Aligned Parallel Corpus of the Vedic Saṁhitās for Machine Translation

Yuzuki Tsukagoshi, Ikki Ohmukai
The University of Tokyo
Tokyo, Japan
{yuzuki, i2k}@l.u-tokyo.jp

Abstract

We introduce a verse-/paragraph-aligned parallel corpus for three Vedic Saṁhitās –the Ṛgveda (ṚV), the Atharvaveda Śaunaka (AVŚ), and the Taittirīya Saṁhitā (TS)– paired with authoritative public-domain translations (Geldner for ṚV, Whitney for AVŚ, and Keith for TS). The source texts are drawn from established digital editions (e.g., TITUS and VedaWeb) and normalized under ISO 15919. Each Sanskrit segment is aligned to exactly one translated unit (verse or paragraph for TS prose), yielding a unified, model-ready format. Using this resource, we fine-tune and evaluate three large language models –GPT-4.1 nano, Gemini 2.5 Flash, and Mitra– on Vedic→German/English translation. Evaluation combines surface and semantic metrics (case-insensitive sacreBLEU and COMET), enabling a balanced assessment of form and meaning. Results show consistent in-domain gains after supervised fine-tuning, but substantial cross-domain degradation when models are tested on unseen Saṁhitās, indicating pronounced stylistic and lexical divergence among ṚV, AVŚ, and TS. These findings motivate domain-aware training and reporting practices for Vedic machine translation. We release the corpus with standardized splits and preprocessing to support reproducibility and future research on historical language modeling, alignment, and translation for low-resource ancient languages.

Keywords: Vedic, Sanskrit, translation, parallel corpus

1. Introduction

Vedic Sanskrit represents the earliest attested stage of Indo-Aryan and differs substantially from Classical Sanskrit in phonology, morphology, syntax, and semantics. Although major Vedic texts have long been available in scholarly translation, this does not mean that they are adequately represented for computational translation research. Most existing translations were produced for philological use rather than for machine learning and they are often not sentence-aligned, not consistently machine-readable, and not easily reusable as parallel data.

This distinction matters because the central bottleneck for neural machine translation (NMT) is not simply whether a text has ever been translated, but whether reliable source–target pairs exist in a form suitable for training, evaluation, and comparison. This problem is especially acute for Vedic Sanskrit, whose archaic grammar, poetic style, and domain-specific vocabulary differ markedly from the Classical Sanskrit texts on which most existing Sanskrit MT research has focused. As a result, the existence of legacy translations does not remove the need for new Vedic MT resources. Rather, it highlights the need to convert philological knowledge into machine-readable data.

To address this gap, this paper presents a new aligned corpus for three core Vedic texts: the Ṛgveda (ṚV), Atharvaveda (AVŚ; Śaunaka recension), and Taittirīya Saṁhitā (TS; one of the Black Yajurveda recensions).

Each Vedic sentence is paired with an academically reliable translation: German translation by Geldner (1951) for the ṚV; English translation by Whitney and Lanman (1905) for the AVŚ; and English translation by Keith (1914) for the TS. All three are public-domain scholarly translations that have informed Vedic studies for over a century. We fine-tune and evaluate three large language models (LLMs) – GPT-4.1-nano, Gemini 2.5 Flash, and Mitra – using BLEU and COMET. We further study cross-domain generalization across ṚV/AVŚ/TS.

Our main contributions are threefold: (i) the first machine-readable, sentence-aligned Vedic Saṁhitā translation corpus (Sanskrit–German and Sanskrit–English); (ii) an empirical analysis of cross-domain generalization across linguistically distinct Vedic sources; and (iii) evidence of substantial domain differences among the three Saṁhitā corpora in both form-based and meaning-based evaluation.

We release the resulting aligned corpus on Hugging Face for reproducible research and future benchmarking.

2. Related Work

2.1. Digitization of Vedic Texts

The digitization of Vedic Sanskrit corpora has progressed through several key initiatives. The TITUS project (Jost Gippert et al., 2016) estab-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3434–3444
11–16 May 2026.

lished foundational machine-readable editions of Indo-European texts, including Vedic materials. General-purpose repositories such as GRETIL (Grünendahl, 2020) and SARIT (Wujastyk et al., 2015) have further aggregated Sanskrit e-texts in standardized formats. GRETIL (Göttingen Register of Electronic Texts in Indian Languages) serves as a general platform for Indic texts, originally providing plain `.txt` files but now including some TEI/XML versions following the Text Encoding Initiative (TEI) guideline (TEI Consortium, 2025). Its collection partially overlaps with TITUS, as both projects digitized similar textual sources. SARIT (Search and Retrieval of Indic Texts), maintained by the TEI community's Special Interest Group for Indic texts, provides TEI-encoded editions of classical and philosophical Sanskrit works, though it does not include Vedic materials.

2.2. Annotated Corpora and Treebanks

Recent work has focused on adding multi-layer linguistic annotations to Vedic texts. The VedaWeb platform (Kölligan et al., 2024) provides the texts, translations, and annotations of the Ṛgveda. All textual data, translations, and annotation layers in VedaWeb are encoded in TEI/XML, ensuring compatibility with digital philological and linguistic standards. In addition, the Atharvaveda Śaunaka corpus distributed via VedaWeb is also provided in TEI/XML format, focusing primarily on philological and editorial information such as variant readings and emendations. The underlying text data originates from the digital edition by Kim (2022).

