MITRA-zh-eval: Using a Buddhist Chinese Language Evaluation Dataset to Assess Machine Translation and Evaluation Metrics

Sebastian Nehrdich[1,3*] **Avery Chen**[1*] **Marcus Bingenheimer**[2*]
Lu Huang[2] **Rouying Tang**[2] **Xiang Wei**[2] **Leijie Zhu**[2] **Kurt Keutzer**[1]

[1]University of California, Berkeley, Berkeley Artificial Intelligence Research (BAIR),
[2]Temple University, Philadelphia, [3]Heinrich Heine University Düsseldorf
*Equal contribution.

Abstract

With the advent of large language models, machine translation (MT) has become a widely used, but little understood, tool for accessing historical and multilingual texts. While models like GPT, Claude, and Deepseek increasingly enable translation of low-resource and ancient languages, critical questions remain about their evaluation, optimal model selection, and the value of domain-specific training and retrieval-augmented generation setups. While AI models like GPT, Claude, and Deepseek are improving translation capabilities for low-resource and ancient languages, researchers still face important questions about how to evaluate their performance, which models work best, and whether specialized training approaches provide meaningful improvements in translation quality. This study introduces a comprehensive evaluation dataset for Buddhist Chinese to English translation, comprising 2,662 bilingual data points from 32 texts that have been selected to represent the full breadth of the Chinese Buddhist canon. We evaluate various computational metrics of translation quality (BLEU, chrF, BLEURT, GEMBA) against expert annotations from five domain specialists who rated 182 machine-generated translations. Our analysis reveals that LLM-based GEMBA scoring shows the strongest correlation with human judgment, significantly outperforming traditional metrics. We then benchmark commercial models (GPT-4 Turbo, Claude 3.5, Gemini), open-source models (Gemma 2, Deepseek-r1), and a domain-specialized model (Gemma 2 Mitra) using GEMBA. Our results demonstrate that domain-specific training enables open-weights models to achieve competitive performance with commercial systems, while also showing that retrieval-augmented generation (RAG) significantly improves translation quality for the best performing commercial models.

1 Introduction

Evaluating machine translation (MT) systems remains a challenging endeavor, especially for literary contexts where a single "correct" translation is often elusive, and interpretation plays a significant role in determining quality. For many years, evaluation relied on string-similarity metrics such as BLEU and chrF, which are not well suited for this scenario (Kocmi et al., 2024). However, the recent advent of deep learning–based methods has sparked a shift toward more sophisticated evaluation techniques, creating what some have aptly termed a "metrics maze" (Kocmi et al., 2024). Although there are large-scale initiatives like the annual WMT evaluation campaign for high-resource languages, comparatively little attention has been devoted to assessing translation quality in literary, premodern, and low-resource domains. In this study, we address the unique challenges of assessing machine translation quality for premodern Buddhist Chinese into modern English, a task that involves bridging considerable cultural and temporal divides. For this, we introduce a novel dataset comprising 2,662 bilingual data points, carefully selected by domain experts to represent the full breadth of the Chinese Buddhist canon. Additionally, we translate a subset of 182 data points using a range of machine translation systems and engage five domain experts to evaluate the quality of these translations. This not only allows us to measure inter-annotator agreement, but also to benchmark various automatic evaluation metrics against expert human judgment. Subsequently, we assess both commercial and open-weight machine translation systems on our dataset to provide an overview of the current performance landscape for this challenging language pair. Finally, we conduct an ablation study to demonstrate how different data augmentation strategies can further enhance the performance of large language

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 129–137
May 3-4, 2025 ©2025 Association for Computational Linguistics

models (LLMs) in this specialized domain. Our contributions can be summarized as follows:

- A novel, comprehensive evaluation dataset for machine translation of premodern Buddhist Chinese, comprising 2,662 bilingual data points.
- A detailed human evaluation of 182 machine-generated translations conducted by domain experts.
- A comparative assessment of automatic evaluation metrics against expert human ratings.
- A comprehensive performance analysis of both commercial and open-weight LLM-based machine translation systems.
- An ablation study highlighting the impact of various data augmentation strategies on LLM performance.

