Less Is More: Distilling Large-Scale Data with LLMs for Chinese-Centric Low-Resource Multilingual Machine Translation

Anonymous ACL submission

Abstract

Neural Machine Translation (NMT) between Chinese and low-resource languages (LRLs) faces significant challenges due to limited, noisy training data. We introduce MERIT, a unified translation framework that transforms the traditional English-centric ALT benchmark into a Chinese-centric evaluation suite for five Southeast Asian LRLs. Our approach integrates Language-specific Token Prefixing (LTP) for effective language conditioning and supervised fine-tuning (SFT). A key innovation, Group Relative Policy Optimization (GRPO) guided by the Score Accuracy Reward (SAR) function, strategically filters training data and optimizes model performance. Experiments with models up to 3 billion parameters (MERIT-3B) confirm the efficacy of our method. Ablation studies demonstrate substantial improvements from SFT-LTP over zero-shot baselines, while GRPO-SAR achieves further significant gains using only 22.8% of the original data, increasing BLEU-chrF scores by 17.4%. MERIT-3B notably surpasses open-source models such as NLLB-200 3.3B by 9.5 BLEU-4 points on Chinese–Indonesian translation and outperforms M2M-100 by 5.1 BLEU-4 points on Chinese–Lao. These findings highlight the pivotal role of targeted data curation and reward-guided training over mere model scaling, advancing multilingual translation in low-resource settings. Code and data are available at https://anonymous.4open.science/r/MERIT-864/.

1 Introduction

The vision of Neural Machine Translation (NMT) is to provide equitable access to information for speakers of over 7,000 languages worldwide. While English–French translation has achieved near-human BLEU scores (Papineni et al., 2002), many low-resource languages—including Chinese–LRL directions such as Tibetan, Lao, or Tagalog—remain virtually untranslatable due to the lack of parallel corpora, standard orthographies, and annotated data (Costa-jussà et al., 2022).

Multilingual pretraining has shown impressive zero-shot gains: mBART-50, mT5, and DeltaLM achieve broad coverage but continue to overlook or underperform on Southeast Asian and Chinese domestic LRLs. NLLB-200 (Costa-jussà et al., 2022) improves on this by expanding coverage to 200 languages and introducing the XSTS metric. However, Chinese–LRL performance still trails behind English-pivoted directions.

Moreover, the lack of publicly available and high-quality evaluation benchmarks hinders objective progress measurement. An ideal benchmark should: (i) cover multiple Chinese→LRL directions, (ii) maintain sufficient balance in scale and domain coverage, and (iii) avoid dependence on English as a pivot. Without these features, model improvements are difficult to reproduce or attribute reliably.

To address the long-standing lack of evaluation benchmarks for Chinese–low-resource language (LRL) directions in Neural Machine Translation (NMT), this paper makes the following three contributions:

- We introduce the first Chinese-centric multilingual benchmark for low-resource languages, derived from the ASEAN Languages Treebank (ALT) through various data filtering (EPDS-DIV) and distillation (QE Agent) techniques. This dataset targets low-resource language scenarios and ensures balanced domain coverage and semantic consistency (Thu et al., 2016).

- We compare multiple metrics through Strict-Overlap and Semantic-Friendly measures on five series of large language models (LLMs). The evaluation is conducted using three approaches: zero-shot, supervised fine-tuning

(SFT), and Group Relative Policy Optimization (GRPO), with the latter utilizing the expert-rated validation set for training.

- We explore the impact of model scale, data adaptation, and reward-based quality filtering on Chinese–LRL translation performance. Our results show that smaller open-source models can match proprietary models when trained with high-quality, strictly filtered data and guided reinforcement. The proposed MERIT framework effectively integrates and leverages these components.

2 Related Work

Early Chinese–LRL Corpora. The CCMT shared tasks released fewer than 200k sentence pairs for ZH–UG and ZH–MN (Liu et al., 2021), while Wiki-based mining typically yields only a few thousand pairs for ZH–LO and ZH–FIL (Artetxe and Schwenk, 2019). The ALT corpus (Thu et al., 2016) extends coverage to 13 ASEAN languages but remains English-centric and lacks direct Chinese–LRL alignment.

Multilingual Pretraining. Models like mBART-50 (Liu et al., 2020), mT5 (Xue et al., 2021), and DeltaLM (Ma et al., 2021) cover 50–101 languages, but still overlook or underperform on languages such as Tibetan and Uyghur. NLLB-200 (Costa-jussà et al., 2022) improves BLEU by 44% on FLORES-200 and adds the XSTS metric, but still underperforms on Chinese→LRL due to data quality gaps.

LLM-based Machine Translation. Instruction-tuned LLMs such as GPT-4o (Huang et al., 2025), Claude-3.5 (Enis and Hopkins, 2024), and Gemini-2.5 (DeepMind, 2025) show strong generalization across many languages. Open-source models like Qwen-2.5 (Cui and et al., 2025) and DeepSeek-v3 (Huang et al., 2025) serve as strong multilingual baselines, though most published evaluations focus on FLORES-200 and overlook Chinese–LRL directions.

Evaluation Limitations. Most multilingual benchmarks pivot through English, mix domain content, or lack human validation, reducing their diagnostic value for Chinesee-LRL MT. Our benchmark addresses these gaps with direct Chinese–LRL alignment, domain balance, and human-validated samples.

3 Methodology

3.1 Dataset

We construct a new test suite based on the ASEAN Languages Treebank (ALT) corpus (Thu et al., 2016). ALT is an English-centric multilingual corpus that already provides sentence-level alignment for several Southeast-Asian languages–Vietnamese (`vi`), Burmese (`my`), Lao (`lo`), Tagalog (`fil`) and Indonesian (`id`). Although Chinese is included as a target aligned with English, no direct Chinese–LRL alignment exists.

We therefore re-index sentences sharing the same `alt_id` and semantic source to form direct Chinese–LRL sentence pairs. In the resulting test set, Chinese can serve either as the source or as the reference language. The benchmark is clean, stylistically consistent and typologically diverse, enabling more controlled and fair evaluation of multilingual NMT systems from a Chinese-centric perspective.

Language Selection. We deliberately focus on five Southeast-Asian low-resource languages (vi, my, lo, fil, id) for four data-driven reasons:

All five appear in ALT with reliable English alignments, so that high-quality Chinese–LRL re-alignment is feasible.

Indonesian rather than Malay is kept because the two belong to the same Malayic subgroup and share 90 % lexical overlap, to the extent that many international surveys treat them as a single "Malay macrolanguage" (Adelaar, 2012; Eberhard et al., 2023b), if both are included simultaneously, it will result in redundancy of the experimental languages. Malay already has over 1 million clean En–Ms

Languages	Speaker Population[1] (Eberhard et al., 2023a)	Filtered Subset[2]	LRL
Chinese (`zh`)	1180M	✗	✗
Hindi (`hi`)	345M	✗	✗
Bengali (`bn`)	234M	✗	✗
Japanese (`ja`)	121M	✗	✗
Vietnamese (`vi`)	86M	10K	✓
Indonesian (`id`)	43M	10K	✓
Burmese (`my`)	33M	10K	✓
Tagalog (`fil`)	24M	10K	✓
Thai (`th`)	20M	✗	✗
Malay (`ms`)	18M	✗	✗
Khmer (`km`)[3]	16M	–	✓
Lao (`lo`)	4.3M	10K	✓

Table 1: Language Statistics from the ALT Corpus for Chinese-Centric Multilingual Translation. ALT corpus statistics sorted by L1 speaker population. All counts refer to L1 speakers and are rounded to the nearest million (M). LRL: Low-resource Language.

sentence pairs (e.g., MT-Wiki and multiple OPUS sub-corpora), pushing it into the mid-resource tier (Duong et al., 2017), whereas Indonesian still lacks sizeable Chinese parallel data (<50 k pairs in total, ALT contributes only 20 k) and is classified as low-resource by FLORES-200 and NLLB benchmarks (Goyal and et al., 2022; Team and AI, 2022).

Thai aggregated corpora exceed one million sentence pairs and the language enjoys dedicated WMT/IWSLT tracks (Lowphansirikul and Chuangsuwanich, 2020), so it no longer fits a strict LRL definition.

Khmer parallel resources are both small and highly noisy, the WMT20 corpus-filtering task emphasised that extensive cleaning is required (Koehn et al., 2020); moreover, unlike the other languages considered, Ethnologue reports virtually no L2 speaker community for Khmer (Simons and Fennig, 2023). Including Khmer would therefore demand a language-specific filtering pipeline and would undermine comparability with the other languages.