In parallel with the development of digital corpora, computational infrastructures for Sanskrit analysis have also been established. A representative example is the Sanskrit Heritage platform (Huet, 2026), which provides tools for lexical lookup, morphological analysis, segmentation, and parsing. Although it is best known as a computational platform rather than as a primary corpus publication venue, it has played an important role in Sanskrit research by providing reusable linguistic analysis tools.

Hellwig (2010–2021) developed a deeply annotated Sanskrit corpus with phonological, morphological, and syntactic labeling, which has become a standard reference for computational Sanskrit studies. The Vedic Treebank (Hellwig et al., 2020) extends syntactic annotation to approximately 4,000 sentences across multiple Vedic layers using the Universal Dependencies framework.

2.3. Machine Translation for Sanskrit

2.4. Pre-Neural Machine Translation

Early work on Sanskrit machine translation (MT) relied on rule-based and hybrid approaches suited to the language's complex morphology and syntax. English-to-Sanskrit Translator and Synthesizer (ETSTS) by Rathod and Sondur (2012) exemplified this phase of research: it used handcrafted transfer rules augmented with example-based techniques, and included a text-to-speech module. Similar rule-driven systems soon followed: English Speech to Sanskrit Speech (ESSS) combined a lexical parser and sandhi (phonological alternations at sound boundaries) rules with text-to-speech synthesis (Pragya Shukla, 2014). Other projects integrated bilingual dictionaries and Sanskrit grammatical rules for reordering and inflection. The ANN-augmented model by Mishra and Mishra (2009) is a notable example, mixing symbolic grammar rules with a neural network to select Sanskrit words by part-of-speech. Around the same time, researchers began exploring statistical MT. Sandeep R. Warhade (2012) built a phrase-based SMT system (using Moses) for English→Sanskrit, demonstrating the viability of statistical decoding for Sanskrit. These early systems underscored the importance of linguistic knowledge (sandhi handling, case endings, etc.) in Sanskrit MT before the advent of large parallel corpora.

2.5. Sanskrit Neural Machine Translation

The shift to neural machine translation (NMT) brought new techniques to Sanskrit MT. Koul and Manvi (2021) proposed one of the first Sanskrit→English NMT models, using an LSTM-based encoder–decoder with attention. To compensate for limited training data, they integrated a two-pronged strategy: a partial Sanskrit–English dictionary for simple words and a machine-learning classifier for complex compound words. Specifically, an SVM classifier was used to identify English equivalents of long Sanskrit compounds, yielding approximately 10% higher accuracy than a naive Bayes baseline. This hybrid neural approach improved translation speed and accuracy over purely statistical baselines. Shortly after, Punia et al. (2020) assembled ~9,000 parallel Sanskrit–English sentences to train a Transformer-based NMT model. They found that augmenting this low-resource model with Sanskrit monolingual data and transfer learning from related Hindi dramatically boosted performance. BLEU improved from only 4.6 (a baseline Transformer) to 18.4 after leveraging Hindi-trained weights. This demonstrates the potential of cross-lingual trans-

fer learning between closely related Indic languages. Around the same time, large parallel corpora for epic texts became available. Itihāsa, released by Aralikatte et al. (2021), provided ~93,000 verse pairs from the Rāmāyaṇa and Mahābhārata. Benchmarking on Itihāsa revealed that even state-of-the-art Transformers struggle on highly inflected Sanskrit poetry, whose translation quality was poor, highlighting the complexity of Sanskrit verse. Nonetheless, Itihāsa filled a critical resource gap for training and evaluating models on classical epic Sanskrit. Neural models began to decisively outperform older systems as data grew.

Sanskrit NMT research also extended into domain-specific translations. For example, Pandey et al. (2022) introduced a gated recurrent neural network model for translating Vedic Sanskrit verses. Their encoder–decoder (3 layers each) was tailored to Vedic hymns and achieved promising BLEU scores ($\simeq$45–46) on this specialized task. Meanwhile, Raulji et al. (2022) revisited Sanskrit$\rightarrow$Gujarati MT with modern tools and unveiled a novel symbolic framework combining grammatical transfer rules with neural components, yielding a BLEU of ~58.

Recently, the Sanskrit NMT landscape has benefited from major dataset releases and initiatives to address low-resource shortcomings. One milestone is Sāmayik, a parallel corpus of contemporary Sanskrit–English prose introduced by Maheshwari et al. (2022). Sāmayik contains about 52,961 sentence pairs drawn from modern domains (educational texts, online tutorials, religious discourse) representing Sanskrit used in recent prose. This is significant because previously available corpora (like the epic-focused Itihāsa) skew heavily toward poetry and archaic literature. In fact, most digitized Sanskrit text collections sum to under 1 million sentences and rarely include modern content. Another notable resource is the Bhasha Parallel Corpus unveiled by Mujadia and Sharma (2025). Bhasha is a massive multilingual dataset of 44 million sentence pairs across 7 Indic languages, intended to support multi-Indic translation research.

