Santham: A Curated Sanskrit–Tamil Dataset with Anvaya and Segmentation for Building and Evaluating Machine Translation

Prasanna Venkatesh T S, Ketaki Mangesh Shetye, Vishnuraj Arjunasamy, Ayush Kumar Sahu, Sriram Krishnan, Parameswari Krishnamurthy
IIIT Hyderabad
{vipranarayan14, sriramk8, ayush.sahu0621}@gmail.com
ketaki.shetye@research.iiit.ac.in, vishnuraj.ka@students.iiit.ac.in
param.krishna@iiit.ac.in

Abstract

Sanskrit-to-Tamil machine translation remains significantly under-researched due to the lack of high-quality parallel corpora and the deep morphological and syntactical divergences between the two languages. To address this gap, we introduce **Santham**, a curated Sanskrit-Tamil parallel dataset comprising over 90,000 pairs drawn from classical texts such as the *Mahābhārata*, *Rāmāyaṇa*, and *Bhagavad Gīta*, as well as modern prose collections. We establish a human-evaluated benchmark to assess translation quality and investigate the efficacy of linguistic preprocessing techniques. Specifically, we explore state-of-the-art segmentation tools to resolve *sandhi* (euphonic combinations) and *samasa* (compounds), and utilize *anvaya* (prose-order reordering) to mitigate the structural complexity of poetic verses. Our experiments involve fine-tuning state-of-the-art Large Language Models (LLMs), such as Gemma and Qwen, which demonstrate better performance compared to specialized multilingual models like IndicTrans2. The results indicate that while segmentation yields performance gains, the use of *anvaya* leads to substantial improvements in translation quality, achieving a relative increase of over 46% in BLEU scores. This work underscores the importance of domain-specific data curation and linguistic preprocessing in advancing low-resource translation for classical languages.

1 Introduction

For centuries, Sanskrit served as the *lingua franca* of the Indian subcontinent, acting as the primary vehicle for discourse on ethics, politics, medicine, and astronomy. This vast repository of literature, collectively forms the rich Indian knowledge tradition. However, as the linguistic landscape shifted, the ability to access these texts became confined to a dwindling number of scholars. To democratise this knowledge and ensure that the wisdom encoded in Sanskrit literature reaches the modern public, translating these works into contemporary Indian regional languages is not merely a scholarly exercise but a cultural necessity.

Historically, the preservation of this knowledge relied on manual translation. While these efforts produced high-quality translations, the current era faces a shortage of bilingual experts proficient in both Sanskrit and modern vernaculars like Tamil. As the global demand for holistic traditional knowledge grows, offering potential solutions for contemporary issues in wellness, linguistics, and philosophy, the gap between the source texts and the audience widens.

LLMs offer a major shift in how we handle text; unlike older systems, they can interpret and explain content rather than just translating it. With a shortage of human experts, these tools are the most practical way to unlock Indian traditions and potentially uncover ancient insights that could solve modern problems.

Translating Sanskrit remains a non-trivial task due to its unique structural complexities:

- **Morphological Richness:** Sanskrit is highly inflectional, where a single word can carry a dense amount of grammatical information.

- **Sandhi and Sāmasa:** The extensive use of *Sandhi* (euphonic combinations) and *Samasa* (compounds) requires a mandatory and sophisticated segmentation preprocessing step to identify word boundaries and improve translation accuracy.
- **Syntactic Flexibility:** A significant portion of Sanskrit literature is composed in verse (*śloka*). These *ślokas* leverage Sanskrit's free word-order nature, creating a vast divergence from the relatively more structured syntax of target languages like Tamil.

Despite the technological progress in machine translation (MT), most research has prioritised Sanskrit-English or Sanskrit-Hindi language pairs. Sanskrit-to-Tamil machine translation remains significantly under-researched, primarily due to the lack of high-quality parallel corpora. Preliminary experiments using Rule-Based Machine Translation (RBMT) have proven unsustainable; the deep morphological and syntactical divergences between Sanskrit (Indo-Aryan) and Tamil (Dravidian) make the manual creation of rules extremely cumbersome.

Furthermore, while existing general-purpose LLMs can perform Sanskrit-to-Tamil translation, they often struggle with accuracy, particularly when translating complex poetic constructions. This paper addresses these limitations by introducing a specialized framework for fine-tuning open-source LLMs on a high-quality, curated dataset.

In this paper, we present **Santham**, a dedicated Sanskrit-Tamil translation dataset. Our primary contributions are: (1) **High-quality curated parallel corpus:** We introduce a large, curated Sanskrit-Tamil parallel dataset designed for MT tasks. (2) **Preprocessing for Translation Quality:** We explore different segmentation models as a preprocessing step to handle the complexities of *sandhi* and *sāmasa* in Sanskrit source texts. We also demonstrate how *anvaya* (prose-order reordering) significantly improve translation quality compared to raw poetry data. (3) **Evaluation:** We provide an evaluation of our two LLM models, Gemma and Qwen, fine-tuned with our datasets, which demonstrates better performance in translating both simple prose and complex poetic verses.

2 Related Work

The systematic development of computational tools for Sanskrit has a long history, recently summarized in the comprehensive survey by Pradeep and Mamidi (2025). While early efforts focused on digital infrastructure and rule-based linguistics, the field has rapidly transitioned toward neural and Large Language Model (LLM) architectures.

