Mitrasaṃgraha: A Comprehensive Classical Sanskrit Machine Translation Dataset

Sebastian Nehrdich[1] **David Allport**[2] **Jivnesh Sandhan**[3] **Manoj Balaji Jagadeeshan**[4]
Sujeet Kumar[4,6] **Sven Sellmer**[5] **Pawan Goyal**[4] **Kurt Keutzer**[2]

[1]Center for Integrated Japanese Studies, Tohoku University [2]University of California, Berkeley
[3]Kyoto University [4]Indian Institute of Technology, Kharagpur
[5]Adam Mickiewicz University in Poznań
[6]B. P. Mandal College of Engineering Madhepura

Abstract

Although machine translation is often considered solved for high-resource languages, it remains challenging for texts involving poetic language, philosophical concepts, and layered metaphors, characteristics that are central to Sanskrit literature. Additionally, Sanskrit's rich morphology, sandhi, and compounding, combined with its multi-millennial and multi-domain textual tradition, highlight the need for large, diverse parallel resources, which are currently scarce. We introduce Mitrasaṃgraha, a large-scale Sanskrit–English machine translation dataset containing 391,548 aligned sentence pairs, over 4 times larger than the previously available Itihāsa corpus (Aralikatte et al., 2021). The dataset spans more than 3 millennia of Sanskrit literature across multiple domains and includes temporally annotated data for studying domain and period effects on translation performance. We release manually corrected development (5,587) and test (5,552) sets. Benchmarks show that fine-tuning models such as NLLB and Gemma substantially improves translation quality, and retrieval-augmented prompting further enhances performance. In a controlled cross-dataset comparison, training on Mitrasaṃgraha outperforms both Itihāsa and web-mined bitext for two model families. Despite these gains, our results reveal persistent challenges in translating complex compounds, philosophical terminology, and metaphorical language in Sanskrit, highlighting the need for further research in this area.

1 Introduction

Sanskrit represents one of the longest continuous literary traditions in human history, spanning more than 3 millennia and encompassing a wide range of textual genres, including ritual hymns, epic narratives, philosophical treatises, poetry, and scientific literature (Kumar and Gaurav, 2025). Despite its historical and cultural importance, Sanskrit remains severely under-resourced in modern natural language processing (NLP) (Sandhan et al., 2023a; Das et al., 2025). In particular, the development of high-quality machine translation (MT) systems for Sanskrit has been limited by the scarcity of large and diverse parallel corpora (Krishna et al., 2020).

Existing Sanskrit–English parallel datasets (Aralikatte et al., 2021; Team et al., 2022; Maheshwari et al., 2024; Nehrdich, 2022) are typically restricted in size, domain coverage, or historical scope. Many available corpora focus primarily on a small number of texts, often epic literature such as the Mahābhārata or Rāmāyaṇa and frequently rely on translations produced by a single translator. As a result, these datasets fail to capture the linguistic diversity, stylistic variation, and domain breadth present in Sanskrit literature. This limitation hinders the development and evaluation of robust machine translation models and restricts broader computational research on Sanskrit texts.

Beyond data scarcity, Sanskrit poses several intrinsic linguistic challenges for machine translation. The language exhibits rich morphology, extensive compounding, and relatively free word order (Nehrdich et al., 2024; Krishna et al., 2019; Sandhan et al., 2023b). In addition, phonological processes such as sandhi merge adjacent words and alter their surface forms, making tokenization and alignment more difficult. Sanskrit texts also frequently contain dense philosophical concepts, metaphorical language, and poetic structures that can lead to multiple valid translations in English. These characteristics complicate both the training and evaluation of machine translation systems.

To address these challenges, we introduce Mitrasamgraha, a large-scale Sanskrit–English parallel corpus designed to support machine translation research across a broad range of historical periods and literary domains. The dataset contains 391,548 aligned sentence pairs, making it the largest publicly available Sanskrit–English machine translation dataset to date. Unlike previous datasets, Mi-

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.

trasamgraha covers texts composed over a span of nearly three millennia and includes diverse genres such as Vedic literature, epic narratives, philosophical treatises, poetry, and religious texts. In addition, we provide manually post-corrected development and test sets to enable reliable evaluation. The domain/time annotations enable controlled evaluation and error analysis across historical strata and genres, and can support more targeted train/dev/test splits or domain-adaptive training.