Beyond data, researchers have turned to pretraining and pipeline architectures tailored for Sanskrit NLP. ByT5-Sanskrit, a byte-level Transformer language model for Sanskrit introduced by Nehrdich et al. (2024) is not a translator per se, but a general pretrained model spanning tasks like word segmentation, lemmatization and parsing. ByT5-Sanskrit was trained on the large Digital Corpus of Sanskrit (Hellwig, 2010–2021) and achieves state-of-the-art results in fundamental Sanskrit NLP tasks. This demonstrates the value of Sanskrit-specific pretrained models in boosting downstream translation quality by addressing the language's complex morphology upfront. Similarly, Isac and Das (2025) proposed an end-to-end framework called SLIP (Sanskrit Linguistic Intelligence Pipeline) to combine traditional linguistic analysis with modern NMT. SLIP incorporates classic Sanskrit grammar processing — morphological feature extraction (verb tenses, case endings), phonological rules (sandhi splitting, syllable patterns), compound detection, etc. — and feeds these into enhanced Transformer and retrieval-augmented generation models. In experiments translating Bhagavad Gītā verses, SLIP showed substantial gains over vanilla models. For instance, a Transformer augmented with SLIP features achieved a +10.2% relative improvement in BERTScore (from 57.63 to 63.51), and using a retrieval-augmented translator within SLIP yielded BLEU-4 increases up to 41.8%. These are huge boosts in translation quality, indicating that injecting Sanskrit-specific linguistic knowledge (morpho-phonological cues, etc.) gives the models a clear advantage. A comparative analysis found SLIP-enhanced models markedly outperformed generic multilingual ones like mBART. Such results reaffirm that linguistically informed architectures can overcome many challenges of Sanskrit MT that pure end-to-end models struggle with.

Despite these advances, most neural MT research still targets Classical Sanskrit prose. Little work has addressed verse alignment in Vedic Saṁhitā or explored cross-domain transfer between Vedic and Classical varieties. Existing benchmarks seldom evaluate whether models trained on epic or prose corpora generalize to archaic Vedic grammar. The lack of parallel corpora for early Vedic texts inhibits the development and evaluation of such systems. Our study addresses this gap by constructing the first parallel corpus of early Vedic verses and establishing translation benchmarks. Importantly, the existence of traditional scholarly translations does not by itself solve this problem, because such translations are rarely distributed as clean, sentence-aligned, machine-readable parallel data.

3. Dataset

3.1. Source Texts

Our parallel corpus covers three major Vedic Saṁhitās: the Ṛgveda (ṚV), the Atharvaveda Śaunaka (AVŚ), and the Taittirīya Saṁhitā (TS) from the Black Yajurveda. To ensure textual fidelity, we rely on established critical editions that have been digitized by modern projects, and we segment the source texts according to their native verse- or paragraph-level structure.

For the Ṛgveda (ṚV), we use the digital edition

of Aufrecht (1877), available through the VedaWeb database (Kölligan et al., 2024). This edition contains 1,028 hymns comprising 10,552 individual *ṛc* (verses). Our dataset follows the standard Śākala recension and segments each hymn into individual verses.

For the Atharvaveda Śaunaka (AVŚ), we use the digital edition provided by TITUS (Jost Gippert et al., 2016). The TITUS digital edition of the AVŚ (Petr and Vavroušek, 1996) is based on the printed editions of Orlandi (1991) and Roth and Whitney (1855) and preserves the canonical structure of the Śaunaka recension.

For the Taittirīya Saṁhitā (TS), which belongs to the Black (Kṛṣṇa) Yajurveda and combines metrical hymns with prose ritual instructions, we use the digital version provided by TITUS. The TITUS edition of TS (Fushimi, 1997) is based on the printed edition of Weber (1871). Because the Black Yajurveda contains both verse and prose, our segmentation scheme distinguishes metrical and prose sections. Metrical hymns are divided by verse, while prose passages are segmented by paragraph, following Keith's translation structure (see below for details).

Across all three Vedic texts we adhere to verse-level granularity wherever possible, reflecting the canonical structure of the source editions. For the prose sections of the TS, paragraph-level grouping ensures semantic coherence. This consistent segmentation facilitates training of sequence-to-sequence models and supports fine-grained alignment between Vedic source texts and their translations.

3.2. Translation Data

For each Saṁhitā we pair the Vedic text with a scholarly translation to create a well-formed parallel corpus. We extract translations from digitized public-domain sources and strip away commentary and notes so that only the translated verses remain. Each translated verse (or paragraph, for TS prose) is aligned with a unique Sanskrit segment, enabling precise sentence-level alignment.

For the Ṛgveda (ṚV), we use the German translation by Geldner (1951), widely regarded as the standard scholarly edition of the text. The VedaWeb database (Kölligan et al., 2024) provides a digitized version of Geldner's translation aligned with other versions including Aufrecht's critical edition. We link each German translation segment to the corresponding Vedic verse.

For the Atharvaveda Śaunaka (AVŚ), we use the English translation by Whitney and Lanman (1905), available via Wikisource[1]. We extract the verse translations from Whitney's extensive notes and commentary and align them with the verse order of Roth and Whitney's critical edition.

