Improving Classical Language Machine Translation using Supervised Fine-Tuning with Philological Commentary

Yuzuki Tsukagoshi, Ikki Ohmukai

Proceedings of the 39th Pacific Asia Conference on
Language, Information and Computation (PACLIC 39)

Emmanuele Chersoni, Jong-Bok Kim (eds.)

2025

Improving Classical Language Machine Translation using Supervised Fine-Tuning with Philological Commentary

Yuzuki Tsukagoshi
The University of Tokyo / Tokyo, Japan
yuzuki@l.u-tokyo.ac.jp

Ikki Ohmukai
The University of Tokyo / Tokyo, Japan
i2k@l.u-tokyo.ac.jp

Abstract

This paper presents a novel approach to improving ancient language translation by integrating scholarly philological commentary into language model training. Using the Ṛgveda with authoritative English translations and annotations from Jamison & Brereton, we employ supervised fine-tuning on five models: GPT-4.1 nano, Gemini 2.5 flash, Llama 3.2 3B, Llama 3.1 8B, and the Sanskrit-specialized Gemma 2 Mitra. Our methodology compares standard fine-tuning against commentary-enhanced training, where models receive both Sanskrit texts and philological commentary as input. Evaluation using BLEU scores demonstrates consistent improvements across four larger models when incorporating scholarly commentary, with particularly strong gains for culturally-specific and morphologically complex translations.

1 Introduction

Sanskrit, the ancient liturgical language of Brahmanism and Hinduism and one of the world's oldest documented languages, presents formidable challenges for machine translation due to its complex morphological system, rich literary heritage, and profound cultural context. Vedic Sanskrit, the earliest attested form preserved in the Vedas, poses additional difficulties through its archaic vocabulary, distinctive accent system, and highly ritualistic contexts that require deep linguistic, cultural, and religious understanding for accurate interpretation.

While recent advances in large language models have demonstrated remarkable capabilities for machine translation across numerous language pairs, their application to low-resource ancient languages remains limited. Existing approaches typically rely on parallel corpora alone, overlooking the wealth of scholarly expertise encoded in philological commentaries that provide essential linguistic, cultural, and religious context.

This paper introduces a methodology that leverages scholarly philological commentary to enhance neural machine translation of Vedic Sanskrit. Our approach recognizes that successful translation of ancient sacred texts requires not merely linguistic competence, but also deep understanding of historical, cultural, and ritualistic contexts –knowledge traditionally preserved in centuries of scholarly commentary.

We make the following contributions: (1) We demonstrate that integrating philological commentary significantly improves Vedic Sanskrit translation quality across multiple large language models; (2) We provide comprehensive evaluation across five diverse architectures, including both general-purpose and Sanskrit-specialized models; (3) We establish a replicable methodology for incorporating scholarly expertise into neural translation of ancient languages.

2 Vedas and Philological Commentary

The Vedas are the oldest sacred texts in the Indian subcontinent, composed in Vedic Sanskrit between about 1500 and 500 BCE. The Ṛgveda (ṚV), the oldest of the four Vedas, is a collection of hymns dedicated to various deities. All hymns in the Ṛgveda comprise complex verses which are often difficult to translate, making it difficult even to identify the correct meaning of individual words. The constraints of metrical structure and the use of archaic phonology, morphonology, and syntax further complicate the translation process.

Vedic texts include not only verses, but also explanatory prose. Saṃhitās, the collections of verses in the Vedas including the Ṛgveda, are highly ritualistic and sophisticated texts, where each hymn is composed with a specific purpose and context. Brāhmaṇas, Āraṇyakas, and Upaniṣads are prose texts that provide detailed explanations of the Saṃhitās about their meanings, their ritual-

istic significance, and their philosophical contexts.

The translations of Vedic texts have been published for centuries, some of which have commentaries that provide essential philological and linguistic insights. These commentaries are crucial for understanding the complex meanings of the hymns. Even one Saṃhitā, the Ṛgveda, has been translated by multiple scholars (Grassmann, 1876; Geldner, 1951; Renou, 1955; Elizarenkova, 1999; Jamison and Brereton, 2014; Witzel et al., 2007, 2013; Dōyama and Gotō, 2022). These academic translations themselves are significant works of philology and linguistics. In addition, the commentaries present the basis of the translators' interpretation. For example, ṚV 8.5.22 [1] is translated and added to the commentary by Jamison and Brereton (2014) as below:

Translation:

> When did the son of Tugra, abandoned in the sea, do reverence to you, o men, so that your chariot would fly with its birds?"

