MITRA-MT: Continued Pretraining, Parallel Corpus Mining, and Multi-Directional Machine Translation for Sanskrit, Tibetan, and Buddhist Chinese

Sebastian Nehrdich[1] **Kurt Keutzer**[2]

[1]Center for Integrated Japanese Studies, Tohoku University,
[2]University of California, Berkeley, Berkeley Artificial Intelligence Research (BAIR)

Abstract

Buddhist literature is preserved in a network of classical languages, above all Sanskrit, Pāli, Buddhist Chinese, and Tibetan, for which machine translation support remains limited, and largely restricted to translation into English. In order to address this shortcoming, we present three connected contributions. First, we release MITRA-parallel v2, a corpus of 1.69 million automatically aligned parallel records (2.34 million segment pairs) between Sanskrit, Tibetan, and Buddhist Chinese, mined with an embedding-based span-mining pipeline. Second, we present MITRA-MT, a domain-adapted large language model built by continued pretraining of Qwen3.5-9B on a 22.6-billion-token corpus of classical Asian languages and related modern material, followed by a lightweight translation fine-tune and preference optimization. Third, we introduce a multi-directional evaluation suite covering 17 translation directions at sentence level (ten of them also at paragraph level), including cross-classical directions (Sanskrit↔Tibetan, Sanskrit↔Chinese, Tibetan↔Chinese) and from Tibetan into seven modern languages beyond English. On this suite, MITRA-MT outperforms all open baselines, including instruction-tuned LLMs up to 122B parameters and dedicated MT systems, on all directions, and matches commercial references on the large majority of directions when evaluated via chrF, with the largest margins for translation into Tibetan, Sanskrit, and Classical Chinese. We release the parallel corpus, the evaluation data, and the model weights.

1 Introduction

The Buddhist tradition has produced one of the largest bodies of premodern literature in existence, preserved primarily in Pāli, Sanskrit, Buddhist Chinese, and Tibetan. Much of this material was translated between these languages historically: from Indic languages into Chinese beginning in the 2nd century CE, and into Tibetan from the 8th century CE onwards. The corpus therefore features a high degree of parallelism across languages, yet these parallels are largely uncataloged. In addition to this, translation into modern languages remains far from complete: for example, only about 10% of the digitally available Buddhist Chinese corpus has an English translation (Nehrdich et al., 2023), and translation coverage into languages other than English is thinner still.

Recent work has shown that domain-adapted LLMs are effective translators for individual classical languages (Nehrdich et al., 2023, 2025), but systematic evaluation has focused almost exclusively on translation into English. For NLP downstream applications and wider use in translation projects, other directions are also very important: Sanskrit→Tibetan and Chinese→Tibetan translation supports the study of the canonical transmission, and translation from Tibetan into languages such as French, Spanish, or Portuguese serves communities that today access this literature primarily through those languages.

We make the following contributions to this set of challenges:

- MITRA-parallel v2, a dataset of 1.69M automatically aligned records (2.34M segment-level pairs) across Sanskrit–Tibetan, Sanskrit–Chinese, and Chinese–Tibetan,[1] mined with an embedding-based pipeline.
- A domain-adapted LLM for classical Asian languages, built by continued pretraining of Qwen3.5-9B on 22.6B tokens.
- A multi-directional MT evaluation suite: 17 sentence-level directions (33,259 pairs), 10 of which are also evaluated at paragraph level

[1]Throughout the paper, "Chinese" without qualification refers to Buddhist Chinese, the classical Chinese of the Buddhist canon. Modern Chinese occurs only as a target language of the Tibetan→modern suite and is marked as such.

(1,908 blocks of up to 10 consecutive sentences), including cross-classical pairs and Tibetan into seven modern languages.

- A detailed evaluation of MITRA-MT against open LLMs (Qwen3.5 at 9B and 122B, Llama-3.1, Gemma-4 up to 31B), dedicated MT systems (NLLB-200, MADLAD-400, TranslateGemma, Tower-Plus), and commercial references (Gemini Flash family), in chrF, BLEURT, and COMET, with zero-/few-shot arms, and a per-collection Tibetan→English breakdown.

The parallel corpus and evaluation data are available at https://github.com/dharmamitra/mitra-parallel; the fine-tuned translation and embedding models are released in the MITRA Qwen3.5 collection at https://huggingface.co/collections/buddhist-nlp/mitra-qwen35-2026. An interactive gateway to the parallel data is provided at https://dharmamitra.org/db.

2 Related Work

Computational detection of textual parallels in Buddhist literature has been studied for Sanskrit (Hellwig, 2013), Tibetan (Klein et al., 2014), and Buddhist Chinese (Nehrdich, 2020), with cross-lingual work on Tibetan–Chinese (Felbur et al., 2022) and Sanskrit–Tibetan (Nehrdich, 2022). Standard bitext-mining approaches based on multilingual sentence embeddings (Schwenk, 2018; Artetxe and Schwenk, 2019a,b; Feng et al., 2022) assume that cosine similarity separates parallel from non-parallel sentences sufficiently. However, in our setting this assumption often does not hold: Buddhist texts are highly formulaic, with frequent occurrences of highly similar but not exactly matching passages, and genuinely parallel passages often have diverging sentence segmentation across languages and can be buried in clusters of very similar but not directly related material, which makes sentence-based cosine similarity alone a weak signal for sentence-pair mining. Our mining pipeline (Section 3.1) instead relies on the document-level geometry of parallel passages rather than on per-sentence similarity thresholds.

For machine translation of these languages, dedicated encoder–decoder systems cover parts of the space: NLLB-200 (Team et al., 2022) includes Sanskrit and Tibetan, and MADLAD-400 (Kudugunta et al., 2023) covers over 400 languages. Domain-specific neural MT for Buddhist Chinese was presented by Nehrdich et al. (2023). Following the recipe of TOWER (Alves et al., 2024), recent work adapts strong open LLMs to translation by continued pretraining on mixed monolingual and parallel data; Nehrdich and Keutzer (2026) applied this recipe to Gemma-2 (Gemma Team, 2024) for this language family. This paper is a continuation of that study and applies it at larger scale to Qwen3.5-9B, extending evaluation from into-English to a multi-directional protocol.

3 MITRA-parallel

3.1 Mining Pipeline

The pipeline described here is a revision of the passage-mining approach introduced with the first MITRA release (Nehrdich and Keutzer, 2026). That version translated all documents into English with a domain-adapted MADLAD-400 model, retrieved candidate regions on the English side with BGE-M3 embeddings, and aligned sentences with BERTALIGN under a moving-average cosine threshold. For MITRA-parallel v2 we drop the English pivot entirely and operate directly on multilingual embeddings, produced by the domain-specific embedding model `gemma-2 mitra-e` released with the previous pipeline;[2] BERTALIGN is replaced by VECALIGN, and the single cosine threshold by an ensemble of lexical and geometric gates and a supervised aligner screen. It covers the three pairs Sanskrit–Tibetan (sa–bo), Sanskrit–Chinese (sa–zh), and Chinese–Tibetan (zh–bo); we use these codes for the three classical languages throughout the paper. Its basic unit is the *segment*, the sentence-level unit of the source editions, delimited by the available punctuation.

The pipeline exploits the observation that true parallels form long *monotone chains* of matching passages across two works, whereas semantically similar but structurally unrelated material does not. In our data, genuinely parallel translated paragraph pairs and merely topically related ones often have nearly indistinguishable cosine similarity, so per-pair thresholds are uninformative and further geometric signal needs to be taken into account. The pipeline proceeds in five stages.