For the Taittirīya Saṁhitā (TS), we employ the English translation by Keith (1914), obtained from the digitized version on Sanskrit Web (Stiehl, 1999-2025). Since Keith's translation includes both verse and prose passages, we segment the prose by paragraph following his own structure to maintain semantic coherence.

During dataset construction we removed annotations, editorial notes and commentary so that only the literal translation remains. We also identified several numbering errors in the Whitney and Keith translations, and manually corrected these to ensure reliable alignment between Vedic text and its translation. Finally, we link each translation segment to the corresponding Vedic sentence ID. This careful curation ensures that our parallel corpus reflects authoritative translations while preserving the verse structure required for machine translation research.

3.3. Normalization and Release Format

All Sanskrit texts are transliterated according to the ISO 15919 standard, following contemporary conventions for Romanized Indic scripts. The released dataset [2] is available in both JSON and CSV formats, containing the fields `text_id`, `source`, `target`, and `target_language`. In addition, we provide a balanced subset in which each corpus contributes an equal number of verses, enabling controlled comparison across textual sources and mitigating corpus size bias.

3.4. Corpus Statistics

Table 1 summarizes the basic statistics of our parallel corpora. "vsn" is an ISO 639-3 language code of Vedic Sanskrit. The combined dataset comprises 18,608 aligned verse or prose pairs drawn from the ṚV, AVŚ, and TS, covering distinct textual traditions, genres, and translation styles. We note that "words" are whitespace-delimited strings, and "tokens" are computed with `tiktoken` for compatibility with downstream evaluation.

The Ṛgveda (ṚV) portion is the largest, with over 10,000 verse-level units. Its structure is predominantly metrical, resulting in relatively short and rhythmically compact sentences. Despite its verse-based segmentation, the ṚV shows high lexical density: 160K Sanskrit words expand to 280K words in the German translation, reflecting both the inflectional compactness of Vedic Sanskrit and

[1] https://en.wikisource.org/wiki/Atharva-Veda_Samhita

[2] The full dataset is publicly available on https://huggingface.co/datasets/yzk/veda-samhita-translation.

Corpus	Lang	#Sentence	Words (Vedic)	Tokens (Vedic)	Words (Target)	Tokens (Target)
ṚV	vsn–deu	10,552	160K	756K	280K	589K
AVŚ	vsn–eng	4,883	70K	361K	125K	212K
TS	vsn–eng	3,173	100K	492K	187K	272K
Total		**18,608**	**332K**	**1.6M**	**594K**	**1.0M**

Table 1: Dataset statistics.

the more analytic syntax of German. Tokenization further reveals that the Sanskrit side (756K tokens) contains substantial morphological variation due to *sandhi*, whereby adjacent sounds undergo phonological changes at morpheme or word boundaries, and compounding, which inflates the token count relative to the word count.

In contrast, the Atharvaveda Śaunaka (AVŚ) corpus, with about 4,800 aligned pairs, exhibits simpler verse structures and a more expository translation style in English. Its average Sanskrit segment length is shorter than the ṚV, but the English translations tend to paraphrase more freely, resulting in a modest expansion ratio (about 1.8$\times$ from 70K to 125K words). The relatively low token count (Vedic 361K vs. target 212K) indicates that the lexical distribution is somewhat more balanced and less morphologically dense than in the ṚV.

The Taittirīya Saṁhitā (TS) diverges markedly in style. Unlike the predominantly metrical ṚV and AVŚ, the TS includes large portions of prose, which we segment by paragraph rather than by metrical unit. This results in fewer total samples (3,173) but significantly longer sentences on average. Although its total Sanskrit word count (100K) is lower than that of the ṚV, it yields nearly half a million tokens (492K) with a much higher token-to-sentence ratio, confirming the narrative and ritual prose character of this text. The English translation is similarly substantially longer (187K words), partly because translators tend to expand syntactically compact Vedic prose into full explanatory clauses.

Across all corpora, the Vedic side totals roughly 332K whitespace-separated words and 1.6M tokenizer-level tokens, while the translations (German and English combined) contain 594K words and 1.0M tokens. The discrepancy between word and token counts is systematic: it largely reflects the segmentation behavior of modern tokenizers, which tend to split Vedic text extensively due to *sandhi* concatenation as well as the presence of accent diacritics. These features increase the apparent token count, whereas English and German, with more stable word boundaries, show a more moderate difference between words and tokens, as their analytic morphology and clearer word boundaries reduce token fragmentation.

In terms of length distribution, Sanskrit segments average 17.9 words (median 15), with translations averaging 31.9 words (median 27). This consistent expansion reflects both the explanatory tendencies of translators and the typological gap between the highly inflected source language and more analytic target langauges. Overall, the corpus provides not only quantitative balance but also stylistic and typological diversity across verse and prose traditions. This diversity is beneficial for both machine translation benchmarking and studies of domain variation within Vedic texts and translations.