2.1 Sanskrit Machine Translation (MT)

Machine translation for Sanskrit has primarily targeted English and Hindi. The *Anusaaraka* systems, part of the *Saṃsādhanī* project, represent the pinnacle of rule-based approaches, utilizing Pāṇinian grammar to bridge Sanskrit with various Indian languages. In the last decade, neural approaches have further expanded this scope; while neural models have been extensively explored for Sanskrit-Hindi (Sethi et al., 2023), recent research has also successfully targeted Sanskrit-Malayalam (Chingamtotattil and Gopikakumar, 2022). More recently, commercial platforms like Google Translate (GT)[1] have introduced Sanskrit support, though academic analysis of its performance on Sanskrit-Tamil remains non-existent. Our preliminary tests suggest that GT's general-purpose training often fails to capture the nuances of classical poetic structures, establishing a clear need for domain-specific models.

2.2 Data Scarcity and Recent Benchmarks

The lack of high-quality parallel data is the primary bottleneck for Sanskrit-Tamil MT. Currently, the NLLB corpus (NLLB Team, 2022) is the only publicly available source for this pair, but its quality is notably low for classical or technical texts.

[1] `https://blog.google/products/translate/24-new-languages`

In contrast, other language pairs have benefited from large-scale data releases. The *Itihasa* corpus (Aralikatte et al., 2021) for Sanskrit-English and *Samayik* (Maheshwari et al., 2024) for Sanskrit-Hindi have set new benchmarks. For Sanskrit-Hindi, recent work by (Kammar et al., 2024) utilizing multimodal neural architectures has further improved translation accuracy. The most significant recent development is the MITRA project (Nehrdich and Keutzer, 2026) and the Mitrasamgraha dataset (Nehrdich et al., 2026), which provide state-of-the-art parallel alignments for Sanskrit, Pāli, and Tibetan, although their primary target remains English and Chinese.

Translating Sanskrit to Tamil is uniquely challenging due to high translation divergence. While Sanskrit-Hindi pairs share an Indo-Aryan lineage, Sanskrit and Tamil belong to different language families i.e. Indo-Aryan and Dravidian, leading to deep morphological and syntactic differences.

Moreover, translating Sanskrit poetry is challenging due to its complex meters, figures of speech, and deep philosophical themes (Lisha C.R, 2024). Existing translation methods generally range from literal approaches, which focus on word-for-word accuracy, to creative adaptations, which prioritize the overall meaning and artistic style. While each method has its strengths, maintaining both the original structure and clear meaning remains difficult. Motivated by these challenges, we focus on two key interventions: applying segmentation to handle complex *sandhi* and *samasa*, and uilizing *anvaya* (prose-order reordering) to normalize poetic syntax before translation.

3 Data Preparation

Sanskrit literature is rich and diverse, encompassing prose, poetry like Rāmāyaṇa, Mahābhārata, and various technical texts, including Āyurveda, Nyāya and Vedānta. Therefore, we are collecting and preparing data from a wide array of Sanskrit texts, including poetry, prose, sastric, and technical literature, to create a large, high-quality, and diverse dataset that covers various styles of Sanskrit text.

3.1 Data Sources

The Sanskrit-Tamil parallel corpus was compiled from a diverse range of sources to ensure linguistic variety and domain coverage. We can categorize our data sources into four primary modalities:

- **Human Translation**: We extracted prose texts from digital repositories, such as the Saṃsādhanī corpus. These data were then translated by human bilingual experts.

- **OCR for Physical Books or Scanned PDF**: Significant volumes of classical texts were sourced from physical books and scanned PDFs, including prominent Gitapress editions of the Bhagavatam. We processed these through an OCR pipeline specifically tuned for Devanagari and Tamil scripts to extract raw text for subsequent alignment.

- **Web-Scraping E-Texts**: Large-scale poetry data, including the Mahābhārata and Rāmāyaṇa, were acquired via custom web-scraping scripts from digital archives (e.g., Arasan.info[2]). These sources provided the bulk of the poetry data used in the corpus.

- **Curating from External Corpora**: We integrated existing parallel data, such as the NLLB Sanskrit-Tamil subset. To ensure consistency with our high-quality benchmarks, these datasets underwent a rigorous curation process involving automated cleaning, noise removal, and re-alignment.

[2] `https://www.arasan.info/p/welcome.html`

3.2 Data Preparation Pipeline

To handle the complexity and volume of data required, we develop a systematic data preparation pipeline. The goal of this pipeline is to create a streamlined and repeatable process that converts raw source materials into clean, aligned, and model-ready training data. This pipeline is for ensuring data quality and consistency, which is the most important factor for building a high-performance NMT model.

The pipeline consists of five major stages:

1. **Source Curation & Acquisition:** This initial stage focuses on gathering high-quality texts. We identify and acquire parallel (Sanskrit-Tamil) and monolingual (Sanskrit) texts from various sources, including digital libraries, scanned books, websites, and existing academic collections. Each source undergoes evaluation based on its domain (classical, modern or technical), style (prose, poetry or mixed), translation quality, and suitability for extraction. Additionally, we acquire and digitise books like Bhagavatam, which includes Tamil translations. To prevent copyright infringement, we obtain necessary permissions from content owners, clearly stating our intended use.

2. **Extraction & Cleaning:** Once acquired, the raw text must be extracted and cleaned. This is a critical step to remove "noise" that would confuse the translation model.

   - **Preprocessing:** Many translation texts use footer extensively. Some of them use it for additional notes and some of them use it for separating source text from translation. Many a times the footer size will be variable. We prepared a script using OpenCV that could correctly detect such variable height footers and remove them, before OCR.
   - **Extraction**: This involves handling various formats, such as converting PDFs to text, scraping web pages, or processing plain text files. For scanned books, Optical Character Recognition (OCR) is employed, which requires a subsequent correction step. We experimented with both Tesseract OCR and Google Vision API, and chose the latter since it was giving better results.
   - **Cleaning**: Automated scripts are used to remove irrelevant artefacts like page numbers, heading information, indexes, and source commentary that are not part of the core translation. But we preserve this information elsewhere for trace back. This stage also standardises encoding and formats.