Commentary:

> The subjunctive pátāt seems to be used in an unusual past prospective sense in this mythological context. This may be an English problem, however. Since the verb of the main clause is injunc. vidhat, this context is not necessarily preterital, but "timeless," and the subjunctive can therefore be expressing pure future modality. The fact that the next verse is also mythological and contains an undoubted present tense form daśasyathaḥ shows that mythological tense is fluid here. Re remarks (ad vs. 23) that the indifference between present and preterite underlines the reflection of the current human situation in the legendary material.

When human experts read closely the original Vedic text, they refer to these commentaries to understand the meaning of the hymns and the thoughts of the translators. Due to the limited number of syllables and the need to maintain a regular rhythm, the word order in Vedic verses is often complex and many elements are omitted, making interpretation difficult. Therefore commentaries represent the invaluable insights of earlier scholars.

[1]The source text is: *kadā́ vāṁ taugriyó vidhat samudré jahitó narā yád vāṁ rátho víbhiṣ pátāt*.

3 Related Works

Recent research in neural machine translation has increasingly recognized the value of incorporating linguistic annotations and contextual information to improve translation quality, particularly for morphologically rich and low-resource languages. This section reviews key developments in annotation-enhanced machine translation, with particular attention to approaches relevant to our work on ancient language processing.

3.1 Morphological and Syntactic Annotations

Early work by Sennrich and Haddow (2016) demonstrated that incorporating explicit linguistic features such as POS tags and morphological information as input features to neural machine translation systems yields improvements in translation quality. Their experiments on English-German and English-Romanian showed lower perplexity and higher BLEU scores compared to word-only baselines, establishing that explicit linguistic annotations provide complementary information to end-to-end neural approaches.

García-Martínez et al. (2016) introduced Factored Neural Machine Translation (FNMT), which generates multiple outputs for each word including lemmas and morphological attributes. This approach proved particularly effective for morphologically rich languages, reducing vocabulary size and handling unknown words more effectively. Similarly, Dalvi et al. (2017) showed that injecting target-language morphological information into the decoder through joint training improved translation accuracy by 0.2-0.6 BLEU points for German and Czech.

For classical languages, Rosenthal (2023) demonstrated the particular value of morphological annotations in low-resource scenarios. Working with Latin-English translation, they achieved a BLEU score of 22.4 by encoding Latin morphology through stem-morpheme splitting, exceeding Google Translate's performance by over 4 BLEU points. This work is especially relevant to our Sanskrit study, as both Latin and Sanskrit are highly inflected classical languages with complex morphological systems.

3.2 Semantic Annotations

The integration of semantic role labeling (SRL) into neural MT has shown promise for preserving meaning relationships. Marcheggiani et al. (2018) were among the first to incorporate predicate-argument structures using Graph Convolutional Networks, achieving improvements from 23.3 to 24.5 BLEU on English-German translation. Their approach encoded PropBank semantic roles as graphs, demonstrating better preservation of "who-did-what-to-whom" relationships.

Rapp (2022) extended this work by annotating Europarl data with semantic roles across multiple language pairs (English to French, German, Greek, and Spanish), showing consistent but modest BLEU improvements of 0.2-0.5 points. While these gains appear small, they were consistent across different runs and language pairs, suggesting that semantic role annotations help capture meaning nuances that purely sequence-based models might miss.

In low-resource settings, semantic annotations have shown larger relative impacts. Wu et al. (2021) reported an average +1.18 BLEU improvement when injecting predicate-argument labels for Chinese, Mongolian, Uyghur, and Tibetan translations, while Nguyen et al. (2020) leveraged Abstract Meaning Representation (AMR) graphs for English-Vietnamese translation in small-data scenarios.

3.3 Cultural and Contextual Annotations

Recent work has explored incorporating cultural and historical knowledge into translation systems. Conia et al. (2024) introduced KG-MT, which retrieves entries from multilingual knowledge graphs to handle culturally nuanced entity references. Their approach achieved dramatic improvements—129% relative BLEU improvement over strong multilingual models and 62% over GPT-4 on culturally sensitive segments—demonstrating the importance of background knowledge for accurate translation.

This line of research is particularly relevant to our work on Vedic Sanskrit, where cultural, ritualistic, and historical contexts are crucial for accurate interpretation. The annotations provided by scholarly works like Jamison & Brereton's Rig Veda translation contain precisely this type of contextual information that has proven valuable in recent MT research.

3.4 Discourse and Pragmatic Annotations

Document-level translation has benefited from discourse structure annotations. Tan et al. (2022) incorporated Rhetorical Structure Theory (RST) annotations to improve coherence and consistency across sentences, achieving +0.9 BLEU and +1.1 METEOR improvements over strong document-level baselines. Style and register control through annotations has also proven effective, with **?** demonstrating that politeness tags can significantly improve translation appropriateness in German output.