4. Experiments with LLMs

4.1. Task and Models

We define the Saṁhitā translation as a sequence-to-sequence generation task, in which a Vedic input sequence is mapped to its corresponding translation in either English or German, depending on the corpus. This setting allows us to evaluate how modern large language models handle the linguistic complexity, archaic morphology, and stylistic variation of Vedic Sanskrit.

To explore the effectiveness of different architectures and fine-tuning strategies in this specialized low-resource domain, we compare three representative large language models with capacities and training paradigms:

- **GPT-4.1 nano** (`gpt-4.1-nano`): A compact, general-purpose model that serves as a strong efficiency-oriented baseline for sequence-to-sequence generation. Despite its small size, it serves as a strong modern baseline for efficient sequence-to-sequence generation.
- **Gemini 2.5 Flash** (`gemini-2.5-flash`): A fast and memory-efficient model from the Gemini series, fine-tuned on randomly sampled subsets of 2,000 examples per corpus. The fine-tuning follows the platform's recommended supervised setup, designed to ensure consistency across domains of varying size.

- **Mitra**[3]: A domain-adapted model based on the Gemma 2 architecture, pretrained on Buddhist and Sanskrit textual data. It is fine-tuned using parameter-efficient Low-Rank Adaptation (LoRA) (Hu et al., 2022), with rank 16, α=16, no dropout, and a learning rate of 3e^{-4}. Optimization uses AdamW (8-bit) with a linear learning-rate scheduler, 10 warmup steps and weight decay 0.01. This configuration balances stability and adaptability when specializing to low-resource, morphologically rich ancient textual domains.

These models are adapted via supervised fine-tuning, enabling controlled comparisons across architectures and training regimes on a morphologically rich historical language.

4.2. Training and Evaluation

We adopt two supervised fine-tuning (SFT) setups, aligned with our evaluation protocols:

- **Single-corpus SFT:** Each model is fine-tuned exclusively on one corpus (ṚV, AVŚ, or TS), capturing corpus-specific linguistic features and stylistic norms. We then evaluate the resulting models both *in-domain* (e.g., ṚV→ṚV) and *cross-domain* (e.g., ṚV→AVŚ/TS), in order to quantify domain transferability across Vedic text types.

- **Mixed SFT:** Each model is fine-tuned jointly on the combined dataset (ṚV+AVŚ+TS), thereby exposing it to the full spectrum of lexical, metrical, and stylistic variation present in the three Saṁhitās. Evaluation is conducted on the merged test set.

Translation quality is evaluated using both surface-level and semantic metrics: we report case-insensitive sacreBLEU (Post, 2018) for lexical accuracy, and COMET (Rei et al., 2022) using the pretrained `Unbabel/wmt22-comet-da` model for semantic adequacy and fluency. This dual-metric evaluation provides a balanced view of performance, capturing both form-based and meaning-based translation quality.

Overall, this experimental design ensures a direct correspondence between training and evaluation conditions. The Single-corpus regime highlights the trade-off between specialization and generalization, while the Mixed regime examines whether multi-domain exposure enhances translation robustness across heterogeneous Vedic corpora.

[3] https://huggingface.co/buddhist-nlp/gemma-2-mitra-it

5. Results

Model	BLEU	COMET
GPT-4.1 nano (before FT)	3.13	0.496
GPT-4.1 nano (after FT)	4.01	0.517
Gemini 2.5 Flash (before FT)	39.7	0.598
Gemini 2.5 Flash (after FT)	19.3	0.611
Mitra (before FT)	0.202	0.364
Mitra (after FT)	14.6	0.568

Table 2: Performance of each model before and after supervised fine-tuning (SFT) on the combined test set (ṚV+AVŚ+TS).

Table 2 presents the performance improvements achieved by models fine-tuned on the combined corpus. Most models exhibit consistent gains across both metrics; however, Gemini shows a substantial decrease in BLEU alongside a small increase in COMET. This divergence is likely due to Gemini's tendency to include source-language words in parentheses within the translations, which negatively impacts n-gram overlap while leaving semantic adequacy largely unaffected.

Table 3 reports the cross-domain generalization results, where models trained on a single corpus are evaluated on the test sets of the remaining corpora. Overall, performance drops considerably when models are applied to unseen Saṁhitā domains, indicating substantial stylistic and lexical divergence across ṚV, AVŚ, and TS. Among the three, the Gemini 2.5 Flash model shows the strongest cross-domain robustness, achieving moderate BLEU and COMET scores even on out-of-domain tests. For example, a model trained on AVŚ still attains BLEU 10.7 and COMET 62.1 on TS. GPT-4.1 nano exhibits the most domain-sensitive behavior: its BLEU scores sharply decline outside the training corpus, suggesting limited transfer of stylistic and lexical knowledge. Mitra, though competitive within-domain (notably on TS, with BLEU 20.2 and COMET 64.2), transfers poorly to other Saṁhitās, reflecting its stronger specialization but narrower generalization capacity. These trends collectively highlight that domain adaptation remains a key challenge for Vedic translation, as stylistic conventions and translational norms vary considerably among the Saṁhitā corpora.