3. **Alignment:** For parallel texts, after cleaning, the source (Sanskrit) and target (Tamil) texts are aligned at the sentence level for prose and verse-level for poetry.

   - **Automated Alignment**: We first use automated scripts to create a baseline alignment based on chapter and verse numbers. This is followed by manual correction wherever needed.
   - **Manual Correction**: This is especially important for our data. Due to the linguistic differences—where a single Sanskrit śloka may translate to multiple Tamil sentences—automated tools often fail. Manual review and correction are essential to ensure the accuracy of the parallel sentences, particularly for our high-quality dataset.

4. **Translation:** For acquired Sanskrit texts that are monolingual (i.e., do not have an existing parallel Tamil text), we commission new translations. This step is essential for much of our prose data, which often lacks pre-existing translations. We employ our team of freelancers to produce the required high-quality, parallel Tamil data.

   - We developed Translation Guidelines to keep the translation neutral, natural and close to the source.
   - Translations were regularly reviewed by language experts, and translation guidelines were updated when needed.

5. **Linguistic Processing & Augmentation:** Given the nature of Sanskrit, we are preparing multiple versions of the source data:
    - **Sandhi Splitting**: The same Sanskrit data will be processed to create two versions: one with sandhi (compounded forms) intact and another that is *unsandhied* (split forms). This allows us to experiment and see if the model learns better from the explicit, split forms.
    - **Anvaya**: For poetry (śloka) texts, the natural word order is often different from prose. We also collect or prepare *anvaya* (prose order) versions of the poetry wherever we can.
    - **Linguistic Features:** We extract linguistic features such as *karaka* relation using our SMT to augment our data in model training.
6. **Dataset Assembly & Splitting:** Finally, all processed and aligned data pairs are compiled. This master dataset is then split into three sets for each text type (prose and poetry):
    - **Training Set**: The largest portion (e.g., 90-95%), used to train the model.
    - **Validation Set**: A smaller portion (e.g., 5%) used during training to check the model's progress and prevent overfitting.
    - **Test Set**: A portion held back for final evaluation. This set is kept separate from the **Benchmark Datasets**, which serve as our final, unbiased measure of model quality on specific text types.

3.3 Human Translation

Human-translated texts represent our highest-quality parallel data. This data is invaluable for fine-tuning and evaluating the NMT model, as it contains accurate and context-aware translations.

To build this dataset, we are actively engaging freelancers, particularly for translating Sanskrit prose texts into Tamil. This effort is focused on creating a robust collection of simple, story-based texts that are ideal for training the model on fundamental language structures.

3.4 Collected Data

Our data collection and preparation efforts have resulted in approximately 2.3 lakh parallel entries and 31.8 lakh monolingual entries, covering a wide range of Sanskrit text styles such as prose, poetry, and Ayurveda.

Our human translation efforts have yielded **21,444** parallel sentences for **Prose** from Saṃsādhanī'scorpus collection[3].

In addition to prose, we have secured high-quality human translations for **Poetry** (651 sentences from the *Bhagavad Gita*).

This process remains a significant bottleneck. Finding experts with deep knowledge of both Sanskrit and Tamil, especially for complex *śāstric* and Ayurvedic domains, is difficult. The complexity of these texts requires more time for accurate understanding and translation, which slows down our data acquisition rate.

Monolingual Corpus: A large monolingual Sanskrit corpus is essential for pre-training and improving the model's understanding of the source language. We have collected, split, and cleaned a significant corpus totalling approximately **31.8 lakh** Sanskrit entries **(21,868,843 words)** sourced from the Sanskrit Monolingual Dataset by Priyanshu et al.

Synthetic Parallel Data: To rapidly expand our dataset, we generated synthetic parallel data by filtering noisy existing corpora and using machine translation to produce new sentence pairs.

- **NLLB Filtered:** We have filtered the publicly available NLLB corpus, resulting in **42,973** higher-quality sentence pairs.

[3] `https://sanskrit.uohyd.ac.in/Corpus/`

Style	Text / Source Name	Method	Size
Prose	8 Modern Prose Collections[a]	Human Translation	21,446
Poetry	Mahābhārata (Arasan.info)	Automatic Aligment	41,683
	Rāmāyaṇa (Arasan.info)	Automatic Aligment	15,069
	Bhagavatam (Gitapress)	Automatic Aligment	12,276
	Narayaneeyam (Anbezhil)	Automatic Aligment	1,024
	Bhagavadgita	Human Translation	651
Anvaya	Rāmāyaṇa (IITK)	Automatic Aligment	18,022
	Bhagavadgita	Automatic Aligment	652

[a]Includes: 130 Stories, Aakhyanvallari, Vetalkatha, Rajkathakunj, Sanskritkathakunj, Balaratnani, Kathanjali, and Tenali stories from Saṃsādhanī corpus (cite)

Table 1: Collected Sanskrit-Tamil parallel data categorized by style and acquisition method.

Dataset	Language Pair	Size (Pairs/Lines)
NLLB	Sanskrit–Tamil	43,000
Bhashaverse	Sanskrit-Tamil	
Saamayik	Sanskrit-Tamil	
Itihasa	Sanskrit–English	93,000
San-Hin MT	Sanskrit–Hindi	6,000
Sanskrit Monolingual	Sanskrit Only	3.18 M

Table 2: Data acquired from existing datasets (used for baseline comparison and model pre-training.)

- **Poetry:** We have generated **93,030** sentence pairs from aligned poetry texts by translating english translations of them.

This gives us a total of **136,003** synthetic sentence pairs.

As noted during curation, this NLLB data is of lower quality and contains mixed languages in the source text.