Our work builds on this foundation by applying commentary-enhanced training specifically to Vedic Sanskrit, where philological expertise is essential for accurate translation. Unlike previous work that primarily used automatically generated annotations, we leverage expert scholarly commentary that provides deep cultural and linguistic insights unavailable to automatic annotation tools.

4 Methodology

4.1 Dataset

Our dataset consists of the Ṛgveda, utilizing the comprehensive English translation by Jamison & Brereton (Jamison and Brereton, 2014) and their detailed scholarly commentaries (Jamison and Brereton, 2015). The translation represents the most authoritative modern complete English translation of the Ṛgveda, while the commentary provides essential philological and linguistic insights that are being progressively published on their website[2].

For the source texts of the Vedic Sanskrit, we employ the electronic text available through TITUS (Thesaurus Indogermanischer Text- und Sprachmaterialien) (Martínez García and Gippert, 1995), which is based on Aufrecht's critical edition (Aufrecht, 1877). The transliteration of the text is converted following the IAST transliteration standard (Royal Asiatic Society of Great Britain and Ireland, 1896). This choice of romanization scheme is motivated by consistency with the English translation and commentary, where Sanskrit words and phrases are consistently referenced using this same IAST transliteration.

For the translation and commentary sources, the English translations are from the published Jamison & Brereton book, providing complete coverage of all 1,028 hymns and over 10,000 verses of

[2] http://rigvedacommentary.alc.ucla.edu/.

the Ṛgveda. The philological commentary is obtained from the authors' website.

The commentary encompasses multiple types of philological information essential for accurate translation: (1) Linguistic explanations including morphological analysis, etymological derivations, and syntactic parsing of complex constructions; (2) Lexical annotations providing semantic clarification of rare or polysemous terms, often with cross-references to other Vedic texts; (3) Cultural and ritualistic explanations elucidating the religious, social, and historical contexts necessary for proper interpretation; (4) Verse-level commentary addressing overall meaning, poetic structure, and intertextual relationships; (5) Background information on mythological references, deity characteristics, and ceremonial practices mentioned in the hymns.

Some hymns contain only commentary for the entire hymn and no verse-level commentary. Although hymn-level commentary provides valuable context for interpreting entire hymns, we restrict our dataset to verse-level commentary only. This decision is motivated by the need to manage input context length for language models, ensuring that each training sample remains within feasible token limits. By focusing on verse-level commentary, we maximize the amount of philological information available for each verse while maintaining compatibility with model context constraints.

This multi-layered commentary structure provides the rich contextual information that distinguishes our approach from previous work relying solely on automatically generated linguistic features. The scholarly commentary represents decades of expert philological analysis, offering insights unavailable to computational annotation tools.

Our final dataset comprises 6,754 total samples split into training (5,282 samples, 78.2%), validation (743 samples, 11.0%), and test (729 samples, 10.8%) sets. Figure 1 provides a box plot visualization of text lengths, highlighting the substantial differences in length and variability between the three text types. The Sanskrit source texts are relatively uniform in length, averaging 117.7 characters (median 129.0, std 27.8) in the training set, with similar statistics across validation and test splits. English translations are consistently longer than their Sanskrit counterparts, averaging 192.6 characters (median 199.0, std 54.6), reflecting the approximately 1.6× expansion typical when translating from Sanskrit to English due to morphological differences and explanatory additions required for clarity.

The philological commentaries show substantially greater length and variability, averaging 1,114.0 characters (median 735.0, std 1,189.4) in the training set, with some commentaries extending beyond 11,000 characters. This high variability reflects the scholarly practice of providing detailed analysis for particularly complex or significant verses while offering more concise notes for straightforward passages. The distribution is right-skewed, indicating that while most commentaries are relatively brief, a significant subset provides extensive philological analysis that substantially exceeds the length of the source texts themselves.

To quantify the lexical overlap between reference and commentary texts, we computed the n-gram Jaccard coefficients (for n = 1, 2, 3, 4). The mean Jaccard coefficients are 0.082, 0.018, 0.0075, 0.0038 for n = 1, 2, 3, 4 respectively. Figure 2 shows the box plots of the n-gram Jaccard coefficients. These low overlap values confirm that the contamination risk from commentary to reference translations is minimal, indicating that the commentary provides largely complementary information rather than redundant content.

We additionally conducted a content characterization of the commentary by randomly sampling 1,000 commentaries and assigning one or more labels from five categories –linguistic, cultural, religious, philosophical, and others– allowing multi-label assignments per commentary The resulting distribution (not mutually exclusive) shows a strong predominance of linguistic content (73.4%), followed by religious (32.6%), philosophical (15.0%), cultural (12.6%), and others (4.0%). At the entry level, 54.4% of commentaries received a single label, while 45.6% were multi-labeled, indicating that nearly half of the commentary have integrated information from multiple aspects.