Beyond assembling the dataset, we conduct a systematic study of several key components required for building reliable machine translation systems for Sanskrit. First, we evaluate sentence alignment approaches for noisy historical texts and identify BERTAlign as the most effective method for Sanskrit–English alignment. Second, we investigate the suitability of different automatic evaluation metrics for Sanskrit translation by comparing them against expert human judgments. Finally, we benchmark a wide range of translation systems, including commercial large language models, open-source multilingual MT models, and retrieval-augmented generation setups, demonstrating how the Mitrasamgraha dataset improves translation performance.

Our key contributions are as follows:

- We introduce *Mitrasamgraha*, the largest publicly available Sanskrit–English MT dataset with 391,548 aligned sentence pairs across diverse domains and historical periods.
- We release manually verified development and test sets with document-level metadata for reliable evaluation and domain analysis.
- We evaluate automatic MT metrics for Sanskrit and show that LLM-based judges align substantially with expert judgments, while string-based and learned neural metrics (BLEURT, COMET) struggle.
- We provide comprehensive MT benchmarks across commercial models, open-source systems, and retrieval-augmented approaches, together with a per-period and per-domain breakdown.
- We report a controlled cross-dataset comparison showing that Mitrasaṃgraha outperforms both Itihāsa and web-mined bitext as training data for two model families.

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 Mitrasamgraha Dataset

3.1 Data Collection Pipeline

We follow a structured multi-step workflow for data collection and preprocessing/postprocessing as il-

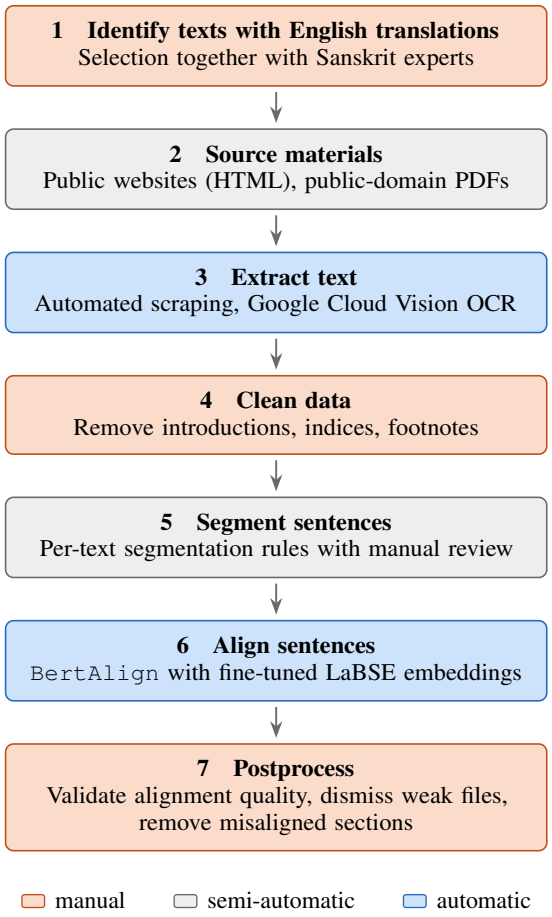

Figure 1: Our Sanskrit–English bitext collection and alignment procedure.

lustrated in Figure 1. 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

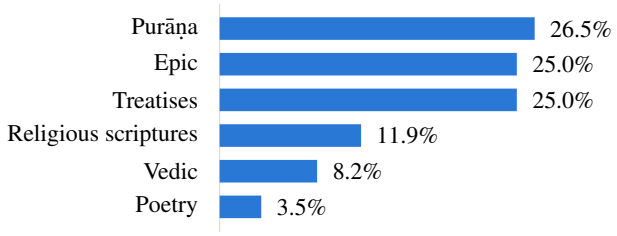

Figure 2: Domain composition of the Mitrasaṃgraha training corpus, as a share of sentence pairs.

insufficient alignment quality. For those that we keep, we do a coarse manual inspection where we remove sections that are obviously misaligned.

3.2 Data Sources

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 D. 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 comprising 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 treatises.

3.3 Sanskrit-English Sentence Alignment

Aligning Sanskrit-English sentence pairs presents unique challenges compared to high-resource lan-

guage 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 Sanskrit–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 B.