3.5 Santham Dataset

Training Datasets: "Gold" data refers to high-quality, human-translated or human-validated parallel sentences, which are essential for fine-tuning the model. We have collected a total of **90,149** gold parallel sentences (Table 3).

This data is obtained by extracting and aligning existing resources, and creating new human translations wherever such data was unavailable. The prose data, which comes from our human translation efforts, is being prepared in an *unsandhied* (split) form.

Style	Size (Pairs)	Description
Prose	20,446	—
Poetry (Total)	69,703	
Anvaya	10,146	Poetry data having corresponding anvaya data.
Non-Anvaya	59,557	Poetry data without anvaya data.
Total	**90,149**	

Table 3: Santham Dataset: Sanskrit-Tamil parallel training data.

Benchmark Datasets: We have developed benchmark datasets for prose, poetry and *anvaya*. These datasets are reserved exclusively for testing and evaluation, not for training. Each

has been human-reviewed and validated to ensure high quality (Table 4).

Style	Size (Pairs)	Description
Prose	1,000	Selected pairs from different prose texts.
Poetry	1,000	Selected pairs from Rāmāyaṇa, Narayaneeyam and Bhagavadgita.
Anvaya	1,000	Selected pairs from Rāmāyaṇa and Bhagavad Gita mapped to translation from poetry data.
Total	**3,000**	

Table 4: Santham Benchmark: Human-reviewed Sanskrit-Tamil parallel data for evaluation.

4 Preprocessing

We investigate whether segmentation as a standalone preprocessing task in our pipeline of the Sanskrit-Tamil Machine Translation helps improve the translation quality. Sanskrit is well known for its rich morphological features, sandhi formation and compounding, together posing challenges during preprocessing and downstream tasks (Krishna and others, 2018). Sanskrit texts are written in sandhied format, requiring segmentation before processing.

Segmentation during preprocessing helps reduce Out-of-Vocabulary (OOV) rates and word error rates (Nehrdich et al., 2024), and drastically improves translation quality (Chaudhari et al., 2024). As the sandhied version differs from its unsandhied (segmented) version, it was a necessary step to obtain two different versions of the same data for comparative analysis of the translation model.

4.1 Segmentation Tools

We shortlist two widely used segmenters for this task: ByT5-Sanskrit (Nehrdich et al., 2024) and Sanskrit Heritage (SH) (Huet, 2009; Krishnan et al., 2024).

ByT5-Sanskrit: A byte-level pretrained language model fine-tuned with a multitask dataset from the Digital Corpus of Sanskrit (DCS), focusing on segmentation, lemmatization and morphosyntactic tagging. This is the current state-of-the-art model for segmentation. The ByT5-Sanskrit segmentation results were accessed using the Dharmamitra API.[4]

SH-segmenter: A lexicon-based segmenter built using finite state automata (Huet, 2009), augmented with a probabilistic ranking mechanism that chooses the best segmentation solution (Krishnan et al., 2024). The output marks compound boundaries with hyphens, allowing for clear identification of constituent words within compounds.

4.2 Preprocessing Pipeline

ByT5-Sanskrit (via Dharmamitra API): The segmentation was performed in two phases. In the first phase, all 69,667 sentences were sent to the API. A simple heuristic was applied for validation: the segmented sentence should contain at least as many tokens as the input. Of the total sentences, 68,565 passed this validation while 1,102 were flagged for reprocessing.

In the second phase, the flagged sentences were split into smaller chunks based on sentence boundary markers (|, ।, ॥) to preserve sandhi integrity. Chunks were limited to approximately 15 space-delimited units before being sent to the API. If a chunk failed due to timeout or network issues, the system retried up to 7 times. Chunks that still failed were further divided into 10-unit and then 5-unit segments. This recursive chunking strategy recovered 933 of the 1,102 rejected sentences.

Since the Dharmamitra API returns output in IAST transliteration, we converted all segmented text back to Devanagari using the `devtrans` library.[5] The final preprocessed corpus contains 69,498 segmented sentences, achieving 99.8% coverage.

[4] `https://dharmamitra.org/`

[5] `https://pypi.org/project/devtrans/`

Sanskrit Heritage Platform: The python-based package[6] is run with a timeout of 30 seconds to extract the segmentations. Certain sentences with longer compounds and with complex sandhi instances require more processing time. All 69,667 sentences are processed, with the segmented output stored alongside the original text. Unlike Dharmamitra, SH performs segmentation locally (with web fallback), making it suitable for large-scale processing without API rate limits. The output marks unrecognized words with a '?' prefix, enabling error detection in our pipeline.

By employing both tools, we obtain two independent segmentations of the same corpus, enabling comparative analysis of the translation model's performance on differently segmented inputs.

4.3 Segmentation Example

Input	निशम्य च बहून्वालान् कृष्णान्पुच्छसमाश्रितान् विषण्णरूपां विनतां कदूर्दास्ये न्ययोजयत्
Output	निशम्य च बहून् वालान् कृष्णान् पुच्छ-समाश्रितान् विषण्ण-रूपाम् विनताम् कदूः दास्ये न्ययोजयत्

Table 5: Example of sandhi segmentation. Hyphens mark compound word boundaries (samāsa).

4.4 Evaluation

We evaluated both segmenters on 3,669 sentences from a manually annotated corpus (START) available at the Saṃsādhanī's START platform (Kumar et al., 2024). This corpus contains simple prose sentences, classical literature texts like Bhagavad Gītā, Saṅkṣepa Rāmāyaṇa and a domain-specific text Aṣṭāṅgahṛdayam.

Metric	SH	ByT5
Precision	**85.87**%	78.45%
Recall	81.09%	**83.96**%
F1-Score	**82.70**%	80.33%
Perfect Match	**25.25**%	19.47%

Table 6: Word-level evaluation on START dataset.