4.2 Model Selection

We evaluate our approach across five diverse large language models to assess generalizability across different architectures, parameter scales, and training paradigms: **GPT-4.1 nano**[3], **Gemini 2.5**

[3] https://platform.openai.com/docs/models/gpt-4.1-nano

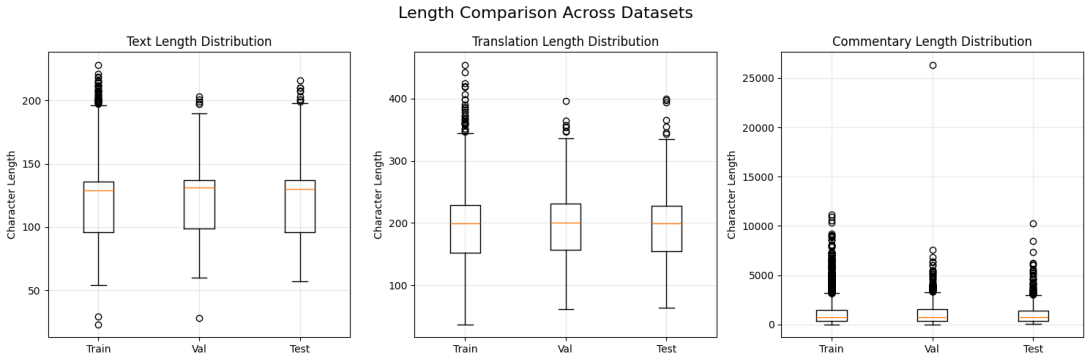

Figure 1: Box plots showing the distribution statistics for text lengths across the three text types. Commentary texts exhibit significantly higher variance and longer tails compared to source texts and translations.

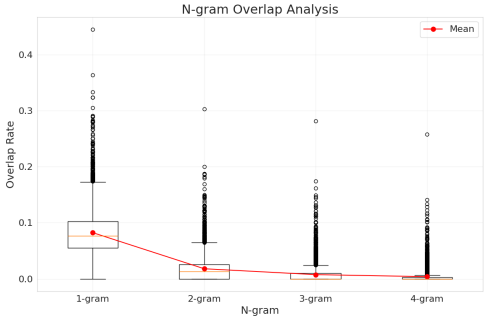

Figure 2: Box plots of n-gram Jaccard coefficients between reference and commentary texts.

flash[4], **Gemma 2 Mitra**[5], **Llama 3.1 8B**[6], and **Llama 3.2 3B**[7]. Gemma 2 Mitra is a specialized model for Sanskrit, developed under the MITRA project and built on Google's Gemma 2 foundation. It is trained on 7B tokens from diverse Buddhist texts, including Sanskrit, Tibetan, English, and Pāli.

4.3 Training Methodology

We formulate the translation enhancement task as conditional text generation, comparing two training conditions:

Standard SFT: Models are fine-tuned using Vedic texts as input to generate English translations.

Commentary-Enhanced SFT: Models are fine-tuned using both Vedic texts and scholarly commentaries in English as input to generate English translations.

Commentary-only SFT: Models are fine-tuned using only the commentaries as input to generate English translations.

For all models, we employed supervised fine-tuning with hyperparameter selection to balance learning efficiency with preservation of pre-trained capabilities. Both proprietary models (GPT-4.1 nano, Gemini 2.5 flash) and open-source models (Llama, Gemma/Mitra) were fine-tuned using platform-appropriate optimization techniques. Detailed training configurations, hyperparameters, and input formatting specifications are provided in Appendix A.

4.4 Evaluation Metrics

Translation quality was assessed using BLEU and COMET scores. The BLEU socores were calculated with SacreBLEU for standardized and reproducible evaluation. We computed BLEU-1 through BLEU-4 scores to capture both local phrase-level and global sentence-level translation quality. COMET scores were calculated using the pre-trained wmt22-comet-da model.

5 Results

5.1 Main Results: BLEU and COMET Scores

Table 1 reports BLEU and COMET scores for all models under three experimental configurations. Across most settings, the integration of philological commentary leads to consistent improvements in both BLEU and COMET, indicating gains in lexical accuracy and semantic alignment. GPT-4.1 nano achieves the highest relative increase in COMET (+0.03), followed by Gemma 2 Mitra (+0.04), while Gemini 2.5 flash maintains overall

[4] https://ai.google.dev/gemini-api/docs/models#gemini-2.5-flash

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

[6] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[7] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

Table 1: BLEU and COMET scores across pretrained models and dataset configurations (src = source text, com = commentary).