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

(1) Embedding. All works across Sanskrit, Tibetan, and Chinese are embedded at paragraph-window granularity. Windows span about 300 characters for Tibetan and Sanskrit and 100 characters for Chinese, are advanced by 200 and 70 characters respectively, and are extended to whole segment boundaries, so that consecutive windows overlap. These window sizes follow the average length ratio of matching translations across these languages. The search space is the canonical trilingual working set of 3.22M windows over 12,348 texts detailed in Appendix A. Each language is stored as an L2-normalized matrix in an exact (flat inner-product) FAISS index (Douze et al., 2024).

(2) Anchors. Margin-scored cross-lingual kNN search (Artetxe and Schwenk, 2019a) produces possible anchor pairs. For each source window we retrieve its k=20 nearest target windows and score every candidate with a ratio margin, $\cos(x, y) \,/\, \left(\frac{1}{2}\,\mathrm{bg}(x) + \frac{1}{2}\,\mathrm{bg}(y)\right)$, where the background similarity $\mathrm{bg}(\cdot)$ of a window is its mean cosine to a fixed random sample of 8,192 same-language windows, with the top 1% trimmed so that genuine co-witnesses and near-duplicates do not inflate it. Margin scoring helps to reduce the impact of formulaic "hub" passages that are very frequently encountered in Buddhist literature, since such passages have a high background similarity. All candidates with cosine $\geq$ 0.30 and margin $\geq$ 1.25 are kept as anchors, not only the top-1 neighbor. Against 951 gold-standard parallel pairs, anchor recall is 95.3% at k=20.

(3) Chains. For each candidate work pair with at least eight anchors, we detect the best monotone chain of anchors in the plane of source and target window positions by dynamic programming with a gap penalty (0.04 per window beyond a free gap of six windows) and a slope constraint (0.1–10); up to 40 disjoint chains of at least eight anchors are extracted iteratively, since a work pair may share several separate passages. Each chain is scored with geometric statistics. *Contiguity* is the number of distinct chain positions divided by the span the chain covers, computed on both sides and reduced to the minimum, i.e., a local two-sided coverage of the chain's bounding box (global coverage of the whole work is recorded but not used for acceptance, since it penalizes short passages shared by large collections): a genuine translation is contiguous on both sides, whereas verses quoted inside a commentary are contiguous only on the short side. The *longest-increasing-subsequence z-score* compares the chain length with the $\sim 2\sqrt{m}$ expected for the longest increasing subsequence of m randomly placed anchors in the chain's bounding box, so it measures how far the chain exceeds chance. The *slope-inlier fraction* is the share of chain anchors within a tolerance band of a RANSAC-fitted line (Fischler and Bolles, 1981), measuring a consistent translation rate. The discriminative value of these statistics was established on a labeled set from the Abhidharmakośa corpus, which is well suited for this purpose because it contains the same material in several closely related forms: as positives the known translations (the Tibetan and Chinese kārikā, D4089 and T1560, and the Tibetan and Sanskrit bhāṣya, D4090 and the Abhidharmakośabhāṣya), and as hard negatives the kārikā contained in the bhāṣya in both languages, the bhāṣya against the bare kārikā, and the sibling Abhidharma text Nyāyānusāra (T1562), i.e., pairs that are topically near-identical but not translations of each other. T1562 is an especially useful hard negative: it reproduces significant portions of the bhāṣya but rearranges them and introduces new commentarial sections, so it is not structurally a matching translation, yet it exposes exactly the kind of challenge that leads to false positives in translation matching. Two-sided contiguity and slope separate the two groups completely (positives reach a contiguity of at least 0.64, the negatives at most 0.13, and the negatives show extreme slopes), whereas mean cosine (about 0.5 for both groups) and the z-score alone do not, since a commentary containing its root text also forms a long continuous chain, and cosine similarity is very sensitive to domain and language effects. The hand-set rule derived from this set (contiguity $\geq$ 0.35, $z \geq 8$, inlier fraction $\geq$ 0.85, at least 10 anchors, mean margin $\geq$ 1.15, slope in $[0.15, 6]$, and no contiguity asymmetry above a factor of 1.8) served as the starting point; the acceptance rule used for the release generalizes it with a per-pair random-forest classifier over 21 bounding-box and trajectory features of a chain, trained on chains that contain at least three sentence pairs attested in our bilingual dictionary alignments as positives and chains from works without any attested alignment as negatives, with the threshold set to 95% recall on the positives under grouped cross-validation, and with chains subsumed by a much larger chain of the same source

work (the root-text-in-commentary case) removed. In the v2 run, 360,347 work pairs yielded 323,591 chains, of which 44,085 were accepted; after removing spans whose bounding boxes overlap another accepted span by at least 80% on both sides, 12,318 spans were passed on to sentence alignment.

(4) Sentence alignment. Accepted spans are aligned at sentence level with VECALIGN (Thompson and Koehn, 2019), restricted to the segment range of the span rather than the whole works, using `gemma-2 mitra-e` as the embedding model with alignments of up to four segments on either side. This is followed by cosine- and lexicon-based trimming of span edges, since author or translator colophons, section banners, etc. are usually not part of the actual translation, do not match across languages, and therefore contribute alignment noise. Leading and trailing segments whose best cross-lingual cosine to any segment group on the other side is below 0.30 are removed, as are edge segments for which fewer than 34% of their bilingual-dictionary terms have a translation attested anywhere on the other side. Trimming never removes more than a third of a side or reduces it below three segments. Both thresholds were set by manually inspecting the trimming decisions on development spans.

(5) Ensemble filter and screening. A final ensemble filter combines three gates, each manually calibrated on genuine and spurious spans. (i) *Bilingual-dictionary diagonality*: for every dictionary term of the source side, we test whether one of its translations occurs in the paired target segment (on the diagonal) or only elsewhere in the span (off the diagonal); the rarity-weighted on-diagonal share must be at least 0.40. (ii) *Alignment density*: at least 45% of the segments of a span must take part in a non-empty alignment, which removes partial or verse-only overlaps and sibling texts. (iii) A *length-ratio gate*: the fraction of aligned pairs whose character-length ratio deviates by more than 2.5 median absolute deviations from the pair-specific ideal ratio must stay below 0.50 (sa–bo, sa–zh) or 0.25 (zh–bo), which removes structurally parallel but untranslated material. Alignments already covered by a larger overlapping span are removed. We further flag texts with low average cosine scores across the alignment records and manually inspect them to remove spurious text pairs where alignment was not successful. It needs to be noted that while we manually inspected low-scoring instances, no systematic human audit has yet been performed for this corpus.

Pair	Records	Seg. pairs	Median chars
sa–bo	555,558	1,005,439	sa 49 / bo 96
zh–bo	836,559	836,559	zh 22 / bo 47
sa–zh	301,613	496,402	sa 48 / zh 20
Total	1,693,730	2,338,400	

Table 1: MITRA-parallel v2 after decontamination and exact deduplication. Each record aligns one source segment to one or more target segments; segment pairs count individual segment-to-segment links (in the sa→bo, zh→bo, sa→zh directions); for zh–bo every record is a single segment pair. Median segment lengths in characters per language side; segmentation granularity differs between pairs, reflecting the source editions.