3.4 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 372,791 sentence pairs (excluding 18,757 pairs with BY-NC-ND restrictions). Detailed licensing information for each text is provided in Appendix D. 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.

4 Experiments

Baselines: 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 benchmark three families of translation systems.

- **Commercial, off-the-shelf LLMs (zero-shot)**: Google Translate, GPT-4o-mini / GPT-4o (OpenAI), Claude 3.5 Haiku (1022) / Claude 3.5 Sonnet (1022) / Claude 3.7 Sonnet (0219) (Anthropic), and Gemini 1.5 Pro / Gemini 2.0 Flash (Google).
- **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 and beam-search. Performance is reported on the held-out test split using chrF (Popović, 2015a), BLEURT (Sellam et al., 2020), and GEMBA (Kocmi and Federmann, 2023a); the latter is the metric that showed the strongest correlation with expert judgments in Section 5.2.

Main Results: We show the results in Table 2.[1] Among the commercial baselines, Google Trans-

[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 477,236 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.

Model	chrF	BLEURT	GEMBA
Commercial baselines			
Google Translate	28.28	47.42	38.52
GPT4o-mini	29.44	50.16	48.58
GPT4o	32.63	53.53	69.76
Claude 3.5 Haiku	31.34	52.45	64.01
Claude 3.5 Sonnet	33.42	**55.28**	76.97
Claude 3.7 Sonnet	**34.58**	54.74	**77.50**
Gemini 1.5 Pro	31.62	51.20	64.96
Gemini 2.0 Flash	34.24	53.87	75.89
Commercial + RAG			
Claude 3.7 Sonnet	**38.83**	**57.54**	**82.13**
Gemini 1.5 Pro	37.07	55.46	74.20
Gemini 2.0 Flash	38.80	56.69	79.92
Open model baselines			
NLLB 600M	15.04	37.17	6.11
NLLB 3.3B	16.90	37.33	5.10
MADlad400-3b-mt	12.89	33.04	4.00
IndicTrans2 1B	16.76	39.35	6.87
Llama-3.1-8B-Instruct	24.35	43.53	21.79
Gemma2 9b it	**27.58**	47.97	35.79
Gemma3 12b it	25.78	47.92	**47.67**
Gemma3 27b it	27.5	**49.13**	47.48
Open models fine-tuned			
NLLB 600M w/o augment	35.53	52.77	54.23
NLLB 600M	35.59	52.88	53.84
NLLB 3.3B	36.06	53.89	60.24
Gemma2 9b	**38.36**	**55.10**	**72.15**

Table 2: Translation performance across different models (chrF, BLEURT, GEMBA on the test set). GEMBA is computed with the Gemini 3.1 Flash Lite judge validated against expert judgments in Section 5.2. Best scores per section in **bold**. W/o augment: trained without the additional license-restricted augmentation data.

late underperforms all chat systems from OpenAI, Anthropic, and Google. Claude 3.7 Sonnet performs best in chrF and GEMBA, Claude 3.5 Sonnet in BLEURT. All commercial models that have in-context learning abilities (which excludes Google Translate) benefit significantly from retrieval augmentation; the effect is most strongly pronounced for Gemini 1.5 Pro (+9.2 GEMBA), while the best overall results are achieved by Claude 3.7 Sonnet with retrieval augmentation. Among the open model baselines, the decoder-only LLMs Llama-3.1-8B-Instruct and Gemma-2 9B significantly outperform the encoder-decoder models NLLB, MADLAD-400, and IndicTrans2, which without fine-tuning frequently generate nonsensical output; as NLLB shows, a higher parameter count does not improve 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 3.5 Haiku, while being somewhat comparable to Google Translate in chrF and BLEURT scores. All open models benefit significantly from full-parameter fine-tuning (last section of Table 2): NLLB 3.3B outperforms NLLB 600M on all metrics, and fine-tuned Gemma-2 9B outperforms the encoder-decoder models. Notably, NLLB 600M fine-tuned without the license-restricted augmentation data performs on par with the augmented version, indicating that the augmentation does not drive the observed trends.