Train Dataset	Test Dataset	BLEU-1	BLEU-2	BLEU-3	BLEU-4	COMET
GPT-4.1 nano						
src+com	src+com	44.0	17.0	7.9	4.2	0.61
src	src	26.2	8.7	3.5	1.6	0.58
src	src+com	26.4	8.7	3.5	1.7	0.59
com	src	20.8	4.7	1.2	0.39	0.52
Gemini 2.5 flash						
src+com	src+com	64.1	35.6	22.0	14.4	0.60
src	src	60.0	31.2	18.2	11.3	0.67
src	src+com	61.4	35.8	21.7	13.6	0.69
com	src	50.9	22.6	11.0	5.1	0.62
Gemma 2 Mitra						
src+com	src+com	49.8	21.4	11.2	6.5	0.63
src	src	42.9	17.5	8.8	4.9	0.59
src	src+com	28.5	9.2	4.2	2.3	0.52
com	src	32.3	10.6	4.2	1.8	0.56
Llama 3.1 8B						
src+com	src+com	48.9	21.8	12.0	7.3	0.62
src	src	47.5	19.4	9.7	5.6	0.63
src	src+com	40.5	17.5	9.2	5.4	0.61
com	src	32.8	8.5	2.8	1.2	0.56
Llama 3.2 3B						
src+com	src+com	41.1	14.5	6.7	3.5	0.59
src	src	44.1	16.0	7.5	3.9	0.60
src	src+com	36.0	12.3	5.5	2.8	0.57
com	src	26.9	4.6	1.0	0.27	0.52

strong performance with COMET values ranging from 0.60 to 0.69. In contrast, Llama 3.2 3B shows a marginal decrease in both BLEU and COMET, suggesting limited benefit from commentary augmentation at smaller model scales.

Table 2 presents the improvements in BLEU-4 scores for each model when using Commentary-Enhanced SFT. GPT-4.1 nano shows the most significant improvement of +155.0%, while Gemini 2.5 flash follows with +45.0%. Large open-source models (Gemma 2 Mitra, Llama 3.1 8B) also demonstrate substantial gains of +30.5% and +31.1% respectively. Llama 3.2 3B, however, shows a decrease of -11.5% in BLEU-4 score with commentary integration, confirming the same trend observed above.

Table 2: Commentary-Enhanced SFT improvement in BLEU-4 scores

Model	Improvement (%)
GPT-4.1 nano	+155.0
Gemini 2.5 flash	+27.4
Gemma 2 Mitra	+30.5
Llama 3.1 8B	+31.1
Llama 3.2 3B	-11.5

5.2 Commentary-Enhanced SFT improvement

We also examine whether the length of the philological commentary (word count) relates to sentence-level BLEU improvement. As visualized in Figure 3, we observe no consistent positive or negative relationship across models: both short and long commentaries can yield gains. This sug-

gests that commentary length per se is not the primary driver of the observed improvements.

5.3 Qualitative Analysis

Our results confirm that philological commentary provides valuable contextual information that significantly enhances Vedic Sanskrit translation quality across diverse model architectures. The consistent improvements observed across four large models –ranging from lightweight 8B parameter models to sophisticated proprietary systems– suggest that the benefits of scholarly commentary integration are robust and generalizable.

Linguistic Insights The complex morphological system in Vedic Sanskrit presents considerable challenges for translation. Especially the language in the Ṛgveda is highly inflected with a slightly different morphological structure compared to later Vedic languages. Philological commentaries frequently provide detailed linguistic explanations: phonological changes, morphological parsing, identifying stems, and inflection patterns. Semantic explanations of ambiguous terms are also provided, enabling models to generate contextually appropriate translations rather than defaulting to the most common or literal meanings.

Cultural Context Most significantly, philological commentary supplies cultural knowledge essential for accurate translation that is unavailable to models trained solely on linguistic data. As shown in Appendix B, the top-5 BLEU improvement examples demonstrate that the linguistic and cultural context in the commentary enables models to better understand and translate Vedic hymns. Vedic hymns assume deep familiarity with Indo-Aryan religious practices, mythological narratives, and social structures. Commentary explicates these assumed contexts, enabling models to generate translations that capture not merely linguistic content but cultural significance.

Model Performance Variations The substantial performance differences between models reveal varying sensitivities to commentary integration. GPT-4.1 nano shows a remarkable improvement, which suggests that smaller, efficient models can particularly benefit from contextual information, potentially compensating for limited parameter capacity through enhanced input quality. By contrast, the performance degradation in Llama 3.2 3B indicates a threshold effect where models below a certain capacity may struggle to effectively utilize complex commentary information.

BLEU Score Interpretation The consistent improvements in BLEU scores across successful models suggest that commentary integration particularly enhances phrase-level and sentence-level coherence rather than merely improving word-level accuracy. This pattern aligns with the nature of philological commentary, which provides contextual and structural insights that facilitate more coherent target language generation.