3.2 Dataset

Table 1 summarizes the released corpus, which covers the three classical pairs Sanskrit–Tibetan, Sanskrit–Chinese, and Chinese–Tibetan. The released files are decontaminated against the evaluation suite of Section 3.3 (normalized substring matching, following the same protocol as the training-data decontamination in Section 4) and exact-deduplicated.

3.3 Evaluation Data

As part of this work we release the held-out evaluation suite used in Section 5. Table 2 shows the composition with 33,259 sentence pairs and 1,908 paragraph blocks over the 17 directions reported below. It contains (i) randomly sampled sentence-level test sets of 2,000 pairs per direction for Sanskrit↔Tibetan, Tibetan↔Buddhist Chinese, Sanskrit↔Buddhist Chinese (all sourced from the manually aligned multilingual editions of the Thesaurus Literaturae Buddhicae (Braarvig, 2007–2022), part of the Bibliotheca Polyglotta of the Norwegian Institute of Philology and the University of Oslo), and 2,594 pairs per direction for Sanskrit↔English (drawn from the Mitrasamgraha dataset (Nehrdich et al., 2026), whose test split is divided into disjoint halves so that the two directions share no pair); (ii) the published MITRA-zh-eval benchmark (Nehrdich et al., 2025) for Buddhist Chinese→English (2,662 pairs drawn from 32 canonical texts); (iii) a Tibetan→X suite based on translations of Lotsawa House,[3] covering trans-

[3] https://www.lotsawahouse.org

Direction	Sent.	Par.	Source
Classical ↔ classical			
sa→bo	2,000	—	TLB
bo→sa	2,000	—	TLB
sa→zh	2,000	—	TLB
zh→sa	2,000	—	TLB
bo→zh (Budd.)	2,000	—	TLB
zh→bo	2,000	—	TLB
Classical ↔ English			
sa→en	2,594	321	Mitrasaṃgraha
en→sa	2,594	—	Mitrasaṃgraha
zh→en	2,662	244	MITRA-zh-eval
bo→en	3,979	400	4 bo–en collections
Tibetan → modern			
bo→fr	2,000	200	Lotsawa House
bo→es	2,000	200	Lotsawa House
bo→de	2,000	200	Lotsawa House
bo→zh (Mod.)	1,660	166	Lotsawa House
bo→pt	950	95	Lotsawa House
bo→it	720	72	Lotsawa House
bo→nl	100	10	Lotsawa House
Total	**33,259**	**1,908**	

Table 2: Composition of the evaluation suite: sentence pairs and paragraph blocks (up to 10 consecutive sentences) actually scored per direction. TLB = Thesaurus Literaturae Buddhicae; MITRA-zh-eval = Nehrdich et al. (2025); Mitrasaṃgraha = Nehrdich et al. (2026), split into disjoint halves for the two directions; the bo→en row pools the four collections of Figure 3 (84000, Khyentse Vision Project, Lotsawa House, Tsadra Foundation).

lation into French, Spanish, German, Portuguese, Italian, Dutch, and Modern Chinese, sized per language by the amount of human-aligned material available, from 2,000 pairs for French down to 100 for Dutch; and (iv) a Tibetan→English suite of 3,979 pairs drawn as 100 blocks of ∼10 consecutive sentences from each of four translation projects (84000,[4] Khyentse Vision Project,[5] Lotsawa House, Tsadra Foundation[6]), enabling a per-collection difficulty breakdown on sentence and paragraph level. Directions with document-contiguous references are additionally evaluated at paragraph level on blocks of up to 10 consecutive sentences (Par. column). A number of further directions whose reference translations cannot be redistributed for licensing reasons were evaluated under the identical protocol but are excluded from the released suite and from the main translation results; six of them are used in the judge study of Section 5.2.

[4] https://84000.co
[5] https://khyentsevision.org
[6] https://tsadra.org

4 Domain-Adaptive Pretraining

4.1 Data Composition

Table 3 gives the composition of the continued-pretraining corpus with a total of 22.56B tokens counted with the Qwen3.5 tokenizer. The three dominant components are monolingual Tibetan (25.6%; Tibetan Unicode text from high-quality manual input as well as OCR-derived and quality-filtered material, sourced among others via the Asian Classics Input Project and Buddhist Digital Resource Center), monolingual Sanskrit (22.1%; manually typed e-texts and quality-gated OCR sources, the latter drawn from the Sangraha corpus of IndicLLMSuite (Khan et al., 2024), in IAST), and domain-specific English (20.8%; OCR-processed academic literature on Buddhist Studies and published translations). Buddhist Chinese (4.7%) is drawn from the CBETA[7] and Kanripo[8] collections. We include a 13.8% multilingual replay component to counteract catastrophic forgetting of general multilingual ability: web text drawn from FineWeb (Penedo et al., 2024) for English (the 10BT sample) and from FineWeb-2 (Penedo et al., 2025) for 16 further languages,[9] with per-language token budgets between 80M and 650M tokens. Parallel data (6.9%) combines the mined MITRA-parallel v2 corpus of Section 3 with human-translation pair collections from translation projects and publishers, used for training under their respective terms and not redistributed. We additionally include the Digital Corpus of Sanskrit (Hellwig, 2010–2021) in CoNLL-U form (2.7%), providing sandhi-split, lemmatized, and morphologically annotated Sanskrit. All Tibetan is in Tibetan Unicode script; Wylie transliteration was normalized out of the corpus. For Sanskrit and for the small Pāli component (about 0.1B tokens of monolingual and Pāli–Sanskrit parallel text; Pāli is used in pretraining only and is part neither of MITRA-parallel v2 nor of the released evaluation suite), we decided to use IAST transliteration as it gives better tokenizer efficiency at comparable baseline performance on the Qwen3.5 model family. For Chinese, we use traditional characters in Unicode.

[7] https://www.cbeta.org
[8] https://www.kanripo.org
[9] Arabic, Chinese (300M tokens traditional, 150M simplified), Dutch, French, German, Hindi, Indonesian, Italian, Japanese, Korean, Portuguese, Russian, Spanish, Thai, Turkish, and Vietnamese.

Component	Tokens	Share
Domain monolingual		
Tibetan	5.77B	25.6%
Sanskrit	4.98B	22.1%
English (Buddhist studies)	4.69B	20.8%
Buddhist Chinese	1.06B	4.7%
Japanese	0.27B	1.2%
Other languages	0.51B	2.2%
Parallel		
Parallel data (all pairs)	1.55B	6.9%
Annotated		
Sanskrit treebank (DCS)	0.61B	2.7%
General multilingual		
Replay mix (17 languages)	3.11B	13.8%
Total (stage one)	22.56B	100%
Stage two (16k context)		
Bidirectional parallel data	≈3.0B	46%
Stage-one replay (en upsampled)	≈2.4B	37%
Instruction data (Tülu 3, Aya)	≈0.6B	9%
Document-level parallel (added)	0.5B	8%
Total (stage two)	6.4B	100%

Table 3: Composition of the continued-pretraining corpus, grouped by data type. The upper block is the stage-one corpus; the lower block gives the second, translation-oriented stage (Section 4), whose base mixture was specified as 50/40/10 over 5.9B tokens before 0.5B tokens of document-level parallel text were added; stage-two shares are given over the 6.4B tokens actually trained. Token counts are exact counts under the Qwen3.5 tokenizer (vocabulary 248k); DCS = Digital Corpus of Sanskrit in CoNLL-U form; stage-one proportions are natural, i.e., no component is up- or down-sampled, whereas the stage-two mixture is imposed by resampling (with English in-domain material upsampled). The replay mix counteracts forgetting of general multilingual ability.