This reconstructed benchmark complements existing resources such as FLORES-200 (Costa-jussà et al., 2022), particularly for Chinese–LRL directions in mainland and maritime Southeast Asia. Unlike pivot-based benchmarks, our test set avoids semantic distortion introduced by intermediate English, thus enabling more realistic, stable, and reproducible evaluation for Chinese-centric multilingual translation systems.

3.2 Model Overview

We evaluate five series representative LLMs, spanning both proprietary and open-source systems:

Qwen-2.5 (Ghosal et al., 2024; Cui and et al., 2025): Chinese-English bilingual models fine-tuned for multilingual transfer, evaluated on several LRLs.

GPT-4o (Huang et al., 2025): OpenAI's flagship model tested on 16 languages, including several low-resource directions such as En–Te and En–Sw.

Claude-3.5 (Enis and Hopkins, 2024): A multilingual LLM from Anthropic, evaluated via MQM metrics on pairs like En–Yoruba and En–Amharic.

Gemini-2.5 (DeepMind, 2025): While lacking peer-reviewed benchmarks on Chinese–LRL tasks, its predecessor covers ultra-low-resource translation (e.g., En–Kalamang).

DeepSeek (Huang et al., 2025; Jiang et al., 2025): A competitive open-source model evaluated in the BenchMAX suite alongside GPT-4o.

Zero-shot prompting was applied exclusively to closed-source LLMs. For the Qwen-2.5 models, both SFT and GRPO-enhanced SFT were applied. Additional tests were conducted with SFT and GRPO-enhanced SFT using enhanced data derived from closed-source LLMs in the zero-shot regime. The GRPO-enhanced SFT regime utilizes a scoring agent trained on expert-rated development sets, optimized with Score Accuracy Reward (SAR) to ensure the selection of only high-quality translations. Model performance is assessed using the following metrics: BLEU-4, sacreBLEU, chrF, ROUGE-L, METEOR, and BERTScore.

3.3 Scoring and Selection

To construct high-quality parallel corpora for low-resource translation, we design a two-stage scoring and filtering pipeline that integrates interpretable statistical features with semantic evaluation, followed by reference-free quality estimation and threshold-based selection.

Stage I: Statistical and Semantic Scoring. We extract key surface-level features such as sentence length ratio and digit proportion difference, along with aggregated indicators for token balance, punctuation consistency, and lexical diversity. These help identify formatting mismatches and alignment noise (Munteanu and Marcu, 2005; Sánchez-Cartagena et al., 2018).

To assess deeper semantic alignment, we incorporate two additional signals: (i) conditional perplexity for fluency estimation, and (ii) instruction-following discrepancy to capture semantic fidelity, inspired by instruction-tuning objectives (Li et al., 2023). All features are normalized and combined through weighted scoring to penalize semantically misaligned pairs (Esplà-Gomis et al., 2020).

Stage II: Quality Estimation and Filtering. A reference-free Quality Estimation (QE) model, trained on human-annotated validation sets, further evaluates translation adequacy and fluency (Rei et al., 2020; Freitag et al., 2021). Sentence pairs surpassing a calibrated threshold are retained. This final stage ensures that the resulting dataset is both scalable and high-quality, suitable for fine-tuning compact models in low-resource scenarios.

[1]Speaker numbers derive from the most recent national censuses or *Ethnologue* reports (2023–2025) and are expressed in millions (M).

[2]Each ALT language contains approximately 20k aligned sentence pairs from a shared English source. See https://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/.

[3]No L2 speaker community (Simons and Fennig, 2023).

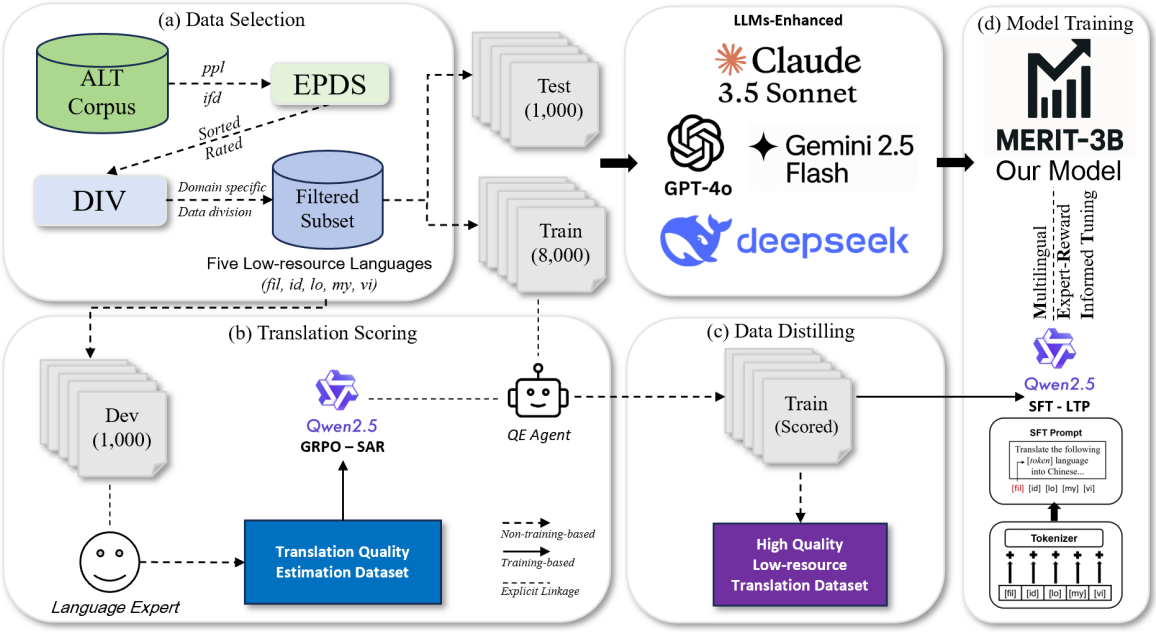

Figure 1: The overall workflow of the MERIT framework for Chinese-centric multilingual translation. **(a) Data Selection**: Sentences from the ALT corpus are filtered and scored using perplexity (ppl) and inverse frequency diversity (ifd) to construct and refine a high-quality multilingual dataset, yielding training and testing subsets for five low-resource Southeast Asian languages. **(b) Translation Scoring**: A Translation Quality Estimation (QE) dataset is created by language experts, followed by training a QE agent to assess the translation quality of additional data. **(c) Data Distilling**: QE Agent trained by Group Relative Policy Optimization (GRPO) with Score Accuracy Reward (SAR) leveraged to further evaluate more training data, yielding more high-quality training samples. **(d) Model Training**: The MERIT-3B model is trained using Supervised Fine-Tuning (SFT), Language-specific Token Prefixing (LTP) with LLM-Enhanced data argmentation to achieve optimal multilingual translation performance.

3.4 Supervised Fine-Tuning

We fine-tune open-source models (Qwen-2.5 0.5B and 3B) on the filtered Chinese–LRL data using supervised fine-tuning (SFT) (Fan et al., 2021). The training objective is to maximize the conditional likelihood of the target sequence $Y = (y_1, \ldots, y_M)$ given the source sequence $X = (x_1, \ldots, x_N)$, using standard sequence-to-sequence formulation with teacher forcing (Sutskever et al., 2014; Williams and Zipser, 1989).

Label smoothing with $\varepsilon = 0.1$ is applied to avoid over-confidence (Szegedy et al., 2016). Fine-tuning is conducted independently for each Chinese–LRL pair to account for language-specific morphology. This strategy, paired with prior quality filtering, yields strong gains over zero-shot performance, echoing findings on targeted adaptation for multilingual NMT (Arivazhagan et al., 2019).

3.5 Language-specific Token Prefixing

To improve language discrimination in one-to-many generation, we adopt Language-specific Token Prefixing (LTP), which prepends a target language token (e.g., `[lo]`) to both the source input and prompt instruction. This token is added to the tokenizer vocabulary and embedded as part of the model input.

For each training sample, the source input is modified as $X' = [\texttt{lang}] \oplus X$, and the final model input becomes a concatenation of the instruction prompt and the language-tagged source sequence:

$$\text{Input} = [\texttt{Prompt} \oplus [\texttt{lang}] \oplus x_1, \ldots, x_n] \quad (1)$$

The training objective minimizes the negative log-likelihood of the target sequence:

$$\mathcal{L}_{\text{MLE}} = -\sum_{t=1}^{m} \log P(y_t \mid y_{<t}, \texttt{Prompt}, X; \theta) \quad (2)$$

This extends target-language prefixing ideas (Johnson et al., 2017) by combining symbolic and prompt-based conditioning for unified multilingual fine-tuning.