Commentary Length and BLEU Improvements The scatter plots in Figure 3 indicate little to no relationship between commentary word count and sentence-level BLEU gains. In practice, both very short and very long commentaries can be beneficial, and performance does not systematically increase with length. We hypothesize that the key factor is information quality –e.g., explicit morphological analyses, disambiguation cues, and culturally specific background– rather than sheer volume. For smaller models, excessively long inputs may even strain context capacity, further weakening any length-gain association.

Implications for Ancient Language Translation Our findings have broader implications for computational approaches to ancient languages. The successful integration of philological commentary into neural machine translation suggests that traditional philological methods remain highly relevant in the age of AI. This commentary enhanced methodology could be extended to other underresourced ancient languages where scholarly traditions provide rich contextual frameworks.

Dataset Limitations Our approach, however, faces several methodological limitations. Notably, there exist authoritative German translations such as Geldner's complete work(Geldner, 1951) and the ongoing series(Witzel et al., 2007, 2013; Dōyama and Gotō, 2022), which represent significant contributions to Vedic translation studies. Comparative analysis with these major translations is essential for a comprehensive evaluation of translation quality.

Scalability Limitations Another important limitation concerns scalability. The integration of extensive philological commentary substantially increases input length and complexity, which can challenge both model context windows and com-

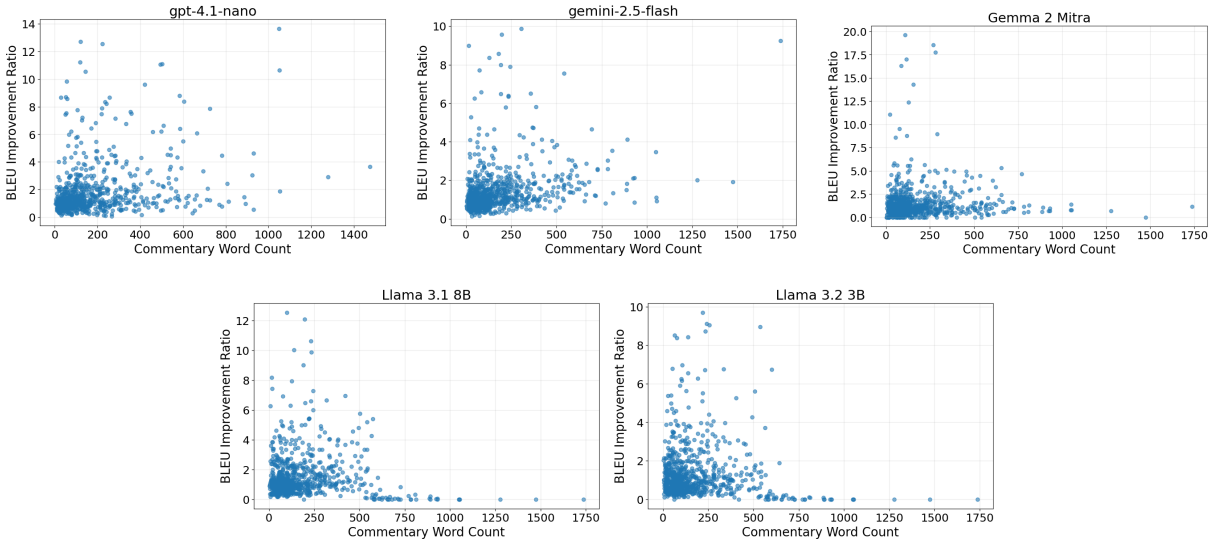

Figure 3: Scatter plots of sentence-level BLEU improvement versus commentary length (word count), grouped by model. Each point represents a verse.

putational resources. As commentary texts often exceed the length of source verses by an order of magnitude, scaling this approach to larger corpora or more comprehensive commentarial traditions may require advanced context management strategies, such as hierarchical encoding or retrieval-augmented generation.

6 Conclusion

This study demonstrates that incorporating scholarly philological commentary significantly improves Vedic Sanskrit translation accuracy across multiple large language model architectures. Our comprehensive evaluation across five diverse models confirms the robustness and generalizability of commentary-enhanced training, with consistent BLEU score improvements observed although this sensitivity varies with model size.

The methodology presented here establishes a replicable framework for leveraging scholarly expertise in neural machine translation of ancient languages. By demonstrating that traditional philological knowledge can effectively enhance modern AI capabilities, this work opens new avenues for computational approaches to ancient language processing and highlights the continued relevance of scholarly commentary in the digital age.

References

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

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16343–16360, Miami, Florida, USA. Association for Computational Linguistics.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, and Stephan Vogel. 2017. Understanding and improving morphological learning in the neural machine translation decoder. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 142–151, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Eijirō Dōyama and Toshifumi Gotō. 2022. *Rig-Veda: das heilige Wissen: sechster und siebter Liederkreis*, 1. aufl edition. Verl. der Weltreligionen, Berlin.