Performance by period and domain: We break down GEMBA performance by source text and historical period in Table 3. A consistent difficulty hierarchy emerges across all system families: epic and narrative literature (*Bhagavadgītā*, *Rāmāyaṇa*) is translated most reliably, while Vedic texts (especially the *Ṛgveda*) and dense philosophical treatises (*Tattvasaṃgraha*) remain the hardest. Notably, retrieval augmentation helps most where zero-shot performance is weakest as is the case for Claude 3.7 Sonnet, which gains +12.4 GEMBA on the *Ṛgveda* and +13.9 on the *Tattvasaṃgraha*.

Text	GT	G2.0F	C3.7	+RAG	NLLB	G2-9b
Vedic						
Ṛgveda	14.3	60.9	64.8	**77.2**	61.7	63.4
Śatapatha-Brāhmaṇa	20.6	69.7	69.3	**74.7**	50.0	66.4
Epic						
Bhagavadgītā	61.8	**92.1**	92.0	91.8	71.9	81.5
Rāmāyaṇa (Itihāsa)	55.4	84.2	85.2	**86.4**	65.4	81.0
Manusmṛti	40.5	83.8	85.8	**91.3**	78.7	77.4
Buddhacarita	43.1	81.4	84.1	**85.3**	58.4	77.8
Vimalakīrtinirdeśa	40.5	76.5	78.3	**83.3**	64.9	76.3
Classical						
Bhāgavata-Purāṇa	37.2	74.1	74.6	**76.7**	46.0	60.1
Skanda-Purāṇa	58.7	84.5	86.7	**88.1**	68.6	79.2
Yogavāsiṣṭha	35.3	70.3	72.1	**73.9**	49.0	64.0
Tattvasaṃgraha	24.9	61.5	62.8	**76.7**	47.6	67.4
Medieval						
Kathāsaritsāgara	44.5	82.5	84.5	**86.6**	66.8	78.4

Table 3: GEMBA scores on the test set by source text and historical period. GT: Google Translate; G2.0F: Gemini 2.0 Flash; C3.7: Claude 3.7 Sonnet; +RAG: Claude 3.7 Sonnet with retrieval augmentation; NLLB: NLLB 3.3B fine-tuned; G2-9b: Gemma2 9b fine-tuned. Best score per text in **bold**.

Cross-dataset comparison: To isolate the contribution of the dataset itself, we fine-tune NLLB-600M and Gemma2-9B with identical, compute-matched recipes on five training corpora: Itihāsa, Mitrasaṃgraha, its publicly released subset, Mitrasaṃgraha combined with Itihāsa, and the 244,367 highest-confidence pairs of the web-mined NLLB bitext (Table 4).

Three findings hold across both model families. First, on the multi-domain test set, training

on Mitrasaṃgraha clearly outperforms training on Itihāsa (NLLB-600M: +3.8 chrF, +7.0 GEMBA; Gemma2-9B: +3.1 chrF, +5.9 GEMBA), whereas on the epic-only Itihāsa test set the single-domain, single-translator Itihāsa corpus retains its expected advantage. Second, the released subset performs comparably to the augmented corpus (within one chrF point for NLLB-600M, slightly ahead for Gemma2-9B), so the gains do not depend on the license-restricted augmentation material. Third, fine-tuning on the web-mined bitext degrades both models below their own zero-shot performance (GEMBA 1.6–2.9): This shows that for classical Sanskrit, curated parallel data is necessary, not merely preferable. Combining both corpora gives the best overall balance, recovering most of the epic-domain gap at essentially no cost on the multi-domain test set. The weaker performance of the combined dataset on the epic-only test set is likely due to the fact that the diverse and to a large extent non-epic material in Mitrasaṃgraha drags the target domain away from the the specific style and preference of the single-translator Itihāsa corpus.

Training corpus	n	Mitrasaṃgraha test		Itihāsa test	
		chrF	GEMBA	chrF	GEMBA
NLLB-600M					
Itihāsa	69k	25.01	24.98	**28.68**	**55.40**
Mitrasaṃgraha (augm.)	477k	**28.85**	31.93	22.61	40.97
Mitrasaṃgraha (rel.)	391k	27.94	28.17	22.91	37.40
Mitrasaṃgraha + Itih.	546k	28.75	**32.33**	26.36	48.51
NLLB web-mined	244k	10.84	2.85	9.57	2.49
Gemma2-9B					
Itihāsa	69k	29.95	53.78	**29.57**	**72.58**
Mitrasaṃgraha (augm.)	477k	31.33	57.08	22.65	60.85
Mitrasaṃgraha (rel.)	391k	**33.09**	59.66	24.13	67.05
Mitrasaṃgraha + Itih.	546k	33.05	**59.89**	26.71	70.17
NLLB web-mined	244k	12.49	1.63	11.21	1.44