We make the datasets and evaluation pipeline used for this study available at https://github.com/dharmamitra/mitra-evaluation.

1.1 Premodern Buddhist Chinese

This paper focuses on the evaluation of machine translations of premodern Buddhist Chinese texts. Premodern Buddhist Chinese is the idiom in which Buddhist texts were written between 150 and 1900 CE, and these texts are read and recited in China, Korea, Japan, and Vietnam until today.

Several thousand of these texts were preserved in canonical editions. Although the language of canonical texts varies greatly depending on time, ideolect, and genre, there are a few features that distinguish Buddhist Chinese from premodern Classical Chinese in general.

Buddhist Chinese has a sizable number of vocabulary terms transliterated or translated from Indian sources. The transmission of this vocabulary from Indian sources was never fully standardized and a many-to-many relationship exists between Indian terms and their Chinese equivalents. Secondly, in part as a result of the presence of these Indian terms, but also because of the occasional adoption of vernacular phrases, Buddhist Chinese tends to have a higher proportion of multisyllabic words than other forms of premodern Chinese, where (ideally) one character equals one word. Thirdly, the translated texts in the Chinese Buddhist corpus often combine prose and verse. While prosimetric literature was common in early India, it is rare in non-Buddhist Chinese at least during the first millennium.

2 Related Work

So far, Buddhist Chinese has received little dedicated attention in NLP research. The first publication that trains and evaluates machine translation for this domain is (Li et al., 2022), but they did not publicly release either their models or their training or evaluation datasets.

Another recent publication discusses the training and evaluation of machine translation systems for Buddhist Chinese (Nehrdich et al., 2023). They released an evaluation dataset consisting of sections of a couple hundred sentence pairs taken from seven different texts. One detailed human-only evaluation compares the MT output of three Buddhist texts from three LLMs (ChatGPT 4, ERNIE Bot 4, and Gemini Advanced) (Wei, 2024).

In the context of Classical poetry, (Chen et al., 2024) provides an evaluation benchmark for Classical Chinese poetical texts, which attempts to assess the poetic "elegance" of machine translations. More distantly related is (Song et al., 2024), which examines how classical Chinese to modern Chinese data influences the process of historical Korean document translation from Hanja to modern Korean and English.

To summarize, in previous publications, the evaluation of machine translation performance for Buddhist Chinese has not played a main role. The only study that provides an evaluation dataset, (Nehrdich et al., 2023), has only used sections from very few texts with very limited domain coverage. So far, there is no study assessing the quality of automatic metrics for machine translation evaluation for this idiom.

3 Dataset

The evaluation dataset we present from Buddhist Chinese to English translation consists of 2,662 ZH-EN data points drawn from 32 Chinese Buddhist texts and their corresponding human translations. The Chinese was taken from the CBETA corpus.[1]

The translations were selected in a way such that they were distributed evenly across the canon

[1] https://github.com/cbeta-org/xml-p5

to prevent bias towards certain sections. Collectors were instructed to move in steps of fifty Taishō[2] numbers and identify a translation close to either side of that number using the "Bibliography of Translations (by human translators) from the Chinese Buddhist Canon into Western Languages" (Bingenheimer Ver 2024-11).[3] Priority was given to translations that are not widely available online, e.g. the open-access translations published by Bukkyō Dendō Kyōkai (仏教伝道協会), to mitigate the influence of data that is overrepresented in web-scraped datasets.[4] For each text, we collected the first 50-100 sentences after the prefaces and introductory paragraph. We cleaned line-end hyphenations, line returns, and deleted notes and note anchors. We use Bertalign for sentence-level alignment of the document pairs (Liu and Zhu, 2022). While the oldest of the English translations date back to 1951, the majority was produced within the last 30 years, ensuring relatively consistent modern English usage across the reference translations

This is the first balanced comprehensive evaluation dataset for Buddhist Chinese. Crucially, it allows control for genre, i.e., it helps us understand whether the output quality is or is not dependent on the type of text that is translated.