4.2 Continued Pretraining

We continue pretraining from Qwen3.5-9B-Base (Qwen Team, 2026), an 8.95B-parameter hybrid-attention model. We chose this model after comparing the zero-shot translation quality of the instruction-tuned Qwen3.5 and Gemma-4 releases on our evaluation suite (Appendix B): at comparable scale, Qwen3.5-9B and Gemma-4-12B perform on par (mean chrF 19.9 vs. 19.9 over the official directions), while Qwen3.5 has a lower parameter count and is therefore more efficient to train given the available compute budget. Documents are shuffled at file level (fixed seed), concatenated with EOS separators, and packed into fixed 8,192-token windows. No component is up- or down-sampled. Training uses 8×H100 with ZeRO-2, bf16, micro-batch 4 with 4-step gradient accumulation (1.05M tokens per optimizer step), peak learning rate 2×10^{-5} with cosine decay, 1% warmup, and weight decay 0.1. The checkpoint used in this paper has seen 27,000 steps ≈ 28.3B tokens (≈1.3 epochs).

A second, shorter pretraining stage (lower block of Table 3) extends the context window to 16,384 tokens and shifts the data distribution toward translation and instruction following. The corpus is repacked into a mixture of 50% bidirectional parallel data, 40% replay of the stage-one distribution with upsampled English in-domain material, and 10% general multilingual instruction data consisting of a uniform random subsample of the Tülu 3 SFT mixture (Lambert et al., 2024) (∼0.6B tokens) together with the Aya dataset (Singh et al., 2024) restricted to our 21 deployment languages (∼19M tokens), rendered as chat-template conversations, supplemented with 0.5B tokens of document-level parallel text created by merging adjacent sentence pairs from the main parallel corpus. This stage runs for 6,145 steps (6.4B tokens, one epoch) at peak learning rate 1×10^{-5} with 5% warmup on the same 8×H100 ZeRO-2 setup; since the sequence length doubles, the micro-batch is reduced to 1 and gradient accumulation raised to 8, so that each optimizer step still sees 64 sequences (1.05M tokens).

4.3 Translation Fine-tuning

Two lightweight steps turn the pretrained model into the translation model MITRA-MT. First, an *interface calibration* fine-tune: 50 optimizer steps (approximately one epoch) on 6,315 plain-prompt translation instructions distilled from Gemini 3 Flash, spanning 15 target languages and using exactly the instruction format of our evaluation protocol (Section 5). In our experiments this stage functions as calibration rather than knowledge injection: translation capability tracks the amount of continued pretraining, while extending instruction tuning beyond ∼50 steps on this dataset yields no measurable gain. We report this checkpoint as the ablation MITRA-SFT.

Second, 100 steps of GRPO (Shao et al., 2024) in the memory-efficient Dr. GRPO variant (Liu et al., 2025): LoRA (rank 16, α=32) at learning rate 5×10^{-6}, KL coefficient 0.01, 8 sampled completions per prompt, on a pool of 8,173 held-out translation prompts in the same plain format. The reward is an unnormalized weighted sum of five terms: chrF against the reference (weight 0.2),

BLEURT-20 (0.4), an LLM-judge adequacy score with Gemini 3.1 Flash-Lite as the judge (0.4), a non-repetition term (0.2), and a target-script check (0.2); the weights sum to 1.4 and were not rescaled. The effect of this stage on judged adequacy and on chrF is quantified in Section 5.2. The GRPO checkpoint is the MITRA-MT evaluated below.

Decontamination. Since pretraining sources (canonical collections) and evaluation sets draw on overlapping literature, we decontaminate training data against evaluation data by string matching: every evaluation sentence pair is normalized (NFC, whitespace-collapsed, lower-cased) and matched as a substring against all parallel training data using an Aho–Corasick automaton, with per-script minimum-length rules so that ubiquitous formulae do not trigger mass deletion (e.g., CJK $\geq$15 characters, Tibetan $\geq$40, Latin/IAST $\geq$25 characters and 5 words). This removes 112,382 training lines. It needs to be emphasized that deduplication of Buddhist digital text collections is very difficult as the frequent repetition of nearly identical passages is a core feature of parts of this literature. We therefore do not deduplicate against the monolingual training data, as this would remove large parts of the corpus.

5 Machine Translation Evaluation

5.1 Setup

Protocol. All systems translate the same inputs at temperature 0.2 with thinking/reasoning modes disabled (LLMs; Gemini 3.6 Flash runs at its minimum thinking budget of 128 tokens, since the API rejects full disabling for this model) or with beam search (dedicated MT models). Instruction-tuned LLMs—including MITRA-MT, with no system-specific prompt engineering—receive the identical instruction "*Translate the following {source} text into {target}. Output only the translation.*" through their chat template; NLLB-200 receives source/target language codes and MADLAD-400 target-language tags, with Sanskrit transliterated IAST→Devanāgarī on input. Because systems differ in output script (Devanāgarī vs. IAST for Sanskrit, Wylie vs. Tibetan script for Tibetan), all outputs for Sanskrit and Tibetan targets are script-normalized (Devanāgarī→IAST, Wylie→Tibetan Unicode) before scoring, for every system identically. A five-shot arm supplies Gemma-4-31B with five in-context exemplar pairs drawn from decontaminated training-side data. Paragraph-level inputs are blocks of up to 10 consecutive sentences translated as a single input.

Metrics. Our primary metric is chrF (Popović, 2015), a deliberate choice driven by the multi-directional setting: it is the only established metric that applies uniformly across all 17 directions, requiring no tokenizer, no language-specific model support, and no assumptions about script or word segmentation. In this we follow Team et al. (2022), who adopt character-level chrF as the primary automatic metric for NLLB-200 in a massively multilingual setting. The learned metrics BLEURT-20 (Sellam et al., 2020) and COMET-22 (Rei et al., 2022) lack domain adaptation for classical Buddhist material and language support for our targets (COMET's encoder does not cover Tibetan and compresses, or even inverts, large chrF differences on Tibetan-target directions; see Appendix C). LLM-as-judge metrics are potentially powerful for this setting as they show strong correlation with human judgment on Sanskrit→English translation (Nehrdich et al., 2026), but their correlation with human judgments on non-English classical language targets is not yet studied and therefore uncertain. Section 5.2 reports a controlled two-judge experiment that motivates this choice and quantifies where judges and chrF disagree.

Baselines. We compare against three groups of systems. *Open instruction-tuned LLMs*: Qwen3.5-9B, Llama-3.1-8B (Grattafiori et al., 2024), Gemma-4-12B/31B (Gemma Team, 2026), and Qwen3.5-122B-A10B (AWQ-quantized mixture-of-experts). *Dedicated MT systems*: NLLB-200-3.3B (Team et al., 2022), MADLAD-400-3B-MT (Kudugunta et al., 2023), TranslateGemma-27B (Finkelstein et al., 2026), and the translation-specialized LLM Tower-Plus-9B (Rei et al., 2025). *Commercial references*: Gemini 3.1 Flash-Lite, 3.5 Flash, and 3.6 Flash. Tables abbreviate these systems as Qwen3.5, Llama-8B, G4-12B, G4-31B (G4-31B-5s for the five-shot arm), Q-122B, NLLB, MADLAD, TG-27B, Tower+, Flash-Lite, 3.5-Flash, and 3.6-Flash; MITRA-SFT denotes our model before GRPO. API costs confine our commercial comparison to a single vendor family. Main-text tables report the most recent model in the family, 3.6 Flash, together with 3.1 Flash-Lite; complete results are given in Appendix C. We do not include the Gemma-2-based model of Nehrdich

and Keutzer (2026), as the deduplication of the evaluation data used in this paper with the training data of that model cannot be established.