3.6 Group Relative Policy Optimization

Group Relative Policy Optimization (GRPO) is a reinforcement learning strategy that refines model

outputs using reward feedback. Inspired by Reinforcement Learning with Human Feedback (RLHF) techniques (Ouyang et al., 2022; Lu et al., 2022), GRPO operates on mini-batches of candidate translations in this task, assigning scalar rewards based on Score Accuracy Reward (SAR) scores. Subsequently, the model learns to maximize the expected reward via policy gradient updates.

Unlike conventional pointwise objective functions, GRPO introduces an intra-batch comparison mechanism and normalizes rewards using a moving baseline. This approach helps to reduce the variance of gradient estimates, thereby enhancing the stability of the training process. Our experiments indicate that GRPO is effective for translation evaluation, enabling the selection and improvement of dataset translation quality.

3.7 Score Accuracy Reward Function

We define the Score Accuracy Reward (SAR) to evaluate model completions based on their ability to accurately reproduce specific numerical scores present in ground-truth answers. This reward function is designed for tasks where precision in extracting or generating key numerical information is critical. SAR rewards model outputs that closely match these target numerical values (representing a form of preference or correctness) while penalizing deviations.

To extract the salient numerical score s_i from the completion content c_i, we first delineate a match-set, $M(c_i)$. This set comprises all integer values identified within c_i through the application of a predefined regular expression, denoted as R. We define a predicate $P_R(m, c_i)$ to be true if and only if m is an integer yielded by matching the regular expression R against the string c_i. The match-set $M(c_i)$ is then formally defined as:

$$M(c_i) = \{m \in \mathbb{Z} \mid P_R(m, c_i)\} \tag{3}$$

The extractor function $E(c_i)$ then determines the score s_i from this match-set. As per the reference, if $M(c_i)$ is not empty, s_i is the minimum integer found; otherwise, s_i is set to -1:

$$s_i = E(c_i) = \begin{cases} \min M(c_i), & \text{if } M(c_i) \neq \emptyset \\ -1, & \text{if } M(c_i) = \emptyset \end{cases} \tag{4}$$

Given a ground-truth integer answer vector $a = (a_1, a_2, \ldots, a_N)$, we define a piecewise reward-mapping function $\phi(d)$ based on the absolute difference $d = |s_i - a_i|$ between the extracted score s_i and the ground-truth answer a_i. This function, detailed in source, is:

$$\phi(d) = \begin{cases} 2.0, & \text{if } d = 0 \\ 1.0, & \text{if } 1 \leq d \leq 10 \\ 0.0, & \text{otherwise} \end{cases} \tag{5}$$

This mapping assigns the highest reward for an exact match, a partial reward for close matches (difference up to 10), and zero reward for larger deviations or mismatches.

Finally, the reward r_i for each instance i is computed based on s_i and $\phi(d)$. If a valid score $s_i \geq 0$ was extracted, the reward is $\phi(|s_i - a_i|)$. If no score was extracted ($s_i < 0$), the reward is 0.0:

$$r_i = \begin{cases} \phi(|s_i - a_i|), & \text{if } s_i \geq 0 \\ 0.0, & \text{if } s_i < 0 \end{cases} \tag{6}$$

The overall outcome is a reward vector $r = (r_1, r_2, \ldots, r_N)$. This SAR mechanism, by focusing on the accuracy of extracted numerical scores against ground-truth values, provides a clear signal for tasks requiring numerical precision. Similar score-alignment objectives, where models are rewarded for matching target scores or preferences, have been successfully adopted in alignment training for various generation tasks (Wu et al., 2023).

4 Experiments and Analysis

4.1 Evaluation Method

We evaluate translation performance using both overlap-based and semantic-aware metrics:

Strict-Overlap: BLEU-4 (Papineni et al., 2002), sacreBLEU (Post, 2018), and ROUGE-L (Lin, 2004) assess lexical match and n-gram precision, which are crucial for evaluating surface-level accuracy and fluency.

Semantic-Friendly: chrF (Popovic, 2015), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang et al., 2020) measure semantic similarity and fluency robustness, capturing aspects that n-gram overlap alone might miss.

Each metric is computed on the reconstructed ALT test suite for five Chinese–LRL pairs. We report averages across directions, comparing zero-shot prompting, SFT, and GRPO-enhanced regimes.

To provide a balanced evaluation that captures both lexical precision and semantic adequacy, we propose a composite metric, BLEU-chrF. This metric integrates insights from both the Strict-Overlap

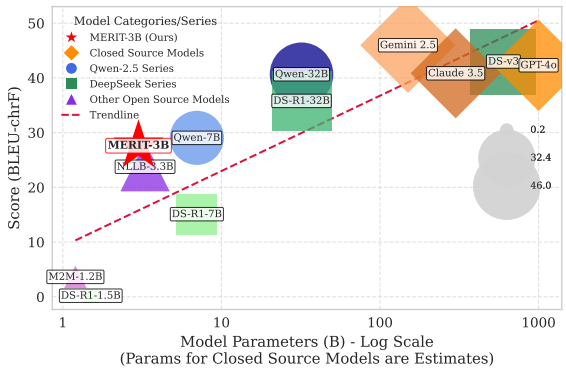

Figure 2: Performance–Scale Trade-offs of MERIT-3B and Baseline Models on Chinese-Centric Multilingual Translation. Comparison of BLEU-chrF scores against model size (log-scale) across MERIT-3B, open-source, and estimated closed-source models.

and Semantic-Friendly categories of evaluation measures by taking the arithmetic mean of the BLEU-4 score and the chrF score:

$$\text{BLEU-chrF} = \frac{\text{BLEU-4} + \text{chrF}}{2} \tag{7}$$

By averaging these two widely-used metrics, one emphasizing n-gram precision and the other character n-gram recall and F-score. We aim to achieve a more holistic assessment of translation quality, particularly for tasks where both lexical fidelity and semantic resemblance are important.

4.2 Main Result

Table 2 presents our evaluation results across five Chinese–LRL directions, categorizing metrics into Strict-Overlap (Papineni et al., 2002; Post, 2018,?; Lin, 2004) and Semantic-Friendly (Popovic, 2015; Banerjee and Lavie, 2005; Zhang et al., 2020).

Among leading closed-source models, Gemini-2.5 Flash consistently achieves top scores in BLEU-4 and chrF across multiple languages, such as Filipino (BLEU-4: 49.26, chrF: 42.68) and Indonesian (BLEU-4: 48.73, chrF: 41.93). Claude-3.5 Sonnet excels in ROUGE-L for Lao (14.00) and Burmese (24.44).

The proposed **M**ultilingual **E**xpert-**R**eward **I**nformed **T**uning (MERIT) framework, demonstrates notable strengths. Specifically, MERIT-3B significantly outperforms the similarly sized open-source NLLB-200 3.3B model across several metrics for Filipino, Indonesian, and Vietnamese. For instance, on Filipino, MERIT-3B achieves 29.71 BLEU-4 and 49.88 METEOR compared to NLLB-200's 25.05 and 46.10, respectively. Furthermore, MERIT-3B shows substantial gains over smaller open-source baselines on particularly challenging low-resource language pairs.

Notably, our MERIT-3B model demonstrates substantial advantages over the DeepSeek-r1 7B. MERIT-3B consistently outperforms DeepSeek-r1 7B on lexical similarity metrics such as BLEU-4 and chrF across all five evaluated languages. Furthermore, when benchmarked against Qwen-2.5 7B, MERIT-3B, with only approximately 42.9% of its parameters, achieves highly competitive translation quality. For instance, on Filipino, MERIT-3B reaches 99.7% of Qwen-2.5 7B's ROUGE-L score (31.91 vs. 31.99) and over 98.3% of its BLEU-4 score (29.71 vs. 30.21). Similar competitiveness is observed for Indonesian, where MERIT-3B attains approximately 95.7% of Qwen-2.5 7B's BLEU-4 score (34.73 vs. 36.28) and 93.2% of its ROUGE-L score (26.63 vs. 28.56).

These results underscore the efficacy of our reward-informed filtering and specialized fine-tuning approach, particularly in enhancing performance for low-resource languages and achieving competitive results within the open-source landscape relative to model scale.