4 Human Evaluation and Computed Metrics

In our evaluation of different metrics for this idiom, five human annotators independently assessed 182 machine-generated translations. These have been generated with machine translation systems of varying quality. We excluded any output of Gemini 2 Flash here, since this LLM is also used as the judge for the GEMBA scoring, and evaluation of its own output could lead to undesired bias. All annotators hold PhDs or are doctoral candidates specializing in Buddhist Chinese texts. The annotators rated each translation on a scale from 1 (worst) to 5 (best), considering the source sentence, the machine translation output, and a reference translation. While we did not conduct specific annotation training, all evaluators worked with identical sets of sentences, allowing us to measure inter-annotator agreement. Table 2 presents these results. The average pairwise Spearman correlation across annotators is 0.4, with considerable variation in agreement between individual pairs. These results suggest that evaluating Buddhist Chinese to English translations is a complex task where applying objective criteria proves challenging. We recognize that more comprehensive annotator training would likely improve inter-annotator agreement.

We evaluated several metrics against the human-annotated reference scores: BLEU, (Papineni et al., 2002), BLEURT (Sellam et al., 2020), chrF (CHaRacter-level F-score) (Popović, 2017), and the LLM-based GEMBA (Kocmi and Federmann, 2023). For GEMBA, we implemented assessment using Gemini 2.0 flash prompting on a scale of 0-100, and additionally tested a reference-free configuration (denoted as GEMBA*). We calculated both Pearson and Spearman correlations against each annotator's scores and present the averaged correlations in Figure 1.

The results reveal weak average correlations for both BLEU and chrF, supporting previous findings (Kocmi et al., 2024) that these metrics are inadequate for evaluating machine translation output across different model types. While BLEURT consistently outperforms BLEU and chrF, both GEMBA variants demonstrate even stronger performance. Notably, the reference-free GEMBA* achieves comparable Spearman correlation to its reference-based counterpart, with only slightly lower Pearson correlation. We attribute this performance pattern to potential issues in automatic sentence alignment and variations in human reference translation quality.

Based on these findings, we recommend using LLM-based metrics, such as GEMBA, for evaluating Buddhist Chinese to English machine translation. Particularly, reference-free LLM-based evaluation proves highly effective, significantly outperforming traditional reference-based systems without needing to rely on costly manual data collection.

5 Model Evaluation

We compare the following different systems against each other: The commercial LLMs Claude

[2] The "Taishō" is the most widely used canonical edition of the Chinese Buddhist canon. Based on earlier editions the Taishō Shinshū Daizōkyō 大正新脩大藏經 was compiled in Japan 1924-1934.

[3] `https://mbingenheimer.net/tools/bibls/transbibl.html`

[4] The Bukkyō Dendō Kyōkai ("Society for the Promotion of Buddhism" `https://www.bdk.or.jp/`) has funded a large number of translations from the Taishō canon into English

Identifier	Full Title	Translation Year	Datapoints
T01n0001	長阿含經	2017	91
T02n0099	雜阿含經	2013	63
T02n0142	玉耶女經	1951	123
T04n0198	義足經	1951	63
T08n0246	仁王護國般若波羅蜜多經	1998	19
T09n0273	金剛三昧經	1989	105
T11n0316	大乘菩薩藏正法經	1976	62
T12n0374	大般涅槃經	1975	89
T13n0417	般舟三昧經	2011	91
T14n0450	藥師琉璃光如來本願功德經	2009	102
T14n0515	如來示教勝軍王經	2024	113
T17n0842	大方廣圓覺修多羅了義經	1997	110
T19n0959	頂輪王大曼荼羅灌頂儀軌	2016	85
T19n1022B	一切如來心祕密全身舍利	2012	165
T20n1060	千手千眼觀世音菩薩廣大圓滿無礙大悲心陀羅尼經	2017	331
T20n1077	七俱胝佛母心大准提陀羅尼經	2012	76
T20n1136	一切諸如來心光明加持普賢菩薩延命金剛最勝陀羅尼經	2021	46
T20n1166	馬鳴菩薩大神力無比驗法念誦儀軌	2015	33
T21n1261	訶利帝母真言經	2019	96
T21n1277	速疾立驗魔醯首羅天說阿尾奢法	2016	56
T21n1305	北斗七星念誦儀軌	2000	23
T21n1394	佛說安宅神經	2023	55
T24n1492	舍利弗悔過經	2012	49
T30n1568	十二門論	1982	96
T32n1666	大乘起信論	2019	62
T34n1725	法華宗要	2012	40
T37n1762	阿彌陀經要解	1997	59
T42n1826	十二門論宗致義記	2015	57
T45n1857	寶藏論	2002	115
T45n1909	慈悲道場懺法	2016	61
T47n1961	淨土十疑論	1992	75
T48n2004	萬松老人評唱天童覺和尚頌古從容庵錄	2005	51
Total			**2689**