Importantly, these systematic differences in cross-domain performance confirm that each Saṁhitā constitutes a distinct translation domain, with characteristic diction, syntax, and translation conventions. This validates our design choice to construct a balanced, verse/paragraph-aligned dataset that explicitly preserves corpus bound-

Training Corpus	Model	RV test		AVŚ test		TS test	
		BLEU	COMET	BLEU	COMET	BLEU	COMET
RV	GPT-4.1 nano	2.97	49.7	2.98	52.4	3.32	54.5
	Gemini 2.5 Flash	14.6	58.1	7.60	57.3	5.01	56.9
	Mitra	6.75	51.6	0.251	44.0	0.092	44.6
AVŚ	GPT-4.1 nano	0.610	36.8	8.96	56.6	3.66	55.0
	Gemini 2.5 Flash	0.374	42.4	28.6	67.3	10.7	62.1
	Mitra	0.226	36.4	11.3	52.6	3.26	46.8
TS	GPT-4.1 nano	1.15	42.6	4.83	52.9	10.7	61.2
	Gemini 2.5 Flash	0.613	40.2	4.00	54.6	17.7	61.0
	Mitra	0.246	40.9	7.38	56.3	20.2	64.2

Table 3: Cross-domain evaluation of models. Each model is trained on one corpus (RV, AVŚ, or TS) and tested on all three datasets. Underlines mark in-domain results.

aries while enabling joint or comparative training. By aligning and normalizing these corpora under a unified format, our dataset provides a reliable foundation for Vedic machine translation tasks.

6. Conclusion

We presented the first verse-aligned parallel corpus of the Vedic Saṁhitās, constructed through careful manual curation, normalization, and cross-textual alignment across three canonical sources: the Ṛgveda, the Atharvaveda Śaunaka, and the Taittirīya Saṁhitā. Using this resource, we benchmarked multiple large language models on Sanskrit→German/English translation and analyzed their cross-domain behavior. The results revealed strong within-domain learning effects but persistent degradation when models were evaluated on unseen Saṁhitā, highlighting the linguistic and stylistic diversity that characterizes early Vedic literature.

Beyond serving as a new resource, our findings emphasize that conventional multilingual fine-tuning does not fully capture morphophonological processes such as sandhi, complex nominal compounding, or the interaction between metrical and syntactic boundaries. Addressing these aspects will require linguistically informed preprocessing and feature integration, including explicit sandhi segmentation, compound decomposition, and meter-aware tokenization strategies. In addition, retrieval-augmented or hybrid symbolic-neural translation frameworks may provide a principled way to handle formulaic repetitions and lexically conservative passages common in ritual and hymnal texts.

Future work will extend this dataset to additional recensions and later ritual layers, explore domain-adaptive fine-tuning and multitask setups that combine translation with linguistic annotation, and systematically assess transfer between Vedic and Classical Sanskrit. All corpus splits are publicly released to facilitate reproducibility and further research in ancient-language modeling, translation, and digital philology.

7. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 25K21518 and by The Nippon Foundation HUMAI Program.

8. Bibliographical References

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

Theodor Aufrecht. 1877. *Die Hymnen des Rigveda*. Bonn: Adolph Marcus.

Makoto Fushimi. 1997. Taittirīya-saṃhitā. Based on Albrecht Weber's edition (*Die Taittirīya-Saṃhitā*, Leipzig: Brockhaus, 1871–1872, *Indische Studien* 11–12). Edited by Makoto Fushimi, Ōsaka; TITUS version prepared by Jost Gippert, Frankfurt a.M., updates 1997–2012.

Karl F. Geldner. 1951. *Der Rig-Veda : aus dem Sanskrit ins Deutsche übersetzt und mit einem laufenden Kommentar versehen*. Number v. 33-36 in Harvard oriental series. Harvard University Press, Oxford University Press, Otto Harrassowitz, Cambridge, Mass., London, Leipzig.

Reinhold Grünendahl. 2020. GRETIL: Göttingen Register of Electronic Texts in Indian Languages. https://gretil.sub.uni-goettingen.de/gretil.html.

Oliver Hellwig. 2010–2021. Dcs - The Digital Corpus of Sanskrit.

Oliver Hellwig, Salvatore Scarlata, Elia Ackermann, and Paul Widmer. 2020. The treebank of vedic Sanskrit. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5137–5146, Marseille, France. European Language Resources Association.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Gérard Huet. 2026. The sanskrit heritage engine reference manual. https://sanskrit.inria.fr/manual.html. Written on February 10th 2026 for Sanskrit Engine Version 3.76.

N. Biraja Isac and Himansu Das. 2025. SLIP: A Sanskrit Linguistic Intelligence Pipeline for Enhanced Neural Machine Translation of Classical Texts. *IEEE Access*, PP(99):1–1. License: CC BY 4.0.

Jost Gippert, Javier Martínez, Agnes Korn, and Roland Mittmann. 2016. Thesaurus Indogermanischer Text- und Sprachmaterialien.

Arthur Berriedale Keith. 1914. *The Veda of the Black Yajus School, entitled Taittiriya Sanhita, translated from the original Sanskrit prose and verse*. Number 18-19 in Harvard Oriental Series. Harvard University Press, Cambridge, Massachusetts.