Table 4: Cross-dataset comparison. Both models are fine-tuned with identical, compute-matched recipes (NLLB: 2,000 steps × 512 effective batch; Gemma2: 500 × 256), differing only in the training corpus. *Augm.* is the corpus behind the fine-tuned models in Table 2; *rel.* is the publicly released subset without license-restricted material; *NLLB web-mined* is the 244,367 highest-confidence pairs of the mined bitext (Team et al., 2022). BLEURT is omitted for space and shows no consistent preference between the curated corpora. Best per column within each family in **bold**.

Impact of decoding strategies: We compare greedy decoding and beam search for two unfine-tuned baselines, NLLB-600M and Gemma2-9B-IT (Table 5). For NLLB-600M, beam search yields small but consistent gains across all three metrics; for Gemma2-9B-IT it produces a metric trade-off (higher GEMBA but slightly lower chrF/BLEURT), indicating that decoding shifts the balance between judge-based adequacy and reference similarity.

Model	chrF	BLEURT	GEMBA
NLLB 600M (Greedy)	15.04	37.17	6.11
NLLB 600M (Beam Search)	**15.28**	**37.56**	**6.17**
Gemma2 9b it (Greedy)	**26.69**	**45.89**	31.48
Gemma2 9b it (Beam Search)	26.15	45.32	**32.12**

Table 5: Ablation results for NLLB 600M and Gemma2 9b it with greedy decoding vs. beam search.

5 Analysis

5.1 Ablations with data cleaning methods

To study the influence of noisy samples on model performance, we conduct an ablation study with different data cleaning methods. Given the alignment results shown in Table 8, 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 whose input-output string length ratio falls outside a predetermined range. We then train a Gemma-2-based MT model on the remainder as an initial reference, and score each sample against its output with BLEU, chrF, and BERTScore, flagging roughly 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.

Model	chrF	BLEURT	GEMBA
NLLB 600M ft (unclean)	35.59	**52.88**	53.84
NLLB 600M ft (clean)	**36.11**	52.85	**59.91**
G1.5 Pro Base	31.62	51.20	64.96
G1.5-Pro + Grammar	32.42	51.52	67.05
G1.5-Pro + RAG	**37.07**	55.46	**74.20**
G1.5-Pro + RAG + Gram	34.57	54.01	73.50
G1.5-Pro + RAG (clean)	36.94	**55.52**	73.74
G1.5-Pro + RAG (G-clean)	35.81	54.83	72.65

Table 6: Results of the ablation study on different data cleaning strategies and augmentation via grammatical annotation.

We present the results in Table 6. The results show that our deterministic filtering leads to performance increases of +6.07 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.

The overall patterns across these ablations are as follows. For encoder–decoder finetuning (NLLB), deterministic cleaning consistently improves chrF and GEMBA relative to the unfiltered baseline. For Gemini 1.5 Pro prompting, adding grammatical information helps in the no-retrieval setting, but when combined with retrieved sentence pairs it yields lower scores, suggesting that the additional grammatical scaffold becomes redundant or distracting once parallel evidence is provided. Finally, for the RAG setup, neither deterministic filtering nor LLM-based cleaning shows a consistent improvement across metrics; given the added cost/-complexity and mixed outcomes, we therefore treat data cleaning as optional rather than a default recommendation in this setting.

5.2 Human-Evaluation of Results

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 evaluation standards. We present the results for the inter-annotator agreement on determining the best and worst translations in Table 7.

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

Table 7: Inter-annotator correlation between Annotator A and Annotator B on "best" vs. "worst" translation rankings, with the average of both.

5.3 What does make Sanskrit-English translation task difficult?

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-

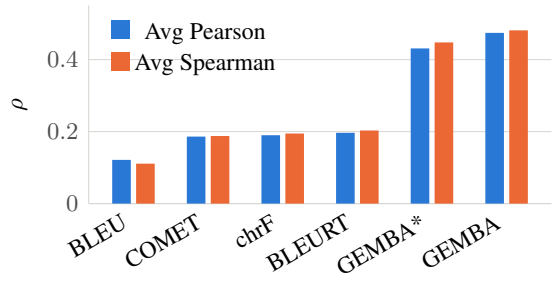

Figure 3: Average Pearson (blue) and Spearman (orange) correlations of automatic metrics vs. human judgments. GEMBA* is the reference-free implementation of GEMBA (Kocmi and Federmann, 2023b), while GEMBA is reference-based.