Table 1: Full title, year of translation, and number of datapoints for each of the evaluation documents. The total number of datapoints across all documents is 2,662.

	1	2	3	4	5
1	-	0.342	0.456	0.452	0.566
2	0.342	-	0.299	0.373	0.384
3	0.456	0.299	-	0.310	0.489
4	0.452	0.373	0.310	-	0.332
5	0.566	0.384	0.489	0.332	-

Table 2: Pairwise Spearman correlations between five different annotators on the machine translation task.

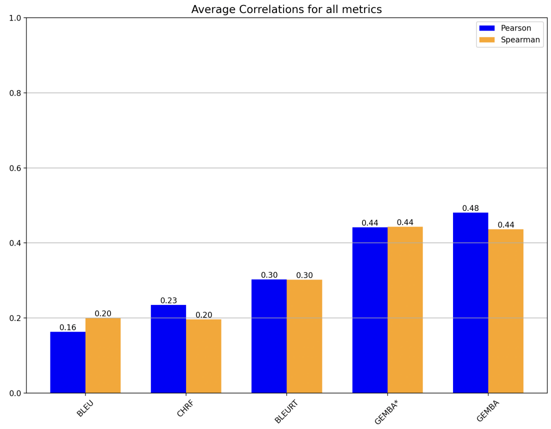

Figure 1: Comparison of evaluation scores for machine-translated Buddhist Chinese texts. For each metric, we give the average Pearson and Spearman correlation with all five human annotators.

Haiku 3.5 and Claude Sonnet 3.5, ChatGPT 4 Turbo, Gemini 1.5 Pro, as well as Gemini 2 Flash. These models were prompted between Jan 15 and Feb 10, 2025. We also evaluate the openly available LLMs DeepSeek-R1-Distill-Qwen-14B (DeepSeek-AI, 2025), Gemma 2 9B IT (Team et al., 2024) as well as Gemma 2 Mitra,[5] which is based on Gemma 2, but utilizes the Buddhist Chinese to English dataset presented in (Nehrdich et al., 2023) together with additional domain-specific monolingual data in a continuous pretraining/fine-tuning setup (publication forthcoming). We further evaluate the two commercial LLMs Gemini (Ver. 2 Flash and Ver. 1.5 Pro) as well as Claude (3.5 Sonnet) in a RAG setup. With RAG setup we mean a setup where the prompt of the LLM is enriched by additional knowledge. In this case, this means retrieving relevant source-target sentence pair examples from bilingual data storage with a semantic embedding model and nearest neighbor search. A recent implementation of such a system that we take inspiration from is (Wang et al., 2024). In our case, we add n=10 k nearest neighbor examples to the prompt from the previously mentioned Chinese-English dataset. Our used prompt template is given in appendix A. We also compare different augmentation strategies for the RAG setup in the ablation study.