SH achieves higher precision and F1-score, while ByT5 shows higher recall. We use both segmentations to compare their impact on translation quality.

5 Experiments

As outlined in Section 5.1, we independently evaluate the quality of the proposed Poetry and Prose benchmark datasets by assessing the performance of machine translation models. In Section 5.2, we fine-tune these models on the Poetry and Prose training sets to quantify performance gains. To assess potential improvements in Sanskrit-Tamil translation, we fine-tune the best-performing models using the segmentation and *anvaya* of Sanskrit verses.

5.1 Baselines

We establish the baselines on the Prose and Poetry benchmark datasets (see Table 4) under zero-shot setting across multiple, multilingual models selected such that they are pretrained on Indic Languages. The models include one encoder-decoder model *IndicTrans2-Indic-Indic-1B*(Gala et al., 2023) (IndicTrans2), and several open-weight LLMs, namely *Gemma-3-4B-IT* (Gemma)(Gemma Team, 2025), *Llama3.2-3B-Instruct* (Llama3.2-3B), *Llama3.1-8B-Instruct* (Llama Team, 2024) (Llama3.1-8B), *Qwen3-8B* (Qwen3-8B) (Qwen Team, 2025).

[6] `https://pypi.org/project/sanskrit-heritage/`

We tailor our baseline experiments to each model's architecture. For the encoder-decoder model, IndicTrans2, we perform direct translation of the Sanskrit poetry or prose to Tamil while for LLMs we use the prompt detailed in Figure 5.1, to guide the model in translating Sanskrit text to Tamil.

Prompt

You are a professional translator. Translate the Sanskrit text provided by the user into Tamil. Output ONLY the Tamil translation, no other text.

5.2 Model Fine-tuning

Hyperparameter	**Value**
LoRA rank (r)	16
LoRA dropout	0.05
Number of epochs	2
Effective batch size	8
Learning rate	2×10^{-4}
Warmup ratio	0.03

Table 7: Hyperparameters used for LoRA-based LLM fine-tuning.

We conduct a series of fine-tuning experiments to evaluate the efficacy of the Santham dataset for Sanskrit-to-Tamil translation of both poetry and prose. All experiments are performed using the same hyperparameter configuration, as mentioned in Table 7 and using the instruction shown in prompt 5.1.

Experiment 1: We fine-tune the baseline models established in Section 5.1 on the combined corpus of 90,149 Prose and Poetry samples (see Table 3). To ensure consistency, we utilize the same prompt template (Figure 5.1) and 10% of the training data for validation. Subsequently, we evaluate these fine-tuned models on the respective Prose and Poetry benchmarks (see Table 4). This experiment serves as a reference point for evaluating the performance of models fine-tuned on raw poetry data without any additional preprocessing.

Experiment 2: In this experiment, we analyze the effect of fine-tuning models on the 69,703 samples of segmented Poetry (Table 3), preprocessed using the two segmentation approaches described in Section 4, namely, Dharmamitra API and Sanskrit Heritage Platform (SH). We select the two best-performing models from Experiment 1 and fine-tune utilizing the prompt shown in Figure 5.1 and with 10% of the training data used for validation. For evaluation, we use the Poetry benchmark (see Table 4), preprocessed using the corresponding segmentation approach for each model. Through this experiment, we aim to assess whether segmentation-based preprocessing improves translation quality compared to the original Sanskrit poetry verses.

Experiment 3: We investigate the impact of utilizing *Anvaya* (the prose order of poetry) as input instead of the original Sanskrit verses. We fine-tune the two best-performing models from Experiment 1 on a subset of the Santham Poetry corpus consisting of 10,146 samples, which comprises anvaya aligned with the corresponding original verses and Tamil translations (see Table 3). From this subset, we use 500 samples each for validation and testing, utilizing the remaining data for training. We use the same prompt as detailed in Figure 5.1. This setup allows us to evaluate whether fine-tuning on syntactically reordered poetic verses leads to improved translation performance.

6 Results and Analysis

This section presents our experimental results and analysis. We evaluate model performance using three standard automated metrics, namely, BLEU (Papineni et al., 2002) which assesses translation quality based on n-gram matching, chrF++ (Popović, 2017), which extends character n-gram overlap with word unigrams and bigrams and COMET (Rei et al., 2022) which utilizes multilingual contextual embeddings to calculate the semantic similarity between predicted and reference translations. First, we establish baselines and compare them against models fine-tuned on the Santham dataset to quantify performance gains. Further, we evaluate the translation quality of Poetry when models are provided with segmented verses and anvaya as inputs as detailed in Experiment 1 and Experiment 2 in Section 5.2.

6.1 Comparison against Baselines

Setup	Dataset	Model	BLEU	chrF++	COMET
Baseline	Prose	IndicTrans2	**4.57**	**34.68**	**81.06**
		Llama3.1-8B	1.45	22.14	62.83
		Llama3.2-3B	0.74	19.93	58.34
		Qwen3-8B	2.06	28.53	71.18
		Gemma-3-4B	<u>2.99</u>	<u>29.75</u>	<u>75.78</u>
	Poetry	IndicTrans2	0.11	15.09	53.45
		Llama3.1-8B	0.06	12.77	49.34
		Llama3.2-3B	0.025	15.30	51.79
		Qwen3-8B	<u>0.42</u>	<u>19.35</u>	<u>56.93</u>
		Gemma-3-4B	**0.46**	**19.62**	**59.91**
Fine-Tuned	Prose	IndicTrans2	4.29	35.28	<u>80.95</u>
		Llama3.1-8B	2.76	27.77	74.76
		Llama3.2-3B	1.69	30.86	62.83
		Qwen3-8B	<u>4.45</u>	<u>38.53</u>	68.59
		Gemma-3-4B	**6.95**	**39.49**	**84.18**
	Poetry	IndicTrans2	0.05	8.88	38.87
		Llama3.1-8B	0.50	15.88	57.58
		Llama3.2-3B	1.84	<u>23.68</u>	<u>66.25</u>
		Qwen3-8B	<u>2.47</u>	21.43	59.72
		Gemma-3-4B	**3.25**	**29.64**	**73.21**

Table 8: Automatic evaluation results for baseline and fine-tuned models on the Prose and Poetry benchmark datasets. The best scores within each group are highlighted in **bold**, and the second-best scores are <u>underlined</u>.

