	4,550
Tattvasaṃgraha	760-780	Jha	Classical	treatises	cc0	981
Tattvasaṃgraha Pañjikā	780-790	Jha	Classical	treatises	cc0	17,186
Triṃśikāvijñaptibhāṣya	550-600	Jacobi	Classical	treatises	cc0	635
Triṃśikāvijñaptikārikā	350-400	Anacker	Classical	treatises	cc0	20
Viṃśatikā	350-450	Anacker	Classical	treatises	cc0	162
Viṣṇu-Purāṇa	400-700	Horace Hayman Wilson	Classical	purana	cc0	3,574
Viṣṇusmṛti	300-600	Jolly	Classical	rel_script	cc0	2,831
Yogavasiṣṭha	900-1300	Mitra	Classical	rel_script	cc0	25,801
					Total sentence pairs:	**391,548**

Table 6: Complete catalog of the documents contained in Mitrasaṃgraha with metadata and sentence pair counts.