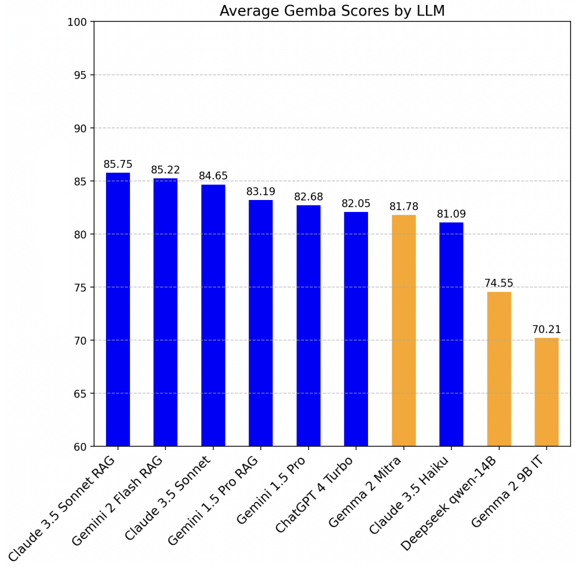

Figure 2: Average GEMBA scores across all documents per model. We present commercial, closed models in blue and open models in orange.

The averaged results per model are presented in Figure 2. The scores for each individual text are given in Figure 3.

Among all models, Claude 3.5 Sonnet RAG shows the best performance, followed by Gemini 2 Flash RAG. We acknowledge that since Gemini 2 Flash is also used as the judge in the GEMBA scoring system, the score might show bias in favor of this system. All the other major commercial LLMs Gemini 1.5 Pro, ChatGPT 4 Turbo, and Claude 3.5 Haiku show very similar performance across all texts with very similar overall trends. The open-source models, except for Gemma 2 Mitra, show a noticeable drop in performance. Among these, Deepseek qwen-14B is doing the best, at times matching the performance of the commercial LLMs. Gemma 2 9B IT is struggling to provide useful quality. The contrast in performance between this model and Gemma 2 Mitra shows that fine-tuning open-source models on an academic budget, even if their base performance is inferior, can lead to competitive performance with the right data selection.

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

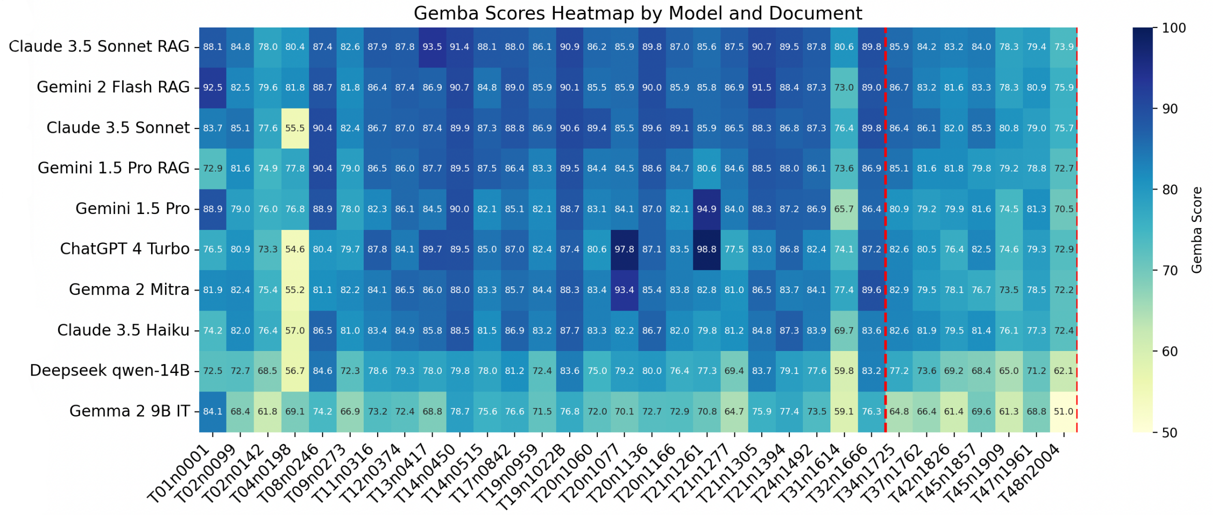

Figure 3: Heatmap of the model performance on the individual texts in GEMBA. Texts on the x-axis are sorted according to their position in the Buddhist Chinese canon. Models are sorted on the y-axis from best performing (top) to weakest (bottom). The breakpoint detected with the PELT method is indicated by the red dashed line.