5.2 Metric Choice and LLM Judges

To check how the choice of chrF as the main metric influences the results compared to LLM judges, we ran a small selection study. For each direction we drew a fixed 50-segment sample from the sources common to all systems and scored every system on those identical items with two judges from two different model families, Gemini 3.1 Flash-Lite and Claude Sonnet 5, using a reference-based GEMBA-style prompt (Kocmi and Federmann, 2023): 23,670 judgments per judge over 21 sentence-level and 12 paragraph-level directions.[10] The scale is deliberately limited, and the results should be read as indicative rather than as a full evaluation.

Table 4 reports the outcome. Agreement with chrF is strong at the system level (Spearman $\rho = 0.93$ for the Gemini judge and 0.96 for Sonnet), and the two judges agree even more closely with each other ($\rho = 0.99$). The disagreement that remains is concentrated at the top of the table: both judges place Gemini 3.5 and 3.6 Flash above MITRA-MT on the aggregate, where chrF places MITRA-MT first (chrF 34.4 vs. 32.5; Gemini judge 80.7 vs. 83.7; Sonnet judge 73.5 vs. 75.3). Because an independent judge reproduces the ranking of the Gemini judge almost exactly, this is not attributable to judge–system family affinity. The effect is also direction-dependent: both judges favor MITRA-MT into Tibetan, Chinese, Korean and Japanese (the latter two being supplementary directions), and favor the commercial systems into Sanskrit and into several European targets, with English→Sanskrit the largest single deficit. Restricting the comparison to English targets (Table 7, Appendix D), where reference-based judging is best established and the judges' own language coverage is strongest, reproduces the same picture on a narrower base: our models lead on chrF by roughly 3.5 points while both judges place 3.5 and 3.6 Flash ahead by around 3. While we adopt chrF as the main metric of our study on cost and applicability grounds, in this limited trial chrF appears to undervalue the Gemini family, and MITRA-MT's aggregate chrF lead over the commercial references should be read as metric-specific and not as a uniform quality claim.

The judge study also quantifies the effect of the preference-optimization stage of Section 4.3. On its held-out validation set, judge-rated adequacy improved from 81.1 to 82.9 and chrF from 39.0 to 39.2, and the same asymmetry appears on the evaluation suite proper: against MITRA-SFT, MITRA-MT gains $+1.3$ GEMBA points at sentence level and $+1.1$ at paragraph level under the Gemini judge (95% CI $+0.5$ to $+2.1$ and $+0.3$ to $+1.9$; better on 18 of 22 and 10 of 13 directions respectively), and $+1.5$ at paragraph level under the independent Sonnet judge, while chrF moves by $+0.03$ and $+0.41$. As the five strongest systems in the study are separated by only 4.4 GEMBA points in total, the preference stage recovers roughly a third of the range spanning MITRA-SFT and the best commercial system. The gain is consistent in judged adequacy and almost entirely invisible to string metrics, which we see as a strong hint that chrF, chosen here for cost and coverage, is insensitive to the kind of improvement our preference optimization targets.

5.3 Sentence-Level Results

Figure 1 presents sentence-level chrF for all 17 directions, where bo→en is the pooled four-collection Tibetan→English benchmark; exact values for every system are given in Appendix C. Three patterns organize the results. First, into classical target languages the gap is structural: MITRA-MT leads every open baseline by wide margins into Tibetan and Sanskrit (sa→bo 55.0 vs. 33.4 for the strongest open baseline, Qwen3.5-122B; zh→sa 39.1 vs. 27.3)—capabilities that neither scale nor instruction tuning provides. Second, against the commercial references MITRA-MT leads Gemini Flash-Lite on 16 of the 17 directions, including sa→en (38.5 vs. 36.6) and zh→en (38.7 vs. 37.3), and leads the stronger Gemini 3.6 Flash on 13 of the 17 directions with a tie on sa→en (38.5 vs. 38.5); the remaining deficits are at most 3.3

[10] The judged directions are the 17 directions of Section 3.3 without bo→en, plus six *supplementary* directions that were evaluated under the identical protocol but whose reference translations cannot be redistributed (sa→ja, sa→de, sa→hi, zh→fr, zh→ko, pa→en); these supplementary directions appear in this paper only within the judge study. The counts 21 and 12 are the directions covered by every system (NLLB-200 has no pa→en run, and paragraph-level zh→en is available for nine systems); the pairwise MITRA-MT vs. MITRA-SFT comparison in Section 4.3 uses all 22 and 13 directions that the two models share. The 23,670 judgments are the sum over systems and directions of the items actually judged; bo→nl, for instance, has only 10 paragraph blocks.

	Sentence			Paragraph		
System	chrF	Gem.	Son.	chrF	Gem.	Son.
MITRA-MT	34.4	80.7	73.5	46.0	89.3	82.4
MITRA-SFT	34.4	79.3	72.7	45.6	87.6	80.1
3.5-Flash	32.5	83.7	75.3	44.9	92.5	86.0
3.6-Flash	31.9	83.4	75.1	43.2	91.5	83.2
Flash-Lite	30.5	82.0	71.7	42.6	91.9	83.4
G4-31B-5s	26.6	65.8	56.2	—	—	—
G4-31B	25.8	65.1	54.3	39.0	82.3	68.6
Q-122B	25.5	68.3	59.0	37.4	81.9	67.6
TG-27B	21.6	48.7	35.6	—	—	—
G4-12B	20.5	44.5	33.5	33.8	58.0	39.3
Qwen3.5	20.2	44.8	32.6	32.9	53.7	34.5
Llama-8B	15.9	28.4	17.0	22.0	22.5	12.5
Tower+	15.5	26.4	16.2	24.8	27.8	15.5
NLLB	12.7	15.4	8.6	—	—	—
MADLAD	9.4	13.4	7.2	—	—	—

Table 4: LLM-judge selection study: chrF against two reference-based GEMBA-style judges (Gem. = Gemini 3.1 Flash-Lite, Son. = Claude Sonnet 5) on a fixed 50-segment sample per direction, identical items for every system (21 sentence-level and 12 paragraph-level directions). Rows are ordered by sentence-level chrF. The judges agree with each other ($\rho = 0.99$) more closely than either agrees with chrF ($\rho = 0.93, 0.96$); the ordering they disagree on is confined to the top of the table. Table 7 reports the same study restricted to English targets.

chrF except for en→sa, where both Gemini systems lead by up to 5.1 chrF (33.3 vs. 38.3). Third, on Tibetan→modern-language directions (Lotsawa House), MITRA-MT leads consistently, with the dedicated MT systems far behind.