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, COMET, BLEURT, and GEMBA perform against the human annotations. BLEU (Papineni et al., 2002) and chrF (Popović, 2015b) capture word n-gram precision/recall and character-level overlap and are calculated on document level. BLEURT (Sellam et al., 2020) ("BLEURT-20" checkpoint) and COMET (Rei et al., 2022) (`wmt22-comet-da`) are learned metrics built on BERT and XLM-R respectively, both widely used in the WMT campaigns; we average their scores across samples. GEMBA (Kocmi and Federmann, 2023b) uses LLMs to rate translation quality; we implement it with the Gemini 3.1 Flash Lite API (August 2026). GEMBA denotes the reference-based variant (source, reference, and candidate) and GEMBA* the reference-free variant (source and candidate only).

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 very suitable to compare machine translation quality across different model types. Strikingly, the learned reference-based metrics BLEURT and COMET perform no better than chrF in our setting, all clustering around correlations of 0.19–0.20. We attribute this to the limited domain-specific coverage of the underlying multilingual encoders and the large legitimate surface variation between English renderings discussed above. Only the LLM-based judges depart from this cluster: Gemini-3.1-flash-lite based GEMBA and the reference-free GEMBA* reach correlations of 0.47–0.48 and 0.43–0.45, respectively. The strong performance of the reference-free GEMBA* is encouraging for domains where reference translations are scarce. In summary, the conventional metric suite, string-based and learned reference-based alike, struggle to track expert judgment for Sanskrit; we therefore report GEMBA results along chrF and BLEURT in our experiments.

Two caveats bound these conclusions. First, LLM judges are known to favour outputs from their own model family, and our judge is Gemini-based. The three systems in the annotation study (GPT-4o-mini, Claude 3.5 Haiku, Claude 3.5 Sonnet) include no Gemini system, so the correlations reported here are not themselves affected; the concern applies to the Gemini rows of Table 2, which should be read with that in mind. A systematic cross-judge study of self-preference bias is left to future work. Second, the annotation study covers 100 sentences and three systems judged by two annotators; its correlations are best read against the agreement between the annotators themselves ($\rho \approx 0.55$, Table 7), which gives a rough sense of the noise floor of the task. The study is exploratory and is intended as a starting point for further research on metric evaluation for Sanskrit–English translation.

6 Conclusion

We introduced Mitrasaṃgraha, the largest publicly available Sanskrit-English parallel corpus with 391,548 bitext pairs — over four times larger than Itihāsa — spanning the Vedic, Epic, Classical, and Medieval periods, with manually verified validation and test sets of roughly 5,500 pairs each, aligned using BertAlign with a fine-tuned LaBSE embedding model. Our metric evaluation showed that LLM-based judges correlate substantially with expert assessments, while string-based and learned neural metrics alike struggle with the domain of Sanskrit to English machine translation. Our benchmarks showed that fine-tuning open mod-

els on Mitrasaṃgraha yields consistent gains, that retrieval from it enhances commercial models, and that in a controlled comparison it outperforms both Itihāsa and web-mined bitext. Mitrasaṃgraha thus provides a foundational resource for Sanskrit NLP and digital humanities.

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, it does not by itself solve deep semantic challenges in Sanskrit MT, such as multi-layered metaphors or complex philosophical concepts; models trained on it may still fail to capture these phenomena. Third, our cross-dataset comparison (Table 4) uses a compute-matched but deliberately moderate training budget; the absolute scores of these controlled arms are therefore lower than those of the fully trained models in Table 2, and the comparison should be read as a relative assessment of the training corpora rather than as best-achievable performance. 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.