4.3 Module Comparison

To investigate the contribution of each component, we conduct an ablation study on Qwen-2.5 0.5B and Qwen-2.5 3B across four distinct setups: zero-shot (serving as our baseline), Supervised Fine-Tuning with Language-Token Prefixing (SFT-LTP), SFT-LTP followed by reward-enhanced tuning using Group Relative Policy Optimization with Score Accuracy Reward (GRPO-SAR), and finally our full SFT-LTP + GRPO-SAR with an additional LLMs-Enhanced (LLME) stage.

As detailed in Table 3, initial SFT-LTP yields substantial improvements in both BLEU-4 and chrF scores over the zero-shot baselines across all languages for both model sizes. For instance, Qwen-2.5 3B sees its overall BLEU-chrF score increase from 9.10 to 15.53 after SFT-LTP. Introducing GRPO-SAR provides further consistent gains. Notably, for the Qwen-2.5 3B model, GRPO-SAR significantly boosts performance on low-resource pairs like Chinese–Lao, improving BLEU-4 from 1.17 (SFT-LTP) to 4.39 and chrF from 2.22 (SFT-LTP) to 5.17. Even with its limited capacity, the Qwen-2.5 0.5B model benefits remarkably from GRPO-SAR, achieving an overall BLEU-chrF score of 4.39, which is nearly a 40-fold increase

Strict-Overlap	BLEU-4					sacreBLEU					ROUGE-L						
Model	fil	id	lo	my	vi	fil	id	lo	my	vi	fil	id	lo	my	vi	FT	OS
GPT-4o	45.26	47.77	28.10	30.48	45.03	43.60	45.81	27.38	29.53	44.50	33.62	31.40	12.33	23.57	28.25	✗	✗
Claude-3.5 Sonnet	42.97	47.35	32.20	30.92	45.14	43.38	46.54	32.65	30.14	44.55	35.03	31.17	**14.00**	**24.44**	28.45	✗	✗
Gemini-2.5 Flash	**49.26**	**48.73**	**35.79**	**36.91**	39.24	**47.48**	**47.20**	**35.03**	**36.07**	38.63	**35.89**	30.83	13.53	23.92	24.62	✗	✗
DeepSeek-v3	46.19	41.83	25.57	26.68	41.29	44.12	41.38	25.25	26.29	40.90	35.15	30.06	12.75	22.95	28.27	✗	✓
Qwen-2.5 32B	43.56	47.87	23.27	20.43	**46.23**	42.01	46.61	22.48	19.50	**44.63**	34.97	**31.85**	11.84	20.61	**29.08**	✗	✓
DeepSeek-r1 32B	37.98	43.29	12.97	8.87	42.57	37.14	42.27	12.71	8.43	41.45	32.40	29.92	10.04	16.31	28.11	✗	✓
Qwen 2.5-7B	30.21	36.28	6.17	6.43	35.22	29.75	36.41	5.67	5.42	35.07	31.99	28.56	9.86	16.41	27.81	✗	✓
DeepSeek-r1 7B	14.77	20.94	0.79	0.37	12.53	15.31	24.17	0.86	0.46	16.16	24.58	23.32	8.67	5.67	18.52	✗	✓
NLLB-200 3.3B	25.05	25.27	15.86	20.83	25.30	24.21	23.64	15.19	20.13	22.97	31.45	25.73	11.49	18.68	25.51	✗	✓
DeepSeek-r1 1.5B	0.07	0.08	0.05	1.06	0.05	0.03	0.06	0.01	0.11	0.02	0.77	0.90	1.80	3.80	0.55	✗	✓
M2M-100 1.2B	2.13	9.53	0.05	0.00	3.33	1.32	9.47	0.01	0.00	3.19	9.88	21.65	4.69	0.00	13.01	✗	✓
MERIT-3B (Ours)	29.71	34.73	5.15	4.56	31.20	27.20	33.16	4.28	3.54	29.25	31.91	26.63	8.40	13.68	21.18	✓	✓
Semantic-Friendly	**chrF**					**METEOR**					**BERTScore**						
Model	fil	id	lo	my	vi	fil	id	lo	my	vi	fil	id	lo	my	vi	FT	OS
GPT-4o	39.36	41.13	24.20	26.76	39.62	67.80	70.34	53.66	56.59	69.44	68.59	70.84	56.27	57.04	70.41	✗	✗
Claude-3.5 Sonnet	38.71	41.30	28.38	28.79	39.65	67.52	70.29	58.53	58.88	69.23	68.28	71.55	60.72	57.68	70.40	✗	✗
Gemini-2.5 Flash	**42.68**	**41.93**	**30.61**	**32.09**	34.66	**70.14**	**70.87**	**60.21**	**61.85**	60.31	**71.67**	**72.77**	**63.93**	**62.39**	60.88	✗	✗
DeepSeek-v3	39.80	36.95	22.37	24.43	36.86	68.04	67.10	51.41	53.66	67.43	70.02	68.81	54.82	54.21	68.87	✗	✓
Qwen-2.5 32B	37.77	41.35	20.36	18.13	**39.80**	66.61	70.57	46.72	43.29	**69.76**	67.73	71.73	50.50	45.09	**71.13**	✗	✓
DeepSeek-r1 32B	33.62	37.62	12.77	9.81	36.97	61.60	66.75	34.07	27.37	66.95	63.32	68.54	36.09	27.36	68.65	✗	✓
Qwen-2.5 7B	27.88	33.68	7.39	7.95	33.28	54.51	62.56	20.84	23.25	63.00	54.95	62.82	22.24	23.35	63.28	✗	✓
DeepSeek-r1 7B	15.43	21.60	2.20	1.35	14.87	36.02	49.35	8.45	6.46	37.61	34.22	49.07	1.46	-4.10	36.54	✗	✓
NLLB-200 3.3B	22.48	23.57	14.92	18.69	22.43	46.10	45.28	36.27	42.28	45.73	48.61	54.34	44.80	48.93	52.59	✓	✓
DeepSeek-r1 1.5B	0.33	0.38	0.36	1.87	0.24	1.99	2.34	2.17	3.80	1.43	-27.97	-26.07	-20.48	-19.68	-29.48	✗	✓
M2M-100 1.2B	5.16	18.21	0.26	0.01	6.49	4.65	14.00	0.57	0.11	6.10	-16.71	-1.68	-10.04	-16.34	-9.51	✓	✓
MERIT 3B (Ours)	25.52	30.22	5.81	5.53	26.98	49.88	56.46	16.55	16.93	50.50	46.70	45.58	11.45	16.12	32.87	✓	✓

Table 2: Evaluation on five Southeast Asian languages. Strict-Overlap metrics include BLEU-4, sacreBLEU, and ROUGE-L. Semantic-Friendly metrics include chrF, METEOR, and BERTScore. For each metric column: **Bold** values indicate the highest score, and Underlined values indicate the second highest score across all models. FT: Fine-tuned; OS: Open Source.