5.4 Paragraph-Level Results

Figure 2 reports paragraph-level results for the directions with document-contiguous data, created by merging up to 10 consecutive source and target sentences on both sides. For the Tibetan→modern directions and bo→en, the paragraph blocks are exactly the sentence-level test items regrouped, so the two granularities score the same references and differ only in input context; for zh→en the blocks cover 2,440 of the 2,662 sentence-level items. For sa→en, by contrast, the 321 blocks are a separate contiguous sample from the Mitrasamgraha pool and share no item with the sentence-level set, so its sentence–paragraph difference reflects the test material as well as context. Paragraph-level scores exceed sentence-level scores for the three evaluated systems MITRA-MT, Gemini 3.6 Flash and Gemma-4 31B by 8–12 chrF, which shows that context is immensely helpful for translating classical Buddhist texts into modern languages.

5.5 Tibetan in Detail

Figure 3 breaks Tibetan→English down by source collection. The ordering is stable across all systems: translations of canonical sūtra material (84000) score highest, material from the Khyentse Vision Project and Lotsawa House sits in between, and the translations sponsored by the Tsadra Foundation are hardest (∼9 chrF below 84000 for MITRA-MT). Pooled over the four collections, MITRA-MT reaches 43.2 chrF, ahead of every baseline including Gemini 3.6 Flash (40.7) and Gemma-4-31B (36.1).

6 Conclusion

We presented a pivot-free mining pipeline and the resulting 1.69M-record trilingual MITRA-parallel v2 corpus; MITRA-MT, a domain-adapted translation model built with a fully documented 22.6B-token continued-pretraining recipe on Qwen3.5-9B and a lightweight translation fine-tune step; and a 17-direction evaluation suite that extends assessment of classical Asian language MT beyond translation into English, including cross-classical directions and Tibetan into seven modern languages. For the chrF metric, MITRA-MT sets the strongest open-model results on most evaluated directions, while LLM-judge-based evaluation indicates that MITRA-MT, although it does not outperform commercial models, closes much of the gap and is therefore a viable open alternative. Corpus, evaluation data, and model weights are released openly via the MITRA Qwen3.5 collection on Hugging Face, with an interactive gateway to the parallel data at `https://dharmamitra.org/db`.

7 Limitations

Corpus quality assessment. The alignment quality of MITRA-parallel v2 is validated by automatic gates, but no manual evaluation of the released alignments was possible within the scope of this work: the corpus has undergone neither a stratified human audit with multiple annotators nor any record-level accuracy assessment, and the manual inspection we did perform was targeted at low-scoring texts rather than drawn as a representative sample. Expert annotators for Sanskrit–Tibetan–Chinese alignment are scarce, and the scale of the corpus makes even a modest stratified sample a substantial annotation effort. Partial and indirect reassurance comes from downstream training. All three cross-classical pairs are learned to

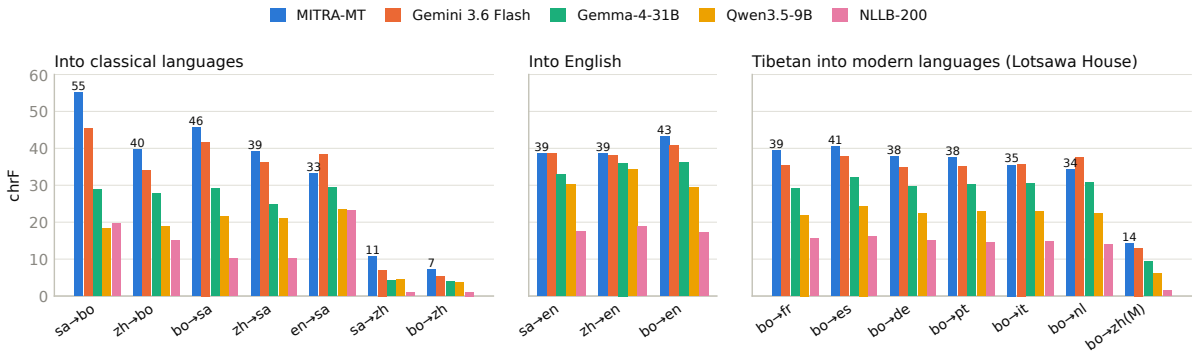

Figure 1: Sentence-level translation quality (chrF) by direction, grouped by target family. MITRA-MT numbers are labeled. bo→en pools the four-collection Tibetan→English benchmark. All systems evaluated under the identical protocol, including output script normalization.

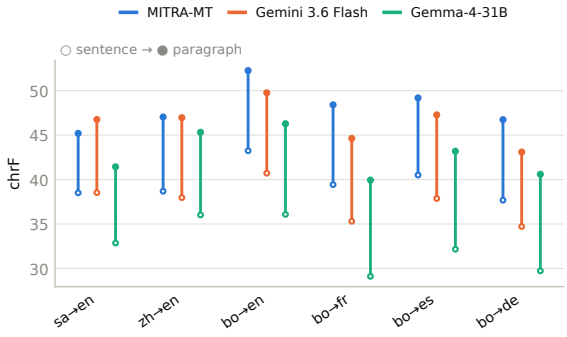

Figure 2: Sentence-level (open marker) vs. paragraph-level (filled marker) chrF. Context helps all LLM systems; the ordering is largely preserved.

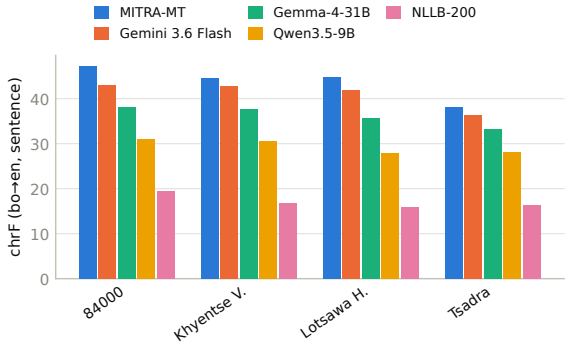

Figure 3: Tibetan→English difficulty by source collection (sentence level). The ordering is stable across systems.

a level that no system trained without this data approaches. Against the strongest open baseline (Qwen3.5-122B, a model an order of magnitude larger), MITRA-MT reaches 55.0 vs. 33.4 chrF on sa→bo, 45.7 vs. 29.5 on bo→sa, 39.6 vs. 29.2 on zh→bo, and 39.1 vs. 27.3 on zh→sa. If the alignment quality had significant flaws, such consistent gains would not be possible. A direct quality estimate nevertheless remains open for future work.

Metrics. Our reliance on chrF is another methodological limitation of this study. Character n-gram overlap is the only metric that applies uniformly across all our directions, but it measures surface form rather than adequacy, under-measures quality for Tibetan and Chinese targets whose writing systems are fundamentally different, and is demonstrably insensitive to differences that LLM judges do register. The preference-optimization stage of Section 4.3 moves judged adequacy while leaving chrF essentially unchanged, and in Section 5.2 chrF and both judges disagree on the ordering at the top of the table. LLM-as-judge is the most promising alternative and the one we would prefer, but it fails on two independent counts. First, its cost quickly becomes prohibitive for an evaluation dataset of this size when comparing even a modest number of systems. Second, its validity beyond English targets is unestablished: agreement with expert human judgment has been shown for Sanskrit→English (Nehrdich et al., 2026), but for Tibetan, Buddhist Chinese, and the modern non-English targets no calibration against human annotation exists.

Acknowledgments

Generative AI coding assistants, including Claude Code using the 4.8 Opus and 5 Fable models, were used during software development for code generation, debugging, refactoring, and development of experimental and visualization scripts. All generated code and experimental results were reviewed and validated by the authors. We thank the Tsadra Foundation for their generous support in making

this study possible.