Acknowledgements

We thank all contributors who supported our data collection efforts, particularly in data scraping and cleaning, including Harsh Gupta (AI Engineer, PredIQt), Abdul Nasir (B.Tech, IIT Kanpur), Yashwant Narsupalli (B.Tech, IIT Kharagpur), Chinmaya H S (Cambridge Institute of Technology, Bengaluru), and Shrey Kamoji and Bibhudutta Pati (IIT Kharagpur); and as interns at the Berkeley AI Research Lab, UC Berkeley: Kush Bhardwaj, Kayshav Bhardwaj, Pragun Seth, Om Chandran, Aarnav Srivastava, Rohan Sarakinti, Varun Rao, Devika Gopakumar, Miranda Zhu, Siya Mehta, Sanjana Srinivasan, Sai Srinivasan, Rhea Mehta, Raj Mehta, Daksh Parikh, and Laksh Patel.

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.

Kunal Kingkar Das, Manoj Balaji Jagadeeshan, Nallani Chakravartula Sahith, Jivnesh Sandhan, and Pawan Goyal. 2025. Still not there: Can LLMs outperform smaller task-specific Seq2Seq models on the poetry-to-prose conversion task? In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 343–357, Mumbai, India. The Asian Federation of Natural Language Processing and The 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.

Amrith Krishna, Bishal Santra, Ashim Gupta, Pavankumar Satuluri, and Pawan Goyal. 2020. A graph-based framework for structured prediction tasks in Sanskrit. *Computational Linguistics*, 46(4):785–845.

Amrith Krishna, Vishnu Sharma, Bishal Santra, Aishik Chakraborty, Pavankumar Satuluri, and Pawan Goyal. 2019. Poetry to prose conversion in Sanskrit as a linearisation task: A case for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1160–1166, Florence, Italy. 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*.

Ravi Kumar and Manish Gaurav. 2025. Sanskrit literary tradition: (re) reading the canon. *The International Journal of Bharatiya Knowledge System*, 2.

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.

Sebastian Nehrdich, Oliver Hellwig, and Kurt Keutzer. 2024. One model is all you need: ByT5-Sanskrit, a unified model for Sanskrit NLP tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13742–13751, Miami, Florida, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jivnesh Sandhan, Anshul Agarwal, Laxmidhar Behera, Tushar Sandhan, and Pawan Goyal. 2023a. SanskritShala: A neural Sanskrit NLP toolkit with web-based interface for pedagogical and annotation purposes. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 103–112, Toronto, Canada. Association for Computational Linguistics.

Jivnesh Sandhan, Yaswanth Narsupalli, Sreevatsa Muppirala, Sriram Krishnan, Pavankumar Satuluri, Amba Kulkarni, and Pawan Goyal. 2023b. DepNeCTI: Dependency-based nested compound type identification for Sanskrit. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13679–13692, Singapore. Association for Computational Linguistics.

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 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).

B Detailed Sentence Alignment Evaluation

B.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.

We present the results of the alignment benchmark in Table 8. 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 reli-

	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 8: Comparison of VecAlign and BertAlign scores across datasets.

able 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.

C Prompt templates

C.1 Open model translation (Llama-3.1-8B-Instruct, Gemma2 9B it, Gemma3 12b it, Gemma3 27b it)

Listing 1: Open model prompt.

```
You are a professional translator. ↩
    ↪Provide only the English ↩
    ↪translation, nothing else.

Translate the following into English:
<x>
Translation:
```

C.2 Commercial LLM translation with retrieval augmentation (Claude 3.7 Sonnet, Gemini 1.5 Pro, Gemini 2.0 Flash)

Listing 2: kNN-RAG translation prompt with optional fields for reference examples and grammatical explanations depending on the evaluation setting.

```
You are a professional translator.
[Optional: Grammar section]
Here is the grammatical analysis of ↩
    ↪the text:

<GRAMMAR_ANALYSIS>


[Optional: Retrieved examples ↩
    ↪section (kNN)]
Here are some similar translation ↩
    ↪examples. Make use of the ↩
    ↪examples to help you ↩
    ↪translate the text:

Source: <SRC_1>
Translation: <TGT_1>

Source: <SRC_2>
Translation: <TGT_2>

...

Now translate the following text to ↩
    ↪English. Provide only the ↩
    ↪translation, without any ↩
    ↪explanation or additional ↩
    ↪information:

<x>
Translation:
```

C.3 Fine-tuned Gemma-2 translation prompt

Listing 3: Fine-tuned Gemma-2 translation prompt.

```
Please translate into English: <x>
Translation:
```

D 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 Tantravartika	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 9: Complete catalog of the documents contained in Mitrasaṃgraha with metadata and sentence pair counts.