Model	BLEU-4					chrF					Overall (BLEU-chrF)
	fil	id	lo	my	vi	fil	id	lo	my	vi	
Qwen2.5-0.5b	0.03	0.03	0.02	0.01	0.01	0.16	0.12	0.40	0.25	0.06	0.11
+ SFT-LTP	$1.86_{\uparrow 1.83}$	$4.02_{\uparrow 4.00}$	$0.25_{\uparrow 0.22}$	$0.15_{\uparrow 0.14}$	$3.12_{\uparrow 3.11}$	$4.85_{\uparrow 4.69}$	$10.38_{\uparrow 10.26}$	$1.41_{\uparrow 1.00}$	$1.07_{\uparrow 0.82}$	$8.55_{\uparrow 8.50}$	$3.57_{\uparrow 3.46}$
+ GRPO-SAR	$2.31_{\uparrow 2.28}$	$4.32_{\uparrow 4.29}$	$0.26_{\uparrow 0.24}$	$0.16_{\uparrow 0.16}$	$6.29_{\uparrow 6.27}$	$5.00_{\uparrow 4.84}$	$10.64_{\uparrow 10.52}$	$1.38_{\uparrow 0.98}$	$1.22_{\uparrow 0.96}$	$12.36_{\uparrow 12.30}$	$4.39_{\uparrow 4.28}$
+ LLME	$0.32_{\uparrow 0.29}$	$0.45_{\uparrow 0.42}$	$0.08_{\uparrow 0.06}$	$0.07_{\uparrow 0.06}$	$0.79_{\uparrow 0.78}$	$1.20_{\uparrow 1.04}$	$1.66_{\uparrow 1.54}$	$0.45_{\uparrow 0.05}$	$1.25_{\uparrow 1.00}$	$2.82_{\uparrow 2.76}$	$0.91_{\uparrow 0.80}$
Avg.	1.13	2.21	0.15	0.10	2.55	2.80	5.70	0.91	0.95	5.95	2.25
Qwen2.5-3b	5.80	14.25	1.11	1.83	17.06	8.83	17.71	2.05	3.03	19.35	9.10
+ SFT-LTP	$23.01_{\uparrow 17.21}$	$26.00_{\uparrow 11.75}$	$1.17_{\uparrow 0.05}$	$2.53_{\uparrow 0.69}$	$27.94_{\uparrow 10.87}$	$20.14_{\uparrow 11.31}$	$24.25_{\uparrow 6.54}$	$2.22_{\uparrow 0.17}$	$3.53_{\uparrow 0.50}$	$24.55_{\uparrow 5.20}$	$15.53_{\uparrow 6.43}$
+ GRPO-SAR	$25.58_{\uparrow 19.78}$	$29.11_{\uparrow 14.86}$	$4.39_{\uparrow 3.28}$	$2.77_{\uparrow 0.93}$	$32.54_{\uparrow 15.48}$	$23.62_{\uparrow 14.78}$	$27.83_{\uparrow 10.12}$	$5.17_{\uparrow 3.12}$	$3.45_{\uparrow 0.42}$	$27.88_{\uparrow 8.53}$	$18.23_{\uparrow 9.13}$
+ LLME	$29.71_{\uparrow 23.91}$	$34.73_{\uparrow 20.48}$	$5.15_{\uparrow 4.04}$	$4.56_{\uparrow 2.73}$	$31.20_{\uparrow 14.14}$	$25.52_{\uparrow 16.69}$	$30.22_{\uparrow 12.51}$	$5.81_{\uparrow 3.76}$	$5.53_{\uparrow 2.50}$	$26.98_{\uparrow 7.63}$	$19.94_{\uparrow 10.84}$
Avg.	21.03	26.02	2.96	2.92	27.19	19.53	25.00	3.81	3.89	24.69	15.70

Table 3: Ablation Study of Qwen-2.5 0.5B and Qwen-2.5 3B on five Southeast Asian languages. All values are rounded to two decimal places. Improvements over the Zero-shot baseline (underlined rows).

(a 3890% relative improvement) over its zero-shot baseline score of 0.11. This underscores the efficacy of reward modeling, consistent with findings in instruction tuning (Ouyang et al., 2022; Wu et al., 2023).

Our proposed LLMs-Enhanced (LLME) stage demonstrates further advancements, particularly for the larger Qwen-2.5 3B model. With LLME, the Qwen-2.5 3B model achieves the highest overall BLEU-chrF score of 19.94, representing a 10.84 absolute point improvement (a 119% relative increase) over its zero-shot baseline. This highlights the synergistic benefits of our full pipeline. While the LLME stage yields more modest gains for the Qwen-2.5 0.5B model in the current setup (overall BLEU-chrF of 0.91), the substantial cumulative improvements from SFT-LTP and GRPO-SAR on this smaller model, and the peak performance achieved by the 3B model with LLME, collectively validate the effectiveness and scalability of our modular tuning strategy in significantly enhancing translation quality.

4.4 Effect of Data Distillation on Performance

We assess the impact of our quality filtering approach by comparing full-scale Supervised Fine-Tuning with Language-Token Prefixing (SFT-LTP) against subsequent reward-informed filtering and tuning via GRPO-SAR, using our MERIT-3B model. Table 4 details the number of retained training instances per language and the corresponding

overall BLEU-chrF scores for these configurations.

The SFT-LTP stage utilizes the full set of 40,000 training instances. In contrast, the GRPO-SAR stage strategically curates this data, drastically reducing the volume to only 9,126 instances. This constitutes an average data reduction of 77.2%, with the most significant reduction observed for Vietnamese, where the training data was cut by 87.8% (from 8,000 to 976 instances). Remarkably, despite this substantial data pruning, the overall BLEU-chrF score not only signifies the efficient retention of highly informative samples but actually improves from 15.53 (achieved with SFT-LTP on 40,000 instances) to 18.23 with GRPO-SAR on the reduced dataset. This represents a relative performance increase of approximately 17.4%.

These findings underscore the efficacy of our reward-based filtering (GRPO-SAR) as a data-efficient strategy that can simultaneously reduce training data requirements and enhance model performance. This offers a compelling alternative to training on larger, potentially noisier, unfiltered datasets. The benefits of leveraging reward signals for targeted data curation align with effective strategies observed in other generative AI tasks, such as summarization and dialogue tuning (Lu et al., 2022; Ouyang et al., 2022).

4.5 Further Discussion

Our work, culminating in the MERIT framework, demonstrates the significant potential of combining data filtering techniques, such as the Score Accuracy Reward (SAR) driven GRPO, with efficient fine-tuning strategies like Language-Token Prefixing (LTP) for multilingual translation, especially into low-resource languages (LRLs). The proposed BLEU-chrF composite metric has also provided a balanced view of lexical and semantic performance. While MERIT-3B exhibits strong performance relative to its scale and against comparable open-source models, several limitations persist and pave the way for future exploration.

First, script-related challenges, particularly for Lao and Burmese, can introduce encoding inconsistencies. These not only affect the performance of QE agents used in SAR but also potentially skew standard evaluation metrics. Future iterations could incorporate more robust character normalization or transliteration techniques at the data preprocessing stage, or develop QE models less sensitive to such variations.

Second, the current reward model underlying GRPO-SAR, while effective, may inadvertently prioritize adequacy (accuracy of content, as captured by our specific SAR function focusing on numerical or key information matching) sometimes at the expense of optimal fluency. This can occasionally lead to subtle grammatical artifacts in some translations. Future work could investigate multi-objective reward functions that explicitly balance adequacy, fluency, and even other aspects like style or register, potentially drawing on more diverse human feedback signals beyond simple ratings.

Moreover, as noted by Zhang et al. (2022), many LLMs, including some baselines we compared against, are often evaluated in zero-shot or few-shot settings for translation. This might not fully reveal their capabilities, which could be significantly enhanced with more sophisticated prompting strategies or in-context learning techniques specifically tailored for translation. Exploring how our data filtering and fine-tuning methods can synergize with advanced prompting for even larger LLMs is a promising direction.

5 Conclusion

This work introduces MERIT, a unified framework combining a reconstructed benchmark and modular training strategies for Chinese–Low-Resource Language (LRL) neural machine translation. We reconstruct a clean and balanced evaluation suite from the ALT corpus, enabling reliable assessment across five Southeast Asian languages. On the training side, MERIT incorporates language-conditioned fine-tuning and reward-guided data selection to improve translation quality efficiently.

Experiments demonstrate that our MERIT-3B model outperforms comparable open-source models and approaches the performance of significantly larger proprietary systems, particularly in LRL scenarios. Ablation results confirm that reward-informed filtering with GRPO-SAR is especially effective, achieving better performance with less data.

Overall, our findings reinforce the importance of strategic data selection and modular fine-tuning over sheer model scale in low-resource settings. Future work will extend MERIT to more languages and explore scaling with stronger reward functions and larger base models.

References

2023. Emnlp 2023 ethics faq. https://2023.emnlp.org/ethics/faq. EMNLP 2023 Conference Website.

Alexander Adelaar. 2012. Indonesian and malay: Are they different languages? *Annual Review of Linguistics*, pages 1–23.

Naveen Arivazhagan and 1 others. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. In *Proceedings of ACL*.

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

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT*.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 11167–11191.

Yiming Cui and et al. 2025. Multilingual machine translation with open large language models at practical scale. *arXiv preprint arXiv:2502.02481*.

Google DeepMind. 2025. Gemini 2.5 technical report: Unlocking multimodal understanding across millions of tokens. Google AI Blog. URL: https://ai.googleblog.com/gemini-25-report.

Long Duong, Thanh-Le Ha, and Le-Minh Phuong. 2017. English–malay parallel corpus construction from wikipedia. In *Proceedings of the 21st Workshop on Asian Language Resources*, pages 29–38.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2023a. *Ethnologue: Languages of the World*, 26 edition. SIL International, Dallas, Texas. L1 speaker estimates for 12 ALT languages; language-specific URLs available in Appendix/Table 1.