References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Jens Braarvig. 2007–2022. Thesaurus literaturae buddhicae (TLB). Bibliotheca Polyglotta, Norwegian Institute of Philology / University of Oslo. https://www2.hf.uio.no/polyglotta/index.php?page=library&bid=2.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. arXiv preprint arXiv:2401.08281.

Rafal Felbur, Marieke Meelen, and Paul Vierthaler. 2022. Crosslinguistic semantic textual similarity of buddhist chinese and classical tibetan. *Journal of Open Humanities Data*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, Markus Freitag, and David Vilar. 2026. TranslateGemma technical report. arXiv preprint arXiv:2601.09012.

Martin A. Fischler and Robert C. Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395.

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

Gemma Team. 2026. Gemma 4 technical report. arXiv preprint arXiv:2607.02770.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. The llama 3 herd of models.

Oliver Hellwig. 2010–2021. DCS – the digital corpus of sanskrit.

Oliver Hellwig. 2013. Googling the Rishi - Graph Based Analysis of Parallel Passages in Sanskrit Literature. In *Recent Researches in Sanskrit Computational Linguistics: Fifth International Symposium IIT Mumbai, India, January 2013 Proceedings*.

Mohammed Safi Ur Rahman Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Suriyaprasaad B, Varun G, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, and Mitesh M. Khapra. 2024. IndicLLMSuite: A blueprint for creating pre-training and fine-tuning datasets for Indian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15831–15879, Bangkok, Thailand. Association for Computational Linguistics.

Benjamin Eliot Klein, Nachum Dershowitz, Wolf Lior, Orna Almogi, and Dorji Wangchuk. 2014. Finding inexact quotations within a Tibetan Buddhist corpus. In *Digital Humanities 2014 Conference Abstracts*, pages 486–488.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203.

Sneha Kudugunta, Isaac Rayburn Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2024. Tülu 3: Pushing frontiers in open language model post-training. arXiv preprint arXiv:2411.15124.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. Understanding R1-Zero-like training: A critical perspective. arXiv preprint arXiv:2503.20783.

Sebastian Nehrdich. 2020. A method for the calculation of parallel passages for buddhist chinese sources

based on million-scale nearest neighbor search. *Journal of the Japanese Association for Digital Humanities*, 5:132–153.

Sebastian Nehrdich. 2022. SansTib, a Sanskrit - Tibetan parallel corpus and bilingual sentence embedding model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6728–6734, Marseille, France. European Language Resources Association.

Sebastian Nehrdich, 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. arXiv preprint arXiv:2601.07314.

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.

Sebastian Nehrdich, Avery Chen, Marcus Bingenheimer, Lu Huang, Rouying Tang, Xiang Wei, Leijie Zhu, and Kurt Keutzer. 2025. MITRA-zh-eval: Using a buddhist Chinese language evaluation dataset to assess machine translation and evaluation metrics. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 129–137, Albuquerque, USA. Association for Computational Linguistics.

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. arXiv preprint arXiv:2601.06400.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems 37 (Datasets and Benchmarks Track)*.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. FineWeb2: One pipeline to scale them all – adapting pre-training data processing to every language. arXiv preprint arXiv:2506.20920.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

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

Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual LLMs. arXiv preprint arXiv:2506.17080.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau

Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved Sentence Alignment in Linear Time and Space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

A Corpus Composition of the Mining Search Space

The span miner of Section 3.1 operates on the corpus collections curated for the DharmaNexus intertextuality system. All texts are segmented into overlapping paragraph windows of consecutive canonical segments and embedded with `gemma-2-mitra-e`. Pair mining for MITRA-parallel v2 searches the canonical trilingual working set, comprising 3.22M paragraph windows over 12,348 texts:

Tibetan (1.91M windows; 5,013 texts). The Derge Kangyur (1,158 texts; 604k windows) and Derge Tengyur (3,431 texts; 1.25M windows) in the Esukhia eKangyur/eTengyur input transcriptions, complemented by 424 further canonical texts catalogued under BDRC scan identifiers (53k windows).

Chinese (936k windows; 6,623 texts). The Taishō Tripiṭaka in the CBETA edition: 5,749 texts of the first Taishō division (740k windows) and 874 texts of the second division (196k windows).

Sanskrit (371k windows; 712 texts). Buddhist scriptures (251 texts; 159k windows) and Buddhist treatises (461 texts; 212k windows), drawn from GRETIL, the Digital Sanskrit Buddhist Canon, and editions digitized within the Dharmamitra project.

B Base-Model Choice

Table 5 compares the zero-shot performance of the instruction-tuned Qwen3.5-9B against the Gemma-3 and Gemma-4 releases on the official evaluation directions, under the protocol of Section 5. At comparable parameter count (9B vs. 12B) no model clearly dominates. While Gemma-4-31B leads most directions, it is 3.4× the size of our base, making finetuning and deployment significantly more challenging.

Direction	Qwen-9B	G3-12B	G4-12B	G4-31B
sa→bo	18.3	26.8	19.6	**28.9**
bo→sa	21.5	21.4	21.3	**29.1**
bo→zh (B)	3.5	3.0	2.9	**4.0**
zh→bo	18.9	26.3	20.4	**27.8**
sa→en	30.3	28.3	28.8	**32.9**
en→sa	23.5	25.1	27.6	**29.4**
sa→zh	**4.6**	4.4	3.0	4.2
zh→sa	21.0	19.5	19.7	**24.7**
zh→en	34.4	32.3	32.7	**36.0**
bo→fr	21.9	21.9	22.1	**29.1**
bo→es	24.4	24.1	25.0	**32.2**
bo→de	22.3	21.4	21.9	**29.7**
bo→pt	22.9	22.4	22.6	**30.3**
bo→it	22.8	22.8	22.5	**30.6**
bo→nl	22.5	21.1	23.3	**30.8**
bo→zh (M)	6.1	5.1	5.6	**9.2**
Mean	19.9	20.4	19.9	**25.6**

Table 5: Zero-shot chrF of instruction-tuned Qwen3.5-9B (Qwen-9B), Gemma-3-12B (G3-12B), and Gemma-4-12B/31B (G4-12B/31B) on 16 of the 17 evaluation directions (bo→en is not included), identical protocol as all other systems (mean over directions available for all listed systems). At comparable scale no model dominates; Gemma-4-31B leads most directions at 3.4× the size of our base.

C All-Systems Overview and Neural-Metric Reference

Figure 4 gives the complete sentence-level chrF picture for every evaluated system; Table 6 provides COMET-22 and BLEURT-20 for the strongest systems.

D LLM-Judge Study on English Targets

Table 7 repeats the judge study of Section 5.2 restricted to the English-target directions, where reference-based judging is best established and the judges' own language coverage is strongest. The ordering disagreement between chrF and the two judges observed on the full direction set reappears on this narrower base.