Tat'jana Jakovrevna Elizarenkova. 1999. *Rigveda*. Number 3 in Literaturnye pamjatniki. Nauka.

Mercedes García-Martínez, Loïc Barrault, and Fethi Bougares. 2016. Factored neural machine translation architectures. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

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.

Hermann Grassmann. 1876. *Rig-Veda : übersetzt und mit kritischen und erläuternden anmerkungen versehen von Hermann Grassmann*. Brockhaus, Leipzig.

Stephanie W. Jamison and Joel P. Brereton. 2014. *The Rigveda: The earliest religious poetry of India*. Oxford University Press, New York.

Stephanie W. Jamison and Joel P. Brereton. 2015. Rigveda translation: Commentary. Center for Digital Humanities, University of California, Los Angeles.

Diego Marcheggiani, Jasmijn Bastings, and Ivan Titov. 2018. Exploiting semantics in neural machine translation with graph convolutional networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 486–492, New Orleans, Louisiana. Association for Computational Linguistics.

Francisco Javier Martínez García and Jost Gippert. 1995. Thesaurus indogermanischer text- und sprachmaterialien.

Long H. B. Nguyen, Viet Pham, and Dien Dinh. 2020. Integrating amr to neural machine translation using graph attention networks. In *2020 7th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 158–162.

Reinhard Rapp. 2022. Using semantic role labeling to improve neural machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3079–3083, Marseille, France. European Language Resources Association.

Louis Renou. 1955. *Études védiques et pāṇinéennes*. Number sér. in-8o ; fasc. 1-2, 4, 6, 9-10, 12, 14, 16-18, 20, 22-23, 26-27, 30 in Publications de l'Institut de civilisation indienne. E. de Boccard, Paris.

Gil Rosenthal. 2023. Machina cognoscens: Neural machine translation for latin, a case-marked free-order language.

Royal Asiatic Society of Great Britain and Ireland. 1896. *Transliteration report*. London : The Society.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Xin Tan, Longyin Zhang, Fang Kong, and Guodong Zhou. 2022. Towards Discourse-Aware Document-Level Neural Machine Translation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4383–4389. International Joint Conferences on Artificial Intelligence Organization.

Michael Witzel, Toshifumi Gotō, Eijirō Dōyama, and Mislav Ježić. 2007. *Rig-Veda : das heilige Wissen Erster und zweiter Liederkreis*. Verlag der Weltreligionen, Frankfurt am Main.

Michael Witzel, Toshifumi Gotō, and Salvatore Scarlata. 2013. *Rig-Veda : das heilige Wissen Dritter bis fünfter Liederkreis*. Verlag der Weltreligionen, Frankfurt am Main.

Nier Wu, Hongxu Hou, Haoran Li, Xin Chang, and Xiaoning Jia. 2021. Semantic Perception-Oriented Low-Resource Neural Machine Translation. In *Machine Translation*, pages 51–62, Singapore. Springer Singapore.

A Training Configuration Details

This section provides comprehensive details on the training configurations, hyperparameters, and input formatting used in our experiments.

A.1 Input Format Specification

The input format for Commentary-Enhanced SFT follows this structure:

```
# Input:
[Original Vedic Sanskrit text]
# Commentary:
[Scholarly philological commentary]
# Translation:
[Target English translation]
```

For Standard SFT, the format is simplified to:

```
# Input:
[Original Vedic Sanskrit text]
# Translation:
[Target English translation]
```

The input format is for training, and the input in inference does not contain the "[Target English translation]" line.

A.2 Proprietary Model Hyperparameters

GPT-4.1 nano Fine-tuning Configuration

- Training epochs: 3
- Batch size: 10
- Learning rate multiplier: 0.1
- Validation split: 10%
- Platform: OpenAI Fine-tuning API

Gemini 2.5 flash Fine-tuning Configuration

- Training epochs: 2
- Learning rate multiplier: 5
- Adapter size: 4
- Platform: Google Vertex AI
- Features: Dynamic batch sizing, automatic optimization

A.3 Open-Source Model Hyperparameters

For Llama and Gemma models, we employed Low-Rank Adaptation (LoRA) for computational efficiency:

LoRA Configuration

- Rank (r): 16
- Alpha: 16
- Dropout: 0.0
- Target modules: All attention and feed-forward layers

Training Parameters

- Learning rate: 3e-4 with linear decay
- Batch size: 8 with gradient accumulation
- Training epochs: 40 with early stopping
- Optimizer: AdamW with 8-bit precision
- Weight decay: 0.01
- Warmup steps: 100
- Gradient clipping: 1.0

B Top-5 Examples with the Largest BLEU-4 Improvements

This appendix presents the five test-set instances with the largest *absolute* gains in BLEU-4 when philological commentary is provided to the model GPT-4.1 nano. For each example we show: the verse number (book, hymn, and verse numbers), the gold/reference translation, the baseline translation *without* commentary, and the translation *with* commentary, together with BLEU-4 before/after.