David M. Eberhard, Gary F. Simons, and Charles D. (eds.) Fennig. 2023b. Malay Macrolanguage. *Ethnologue: Languages of the World*, 26th ed. https://www.ethnologue.com/macrolanguages.

Can Enis and Mark Hopkins. 2024. From llm to nmt: Advancing low-resource machine translation with claude. *arXiv preprint arXiv:2404.13813*.

Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena, Jeff Zaragoza-Bernabeu, and Felipe Sánchez-Martínez. 2020. Bicleaner at WMT 2020: Universitat d'Alacant–Prompsit's submission to the parallel corpus filtering shared task. In *Proceedings of the 5th Conference on Machine Translation (WMT 2020)*, Online. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, and Yihyung C. Aharoni. 2021. Beyond english-centric multilingual machine translation. In *Proc. of ACL*.

Markus Freitag, Yaser Al-Onaizan, and 1 others. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. In *Proceedings of ACL*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Sarthak S Ghosal, Ritam Roy, Sumanth Doddapaneni, and 1 others. 2024. Promptrefine: Enhancing few-shot performance on low-resource indic languages with example selection from related example banks. *arXiv preprint arXiv:2412.05710*.

Naman Goyal and et al. 2022. The flores-200 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, pages 855–872.

Yutong Huang, Minghao Wang, Yujie Xie, and 1 others. 2025. Benchmax: A comprehensive multilingual evaluation suite for large language models. *arXiv preprint arXiv:2502.07346*.

Zhengyong Jiang, Lingfan Zhang, Ruobing Guo, and 1 others. 2025. Deepseek v3 vs o3-mini: How well can reasoning llms evaluate mt and summarization? *arXiv preprint arXiv:2504.08120*.

Melvin Johnson, Mike Schuster, Quoc V Le, and 1 others. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. In *Transactions of the Association for Computational Linguistics*.

Philipp Koehn and 1 others. 2020. Findings of the 2020 wmt shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–746.

Mingxu Li, Yuxian Zhang, Shuai He, Zhipeng Li, Haoran Zhao, Junkai Wang, Ning Cheng, and Tianxiang Zhou. 2023. From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning. *arXiv preprint arXiv:2308.12032*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*.

Ming Liu, Jingjing Xu, Jiqiang Zhang, and Yidong Chen. 2021. Research on uyghur–chinese neural machine translation based on ccmt shared tasks. In *Proceedings of the China Conference on Machine Translation (CCMT)*, pages 120–130.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Lalida Lowphansirikul and Ekaterina Chuangsuwanich. 2020. Scb-mt en–th: A large, clean parallel corpus for english–thai machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3612–3622.

Zichao Lu and 1 others. 2022. Quark: Controllable and compositional generation of text using reinforcement learning. In *Proceedings of ACL*.

Shuming Ma, Li Dong, Shaohan Huang, Saksham Singhal, Shuming Li, Long Zhou, Yaru Wang, Kaitao Song, Yuxiang Zhang, and Furu Wei. 2021. Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 9064–9078.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2018. Model cards for model reporting. *arXiv preprint arXiv:1810.03993*.

Dragos Ştefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Long Ouyang, Jeffrey Wu, Xu Jiang, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

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 of the Association for Computational Linguistics*, pages 311–318.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation*.

Ricardo Rei, Andre F. T. Martins, Ana B. Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

ACL Rolling Review. 2023. Responsible nlp research checklist. https://aclrollingreview.org/responsibleNLPresearch/.

Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez-Sánchez. 2018. Prompsit's submission to WMT 2018 parallel corpus filtering shared task. In *Proceedings of the 3rd Conference on Machine Translation (WMT 2018)*, Brussels, Belgium. Association for Computational Linguistics.

Gary F. Simons and Charles D. (eds.) Fennig. 2023. Central Khmer. *Ethnologue: Languages of the World*, 26th ed. https://www.ethnologue.com/language/khm.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proc. of NIPS*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proc. of CVPR*.

NLLB Team and Meta AI. 2022. No language left behind: Building a complete world-wide machine translation system. Technical Report arXiv:2207.04672, Meta AI Technical Report.

Ye Kyaw Thu, Pa Pa Win, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Introducing the asian language treebank (alt). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 1574–1578.

Ronald J. Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280.

Yuntao Wu and 1 others. 2023. Fine-grained alignment of language models with preferences. In *arXiv preprint arXiv:2305.18290*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Susan Zhang, Stephen Roller, Naman Goyal, and 1 others. 2022. Opt: Open pre-trained transformer language models. In *arXiv preprint arXiv:2205.01068*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Appendix

A.1 Ethics Statements

This work presents a Chinese-centric multilingual translation benchmark targeting five Southeast Asian low-resource languages (LRLs), constructed from publicly available corpora and evaluated under reproducible protocols. We aim to support responsible research in multilingual NLP by releasing rigorous evaluation resources while proactively addressing ethical concerns related to data provenance, model fairness, environmental impact, and potential misuse.

Data Privacy and Consent All data is derived from the publicly available ASEAN Languages Treebank (ALT), which includes multilingual translations of government and news texts. While the dataset is openly licensed, the original collection did not explicitly document consent procedures or personal identifiable information (PII) removal. To mitigate this, we apply a multi-stage filtering process to exclude named entities, explicit language, and potentially sensitive content. Nonetheless, due to the limitations of automated and manual filtering, some residual risk may remain. We follow the data statements framework (Bender and Friedman, 2018) and document licensing, provenance, and usage constraints in the appendix.

Bias and Fairness Despite the use of a three-stage filtering pipeline and expert-rated supervision, the training data may still encode latent cultural, linguistic, or regional bias—particularly due to its English-pivoted design and limited coverage of dialectal variations or non-standard orthographies. Annotators are bilingual graduate students, and while they are experienced, demographic diversity is limited. Future work will prioritize the inclusion of more diverse annotators and typologically broader sources to mitigate such representational imbalances. Our work aligns with global AI ethics principles of fairness, transparency, and non-maleficence (Gebru et al., 2018).

Environmental Impact Model training and inference were conducted on a single-node NVIDIA A100 80GB GPU. We log training FLOPs and wall-clock runtime for both the SFT and GRPO stages. While the GRPO procedure improves data efficiency through reward-based filtering, it introduces additional computational cost. We estimate that the total training corresponds to a typical single-node compute workload and plan to explore more lightweight reward models or compute-efficient alternatives to reduce carbon impact in future iterations (emn, 2023).

Intended Use and Misuse Risks The benchmark is designed to support objective evaluation and supervised training for Chinese–LRL translation tasks. It is intended for academic research and language technology development, particularly in regions underrepresented in NLP. However, misuse is possible—such as generating misinformation or content targeting marginalized communities. We explicitly discourage such applications and recommend that any downstream use include fairness auditing, risk controls, and human oversight (Mitchell et al., 2018).

Transparency and Reproducibility We adhere to the ACL Responsible NLP Research Checklist (Review, 2023) and release all code, data, and model checkpoints under a CC BY-NC 4.0 license. All filtering procedures, model configurations, and hyperparameter settings are fully documented. Some language-specific heuristics (e.g., token ratio thresholds) were empirically selected and may not generalize across domains; future validation is necessary to ensure robustness.

A.2 Limitations

Despite the encouraging results achieved by our proposed framework and the MERIT-3B model, this work has several limitations that warrant discussion and offer avenues for future improvement.

First, while our evaluation benchmark improves upon existing English-pivoted resources by constructing direct Chinese–LRL sentence pairs, its current scope is confined to five Southeast Asian languages. Other significant low-resource languages, including domestic Chinese minority languages such as Tibetan, Uyghur, and Kazakh, remain unaddressed due to the scarcity of high-quality aligned corpora. Expanding the linguistic diversity of our benchmark is crucial for assessing broader generalizability.

Second, the ALT-based test suite, although semantically aligned through shared `alt_id` indexing, is fundamentally constrained by its original English-centric design. While our realignment efforts aim to mitigate semantic drift when adapting it for Chinese–LRL evaluation, some residual domain-specific or stylistic artifacts originating

from the English-centric source may persist and subtly influence translation assessment.

Third, although our methodology employs a QE agent and the statistical-semantic Score Accuracy Reward (SAR) function for automatic data filtering, the scale of human validation for these components is currently limited. The expert-rated set used for developing or validating the reward model is modest in size. This might restrict the overall robustness and generalizability of the SAR model, particularly its alignment with nuanced human preferences across diverse linguistic phenomena. Future work should prioritize the integration of more extensive and varied human annotations.