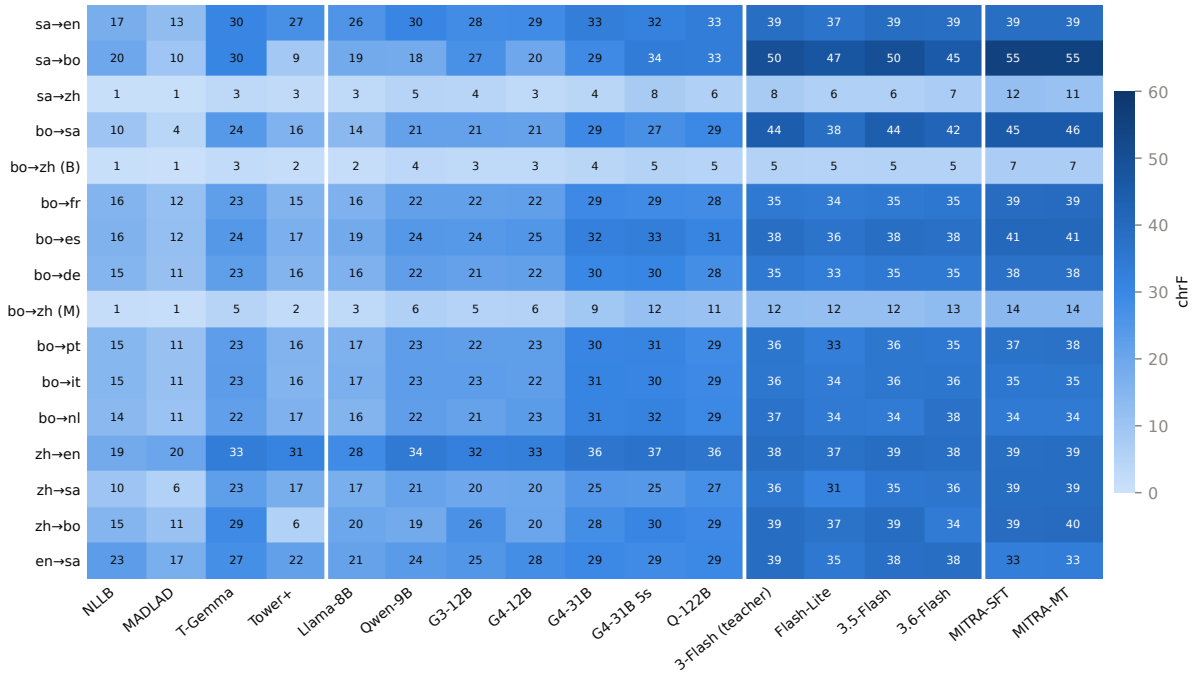

Figure 4: Sentence-level chrF for all systems (columns, grouped left to right: dedicated MT, open LLMs, commercial references, ours) across 16 of the 17 directions (rows; bo→en, scored per collection, is not included). Column abbreviations: NLLB = NLLB-200-3.3B; MADLAD = MADLAD-400-3B-MT; T-Gemma = TranslateGemma-27B; Tower+ = Tower-Plus-9B; Llama-8B = Llama-3.1-8B-Instruct; Qwen-9B = Qwen3.5-9B; G3-12B = Gemma-3-12B; G4-12B/31B = Gemma-4-12B/31B; G4-31B 5s = Gemma-4-31B with five in-context exemplars; Q-122B = Qwen3.5-122B-A10B (AWQ); 3-Flash (teacher) = Gemini 3 Flash Preview, the SFT distillation teacher; Flash-Lite / 3.5-Flash / 3.6-Flash = Gemini 3.1 Flash-Lite / 3.5 Flash / 3.6 Flash; MITRA-SFT = our model without GRPO; MITRA-MT = our full model. Row labels: (B) = Buddhist Chinese, (M) = Modern Chinese; the Tibetan→modern rows use Lotsawa House data.

	COMET-22					BLEURT-20				
Direction	MITRA-MT	Flash-Lite	3.5-Flash	3.6-Flash	G4-31B	MITRA-MT	Flash-Lite	3.5-Flash	3.6-Flash	G4-31B
sa→bo	0.90	**0.91**	0.90	0.86	0.81	**0.66**	0.61	0.62	0.60	0.47
zh→bo	**0.87**	0.87	0.85	0.81	0.78	**0.57**	0.54	0.54	0.52	0.46
bo→sa	**0.81**	0.78	0.80	0.78	0.77	**0.75**	0.72	0.74	0.73	0.68
zh→sa	**0.80**	0.76	0.80	0.79	0.77	**0.69**	0.67	0.69	0.69	0.65
en→sa	0.77	0.77	**0.78**	0.78	0.76	0.69	0.69	0.69	**0.70**	0.66
sa→zh	**0.65**	0.62	0.60	0.62	0.57	**0.58**	0.52	0.50	0.52	0.47
bo→zh (B)	**0.65**	0.63	0.63	0.63	0.60	**0.49**	0.46	0.47	0.46	0.43
sa→en	0.66	0.65	0.66	**0.66**	0.62	0.57	0.56	0.57	**0.57**	0.52
zh→en	**0.69**	0.68	0.69	0.68	0.67	**0.60**	0.59	0.59	0.58	0.57
bo→fr	**0.64**	0.60	0.61	0.61	0.58	**0.46**	0.41	0.42	0.42	0.35
bo→es	**0.66**	0.63	0.63	0.63	0.60	**0.51**	0.47	0.48	0.48	0.42
bo→de	**0.64**	0.61	0.62	0.61	0.59	**0.55**	0.52	0.53	0.53	0.49
bo→pt	**0.68**	0.65	0.66	0.66	0.62	**0.50**	0.46	0.47	0.47	0.42
bo→it	**0.65**	0.63	0.64	0.64	0.61	**0.51**	0.49	0.51	0.50	0.44
bo→nl	0.64	0.62	0.63	**0.65**	0.60	0.51	0.49	0.51	**0.53**	0.46
bo→zh (M)	**0.73**	0.72	0.71	0.72	0.69	**0.62**	0.60	0.61	0.61	0.56

Table 6: COMET-22 and BLEURT-20 on 16 of the 17 directions (bo→en, scored per collection, is not included) for MITRA-MT, the Gemini references, and the strongest mid-size open baseline (best per metric block in bold). Where these metrics have language support they corroborate the chrF rankings; for Tibetan-target directions COMET lacks coverage and compresses or inverts large chrF differences (e.g. sa→bo), and both metrics are unvalidated for classical Buddhist registers.

	Sentence			Paragraph		
System	chrF	Gem.	Son.	chrF	Gem.	Son.
MITRA-SFT	43.1	81.5	79.5	50.4	88.7	82.9
MITRA-MT	42.9	81.3	79.0	50.4	89.3	84.1
3.5-Flash	39.5	84.5	80.7	48.8	94.0	88.6
3.6-Flash	39.3	84.7	81.8	47.5	93.7	87.5
Flash-Lite	37.9	81.5	77.0	45.9	93.6	86.6
Q-122B	35.6	77.3	71.8	43.6	88.2	78.1
G4-31B	34.7	70.7	66.6	43.3	85.2	73.6
G4-31B-5s	34.4	68.8	64.3	—	—	—
Qwen3.5	32.1	60.7	53.7	40.2	67.7	54.6
G4-12B	29.8	53.1	44.4	—	—	—
TG-27B	29.4	52.2	42.3	—	—	—
Tower+	27.9	44.1	35.4	—	—	—
Llama-8B	26.2	40.1	30.7	—	—	—
MADLAD	14.4	14.3	9.1	5.7	2.0	1.3

Table 7: As Table 4, restricted to the English-target directions (sa→en, zh→en, and the supplementary, non-released direction pa→en): 3 sentence-level and 3 paragraph-level directions. Averages are not comparable with Table 4, whose common direction set excludes pa→en. Systems not covering every direction here are shown as —.