1. **BLEU-4 gain:** 51.0270 (2.6316 → 53.6585)
   Verse number: ṚV 6.2.7

   Reference translation (gold)
   For now you are to be reverently invoked among the clans, as our dear guest, to be protected like a son who brings delight to the home, (but also) to be enflamed like a battle-lusty (warrior) in a stronghold.

   Translation without commentary (BLEU-4: 2.6316)
   For here you are, o widely spread (flame), the fire to be kindled, dear guest to us, delighting like an old man in his fortress, like a son who is to be protected.

   Translation with commentary (BLEU-4: 53.6585)
   For you are eager for battle, dear guest; you are to be protected like a son who brings delight (to the home), (but also) to be enflamed like a battle-lusty (warrior) in a stronghold.

2. **BLEU-4 gain:** 33.0282 (1.7544 → 34.7826)
   Verse number: ṚV 9.73.1

   Reference translation (gold)
   In the jaw (of the pressing stones) while the droplet was blowing its blast, they sounded in unison. The ties of lineage have joined together in the womb of truth. The lord [=Soma] has made his three heads to be seized. The boats of the trusty one [=Soma] have brought the good (ritual) performer across to the other shore.

   Translation without commentary (BLEU-4: 1.7544)
   Streams of soma roar as they cross the filter together, the womb of truth together with the navel of truth, the three heads of the lord of the night (being) joined together. The boat of truth, well made, has been set in motion. The trusty boat has been set in motion.

   Translation with commentary (BLEU-4: 34.7826)
   The streams of soma roar together as they cross the filter, and the ties of lineage have joined together in the womb of truth. The demon made his three heads to be seized. The boat of truth, well made, has been set forth.

3. **BLEU-4 gain:** 26.4069 (19.0476 → 45.4545)
   Verse number: ṚV 8.75.11

   Reference translation (gold)
   Surely you will toil for wealth for us, for our quest for cattle, Agni? Make a wide (way) for us, you wide-maker.

   Translation without commentary (BLEU-4: 19.0476)
   Make for us a dwelling place, Agni, for our quest for cattle, for wealth. Make us strong, strong.

   Translation with commentary (BLEU-4: 45.4545)
   O Agni, you will toil for wealth for us, for our quest for cattle. Make it broad and broad for us.

4. **BLEU-4 gain:** 25.2519 (3.3195 → 28.5714)
   Verse number: ṚV 9.65.8

   Reference translation (gold)
   Him whose tawny hue, dripping with honey, they impel (like) a fallow bay with the stones, the drop for Indra to drink,

   Translation without commentary (BLEU-4: 3.3195)
   The tawny color of which they impel the honeyed (soma) with the stones, the tawny (color) they impel for Indra to drink, – Commentary: The first hemistich is somewhat tricky. The relative clause yāsyā (whose) must refer to soma, with the relative clause hanging off (the first hemistich) (see v. 7). In pāda a, the relative yāsyā qualifies the color (vārṇam), which serves as the object of impel (hinvanti) in b. "They impel his color" is a slightly odd locution, and it becomes odder in b, with the accusative hārim. The stem hār- is a color term and could qualify vārṇam ("tawny color"), but it also is regularly applied directly to soma (e.g., in vss. 12 and 25 of this hymn) and also identifies Indra's horse(s). Ge (note 8b) cleverly suggests that hārī- is to be read twice, with the second reading an unmarked simile referring to a horse of the appropriate color. This provides a more appropriate object for impel ("they impel (like) a fallow bay (horse)"), while connecting the putative horse with the "tawny color" (vārṇam … hārim) that is the first object of the verb.

   Translation with commentary (BLEU-4: 28.5714)
   The tawny color of which they impel (like) a fallow bay (horse) with stones, for Indra to drink.

5. **BLEU-4 gain:** 24.2857 (2.3810 → 26.6667)
   Verse number: ṚV 8.21.13

Reference translation (gold)
You are without rival, but by the same token, without friend, Indra, by birth and from of old. Only in battle do you seek friendship.

Translation without commentary (BLEU-4: 2.3810)
Nephew, rival, you are not, Indra. You have been from of old, and you seek to be friend.

Translation with commentary (BLEU-4: 26.6667)
You are without rival, Indra, by the same token, without friend, but from of old you have been there. You will fight if you want to fight.