Fourth, while we have evaluated a range of LLMs under zero-shot, SFT, and GRPO regimes, the decoding strategies (e.g., beam size, sampling temperature) and specific prompt formats were kept fixed across these models for controlled comparison. These settings can significantly influence translation behavior and perceived quality, especially for proprietary models whose internal mechanisms are opaque. A more exhaustive exploration of model-specific optimal decoding parameters and prompt engineering could reveal further performance variations.

Finally, due to computational resource constraints, our current experiments, including the development and evaluation of the MERIT-3B model and its associated fine-tuning framework (SFT-LTP, GRPO-SAR), have been conducted on models up to the 3B parameter scale. We have not yet extended this framework to significantly larger parameter models (e.g., 7B+, or state-of-the-art models in the tens or hundreds of billions of parameters). Applying and evaluating our data filtering and reward-informed tuning strategies on such larger-scale models is an important next step to ascertain their scalability and potential for even greater performance gains, though this would require substantial additional computational resources. Furthermore, a detailed efficiency analysis, including training time, inference latency, and computational cost of the filtering and fine-tuning stages, was not conducted and would be valuable for assessing practical deployability.

A.3 Experimental Setup

All experiments were conducted on a local workstation equipped with two NVIDIA RTX 3090 GPUs (24 GB). Under a 2×2 parallel configuration, the per-GPU batch size was set to 8 with a gradient accumulation step of 2, resulting in an effective total batch size of 32. The maximum input sequence length was set to 1024 tokens, and the initial learning rate was configured as `2e-4`. The system environment included Ubuntu 20.04, CUDA 12.1, and Python 3.10, with PyTorch 2.1 and Transformers v4.49 as the core libraries.

All training was performed using standard mixed-precision (fp16) computation via custom training scripts. Due to hardware limitations, the batch size was carefully adjusted to fit within the available GPU memory, and no experiments were conducted using larger-parameter models. To ensure reproducibility, all random seeds were fixed, and detailed runtime logs were maintained for each experiment.

A.4 Reward Function

In this study, we introduce the Score Accuracy Reward (SAR) function as a key component of the reward mechanism for evaluating the numerical accuracy of generated outputs. Specifically, we design a step-wise reward strategy, which operates as follows:

- A reward of 2.0 is assigned if the model's output exactly matches the reference answer.
- A reward of 1.0 is assigned if the output deviates from the reference by a small margin.
- A reward of 0.0 is assigned if the deviation exceeds the acceptable threshold.

This step-wise reward formulation differs from conventional binary reward functions commonly used in correctness or format-checking tasks. It is motivated by two main considerations:

(1) Task-specific suitability. SAR is designed for numerical question answering and tasks that require precise arithmetic reasoning. In such settings, the reference answer is often a numeric value with an acceptable tolerance range rather than a single exact string. Therefore, outputs that are numerically close to the reference can be considered approximately correct. Compared to rigid binary matching, the step-wise reward better aligns with the intrinsic characteristics of these tasks.

(2) More informative training signal. Unlike traditional 0/1 rewards, step-wise rewards provide finer-grained feedback, allowing the model to receive gradient signals that vary with the degree of deviation. This facilitates smoother optimization

and more stable convergence during training, enabling the model to gradually improve its numeric prediction capabilities. In contrast, overly rigid reward mechanisms may lead to sparse or uninformative training signals and hinder early-stage learning.

A.5 Recruitment And Payment

To ensure the accuracy and objectivity of human evaluation, we recruited ten annotators with academic backgrounds in the target Southeast Asian languages. All annotators were either language instructors or graduate students from relevant universities. For each target language, two annotators were assigned, and a cross-review protocol was adopted to enhance annotation quality and consistency.

All participants had formal training in translation or linguistics and possessed strong language comprehension and evaluative capabilities. Annotators were compensated at a rate of 1 RMB per evaluated sample. Before the evaluation began, all participants received detailed instructions and training on annotation guidelines. Participation was voluntary, and compensation was provided proportionally based on the amount of completed work.

Since the dataset contains no personally identifiable information (PII) and the task involves only linguistic quality assessment, the annotation process entails no ethical risks and does not require institutional ethics approval.

Algorithm 1 Elite Parallel Data Sampler

Input: XML dataset $\mathcal{D}_{\text{xml}}$;
Target sizes $\{T_{\text{train}}, T_{\text{dev}}, T_{\text{test}}\}$;
Domain set $\mathcal{D}$
Output: Datasets $\{D_{\text{train}}, D_{\text{dev}}, D_{\text{test}}\}$

Stage 1: Domain Balancing Global pool $M^* \leftarrow \emptyset$
foreach *domain* $d \in \mathcal{D}$ **do**
  $M_d \leftarrow \{\, f \in \mathcal{D}_{\text{xml}} \mid \text{domain}(f) = d \,\}$ $M^* \leftarrow M^* \cup M_d$
$N \leftarrow \sum_d |M_d|$; $P_d \leftarrow |M_d|/N$ $(\forall d)$
Stage 2: Proportional Split Generation **foreach** *split* $t \in \{\textit{train}, \textit{dev}, \textit{test}\}$ **do**
  $S_t \leftarrow \emptyset$ **foreach** *domain* $d \in \mathcal{D}$ **do**
    $n_t(d) \leftarrow \lfloor T_t \cdot P_d \rfloor$ $S_t \leftarrow S_t \cup \textsc{Shuffle}(M_d)[: n_t(d)]$
  $\Delta \leftarrow T_t - |S_t|$ **if** $\Delta > 0$ **then**
    $S_t \leftarrow S_t \cup \textsc{Sample}(M^* \setminus S_t, \Delta)$
  else
    if $\Delta < 0$ **then**
      $S_t \leftarrow \textsc{Head}(S_t, T_t)$
  $M^* \leftarrow M^* \setminus S_t$
Stage 3: Domain Proportion Verification **foreach** *domain* $d \in \mathcal{D}$ **do**
  $\varepsilon_d \leftarrow \left| |S_{\text{train}} \cap M_d|/|S_{\text{train}}| - P_d \right|$ **if** $\varepsilon_d \geq 0.05$ **then**
    $\textsc{GlobalShuffle}(\{S_t\})$; **goto** Stage 2
Stage 4: Output Generation **foreach** *split* $t \in \{\textit{train}, \textit{dev}, \textit{test}\}$ **do**
  $\textsc{Shuffle}(S_t)$
return $\{S_{\text{train}}, S_{\text{dev}}, S_{\text{test}}\}$

Method	# Training Size fil	id	lo	my	vi	Overall (Size / BLEU-chrF)
MERIT-3B						
+ SFT-LTP	**8,000**	**8,000**	**8,000**	**8,000**	**8,000**	**40,000 / 15.53**
+ GRPO-SAR	$1{,}851_{\downarrow 76.9\%}$	$1{,}779_{\downarrow 77.7\%}$	$2{,}058_{\downarrow 74.2\%}$	$2{,}462_{\downarrow 69.2\%}$	$976_{\downarrow 87.8\%}$	$9{,}126_{\downarrow 77.2\%}$ / $18.23_{\uparrow 17.4\%}$
+ LLME	$2{,}891_{\downarrow 63.9\%}$	$3{,}104_{\downarrow 61.2\%}$	$3{,}300_{\downarrow 58.8\%}$	$3{,}764_{\downarrow 53.0\%}$	$2{,}193_{\downarrow 72.6\%}$	$15{,}252_{\downarrow 61.9\%}$ / $19.94_{\uparrow 28.4\%}$

Table 4: Training size comparison across five low-resource languages for MERIT-3B. Overall column shows total training data (with percentage reduction relative to initial 40,000) and BLEU-chrF score (with percentage improvement relative to the SFT-LTP stage).

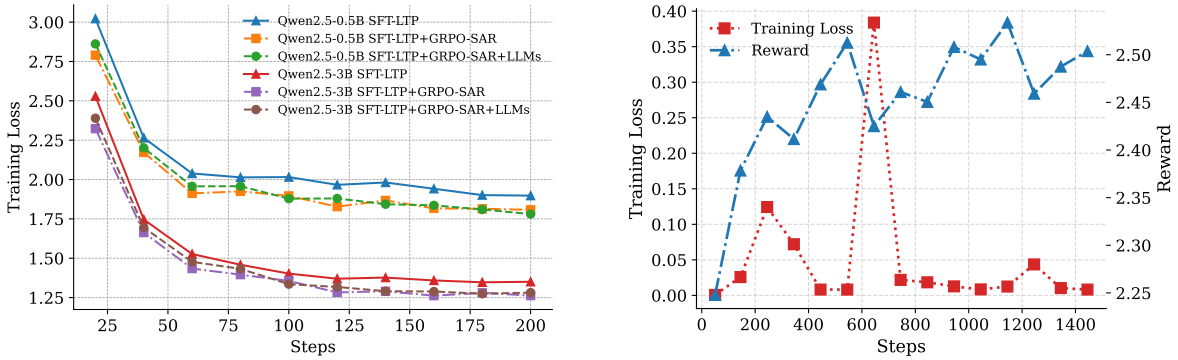

Figure 3: Training loss and reward evolution across SFT and GRPO strategies.