The RAG setup improves the performance of both Gemini 1.5 Pro and Claude, with the improvement being more pronounced in the case of Gemini 1.5 Pro. For both LLMs, the improvements are more pronounced for the earlier sections of the corpus, which are also better represented in the dataset.

To identify significant changes in GEMBA Score trends across documents, we applied a change point detection algorithm based on the Pruned Exact Linear Time (PELT) method (Killick et al., 2012) with respect to the scores for all models, treating the documents as a time series. In our analysis, we set the penalty parameter (pen=1) to ensure that additional change points would only be introduced if they led to a substantial reduction in the overall cost. Notably, the single change point detected with this penalty setting occurs after T1725, between categories 32 and 34: categories 1-32 comprise material presented traditionally as translated from Indic sources into Chinese. In contrast, categories 34 and onward consist of original compositions and commentaries composed in China that are not presented as direct translations from Indian source texts. The BLEURT scores too start to decline after T1725. The detected change point, combined with the lower values observed for categories 34 and higher, suggests that less suitable data is available to the LLMs for training on this type of data. This hints at a generally reduced translation activity and scholarly attention in texts not explicitly claimed to be based on Indian originals.

5.1 Qualitative Discussion

Some texts present particular challenges to all advanced models. All scores register a performance drop for T31n1614 (*Da sheng bai fa ming men lun* 大乘百法明門論). Taishō 1614 is a short list of hundred *dharma*, doctrinal concepts which Xuanzang translated in the 7th century. Here the ZH-EN data points are not full sentences but items in a group of numbered lists. The LLM-based metrics as well as the purely statistical chrF highlight a stronger than usual difference between the LLM MT output and the reference translation. This is not due to an inherent textual difficulty, but simply reflects the list-like nature of the original, where single items without syntactic context can be translated very differently. It proves, to a degree, that the metrics work and indeed pick up a larger than usual variance between MT output and reference translation.
One text for which LLMs seem to produce comparatively less reliable results is T48n2004 (*Wan song lao ren ping chang tian tong jue he shang song gu cong rong an lu* 萬松老人評唱天童覺和尚頌古從容庵錄). This 13th-century Chan Buddhist work, the recorded sayings of Xingxiu 行秀 (1166-1246), presents unique linguistic challenges due to its intentionally poetic and obscure nature. The text is characterized by antinomic expressions, vernacular language elements, and non-sequential narrative structure. The text's deliber-

ate ambiguity poses challenges for LLMs, which are optimized for generating coherent prose. This limitation is evident e.g. in the translation of “一段真風見也麼”. While the human translator rendered it as “Do you see the true manner of the primal stage?”, the models showed varying degrees of comprehension. Claude Sonnet 3.5 came closest with “Is this a glimpse of true reality?”, while other models struggled significantly. Claude Haiku produced the incorrect “A paragraph of true Kazami, indeed,” GPT-4 Turbo was unable to process the text and flagged it as incorrect Chinese, and Gemma 2 produced an incorrect literal translation: “A genuine gust of wind.”

A major problem with sentence-based evaluation metrics becomes obvious in the low scores for T04n0198. As the machine translations are produced sentence by sentence, the context is lost. The human reference translation has an unfair advantage in that it usually gets the subject and numerus right, which in Chinese is often omitted. Also, Chinese characters arguably have a higher semantic variance than most English words thus context beyond the sentence level is even more important. Thus the MT output is often a possible, correct rendering of the out-of-context sentence, but at the same time quite wrong in-context, and consequently the MT differs significantly from the human reference translation and receives a low score. Thus “當亡棄法” in T04n0198, which in context means “Things that are bound to perish”, is plausibly translated as “When the Dharma is abandoned” (Claude 3.5 Sonnet) or “When abandoning the law” (GPT 4 Turbo). The low scores of T04n198 are probably due to a larger than usual number of such cases, where translations that are correct on the sentence level, are flagged as mistakes when compared to the human reference which was done with the paragraph in mind. Such findings suggest that paragraph-based evaluation might result in higher scores.