In the baseline setting, for the Prose benchmark, as reported in Table 6.1, the encoder-decoder model, IndicTrans2, outperforms all LLMs, achieving a BLEU score of 4.57 and a COMET score of 81.06. This demonstrates that a specialised architecture trained explicitly for Indic language translation remains superior to general-purpose LLMs. Among the LLMs, Qwen3-8B and Gemma-3-4B perform competitively, surpassing both Llama 3.1-8B and Llama-3.2-3B.

This performance gap suggests that Qwen3-8B and Gemma-3-4B benefit from pre-training on large multilingual datasets containing Indic languages, whereas the Llama-3 models lack robust internal representations for Sanskrit and Tamil.

In contrast, performance on the Poetry benchmark drops significantly across all models, highlighting the inherent complexity of translating Sanskrit verses. Unlike prose, Sanskrit poetry frequently has non-standard word order as well as dense metaphorical and morphological structures. Consequently, the specialised IndicTrans2 model, which relies on consistent syntactic patterns, struggles in this domain. Gemma-3-4B and Qwen3-8B achieve superior baseline scores compared to IndicTrans2, suggesting that their extensive multilingual pre-training and generalised reasoning capabilities allow them to translate the stylistic irregularities and flexible syntax of poetic text.

We observe that post-fine-tuning, all evaluated models outperform IndicTrans2 on both the Prose and Poetry benchmarks. Notably, Qwen3-8B and Gemma-3-4B are the two best-performing models, suggesting that general-purpose LLMs demonstrate superior domain adaptability compared to the specialised pre-trained translation model in this context. However, performance across all models remains lower for Poetry than for Prose, which highlights the inherent complexity of translating poetic verses compared to standard prose.

6.2 Analysing Poetry Translation: Segmentation and Anvaya

As discussed in Section 6.1, the baseline performance on the Poetry benchmark remains significantly lower than that of Prose. To address this disparity and analyze potential quality gains in Sanskrit-to-Tamil poetry translation, we investigate two linguistic preprocessing techniques, namely, Segmentation and Anvaya (See Experiment 2 and Experiment 3 in Section 5.2).

6.2.1 Impact of Word Segmentation

As reported in Table 9, segmentation yields performance gains by providing approximately 8% increase in BLEU for Gemma-3-4B compared to a 6.28% average increase in BLEU for Qwen3-8B. These results demonstrate that segmentation effectively aids in translating complex Sanskrit poetry.

Comparing the segmentation techniques, Dharmamitra (ByT5) proves to be the more robust than Sanskrit Heritage (SH) as it consistently improved performance over the unsegmented baseline across all metrics. This suggests that a segmenter with higher recall (like Dharmamitra, see Table 6) is more effective at simplifying morphologically rich poetic verses than a high-precision, rule-based SH segmenter. By correctly identifying more words that require splitting, Dharmamitra better supports translation to Tamil (see Table 10). However, it is important to note that while beneficial, the overall performance gain from segmentation remains marginal compared to using the original unsegmented verse.

6.2.2 Impact of Anvaya

As reported in Table 11, we observe consistent performance improvements when models are fine-tuned on *Anvaya* (the prose order of verses) compared to the original poetry. Specifically, both Gemma-3-4B and Qwen3-8B achieve a relative improvement of over 46% in BLEU and 2.3% in COMET. These results indicate that simplifying the input structure allows the models to generate Tamil translations that are semantically closer to the reference. Furthermore, this confirms that the non-standard word order inherent in Sanskrit poetic verses constitutes a significant barrier to accurate translation, which can be effectively mitigated by syntactic reordering.

Despite the relative improvements observed, the absolute scores for all models are low. This indicates a significant gap in current model capabilities and highlights the substantial scope for improvement in learning representations for low-resource and classical language pairs like Sanskrit and Tamil.

Segmentation Technique	Model	BLEU	CHRF++	COMET
No Segmentation (Original)	Qwen3-8B	2.47	21.43	59.72
	Gemma-3-4B	3.25	**29.64**	73.21
Sanskrit Heritage	Qwen3-8B	2.56	21.48	59.28
	Gemma-3-4B	<u>3.32</u>	<u>29.52</u>	<u>73.38</u>
Dharmamitra (ByT5)	Qwen3-8B	2.69	21.96	60.08
	Gemma-3-4B	**3.51**	29.37	**74.37**

Table 9: Impact of segmentation on Sanskrit-to-Tamil Poetry translation. We compare the original verse (No Segmentation) against two segmentation techniques: Sanskrit Heritage and Dharmamitra (ByT5) for translation from Sanskrit to Tamil. Best scores across all settings are highlighted in **bold** and second-best are <u>underlined</u>.

Original Verse: आत्मौपम्येन सर्वत्र समम्पश्यति यो अर्जुन सुखं वा यदि वा दुःखं स योगी परमो मतः
Target Translation: அர்ஜுனா ! (அதனால்) சுகம் (இருக்கட்டும்) அல்லது (அதனால்) துன்பம் (என்றாலும்) எவன் தன்னைப்போலவே எல்லா உயிரினங்களையும், சமமாக பார்கிறானோ, அந்த யோகி உயர்ந்தவனாக மதிக்கப்படுகிறான்.