Jeong-Soo Kim. 2022. *Atharvavedasaṃhitā der Śaunakaśākhā : Eine neue Edition unter besonderer Berücksichtigung der Parallelstellen der Paippalādasaṃhitā*. Universität Würzburg.

Nimrita Koul and Sunilkumar S. Manvi. 2021. A proposed model for neural machine translation of Sanskrit into English. *International Journal of Information Technology*, 13(1):375–381.

Daniel Kölligan, Claes Neuefeind, Uta Reinöhl, Patrick Sahle, Antje Casaretto, Anna Fischer, Börge Kiss, Natalie Korobzow, Jürgen Rolshoven, Jakob Halfmann, and Francisco Mondaca. 2024. VedaWeb. Online Research Platform for Old Indic Texts.

Ayush Maheshwari, Nikhil Singh, Amrith Krishna, and Ganesh Ramakrishnan. 2022. A benchmark and dataset for post-OCR text correction in Sanskrit. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6258–6265, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vimal Mishra and R. B. Mishra. 2009. Ann and rule based model for english to sanskrit machine translation. *INFOCOMP Journal of Computer Science*, 9(1):80–89.

Vandan Mujadia and Dipti Misra Sharma. 2025. Bhashaverse : Translation ecosystem for indian subcontinent languages.

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.

Chatia Orlandi. 1991. *Gli inni dell'Atharvaveda (Śaunaka): traslitterazione*. Giardini, Pisa. Transliteration of the Atharvaveda (Śaunaka) recension.

Mrinal Pandey, Rashmikiran Pandey, and Alexey Nazarov. 2022. Machine translation of vedic sanskrit using deep learning algorithm. In *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, pages 1477–1480.

Vladimir Petr and Petr Vavroušek. 1996. Atharvaveda saṃhitā (Śaunaka recension). Based on the editions by Chatia Orlandi (Pisa 1991) and Rudolf Roth & William D. Whitney (Berlin 1856). Entered by Vladimir Petr and Petr Vavroušek, Praha 1996; TITUS version prepared by Jost Gippert, Frankfurt a.M., updates 1997–2012.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Akanksha Shukla Pragya Shukla. 2014. English speech to sanskrit speech (esss) using rule based translation. *International Journal of Computer Applications*, 92(10):37–42.

Ravneet Punia, Aditya Sharma, Sarthak Pruthi, and Minni Jain. 2020. Improving neural machine translation for Sanskrit-English. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 234–238, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Sarita G. Rathod and Shanta Sondur. 2012. English to sanskrit translator and synthesizer (etsts). *International Journal of Emerging Technology and Advanced Engineering*, 2(12):379–382.

Jaideepsinh K. Raulji, Jatinderkumar R. Saini, Kaushika Pal, and Ketan Kotecha. 2022. A novel framework for sanskrit-gujarati symbolic machine translation system. *International Journal of Advanced Computer Science and Applications*, 13(4).

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.

Rudolf Roth and William Dwight Whitney. 1855. *Atharva Veda Sanhita*. Friedrich Dümmlers Buchhandlung, Berlin.

Prakash R. Devale Sandeep R. Warhade, Suhas H. Patil. 2012. English-to-sanskrit statistical machine translation with ubiquitous application. *International Journal of Computer Applications*, 51(1):41–46.

Ulrich Stiehl. 1999-2025. Sanskrit-Web.

TEI Consortium. 2025. TEI P5: Guidelines for Electronic Text Encoding and Interchange.

Albrecht Weber. 1871. *Die Taittirīya-Saṃhitā*. F. A. Brockhaus, Leipzig. Published in *Indische Studien*, Vols. 11–12 (1871–1872).

William Dwight Whitney and Charles Rockwell Lanman. 1905. *Atharva-veda Saṁhitā / translated with a critical and exegetical commentary, by William Dwight Whitney; revised and brought nearer to completion and edited by Charles Rockwell Lanman.* Number 2 v. in Harvard oriental series ;7-8. Harvard University, Cambridge, Mass.

Dominik Wujastyk, Birgit Kellner, and Sheldon Pollock. 2015. SARIT: Search and Retrieval of Indic Texts. TEI-encoded digital library of Sanskrit and Indic texts. Established 2013.

A. Reproducibility of Preprocessing and Postprocessing

This appendix describes the preprocessing and postprocessing procedures used in the experiments. The goal is to make explicit which transformations are applied to the data before training and evaluation, and which are not. All steps are deterministic.

A.1. Source Data

The experiments use three parallel corpora, referred to as `avs`, `rv`, and `ts`. Each record consists of a verse identifier (`text_id`), a Sanskrit source string (`source`), and a reference translation (`target`).

A.2. Data Partitioning

Each corpus is partitioned independently into training, validation, and test sets. The split ratio is 70%/15%/15%.

In addition to corpus-specific splits, the training, validation, and test portions of the three corpora are concatenated to create combined splits used in experiments on the full dataset collection.

A.3. Construction of the 2k Subsets

For corpus-specific experiments, an additional capped setting is constructed for each corpus. When a corpus contains more than 2,000 examples, exactly 2,000 instances are sampled using random sampling. Otherwise, the full corpus is retained. The resulting subset is then partitioned by 70%/15%/15%.