Algorithm 2 Data Integrity Validation

Input: Candidate splits $\{S_t\}$;
Anomaly threshold $\tau_{\text{anom}} = 0.7$;
Validity threshold $\tau_{\text{valid}} = 0.8$;
PCA dimension k; Regularization λ.
Output: Validation $\in \{\texttt{True}, \texttt{False}\}$

Stage 1: Feature Extraction $F_{\text{outlier}} \leftarrow \emptyset$ $X \leftarrow \emptyset$
foreach *sample* $s \in S_{train}$ **do**
  $\mathbf{e} \leftarrow \text{CNNEncoder}(s)$ $\mathbf{c} \leftarrow \text{BERTEmbed}(s)$ $\mathbf{v} \leftarrow \text{L2Normalize}([\mathbf{e}\|\mathbf{c}])$ $X \leftarrow X \cup \{\mathbf{v}\}$
StandardizeX(X) PCA(X, k)
Stage 2: Cluster Anomaly Discovery
$C \leftarrow \text{HDBSCAN}(X)$ **foreach** *cluster* $C_k \in C$ **do**
  $\rho_k \leftarrow |C_k|/|X|$ $\Sigma \leftarrow \text{CosineSimilarity}(C_k)$ **foreach** *sample* $x_i \in C_k$ **do**
    $s_i \leftarrow 1 - \frac{1}{|C_k|}\sum_{j \in C_k} \Sigma_{ij}$ **if** $s_i > \tau_{anom}$ **then**
      $F_{\text{outlier}} \leftarrow F_{\text{outlier}} \cup \{x_i\}$
  $w_k \leftarrow \rho_k \cdot \left(1 - |F_{\text{outlier}} \cap C_k|/|C_k|\right)$
Stage 3: Composite Validity Score $V_{\text{geo}} \leftarrow \prod_k w_k^{\rho_k}$ $V_{\text{pen}} \leftarrow e^{-\lambda |F_{\text{outlier}}|}$ $V \leftarrow V_{\text{geo}} \cdot V_{\text{pen}}$
Stage 4: Validation Check **if** $V < \tau_{valid}$ **then**
  is Valid ← False; PurgeCorruptedData(S_t)
else
  is Valid ← True

A.6 Feature Extraction

Let H', W' denote the height and width of the final CNN feature map after L layers. Let d_e be the number of CNN output channels, and let d_c be the BERT hidden dimension (which equals the token embedding size d_m).

Let the training set be $\mathbf{S}_{\text{train}} = \{\mathbf{s_i}\}_{\mathbf{i=1}}^{\mathbf{N}}$. Each sample s_i comprises an image s_i^{img} and text s_i^{text}, processed as follows:

1. **CNN Encoding:**

$$F_i^{(0)} = s_i^{\text{img}} \tag{8}$$

$$F_i^{(l)} = \text{ReLU}\big(W^{(l)} * F_i^{(l-1)} + b^{(l)}\big), \tag{9}$$

$$F_i^{(L)} \in \mathbb{R}^{d_e \times H' \times W'} \tag{10}$$

$$e_i^j = \frac{1}{H'W'} \sum_{h=1}^{H'} \sum_{w=1}^{W'} F_i^{(L)}(j,h,w) \tag{11}$$

$$\mathbf{e}_i = [e_i^1, \ldots, e_i^{d_e}]^\top \in \mathbb{R}^{d_e} \tag{12}$$

2. **BERT Embedding:**

$$X_i^{(0)} = \big[E_{\text{tok}}(w_t) + E_{\text{pos}}(t)\big]_{t=1}^{T} \tag{13}$$

$$X_i^{(\ell)} = \text{TransformerLayer}^{(\ell)}\big(X_i^{(\ell-1)}\big), \tag{14}$$

$$\mathbf{c}_i = X_i^{(L')}[1] \in \mathbb{R}^{d_c} \tag{15}$$

3. **Concatenation & Normalization:**

$$\mathbf{u}_i = \begin{bmatrix} \mathbf{e}_i \\ \mathbf{c}_i \end{bmatrix} \in \mathbb{R}^{d_e+d_c} \tag{16}$$

$$\mathbf{v}_i = \frac{\mathbf{u}_i}{\|\mathbf{u}_i\|_2} \tag{17}$$

4. **Matrix Assembly:**

$$X = \begin{pmatrix} \mathbf{v}_1^\top \\ \vdots \\ \mathbf{v}_N^\top \end{pmatrix} \in \mathbb{R}^{N\times(d_e+d_c)} \tag{18}$$

5. **PCA Reduction:**

$$Z = \mathrm{PCA}_k(X) = \begin{pmatrix} \mathbf{z}_1^\top \\ \vdots \\ \mathbf{z}_N^\top \end{pmatrix} \in \mathbb{R}^{N\times k} \tag{19}$$

A.7 Cluster Anomaly Detection

1. **DBSCAN Clustering:**

$$\mathcal{N}_\epsilon(\mathbf{z}_i) = \{\mathbf{z}_j : \|\mathbf{z}_j - \mathbf{z}_i\|_2 \le \epsilon\} \tag{20}$$

$$\mathbf{z}_i \iff |\mathcal{N}_\epsilon(\mathbf{z}_i)| \ge \mathrm{MinPts} \tag{21}$$

(22)

$$\mathbf{z}_i \rightsquigarrow \mathbf{z}_j \iff \exists\,(\mathbf{z}_{i_0}, \ldots, \mathbf{z}_{i_m}) \tag{23}$$

$$\mathbf{z}_p, \mathbf{z}_q \Leftrightarrow \iff \exists\,\mathbf{z}_o :\ \mathbf{z}_p, \mathbf{z}_q \rightsquigarrow \mathbf{z}_o \tag{24}$$

2. **Clustering Procedure:**

   (a) Mark all $\mathbf{z}_i$ unvisited, set cluster counter $c \leftarrow 0$.

   (b) For each unvisited $\mathbf{z}_i$:

      i. Mark $\mathbf{z}_i$ visited; let $N \leftarrow \mathcal{N}_\epsilon(\mathbf{z}_i)$.

      ii. If $|N| < \mathrm{MinPts}$, label $\mathbf{z}_i$ as noise.

      iii. Else:

$$c \leftarrow c+1, \quad C_c \leftarrow \{\mathbf{z}_i\} \tag{25}$$

   (c) **expand**(C, N):

$$\mathbf{z}_j :\ \begin{cases} \text{if unvisited: mark visited} \\ N' \leftarrow \mathcal{N}_\epsilon(\mathbf{z}_j) \\ \text{if } |N'| \ge \mathrm{MinPts}, \quad N \leftarrow N \cup N' \end{cases} \tag{26}$$

$$\text{if } \mathbf{z}_j \notin C, \text{ then } C \leftarrow C \cup \{\mathbf{z}_j\} \tag{27}$$

   (d) Resulting clusters: $C_1, \ldots, C_K$

3. **Cosine Similarity:**

$$\Sigma_{ij} = \frac{\mathbf{z}_i^\top \mathbf{z}_j}{\|\mathbf{z}_i\|_2\,\|\mathbf{z}_j\|_2} \tag{28}$$

4. **Anomaly Score:**

$$s_i = 1 - \frac{1}{|C_k|}\sum_{j\in C_k} \Sigma_{ij} \tag{29}$$

5. **Outlier Set:**

$$\mathcal{F}_{\text{outlier}} = \{\, i \mid s_i > \tau_{\text{anom}}\} \tag{30}$$

6. **Cluster Weighting:**

$$\rho_k = \frac{|C_k|}{N} \tag{31}$$

$$\delta_k = 1 - \frac{|\mathcal{F}_{\text{outlier}} \cap C_k|}{|C_k|} \tag{32}$$

$$w_k = \rho_k\,\delta_k \tag{33}$$