Technique	Segmented Input	Predicted Translation
Dharmamitra	आत्म औपम्येन सर्वत्र समम्पश्यति यः अर्जुन सुखम् वा यदि वा दुःखम् स योगी परमः मतः	ஓ! அர்ஜுனா, தன்னைப் போன்றவர்கள் அனைவரையும் கண்டு மகிழ்ச்சியையோ, துன்பத்தையோ அடைவதில் வெல்பவன் உயர்ந்த யோகியாகக்
Sanskrit Heritage	आत्म औपम्येन सर्वत्र समम् पश्यति यः अर्जुन सुखम् वा यदि वा दुःखम् स योगी परमः मतः	ஓ! ஜனார்த்தனா, அனைத்து உயிரினங்களிடமும் நீ கொண்ட அன்பு மற்றும் அன்புக்குரிய பொருளாக இருப்பதால் நீ அனைத்து உயிரின...

Table 10: Qualitative comparison of translation outputs using different segmentation techniques. The Dharmamitra segmentation preserves the compound *samampashyati*, leading to a translation closer to the context, while the Sanskrit Heritage segmentation splits it, resulting in a divergent translation.

Model	Input Type	BLEU	chrF++	COMET
Qwen3-8B	Original	2.62	25.39	62.87
	Anvaya	**3.83**	**27.79**	**64.37**
Gemma3-4B	Original	3.12	30.38	74.55
	Anvaya	**4.83**	**32.83**	**76.33**

Table 11: Performance comparison of models fine-tuned on the original Poetry (Original) verses against those fine-tuned on the syntactically simplified Anvaya. Best scores for each model are highlighted in **bold**.

6.3 Human Evaluation

Recognizing the inherent limitations of automated metrics in capturing linguistic nuances of classical languages, we conduct human evaluation to assess semantic consistency and grammatical accuracy of translations across prose, segmented poetry, and poetry in *Anvaya* form. We randomly select 15 Sanskrit prose samples and 15 poetry verses for each experimental configuration, specifically Sanskrit Heritage (SH) segmentation, Dharmamitra (ByT5) segmentation, and *Anvaya*. The predicted Tamil translations for these inputs are evaluated for our two best-performing fine-tuned models, Qwen3-8B and Gemma3-4B. A total of 115 unique samples were independently assessed by two bilingual experts. Adhering to the guidelines, annotators rated translations on a 1–5 Likert scale across four distinct categories:

1. **Adequacy:** Preservation of the original Sanskrit meaning in the Tamil output.
2. **Fluency:** Adherence to the grammatical and stylistic norms of contemporary Tamil.
3. **Syntactic Structure and Flow:** Effectiveness in navigating complex Karaka (case) relations and narrative continuity.
4. **Technical Consistency:** Proper retention of specialized terms (e.g., *Sandhyopasana*) and maintenance of contextual references.

Model	Experiment	Adequacy	Fluency	Syn. Flow	Tech. Cons.
Qwen3-8B	Prose	4.111	4.278	4.167	4.500
	Segmentation-SH	**3.029**	**3.088**	**3.294**	**3.353**
	Segmentation-ByT5	2.292	2.562	2.375	2.771
	Anvaya	<u>2.357</u>	<u>2.643</u>	<u>2.429</u>	<u>3.071</u>
Gemma3-4B	Prose	3.400	4.000	3.700	4.300
	Segmentation-SH	3.188	**3.688**	3.406	**3.781**
	Segmentation-ByT5	2.635	<u>3.673</u>	2.827	<u>3.442</u>
	Anvaya	**3.417**	3.250	**3.583**	3.250

Table 12: Comparative human evaluation of Sanskrit-to-Tamil machine translation (1–5 scale) across Qwen3-8B and Gemma3-4B models for prose and poetry, utilizing Sanskrit Heritage (SH) segmentation, Dharmamitra (ByT5) segmentation and *Anvaya*. The evaluation criteria consist of Adequacy, Fluency, Syntactic Structure and Flow (Syn. Flow), and Technical Consistency (Tech. Cons.). The best scores within each model for Poetry translation are in **bold**, and second-best are <u>underlined</u>.

We observe substantial agreement in Adequacy ($\kappa = 0.683$) and moderate-to-substantial agreement across Fluency ($\kappa = 0.563$), Syntactic Structure and Flow ($\kappa = 0.594$), and Technical Consistency ($\kappa = 0.565$). These high Weighted-Kappa scores, combined with an Adjacent Agreement of over 80% in Adequacy, validate the reliability of our expert annotations.

As reported in Table 12, performance trends indicate that translations of Sanskrit prose consistently outperform poetry across all evaluation criteria. This suggests that narrative structures facilitate superior preservation of original meaning and grammatical integrity.

Within the domain of poetry translation, the results reveal distinct performance trends. While *Anvaya* often performs better in terms of automatic lexical metrics than segmentation, its semantic performance, as reflected by the human evaluation criteria, varies. Specifically, *Anvaya* results in lower adequacy for Qwen3-8B, whereas for Gemma3-4B, it performs better to segmentation-based experiments. Furthermore, it is consistently observed that *Anvaya* performs lower than or nearly equivalent to segmentation in terms of preserving syntactic structure and fluency. This highlights that while syntactic reordering through *Anvaya* aids in decoding the source, the resulting semantic accuracy and grammatical naturalness in Tamil may be slightly

reduced when compared to direct translation from segmented verses. We observe that among the automatic metrics, COMET exhibits the strongest correlation with human judgment as it reflects semantic adequacy and structural flow observed in the human evaluation (Table 12).