This design ensures that the smaller experimental condition is also fully reproducible.

A.4. Language Labels and Prompt Instantiation

Each processed example is assigned an explicit target-language label. The `rv` corpus is labeled as `German`, whereas `avs` and `ts` are labeled as `English`. This label is used only to instantiate the translation prompt and does not otherwise alter the source or reference text.

All models use the same prompt structure:

```
Please translate the following
Sanskrit text to {target_language}.

### Input:
{source}

### Translation:
```

At inference and evaluation time, `{target_language}` is replaced with the language label associated with the example, and `{source}` is replaced with the original Sanskrit input string. No few-shot examples, retrieval-based augmentation, auxiliary metadata, or source-side reformulation are added to the prompt.

A.5. Inference-Time Postprocessing

Postprocessing is intentionally minimal.

For Mitra, the generated text is first decoded. If the decoded output begins with the prompt verbatim, that prompt prefix is removed, and the remaining text is trimmed for leading and trailing whitespace. If the output does not begin with the prompt exactly, no internal modification is made and only outer whitespace is removed. This step is intended solely to recover the predicted translation when the model reproduces the prompt in its output.

For OpenAI models, the prediction is taken directly from the returned assistant message, after trimming leading and trailing whitespace. No further normalization is applied.

For Gemini models, the prediction is taken from the returned text field when a valid candidate is available. Responses that are blocked or malformed are excluded from scoring. Apart from this validity check, no additional normalization is performed.

Across all model families, the pipeline does not apply sentence segmentation, detokenization, punctuation repair, case normalization, or manual post-editing before evaluation.

A.6. Scope of Reproducibility

In summary, reproducibility in this study depends on four explicit design choices: a shared prompt template with target-language labels, minimal and model-specific output extraction rules, and BLEU calculation with `sacrebleu` and COMET with Unbabel's `comet`. No hidden normalization, undocumented filtering, or manual correction stage is introduced outside these procedures.

B. Examples of Generated Translations

This section presents representative examples from each corpus. For the Ṛgveda, the reference translations used in this study are in German, whereas those for Atharvaveda Śaunaka and Taittirīya Saṁhitā are in English.

B.1. ṚV Example

The following example from the Ṛgveda is short and self-contained, and the qualitative differences among the model outputs are easy to observe.

Text ID RV 5.82.4

Source *adyā́ no deva savitaḥ prajā́vat sāvīḥ saúbhagam párā duṣvápnyaṁ suva*

Reference Mögest du, Gott Savitr, uns heute kinderreiches Glück zuweisen. Weise üblen Traum ab!

Gemini Weise uns heute, Gott Savitr, kinderreiches Glück zu, halte den bösen Traum fern!

GPT Heute, o Gott Savitr, soll uns dein Gunst, die uns wie eine Nachkommenschaft zuteil wird, von dem schlechten Glück befreien, das uns umgibt.

Mitra "Heute gib uns, Gott Savitr, die Erlabung, das Glück, die (guten) Träume zu gewinnen, du guter Seher! –"

B.2. AVŚ Example

The following Atharvaveda Śaunaka example uses an English reference translation and clearly illustrates differences in how the systems handle negation and syntactic structure.

Text ID AVS 5.30.6

Source *ihaídhi puruṣa sárveṇa mánasā sahá dūtaú yamásya mānu gā ádhi jīvapurā́ ihi*

Reference Be thou here, O man, together with thy whole mind; go not after Yama's (two) messengers; go unto the strongholds of the living.

Gemini Be thou here, O man, with all the mind together; go not after Yama's messengers; go unto abodes of the living.

GPT Here, O man, go thou with all thy mind along with the messenger of Yama; go thou over the life-places.

Mitra 1. Come thou hither, O man, with all mind; in company with Yama's two messengers, O living one, go thou up to the city of the Fathers.

In this example, the output of Gemini preserves the negation correctly and remains relatively close to the reference translation, whereas the outputs of GPT and Mitra show more substantial semantic deviations.

B.3. TS Example

The following Taittirīya Saṁhitā example also provides a clear contrast among the systems. Gemini is nearly identical to the reference, whereas the outputs of GPT and Mitra drift to an unrelated passage.

Text ID TS_7_4_20=b

COMET (Gemini) 0.879

Source *vásavas tvā chrndantu gāyatréṇa chándasāṅgirasvád rudrās tvā chrndantu traíṣṭubhena chándasāṅgirasvád ādityās tvā chrndantu jā́gatena chándasāṅgirasvád*

Reference Let the Vasus anoint thee with the Gayatri metre. Let the Rudras anoint thee with the Tristubh metre. Let the Adityas anoint thee with the Jagati metre.

Gemini Let the Vasus anoint thee with the Gayatri metre; let the Rudras anoint thee with the Tristubh metre; let the Adityas anoint thee with the Jagati metre.

OpenAI From the stabs (the pointed arrow) the sky, from the stabs the atmosphere, The measure of the earth, all worlds were filled; All the worlds are the vows of Varuna.

Mitra 3: The bull hath established sky and atmosphere; He hath meted the greatest extent of earth; He hath sat down as lord of all these worlds; All these are in the ordinances of Varuna.