For all the slight differences in the evaluation of individual texts by different metrics, all metrics show superior performance for the commercial models and Gemma 2 Mitra as compared to the open-access models DeepSeek-R1-Distill-Qwen-14B and Gemma 2 9B IT. As Gemma 2 Mitra is based on Gemma, this shows that research communities still can benefit significantly from developing their own, domain-specific machine translation systems.

Model	BLEU	chrF	BLEURT	GEMBA
Base	9.01	33.49	0.558	82.8
+Dict	9.05	33.85	0.555	81.3
+En	**11.06**	**35.41**	**0.583**	**83.7**
+Ko	9.93	34.62	0.563	83.6
+Zh	9.38	34.41	0.567	81.5
+En +Ko	10.72	35.03	0.574	83.5
+En +Ko +Dict	10.28	34.70	0.566	82.9

Table 3: Translation performance of Gemini 1.5 Pro with different additional data sources used for retrieval augmentation.

6 Ablation Study

To investigate the impact of different data sources on RAG translation performance for this evaluation dataset, we conducted an ablation study with the Gemini-pro model across multiple configurations:

- A baseline without additional data (Base)
- Augmented with Buddhist dictionary entries taken from the Digital Dictionary of Buddhism[6] (+Dict)
- Enhanced with Buddhist Chinese-English parallel data (Nehrdich et al., 2023) as k nearest neighbor retrieval examples (+En)
- Supplemented with Buddhist Chinese-Korean parallel data (Nehrdich et al., 2023) (+Ko)
- Enriched with Classical-Modern Chinese parallel data from the NiuTrans project[7] (+Zh)
- Combination of Chinese-English and Chinese-Korean parallel data (+En +Ko)
- Korean, English, as well as dictionary entries combined (+En +Ko +Dict)

For all augmentation settings, we used semantic embeddings and nearest neighbor search to retrieve a fixed number of 10 samples that are most closely relevant to the translation query segment. We show the results in Table 3. The findings reveal several key patterns. First, the addition of dictionary entries (+Dict) yields minimal improvement over the baseline. In contrast, incorporating

[6] http://www.buddhism-dict.net/

[7] https://github.com/NiuTrans/Classical-Modern

Buddhist Chinese-English parallel data (+En) produces the most substantial gains across all metrics, establishing the best-performing configuration. Both Buddhist Chinese-Korean (+Ko) and Classical-Modern Chinese (+Zh) data contribute slight improvements across all evaluation metrics. This mirrors observations made in (Song et al., 2024) where incorporating the Classical-Modern Chinese dataset yields minimal or non-significant improvements for Hanja document machine translation. Notably, combining Chinese-English and Chinese-Korean parallel data (+En +Ko) slightly degrades performance compared to using Chinese-English data alone (+En). This performance deterioration becomes more pronounced when dictionary entries are added to this combination (+En +Ko +Dict). In conclusion, we recommend the augmentation of commercial LLMs with Buddhist Chinese-English data for best performance, as this yields significant improvements.

7 Conclusion and Future Work

We have presented a comprehensive and balanced manually assembled dataset for the benchmarking of machine translation of Buddhist Chinese material into English. We further conducted a manual evaluation of automatically generated translations against their reference data, which enabled us to benchmark different evaluation scores, establishing GEMBA as the best-performing automatic evalation method. Strikingly, we could show that the reference-free GEMBA* performs almost as good as reference-based GEMBA, which means that reliable evaluation of Buddhist Chinese to English machine translation is possible even when no dedicated reference data is collected. This is significant, since collecting domain-specific evaluation data is time-intensive and not many annotation experts exist who can do this type of work.

We then conducted an evaluation of commercial as well as open-source LLMs on this dataset, mapping out the current performance landscape for this task. Our results show that even high-performing commercial LLMs significantly benefit from data augmentation using curated domain-specific datasets, highlighting that dedicated data collection efforts are still crucial for optimal performance.