7 Conclusion

In this work, we introduce **Santham**, a curated, novel, Sanskrit-Tamil parallel dataset comprising over 90,000 pairs drawn from classical texts such as the Mahābhārata, Rāmāyaṇa, Bhagavatam, Narayaneeyam, and Bhagavad Gita, as well as eight modern prose collections. We establish a human-evaluated benchmark to rigorously assess translation quality for both genres. Additionally, we evaluate a segmentation preprocessing pipeline designed to address the linguistic complexities like Sandhi and Samasa in Sanskrit. Our benchmarking of translation models and state-of-the-art LLMs reveal that fine-tuning on Santham yields significant performance improvements across various metrics which demonstrates the efficacy of our dataset. Furthermore, we observe that linguistic interventions, specifically segmentation and the use of Anvaya, lead to improvements in Tamil translation quality. While absolute scores remain low, highlighting the inherent difficulty of translating Sanskrit prose and poetry, our results underscore the substantial scope for future research in this domain. We aim for Santham to bridge this gap and facilitate low-resource language development.

Acknowledgements

We gratefully acknowledge the financial support provided by the BHASHINI Initiative of the Ministry of Electronics and Information Technology (MeitY) for this research under the Sanskrit Knowledge Accessor (SKA) project. We also thank the Saṃsādhanī team for providing their curated corpora of both poetry and prose. We extend our thanks to Ms. Radhika, the Senior Language Editor for the project, for her assistance with translation and review. Furthermore, we recognise the freelancers who contributed high-quality prose translations and participated in the human evaluation process. Finally, we are grateful to Mr. Soumyadip and Mr. Abhinav (LTRC, IIITH) for their work on the initial model experiments, and to Mr. Nagaraju for his invaluable technical support.

References

Rahul Aralikatte, Miryam de Lhoneux, Anoop Kunchukuttan, and Anders Søgaard. 2021. Itihasa: A large-scale corpus for Sanskrit to English translation. In Toshiaki Nakazawa, Hideki Nakayama, Isao Goto, Hideya Mino, Chenchen Ding, Raj Dabre, Anoop Kunchukuttan, Shohei Higashiyama, Hiroshi Manabe, Win Pa Pa, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, Katsuhito Sudoh, Sadao Kurohashi, and Pushpak Bhattacharyya, editors, *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 191–197, Online, August. Association for Computational Linguistics.

Rhugved Pankaj Chaudhari, Bhakti Jadhav, Pushpak Bhattacharyya, and Malhar Kulkarni. 2024. SansGPT: Advancing generative pre-training in Sanskrit. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 432–441. NLP Association of India (NLPAI).

Rahul Chingamtotattil and Rajamma Gopikakumar. 2022. Neural machine translation for Sanskrit to Malayalam using morphology and evolutionary word sense disambiguation. *Indonesian Journal of Electrical Engineering and Computer Science*, 28(3):1709–1719, December.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul 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*.

Gemma Team. 2025. Gemma 3 technical report.

Gérard Huet. 2009. Sanskrit Segmentation. In *Proceedings of the South Asian Languages Analysis Roundtable XXVIII*, October.

Prashanth Kammar, Parashuram Baraki, Sunil Kumar Ganganayaka, Manjunath Swamy Byranahalli Eraiah, and Kolakaluri Lakshman Arun Kumar. 2024. Translating Sanskrit to Hindi Language using Recurrent Neural Network (RNN)-L2 Regularization. *International Journal of Intelligent Systems and Applications in Engineering*, 12(3):199–208, March.

Amrith Krishna et al. 2018. Free as in free word order: An energy based model for word segmentation and morphological tagging in sanskrit. In *Proceedings of EMNLP 2018*, pages 2550–2561.

Sriram Krishnan, Amba Kulkarni, and Gérard Huet. 2024. Normalized dataset for sanskrit word segmentation and morphological parsing. *Language Resources and Evaluation*, 59(2):1279–1330, Aug.

Anil Kumar, Amba Kulkarni, and Nakka Shailaj. 2024. START: Sanskrit teaching; annotation; and research tool. In *Proceedings of the 7th International Sanskrit Computational Linguistics Symposium*, pages 113–124. Association for Computational Linguistics, February.

Lisha C.R. 2024. Techniques in Translating Sanskrit Poetry: From Literal to Creative Approaches. *International Journal of Innovative Research in Technology*, 11(4):241–248, August.

Llama Team. 2024. The llama 3 herd of models.

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 Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14298–14304, Torino, Italia, May. ELRA and ICCL.

Sebastian Nehrdich and Kurt Keutzer. 2026. Mitra: A large-scale parallel corpus and multilingual pretrained language model for machine translation and semantic retrieval for pāli, sanskrit, buddhist chinese, and tibetan.

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. Association for Computational Linguistics, November.

Sebastian Nehrdich, David Allport, Sven Sellmer, Jivnesh Sandhan, Manoj Balaji Jagadeeshan, Pawan Goyal, Sujeet Kumar, and Kurt Keutzer. 2026. Mitrasamgraha: A comprehensive classical sanskrit machine translation dataset.

NLLB Team. 2022. No language left behind: Scaling human-centered machine translation.

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

Maja Popović. 2017. chrF++: words helping character n-grams. In Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer, editors, *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark, September. Association for Computational Linguistics.

Anagha Pradeep and Radhika Mamidi. 2025. Sandarśana: A Survey on Sanskrit Computational Linguistics and Digital Infrastructure for Sanskrit. *ACM Computing Surveys*, May. Publisher: ACM-PUB27New York, NY.

Qwen Team. 2025. Qwen3 technical report.

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 Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata,

Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Nandini Sethi, Amita Dev, and Poonam Bansal. 2023. A Novel Neural Machine Translation Approach for low-resource Sanskrit-Hindi Language pair. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, April. Just Accepted.