The results also demonstrated that domain-specific fine-tuned models such as Gemma 2 Mitra vastly outperform other open-weight models and show competitive performance with commercial models, highlighting that fine-tuning such models can be very worthwhile for research communities. One research question was whether genre plays a role in translation performance. Our experiments show no clear difference regarding the type of text. Although the evaluation dataset is a cross-section of the canon, no genre stands out as particularly easy or difficult for current MT systems. The notable exception here is the divide between categories 1-32, which all models handle better, and 34 onwards, which all models handle worse, indicating that the autochthonous sections of the Buddhist Chinese canon are likely less represented in the training data of these models.

8 Limitations

This study has a number of important limitations to consider. First, while 32 texts selected evenly across the Buddhist Chinese canon is considerable, they only reflect a small portion of about 1.4% of the total 2,437 texts present in the digital CBETA collection. Also, the selected passages are from the beginning of the texts, which might not capture the full possible variation in content, language, and style of the works.

The human evaluation, while conducted by 5 different domain experts, was limited to a rather small sample size of 182 sentences. With a relatively low inter-annotator agreement with an average 0.4 pairwise Spearman correlation, we have to ask ourselves whether more structured annotation guidelines and training or a larger number of evaluators could lead to better agreement.

In the metric evaluation, we relied on GEMBA with Gemini 2 Flash as the LLM judge. We acknowledge that this might lead to bias in the scoring, and repeated experiments with different LLMs are necessary in order to evaluate the impact of the LLM selection for this metric type. This is especially relevant for the comparative evaluation of the different LLMs presented in Figure 2 as well as Figure 3, wherein the current setup Gemini 2 Flash RAG is judged by the Gemini 2 Flash based metric GEMBA.

In the ablation study, we focused on just one LLM, Gemini 1.5 Pro. The impact of the data augmentation strategies on different LLM types might vary. More extensive testing across different LLM types is therefore very desirable to see if the observed patterns are consistent. Also, we only used one retrieval stategy here, nearest neighbor retrieval based on semantic similarity embedding. We acknowledge that further comparison of different retrieval methods as well as other in-context-learning strategies for few-shot machine translation is very desirable.

References

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024. Large language models for classical chinese poetry translation: Benchmarking, evaluating, and improving.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Rebecca Killick, Paul Fearnhead, and I.A. Eckley. 2012. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107:1590–1598.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *European Association for Machine Translation Conferences/Workshops*.

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

Denghao Li, Yuqiao Zeng, Jianzong Wang, Lingwei Kong, Zhangcheng Huang, Ning Cheng, Xiaoyang Qu, and Jing Xiao. 2022. Blur the Linguistic Boundary: Interpreting Chinese Buddhist Sutra in English via Neural Machine Translation . In *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 228–232, Los Alamitos, CA, USA. IEEE Computer Society.

Lei Liu and Min Zhu. 2022. Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities*.

Sebastian Nehrdich, Marcus Bingenheimer, Justin Brody, and Kurt Keutzer. 2023. MITRA-zh: An efficient, open machine translation solution for buddhist Chinese. In *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pages 266–277, Tokyo, Japan. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *ACL*.

Seyoung Song, Haneul Yoo, Jiho Jin, Kyunghyun Cho, and Alice Oh. 2024. When does classical chinese help? quantifying cross-lingual transfer in hanja and kanbun.

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

Jiaan Wang, Fandong Meng, Yingxue Zhang, and Jie Zhou. 2024. Retrieval-augmented machine translation with unstructured knowledge. *ArXiv*, abs/2412.04342.

Xiang Wei. 2024. The use of large language models for translating buddhist texts from classical chinese to modern english: An analysis and evaluation with chatgpt 4, ernie bot 4, and gemini advanced. *Religions*.

Appendix

A RAG Translation Prompt Template

You are an expert translator of classical Asian languages.
{dictionary_entries}
{example_sentence$_{pairs}$}
Now translate the following text to English. Make use of the provided examples. Provide only the translation, without any explanation or additional information: