Can Reward-Guided Optimization Bridge the Gap in Chinese-Centric Low-Resource Translation?

Anonymous ACL submission

Abstract

Despite rapid progress in large language models (LLMs), direct translation from low-resource languages (LRLs) into Chinese remains constrained by scarce high-quality parallel data, noisy mined corpora, and English-centric evaluation resources. This challenge is particularly pronounced for Southeast Asian LRLs, where existing multilingual models often perform unsteadily without relying on English as a pivot. To support direct evaluation, we introduce CALT, a Chinese-centric LRL→Chinese benchmark reconstructed from ALT and covering five Southeast Asian languages. Building on CALT, we propose **M**ultilingual **E**xpert-**R**eward **I**nformed **T**uning (**MERIT**), a reward-informed framework for improving Chinese-centric low-resource translation. MERIT combines language-specific prefixing, supervised fine-tuning, expert-aligned quality estimation, semantic alignment reward, and reward-informed data distillation to better exploit limited training data. Experiments across five LRL→Chinese directions show that MERIT consistently improves translation quality. In particular, MERIT-32B achieves the strongest overall performance among evaluated open-source systems and obtains statistically significant average gains over GPT-4o, indicating that carefully aligned reward signals and data curation can narrow the Chinese-centric low-resource translation gap more effectively than relying on parameter scaling alone. Our code is available at: https://anonymous.4open.science/r/MERIT-3E76.

1 Introduction

Large Language Models (LLMs) and Neural Machine Translation (NMT) have substantially improved multilingual access, yet their benefits remain uneven across language pairs. Most progress is concentrated in high-resource or English-centric settings, while direct translation from low-resource languages (LRLs) into Chinese remains underexplored. This gap is particularly pronounced for Southeast Asian languages, where direct LRL-Chinese parallel corpora are scarce and existing multilingual pipelines are often organized around English as a pivot. Consequently, progress in Chinese-centric low-resource translation remains difficult to measure, optimize, and reproduce, even as massively multilingual models continue to scale (Liu et al., 2020; Xue et al., 2021; NLLB Team et al., 2024).

In this work, we focus on five Southeast Asian LRLs: Filipino/Tagalog (FIL), Indonesian (ID), Lao (LO), Burmese (MY), and Vietnamese (VI). Translating these languages directly into Chinese introduces several compounding challenges, including limited direct supervision, script and orthographic variation, insufficient expert annotation, and domain mismatch in mined corpora. Although general-purpose multilingual models provide broad cross-lingual coverage, their performance in direct LRL→Chinese settings remains unstable without high-quality aligned data and reliable evaluation. This suggests that model capacity alone is insufficient when the underlying data are noisy and the evaluation protocol is not well aligned with the target translation direction.

A central obstacle is the lack of direct Chinese-centric benchmarks for Southeast Asian LRLs. The Asian Language Treebank (ALT) provides valuable multilingual sentence-level alignments (Thu et al., 2016), but its construction is organized around English, making direct LRL→Chinese evaluation nontrivial. To build a more suitable testbed, we reconstruct ALT by aligning shared sentence identifiers and introduce CALT, a Chinese-centric benchmark covering five LRL→Chinese directions. CALT avoids runtime English pivoting and provides balanced train, development, and test splits, enabling more reproducible evaluation of Chinese-centric low-resource translation systems.

Equipped with this benchmark, we ask whether

reward-guided optimization can improve Chinese-centric low-resource translation beyond direct scaling of model size. We propose MERIT, short for Multilingual Expert-Reward Informed Tuning, a unified framework tailored to LRL→Chinese translation. Rather than treating all aligned sentence pairs as equally reliable, MERIT explicitly models data quality and uses expert-aligned reward signals to guide both data curation and model optimization. The framework combines Language-specific Token Prefixing (LTP) and Supervised Fine-Tuning (SFT) to establish language-aware translation capability, expert-aligned quality estimation to identify noisy or weak pairs, Group Relative Policy Optimization guided by Semantic Alignment Reward (GRPO-SAR) to optimize reward-aligned preferences, and LLM-Enhanced (LLME) data distillation to further improve the training signal.

We evaluate MERIT against representative open-source and closed-source systems, including GPT-4o, using both lexical metrics such as sacreBLEU and semantic metrics such as BERTScore. The results show that MERIT consistently improves Chinese-centric low-resource translation under our evaluation protocol. In particular, MERIT-32B achieves the strongest overall performance among evaluated open-source systems and obtains statistically significant average gains over GPT-4o. Ablation studies further show that the proposed reward-informed pipeline improves an 8B backbone as well, indicating that the framework is effective beyond a single model scale. Our contributions are summarized as follows:

- **Chinese-centric Benchmark.** We introduce CALT, a reconstructed LRL→Chinese benchmark spanning five Southeast Asian languages. CALT reduces dependence on English-pivot evaluation and supports more standardized Chinese-centric low-resource MT evaluation.

- **Reward-informed Tuning Framework.** We propose MERIT, a unified framework that combines SFT, language-specific token prefixing, expert-aligned quality estimation, GRPO-SAR, and LLM-enhanced data distillation to improve translation quality under severe data scarcity.

- **Comprehensive Empirical Evaluation.** We evaluate MERIT across representative open-source and closed-source systems. Results show that MERIT-32B achieves the strongest overall open-source performance under our evaluation protocol, while ablation studies confirm the progressive contribution of each module.

2 Methodology

2.1 Overview

MERIT is designed for direct low-resource Southeast Asian languages → Chinese translation. Given a source sentence $X = (x_1, \ldots, x_N)$ in a low-resource language and a Chinese target sentence $Y = (y_1, \ldots, y_M)$, the goal is to improve translation quality under limited and noisy supervision.

As shown in Figure 1, MERIT consists of four stages. First, we reconstruct a Chinese-centric corpus from ALT and apply initial statistical-semantic filtering. Second, we train an expert-aligned quality estimation (QE) agent to score candidate translations. Third, we optimize the QE agent with group relative policy optimization guided by semantic alignment reward (GRPO-SAR). Finally, we use reward-informed distilled data to optimize translation backbones with supervised fine-tuning and language-specific token prefixing (SFT-LTP), optionally enhanced by LLM-generated candidates in the LLME stage.

2.2 CALT Benchmark Construction

We construct CALT from the Asian Language Treebank (ALT) corpus (Thu et al., 2016). ALT is an English-centric multilingual corpus containing sentence-level alignments for several Southeast Asian languages. Although Chinese is included in ALT, direct LRL→Chinese evaluation is not its

Language	Code	L1 Speakers	Selected
Chinese	zh	1180M	✗
Vietnamese	vi	86M	✓
Indonesian	id	43M	✓
Burmese	my	33M	✓
Tagalog/Filipino	fil	24M	✓
Lao	lo	4.3M	✓
Thai	th	20M	✗
Malay	ms	18M	✗
Khmer	km	16M	✗

Table 1: Language coverage considered in CALT construction. Speaker statistics follow recent census or *Ethnologue* reports (Eberhard et al., 2026). We select five Southeast Asian LRLs with reliable ALT alignments and comparable LRL→Chinese reconstruction quality.

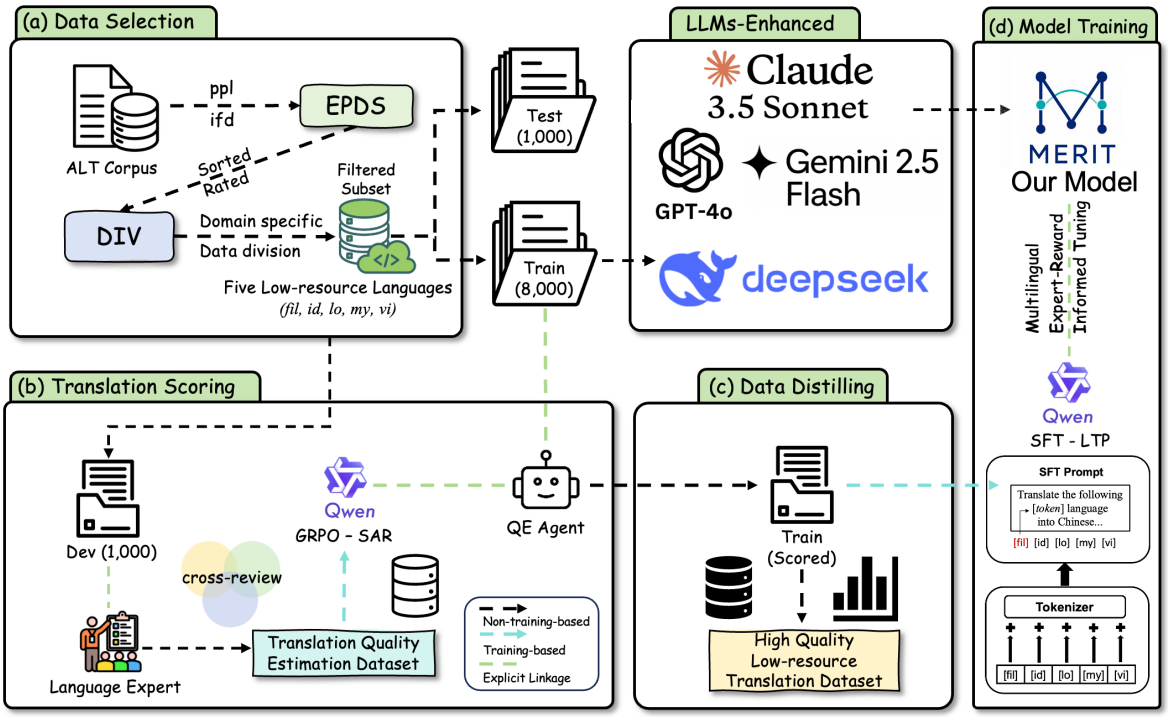

Figure 1: Overview of the MERIT framework. MERIT first reconstructs a Chinese-centric LRL→Chinese corpus and applies statistical-semantic filtering. It then trains an expert-aligned QE agent with semantic alignment reward, uses the agent for reward-informed data distillation, and finally optimizes the translation model with SFT-LTP and LLME-enhanced distilled data.

primary design objective. We therefore reconstruct direct LRL→Chinese pairs by aligning sentences with shared `alt_id` values. For each selected language $\ell \in \mathcal{L}$, where $\mathcal{L} = \{\texttt{fil}, \texttt{id}, \texttt{lo}, \texttt{my}, \texttt{vi}\}$:

$$\mathcal{D}_{\ell \to zh} = \{(s_i^{\ell}, s_i^{zh}) : \mathrm{id}(s_i^{\ell}) = \mathrm{id}(s_i^{zh})\}. \quad (1)$$

This reconstruction avoids runtime English pivoting during evaluation while preserving the shared semantic structure of ALT.

2.3 Language Selection

We focus on Filipino/Tagalog, Indonesian, Lao, Burmese, and Vietnamese. These languages are selected because they are covered by ALT with reliable sentence-level alignments and represent diverse scripts, typological properties, and resource conditions. Indonesian is retained instead of Malay because the two belong to the same Malayic subgroup and including both would introduce redundancy (Lim and Poedjosoedarmo, 2016). Thai is excluded because it has substantially larger parallel resources and dedicated MT benchmarks (Lowphansirikul et al., 2021). Khmer is excluded from the main evaluation due to noisier available resources and the need for a language-specific cleaning pipeline (Koehn et al., 2020). This selection provides a controlled testbed for Chinese-centric low-resource translation while complementing broader multilingual resources such as FLORES-200 (Goyal et al., 2022; NLLB Team et al., 2024).

2.4 Initial Data Scoring

Given an initial noisy corpus $\mathcal{D}_{raw} = \{(x_i, y_i)\}_{i=1}^{N}$, we first compute a feature-based score for each sentence pair. The score integrates surface-level alignment features such as length ratio, token ratio, punctuation divergence, digit divergence, and lexical diversity difference:

$$S_{\text{base}}(x_i, y_i) = \sum_{k=1}^{K} w_k \phi_k(f_i^{(k)}). \quad (2)$$

These features are commonly used for noisy parallel data detection and corpus filtering (Munteanu and Marcu, 2005; Sánchez-Cartagena et al., 2018). We further incorporate conditional perplexity and instruction-following discrepancy signals to capture fluency and semantic consistency:

$$S_{\text{final}} = \alpha S_{\text{base}} + \beta S_{\text{PPL}} + \gamma S_{\text{IFD}}. \quad (3)$$

The weights (α, β, γ) are calibrated on the development set. High-scoring pairs are retained by EPDS,

and DIV is applied to enforce split size and domain consistency. Additional filtering details are provided in Appendix E.

2.5 Language-Specific Token Prefixing

We use Language-specific Token Prefixing (LTP) to improve source-language discrimination in many-to-one multilingual generation. For a source language ℓ, we introduce a special token $[\ell]$ and prepend it to the source:

$$X' = [\ell] \oplus X. \tag{4}$$

The model input is then:

$$\widetilde{X} = (P_\ell, [\ell], x_1, \ldots, x_N), \tag{5}$$

where P_ℓ is an instruction such as:

$$P_\ell = \text{“Translate the following } \ell \text{ sentence into Chinese:”}. \tag{6}$$

This strategy extends target-language control in multilingual NMT (Johnson et al., 2017) to a Chinese-centric many-to-one setting. The target language is fixed as Chinese, while the prefix identifies the source language.

2.6 Supervised Fine-Tuning with LTP

Given the prefixed input $\widetilde{X}$ and target sequence Y, we train the translation model with teacher forcing (Williams and Zipser, 1989). The sequence-level SFT loss is:

$$\mathcal{L}_{\text{SFT}} = -\frac{1}{M}\sum_{t=1}^{M} \log p_\theta(y_t^* \mid \widetilde{X}, y_{<t}). \tag{7}$$

We use label smoothing (Szegedy et al., 2015). For token y_t^*, the smoothed target distribution is:

$$q'(k) = (1-\varepsilon_{\text{ls}})\mathbb{I}(k = y_t^*) + \frac{\varepsilon_{\text{ls}}}{|\mathcal{V}|}. \tag{8}$$

In practice, Eq. 7 is computed with q' over the vocabulary. SFT-LTP serves as the base translation optimization stage before reward-informed data distillation.

2.7 Expert-Aligned Quality Estimation

MERIT uses a QE agent to evaluate candidate translations without requiring reference translations at filtering time. The QE agent is trained on expert-annotated development examples, where each candidate translation receives a scalar quality score reflecting adequacy, fluency, and semantic faithfulness. This design follows the motivation of reference-free MT evaluation methods such as COMET-QE (Rei et al., 2020; Freitag et al., 2021).

For a source sentence x_i and candidate translation $\hat{y}_i$, the QE agent generates an evaluation log c_i and a scalar score $s_i \in [0, 100]$. To avoid ambiguity, we require an explicit score field:

$$s_i = E(c_i). \tag{9}$$

The extraction function is:

$$s_i = E(c_i) = \begin{cases} \min M(c_i), & \text{if } M(c_i) \neq \emptyset \\ -1, & \text{if } M(c_i) = \emptyset \end{cases} \tag{10}$$

where $m \in [0, 100]$. Invalid or unparsable outputs are assigned -1.

2.8 Semantic Alignment Reward

Semantic Alignment Reward (SAR) aligns QE predictions with expert judgments. Let a_i denote the expert score and s_i the extracted QE score. We first compute:

$$d_i = |s_i - a_i|. \tag{11}$$

The reward mapping is:

$$\phi(d_i) = \begin{cases} 2.0, & d_i = 0, \\ 1.0, & 1 \leq d_i \leq 10, \\ 0.0, & d_i > 10. \end{cases} \tag{12}$$

The final SAR reward is:

$$R_i = \begin{cases} \phi(d_i), & s_i \geq 0, \\ 0.0, & s_i < 0. \end{cases} \tag{13}$$

This reward encourages the QE agent to approximate expert preferences rather than rely only on surface-level heuristics (Xu et al., 2024).

2.9 GRPO-SAR Optimization

We optimize the QE agent with GRPO-SAR. For each input x_i, the policy samples G candidate evaluation outputs $\{c_{i,j}\}_{j=1}^{G}$ and receives rewards $\{R_{i,j}\}_{j=1}^{G}$. The group mean and standard deviation are:

$$\mu_i = \frac{1}{G}\sum_{j=1}^{G} R_{i,j}, \quad \sigma_i = \sqrt{\frac{1}{G}\sum_{j=1}^{G}(R_{i,j} - \mu_i)^2}. \tag{14}$$

The normalized advantage is:

$$A_{i,j} = \frac{R_{i,j} - \mu_i}{\sigma_i + \epsilon_r}. \tag{15}$$

Let π_θ be the current QE policy and $\pi_{\theta_{\text{old}}}$ the behavior policy. The probability ratio is:

$$\rho_{i,j}(\theta) = \frac{\pi_\theta(c_{i,j} \mid x_i)}{\pi_{\theta_{\text{old}}}(c_{i,j} \mid x_i)}. \tag{16}$$

We define the clipped surrogate:

$$g_{i,j}(\theta) = \min\left(\rho_{i,j} A_{i,j}, \bar{\rho}_{i,j} A_{i,j}\right), \tag{17}$$

where

$$\bar{\rho}_{i,j} = \mathrm{clip}(\rho_{i,j}, 1-\epsilon, 1+\epsilon). \tag{18}$$

The GRPO-SAR objective is:

$$\mathcal{J}_{\mathrm{GRPO}} = \mathbb{E}_{i,j}\left[g_{i,j}(\theta) - \beta D_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid x_i) \| \pi_{\mathrm{ref}}(\cdot \mid x_i)\right)\right]. \tag{19}$$

This objective encourages the QE agent to produce expert-aligned scores while constraining policy drift, following the general principle of preference-based optimization (Ouyang et al., 2022).

2.10 Reward-Informed Data Distillation

After GRPO-SAR optimization, the QE agent rescored candidate translation pairs from the filtered training pool and the LLME-enhanced candidate pool. For each pair $(x_i, \hat{y}_i)$, the QE agent assigns a score s_i. We retain high-quality pairs using a threshold τ calibrated on the development set:

$$\mathcal{D}_{\mathrm{distill}} = \{(x_i, \hat{y}_i) : s_i \geq \tau\}. \tag{20}$$

The distilled data are used for final SFT-LTP optimization of the translation model. This stage links expert-aligned quality estimation with data selection: instead of increasing data quantity indiscriminately, MERIT prioritizes translation pairs with high semantic adequacy and fluency.

3 Experiments and Analysis

3.1 Evaluation Method

We evaluate all systems on five LRL→Chinese directions in CALT: Filipino/Tagalog (FIL), Indonesian (ID), Lao (LO), Burmese (MY), and Vietnamese (VI). Following standard MT evaluation practice, we report both lexical-overlap and semantic-aware metrics. The strict-overlap group includes BLEU-4 (Papineni et al., 2002), sacreBLEU (Post, 2018), and ROUGE-L (Lin, 2004). The semantic-friendly group includes chrF (Popović, 2015), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang et al., 2020). These complementary metrics are used because n-gram overlap alone may miss semantic adequacy, especially in low-resource translation, while neural or semantic metrics have been shown to correlate more closely with expert judgments in MT evaluation (Rei et al., 2020; Freitag et al., 2021).

All metric values are reported on a unified 0-100 scale. Metrics whose native range is $[0, 1]$, such as METEOR and BERTScore, are multiplied by 100 before reporting and aggregation. To summarize performance across languages, we additionally report an Overall Score based on sacreBLEU and BERTScore:

$$\mathrm{Overall} = \frac{1}{2}\left(\overline{\mathrm{sacreBLEU}} + \overline{\mathrm{BERTScore}}\right), \tag{21}$$

where

$$\overline{m} = \frac{1}{|\mathcal{L}|}\sum_{\ell \in \mathcal{L}} m_\ell, \quad |\mathcal{L}| = 5. \tag{22}$$

This definition combines standardized lexical matching and semantic similarity after both metrics are scaled to the same 0-100 range. We use the same Overall Score in the scaling analysis, ablation study, and significance analysis. (see Table 5 and Table 6).

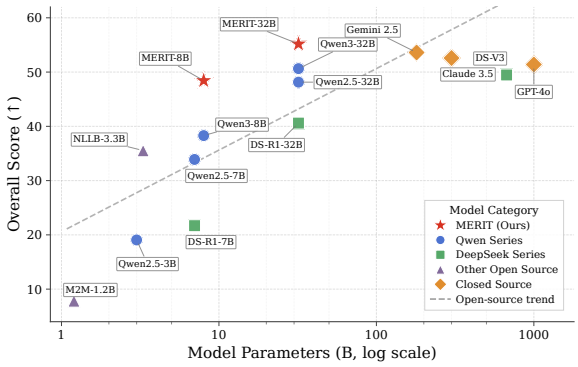

Figure 2: Scaling behavior of representative multilingual LLMs on five Southeast Asian languages. The x-axis shows model size in billions of parameters on a log scale, and the y-axis reports Overall Score, computed as the average of mean sacreBLEU and mean BERTScore across five languages. Parameter counts for closed-source models are estimated.

3.2 Main Result

Table 2 presents the full evaluation results across five LRL→Chinese directions. Overall, MERIT-32B delivers the strongest and most balanced performance among the evaluated systems. It achieves the best scores on all ROUGE-L columns and obtains top or second-best results on most sacreBLEU, METEOR, and BERTScore columns. This indicates that MERIT improves both lexical consistency and semantic adequacy, rather than optimizing for a single metric family.

Compared with its Qwen-3 32B backbone, MERIT-32B yields substantial gains on the most difficult directions. For Lao, sacreBLEU improves from 27.31 to 33.84 and BERTScore from 54.18 to 62.14. For Burmese, sacreBLEU improves from 24.88 to 34.24 and BERTScore from 50.12 to 61.43.

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-3 32B	45.62	48.14	28.47	26.19	46.88	44.06	46.92	27.31	24.88	45.24	35.34	32.14	13.11	24.16	29.46	✗	✓
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-3 8B	35.41	42.67	13.12	12.34	40.34	34.23	41.15	12.44	11.27	39.12	32.91	29.87	10.42	18.74	28.22	✗	✓
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	✗	✓
Qwen-2.5 3B	15.83	24.24	1.16	1.82	17.09	18.54	24.12	1.05	1.88	19.34	11.42	19.31	2.84	3.67	22.18	✗	✓
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	✓	✓
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-32B	47.82	**49.54**	34.12	33.46	**50.42**	45.32	**48.16**	33.84	34.24	**48.24**	**36.24**	**32.84**	**14.28**	**25.34**	**31.14**	✓	✓
MERIT-8B	43.17	48.92	26.83	24.56	49.13	40.89	47.34	25.61	22.43	46.93	34.72	32.36	12.91	23.12	29.83	✓	✓
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-3 32B	39.66	41.82	24.13	22.17	40.43	68.12	70.94	51.34	49.73	70.62	69.24	72.24	54.18	50.12	72.16	✗	✓
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-3 8B	31.12	36.14	11.04	11.31	35.81	57.13	64.63	26.22	28.06	65.14	57.91	64.92	27.34	29.11	65.43	✗	✓
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	✗	✓
Qwen-2.5 3B	8.83	17.71	2.05	3.03	19.35	36.18	34.52	9.14	11.23	26.84	38.42	36.15	11.45	16.12	23.56	✗	✓
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	20.48	19.68	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	✓	✓
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	21.68	10.04	6.34	9.51	✓	✓
MERIT-32B	41.16	**42.44**	29.42	30.18	**42.16**	69.42	**71.86**	59.34	60.12	**72.34**	70.82	**73.84**	62.14	61.43	**73.68**	✓	✓
MERIT-8B	35.84	42.12	21.36	19.42	40.92	63.48	71.32	39.84	43.21	71.14	64.72	73.12	45.92	44.67	72.62	✓	✓

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. **Bold** values indicate the highest score, and underlined values indicate the second-highest score across all models. FT: Fine-tuned; OS: Open Source.

These large gains suggest that reward-informed data curation is particularly useful when direct supervision is sparse and noisy. In contrast, the improvements on higher-resource or better-aligned directions such as Indonesian and Vietnamese are smaller but still consistent.

Closed-source models remain strong competitors. Gemini-2.5 Flash achieves the best BLEU-4 and chrF scores on several languages, especially Filipino, Lao, and Burmese, showing that proprietary LLMs retain strong zero-shot lexical generation ability. However, MERIT-32B provides stronger average behavior under the Overall Score and significantly improves over GPT-4o in paired bootstrap testing. This suggests that Chinese-centric reward-informed optimization can narrow the gap to strong proprietary systems under the evaluated LRL→Chinese protocol.

MERIT-8B further confirms that the framework is not effective only at 32B scale. Compared with Qwen-3 8B, MERIT-8B improves the Overall Score from 38.29 to 48.43, corresponding to a gain of 10.14 points. The improvement is especially large for Lao and Burmese, where semantic metrics increase sharply. This pattern indicates that MERIT benefits smaller backbones by improving data quality and reward alignment, rather than relying solely on parameter scaling.

3.3 Module Ablation

Table 3 analyzes the contribution of each MERIT component. For Qwen-3 8B, SFT-LTP raises the Overall Score from 38.29 to 41.59, showing that language-aware supervised fine-tuning already provides a stable base for LRL→Chinese translation. Adding GRPO-SAR further improves the score to 45.06, indicating that reward-aligned optimization helps the model better exploit expert-informed quality signals. The full LLME stage achieves 48.43, yielding a total gain of 10.14 points over the baseline.

The same progressive trend is observed for the 32B backbone. Qwen-3 32B starts from an Overall Score of 50.64. SFT-LTP improves it to 52.09, GRPO-SAR further improves it to 53.62, and the full LLME stage reaches 55.17. Although the absolute gain is smaller than that of the 8B model, the improvement is still consistent across all five languages. This suggests that MERIT remains beneficial even when the base model is already strong.

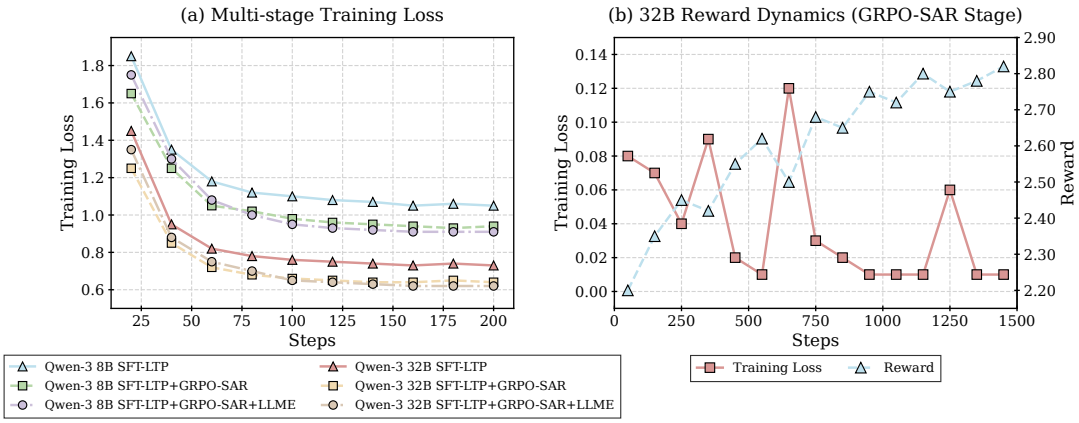

Figure 3: Training dynamics of MERIT. (a) Multi-stage training loss for Qwen-3 8B and 32B under SFT-LTP, GRPO-SAR, and LLME. (b) Training loss and group-summed SAR return during the Qwen-3 32B GRPO-SAR stage. Reward is computed by aggregating SAR rewards over the sampled candidate group.

Model	sacreBLEU					BERTScore					Overall Score
	fil	id	lo	my	vi	fil	id	lo	my	vi	
Qwen-3 8B	34.23	41.15	12.44	11.27	39.12	57.91	64.92	27.34	29.11	65.43	38.29
+ SFT-LTP	36.35$_{\uparrow 2.12}$	43.47$_{\uparrow 2.32}$	16.69$_{\uparrow 4.25}$	15.11$_{\uparrow 3.84}$	41.57$_{\uparrow 2.45}$	60.06$_{\uparrow 2.15}$	67.34$_{\uparrow 2.42}$	33.46$_{\uparrow 6.12}$	34.15$_{\uparrow 5.04}$	67.74$_{\uparrow 2.31}$	41.59$_{\uparrow 3.30}$
+ GRPO-SAR	38.58$_{\uparrow 4.35}$	45.58$_{\uparrow 4.43}$	21.17$_{\uparrow 8.73}$	18.78$_{\uparrow 7.51}$	44.28$_{\uparrow 5.16}$	62.42$_{\uparrow 4.51}$	70.53$_{\uparrow 5.61}$	39.68$_{\uparrow 12.34}$	39.34$_{\uparrow 10.23}$	70.24$_{\uparrow 4.81}$	45.06$_{\uparrow 6.77}$
+ LLME	40.89$_{\uparrow 6.66}$	47.34$_{\uparrow 6.19}$	25.61$_{\uparrow 13.17}$	22.43$_{\uparrow 11.16}$	46.93$_{\uparrow 7.81}$	64.72$_{\uparrow 6.81}$	73.12$_{\uparrow 8.20}$	45.92$_{\uparrow 18.58}$	44.67$_{\uparrow 15.56}$	72.62$_{\uparrow 7.19}$	48.43$_{\uparrow 10.14}$
Qwen-3 32B	44.06	46.92	27.31	24.88	45.24	69.24	72.24	54.18	50.12	72.16	50.64
+ SFT-LTP	44.47$_{\uparrow 0.41}$	47.35$_{\uparrow 0.43}$	29.45$_{\uparrow 2.14}$	28.00$_{\uparrow 3.12}$	46.29$_{\uparrow 1.05}$	69.75$_{\uparrow 0.51}$	72.78$_{\uparrow 0.54}$	56.52$_{\uparrow 2.34}$	53.57$_{\uparrow 3.45}$	72.68$_{\uparrow 0.52}$	52.09$_{\uparrow 1.45}$
+ GRPO-SAR	44.90$_{\uparrow 0.84}$	47.74$_{\uparrow 0.82}$	31.56$_{\uparrow 4.25}$	31.12$_{\uparrow 6.24}$	47.36$_{\uparrow 2.12}$	70.29$_{\uparrow 1.05}$	73.36$_{\uparrow 1.12}$	59.30$_{\uparrow 5.12}$	57.36$_{\uparrow 7.24}$	73.18$_{\uparrow 1.02}$	53.62$_{\uparrow 2.98}$
+ LLME	45.32$_{\uparrow 1.26}$	48.16$_{\uparrow 1.24}$	33.84$_{\uparrow 6.53}$	34.24$_{\uparrow 9.36}$	48.24$_{\uparrow 3.00}$	70.82$_{\uparrow 1.58}$	73.84$_{\uparrow 1.60}$	62.14$_{\uparrow 7.96}$	61.43$_{\uparrow 11.31}$	73.68$_{\uparrow 1.52}$	55.17$_{\uparrow 4.53}$

Table 3: Ablation studies of progressive stages (SFT-LTP, GRPO-SAR, and LLME) across Qwen-3 8B and 32B backbones evaluated on five Southeast Asian languages. Subscripts indicate gains over the corresponding baseline. Overall Score is computed as the average of sacreBLEU and BERTScore across the five languages.

The largest ablation gains occur on Lao and Burmese. For the 8B backbone, the full MERIT pipeline improves BERTScore by 18.58 points on Lao and 15.56 points on Burmese. For the 32B backbone, MERIT improves sacreBLEU by 6.53 points on Lao and 9.36 points on Burmese, while BERTScore increases by 7.96 and 11.31 points, respectively. These results support the hypothesis that reward-informed data distillation is most beneficial for the most challenging low-resource directions.

3.4 Training Dynamics

Figure 3 provides additional evidence for the stability and effectiveness of the progressive MERIT training pipeline. As shown in Figure 3(a), both Qwen-3 8B and Qwen-3 32B exhibit clear loss reduction as training proceeds, but the 32B backbone consistently converges to a lower loss region. For the 8B model, SFT-LTP reduces the loss rapidly in the early stage and gradually stabilizes around 1.05, while the addition of GRPO-SAR and LLME further lowers the converged loss to approximately 0.94 and 0.91, respectively. For the 32B model, the same progressive pattern is more pronounced: SFT-LTP converges around 0.73, GRPO-SAR around 0.64, and the full LLME stage around 0.62. This monotonic improvement across stages is consistent with the ablation results in Table 3, suggesting that each MERIT component contributes to more stable optimization rather than producing isolated metric gains.

Figure 3(b) further illustrates the behavior of the GRPO-SAR stage on the 32B backbone. Although the training loss exhibits small fluctuations due to policy exploration, the reward signal increases steadily from roughly 2.20 to above 2.80 over the course of training. This divergence between short-term loss oscillation and long-term reward improvement is expected in reward-based optimization, where exploration may temporarily increase loss while improving alignment with the reward objective. The overall upward trajectory of the group-summed SAR return indicates that GRPO-SAR successfully guides the QE agent toward expert-aligned scoring behavior, which in turn supports the reward-informed data distillation used in the final MERIT pipeline.

3.5 Reward-Informed Data Distillation

To assess the effect of data distillation, we compare the retained training data size and the resulting

Method	Training Size (Parallel Sentences)					Overall Score (Total Size / Score)
	fil	id	lo	my	vi	
+ SFT-LTP	8,000	8,000	8,000	8,000	8,000	40,000 / 52.09
+ GRPO-SAR	2,423$_{\downarrow 69.7\%}$	2,148$_{\downarrow 73.2\%}$	2,845$_{\downarrow 64.4\%}$	3,411$_{\downarrow 57.4\%}$	1,848$_{\downarrow 76.9\%}$	12,675$_{\downarrow 68.3\%}$ / 53.62$_{\uparrow 2.9\%}$
+ LLME	4,213$_{\downarrow 47.3\%}$	3,921$_{\downarrow 51.0\%}$	4,853$_{\downarrow 39.3\%}$	5,429$_{\downarrow 32.1\%}$	3,377$_{\downarrow 57.8\%}$	21,793$_{\downarrow 45.5\%}$ / 55.17$_{\uparrow 5.9\%}$

Table 4: Training size comparison and corresponding Overall Scores for **MERIT-32B**. Language-wise reduction percentages are computed relative to the corresponding 8,000-sentence training subset for each language, while total-corpus reduction percentages are computed relative to the full 40,000-sentence multilingual pool. Overall Score improvements are computed relative to the SFT-LTP stage.

Overall Score across progressive stages. Data filtering and selection have long been effective for improving noisy parallel corpora (Moore and Lewis, 2010; Junczys-Dowmunt, 2018). MERIT extends this idea by using expert-aligned reward signals to select semantically reliable training instances.

As shown in Table 4, the initial SFT-LTP stage uses 40,000 training instances and obtains an Overall Score of 52.09. GRPO-SAR retains only 12,675 instances, corresponding to 31.7% of the original pool, yet improves the Overall Score to 53.62. This demonstrates that reward-informed selection can remove noisy or low-utility pairs while retaining examples that contribute more effectively to translation quality. The LLME stage expands the candidate pool with strong LLM-generated translations and then applies the same QE-based filtering mechanism. Although the retained data size increases to 21,793 instances, it remains substantially smaller than the original SFT-LTP pool and achieves the best Overall Score of 55.17. This indicates that the quality of selected examples matters more than raw corpus size, and that LLM-enhanced candidates are useful only when filtered by expert-aligned quality estimation.

4 Discussion

Our results highlight the need for direct Chinese-centric evaluation in low-resource translation. Although massively multilingual models offer broad language coverage, their performance remains uneven in LRL→Chinese settings, especially for Lao and Burmese. By reducing dependence on English-pivot evaluation, CALT provides a more suitable testbed for analyzing cross-lingual transfer into Chinese. The progressive gains from MERIT show that reward-informed data curation effectively complements model scaling. GRPO-SAR and LLME improve performance even after pruning the training pool, suggesting that data quality and expert alignment are critical in noisy low-resource regimes. By linking expert-aligned quality estimation with data distillation and model optimization, MERIT consistently improves both 8B and 32B backbones.

These findings also point to several limitations and future directions. Lao and Burmese still pose script-specific challenges, including tokenization instability, encoding variation, and orthographic inconsistency, which may affect both training and evaluation. The current SAR objective mainly aligns QE predictions with expert adequacy scores, leaving fluency, terminology consistency, hallucination avoidance, and Chinese naturalness to be modeled more explicitly in future multi-objective rewards. In addition, closed-source models are evaluated under a controlled zero-shot protocol for comparability, and stronger prompting or in-context learning may further improve their performance. Extending CALT and MERIT to more typologically diverse languages remains an important direction. Further analysis is provided in Appendix A.

5 Conclusion

This work introduces MERIT, a Chinese-centric framework for low-resource Southeast Asian languages → Chinese translation. By reconstructing CALT for direct evaluation and integrating language-specific prefixing, supervised fine-tuning, GRPO-SAR, and reward-informed data distillation, MERIT improves both data quality and model alignment under limited supervision. Experiments across five LRL→Chinese directions show that MERIT achieves strong and consistent gains, with MERIT-32B obtaining the best overall open-source performance and statistically significant average improvements over GPT-4o. These results demonstrate that expert-aligned data curation and reward-guided optimization provide an effective and reproducible path for improving Chinese-centric low-resource translation beyond model scaling alone.

Limitations

Despite the encouraging results achieved by our proposed framework and the MERIT-32B model, this work has several limitations that warrant discussion and offer avenues for future improvement.

Limited Linguistic Coverage. CALT currently covers five Southeast Asian LRL→Chinese directions: Filipino/Tagalog, Indonesian, Lao, Burmese, and Vietnamese. This scope enables controlled comparison across languages with reliable ALT alignments, but it does not cover all low-resource languages relevant to Chinese-centric translation. Future work should extend the benchmark to additional Southeast Asian languages such as Khmer, as well as broader Chinese-facing low-resource directions such as Tibetan, Uyghur, and Kazakh, when high-quality parallel data and language-specific preprocessing resources become available.

Residual English-Centric Bias in the Test Suite. Although the ALT-based test suite has been realigned for LRL-Chinese evaluation using shared `alt_id` indexing, it remains inherently constrained by its original English-centric design, potentially retaining subtle domain-specific or stylistic artifacts that affect translation assessment.

Insufficient Scale of Human Validation: While the QE agent and statistical-semantic SAR function rely on automatic filtering, the volume of human validation, particularly the expert-annotated data employed in reward model training, remains substantially limited. This constraint may compromise the robustness and alignment of the SAR model with diverse human preferences.

Fixed Decoding and Prompting Strategies. Although multiple LLMs were evaluated under zero-shot, SFT, and GRPO settings, decoding hyperparameters and prompt formats were uniformly fixed for fair comparison, potentially masking model-specific performance differences that could be uncovered through more extensive hyperparameter and prompt engineering.

Reward Scope. The current SAR objective mainly aligns QE predictions with expert adequacy-oriented quality scores. Although this provides a stable reward signal for data selection, it may not fully capture fluency, stylistic naturalness in Chinese, terminology consistency, or hallucination avoidance. Extending SAR to a multi-objective reward would provide a more comprehensive approximation of human translation preferences.

Ethics Statement

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 Asian Language 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 procedures for removing personally identifiable information (PII). 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., 2021).

Environmental Impact. Model training and inference were conducted on four NVIDIA A100 GPUs (40 GB). 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 (EMNLP 2023 Program Committee, 2023; Li and Lu, 2026).

Intended Use and Misuse Risks. The benchmark is designed to support objective evaluation and supervised training for LRL→Chinese 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., 2019).

References

Mikel Artetxe and Holger Schwenk. 2018. 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 Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

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.

DeepMind Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical Report v2.5, Google DeepMind. Accessed: 2025-12.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36:10088–10115.

David M. Eberhard, Gary F. Simons, and Alison J. Robinson, editors. 2026. *Ethnologue: Languages of the World*, 29 edition. SIL Global, Dallas, Texas. Online version.

EMNLP 2023 Program Committee. 2023. Emnlp 2023 ethics faq. `https://2023.emnlp.org/ethics/faq`. Conference ethics guidelines, accessed May 2024.

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

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*, pages 878–891. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Preprint*, arXiv:1803.09010.

Soumya Suvra Ghosal, Soumyabrata Pal, Koyel Mukherjee, and Dinesh Manocha. 2025. PromptRefine: Enhancing few-shot performance on low-resource Indic languages with example selection from related example banks. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 351–365, Albuquerque, New Mexico. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025. Benchmax: A comprehensive multilingual evaluation suite for large language models. *Preprint*, arXiv:2502.07346.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Preprint*, arXiv:1611.04558.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online. Association for Computational Linguistics.

Daniil Larionov, Sotaro Takeshita, Ran Zhang, Yanran Chen, Christoph Leiter, Zhipin Wang, Christian Greisinger, and Steffen Eger. 2025. Deepseek-r1 vs. o3-mini: How well can reasoning llms evaluate mt and summarization? *Preprint*, arXiv:2504.08120.

Linxiao Li and Zhixiang Lu. 2026. Ecothink: A green adaptive inference framework for sustainable and accessible agents. *Preprint*, arXiv:2603.25498.

Beng Soon Lim and Gloria R. Poedjosoedarmo. 2016. Bahasa indonesia and bahasa melayu: Convergence and divergence of the official languages in contemporary southeast asia. In Gerhard Leitner, Azirah Hashim, and Hans-Georg Wolf, editors, *Communicating with Asia: The Future of English as a Global Language*, pages 170–187. Cambridge University Press, Cambridge.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Preprint*, arXiv:2001.08210.

Lalita Lowphansirikul, Charin Polpanumas, Attapol T. Rutherford, and Sarana Nutanong. 2021. A large english–thai parallel corpus from the web and machine-generated text. *Language Resources and Evaluation*, 56(2):477–499.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229. ACM.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224.

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

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, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

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

Maja Popović. 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.

Matt Post. 2018. A call for clarity in reporting bleu scores. *Preprint*, arXiv:1804.08771.

Ricardo Rei, Craig Stewart, Ana C 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, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez. 2018. Prompsit's submission to WMT 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 955–962, Belgium, Brussels. Association for Computational Linguistics.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *Preprint*, arXiv:1512.00567.

Ye Kyaw Thu, Win Pa Pa, 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'16)*, pages 1574–1578, Portorož, Slovenia. European Language Resources Association (ELRA).

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

Dehong Xu, Liang Qiu, Minseok Kim, Faisal Ladhak, and Jaeyoung Do. 2024. Aligning large language models via fine-grained supervision. *Preprint*, arXiv:2406.02756.

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. *Preprint*, arXiv:2010.11934.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Appendix

A Further Discussion

Script-specific Challenges. The main results show that Lao and Burmese remain among the most challenging directions, even after reward-informed data distillation. One reason is that these languages introduce script-specific issues, including tokenization instability, Unicode normalization differences, inconsistent segmentation, and orthographic variation. Such issues may affect both model training and automatic evaluation. Future work could incorporate language-specific normalization, transliteration, or segmentation-aware preprocessing before quality estimation and fine-tuning. Another promising direction is to design QE models that are less sensitive to superficial encoding differences while remaining sensitive to genuine semantic mismatches.

Reward Design. The current Semantic Alignment Reward is designed to align QE predictions with expert quality scores. This provides a stable and interpretable reward signal, but it may emphasize adequacy more strongly than other aspects of translation quality. In some cases, a translation may preserve the core meaning while still exhibiting fluency issues, terminology inconsistency, or unnatural Chinese style. Future work could extend SAR into a multi-objective reward that separately models adequacy, fluency, terminology consistency, hallucination avoidance, and stylistic naturalness. Such a reward could better reflect human evaluation practices and reduce the risk of optimizing for a single quality dimension.

Closed-source Model Prompting. Our comparison with closed-source LLMs follows a controlled zero-shot protocol to ensure reproducibility and avoid prompt overfitting. However, this setting may underestimate the full capability of proprietary models. More advanced prompting strategies, including few-shot demonstrations, retrieval-augmented prompting, chain-of-thought-style translation planning, or terminology-constrained prompts, may improve their LRL→Chinese performance. Exploring the interaction between MERIT-style data distillation and stronger prompting strategies is a promising future direction.

Generalization Beyond Southeast Asian LRLs. MERIT is evaluated on five Southeast Asian LRL→Chinese directions in CALT. Although these languages cover diverse scripts and resource conditions, they do not exhaust the typological diversity of low-resource translation. Extending CALT to additional languages, such as Khmer or other underrepresented regional languages, would provide a stronger test of MERIT's generality. This extension may require language-specific filtering and normalization pipelines, especially for languages with noisier alignment resources.

Data Quality versus Data Quantity. The data distillation results in Table 4 suggest that reducing the training pool can improve performance when filtering is guided by expert-aligned rewards. This finding does not imply that smaller datasets are universally better, rather, it shows that in noisy low-resource settings, unfiltered data may introduce harmful supervision. The effectiveness of MERIT depends on whether the QE agent can reliably identify examples with high semantic adequacy and fluency. Future work should further analyze threshold sensitivity, reward calibration, and the trade-off between recall and precision in distilled data selection.

B Implementation Details

All experiments were conducted on a local server equipped with four NVIDIA A100 GPUs (40 GB each). We used distributed data parallel training with four-way GPU parallelism. Unless otherwise specified, the per-GPU batch size was set to 4 and gradient accumulation was set to 2, resulting in an effective batch size of 32. The maximum input sequence length was set to 1024 tokens, and the initial learning rate was set to `2e-4`. The training environment used Ubuntu 20.04, CUDA 12.1, Python 3.10, PyTorch 2.1, and Transformers v4.49.

All training stages were performed with mixed-

precision computation. For larger backbones, including the 32B setting, we used memory-efficient training configurations with gradient checkpointing and parameter-efficient adaptation where applicable to fit the model within the available GPU memory. Batch size and accumulation steps were adjusted when necessary to keep the effective batch size consistent across model scales. For reproducibility, we fixed random seeds for data shuffling, model initialization, and decoding, and maintained detailed runtime logs for each training run, including hardware configuration, hyperparameters, retained data size, and evaluation outputs.

To address heavy-duty scaling constraints and fit the larger backbones (including the 32B variant) within the available 4× NVIDIA A100 (40 GB) infrastructure, we employ Parameter-Efficient Fine-Tuning (PEFT) via Quantized Low-Rank Adaptation (QLoRA) (Dettmers et al., 2023).

The base Transformer backbones are frozen and quantized into 4-bit NormalFloat (NF4) format with double quantization enabled, while the standard computation is conducted using `bfloat16` precision to maintain training stability. We inject trainable rank-stabilized LoRA adapters into all major linear projection modules across the Transformer architecture, specifically targeting the query (`q_proj`), key (`k_proj`), value (`v_proj`), and output (`o_proj`) layers within the self-attention block, as well as the intermediate layers of the feed-forward networks (`gate_proj`, `up_proj`, and `down_proj`).

The QLoRA hyperparameters are uniformly set across both the 8B and 32B scales: the adapter rank r is set to 16, the scaling factor α is set to 32, and the LoRA dropout rate is fixed at 0.05. This parameter-efficient setup restricts the total number of trainable parameters to approximately 0.85% of the total backbone capacity, drastically reducing the GPU VRAM footprint while preventing over-fitting on our localized data distribution. Furthermore, gradient checkpointing is activated throughout all SFT and GRPO training stages to further optimize memory efficiency.

C Reward Function

We introduce the Semantic Alignment Reward (SAR) to align the Quality Estimation (QE) agent's predictions with human expert judgments. Given an extracted QE score s_i and the ground-truth expert score a_i, we define the absolute deviation as:

$$d_i = |s_i - a_i|. \tag{23}$$

The SAR function maps this deviation to a stepwise reward:

$$R_i = \begin{cases} 2.0, & d_i = 0, \\ 1.0, & 1 \le d_i \le 10, \\ 0.0, & d_i > 10 \text{ or } s_i < 0. \end{cases} \tag{24}$$

Here, $s_i < 0$ indicates that the QE output does not contain a valid `Score:` field and is therefore treated as invalid. This design is motivated by two considerations. First, quality estimation is inherently subjective: different expert annotators may assign slightly different scalar scores to translations of similar quality. The 10-point tolerance window therefore avoids over-penalizing minor score deviations that do not reflect substantial quality differences. Second, compared with sparse exact-match rewards, SAR provides a denser supervision signal by assigning partial credit to approximately aligned predictions. This helps stabilize GRPO optimization and encourages the QE agent to approximate expert preference distributions rather than overfit to rigid scalar values.

The SAR reward in Eq.(24) is defined for a single candidate evaluation output. In Figure 3(b), we report the batch-averaged group-summed SAR return used for monitoring GRPO-SAR training:

$$\bar{R}_t^{\text{group}} = \frac{1}{|B_t|} \sum_{i \in B_t} \sum_{j=1}^{G} R_{i,j},$$

where B_t denotes the training batch at step t, and G is the number of sampled evaluation outputs for each input. This quantity is used only as a diagnostic curve in Figure 3(b). Since it sums SAR rewards over the G sampled outputs before batch averaging, its range is $[0, 2G]$, rather than the single-candidate SAR range of $[0, 2]$. In our experiments, we set $G = 2$ for GRPO-SAR.

D 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, three annotators were assigned, and a cross-review protocol was adopted to enhance annotation quality and consistency. All participants had formal training in

Combination	$\rho \uparrow$	$\sigma^2 \downarrow$
$(\text{sacreBLEU} + \text{BERTScore})/2$ (equal)	**0.988**	**1.05**
$\sqrt{\text{sacreBLEU} \times \text{BERTScore}}$ (geometric)	0.986	1.18
$0.4 \cdot \text{sacreBLEU} + 0.6 \cdot \text{BERTScore}$	0.987	1.12
$0.6 \cdot \text{sacreBLEU} + 0.4 \cdot \text{BERTScore}$	0.985	1.25

Note: ρ denotes Spearman's rank correlation between each combined metric and the reference system-level ranking. σ^2 denotes the variance of ranking deviations across the five languages. sacreBLEU is computed with version 2.4.2, and BERTScore uses the multilingual *xlm-roberta-large* checkpoint. All scores are normalized to the same 0-100 scale before aggregation.

Table 5: Comparison of different combination strategies for aggregating sacreBLEU and BERTScore into an overall score. The equal-weighted arithmetic mean is used as the default Overall Score in our main analysis.

Metric	Mean Δ (vs. GPT-4o)	p-value
BLEU-4	+3.74	**0.014**
sacreBLEU	+3.80	**0.012**
ROUGE-L	+2.13	**0.045**
chrF	+2.86	**0.021**
METEOR	+3.05	**0.032**
BERTScore	+3.75	**0.015**

Note: Mean Δ represents the score difference between **MERIT-32B** and GPT-4o, averaged across all five Southeast Asian test sets. p-values are obtained using paired bootstrap resampling over sentence-level outputs with 1,000 iterations. **Bold** p-values indicate statistically significant improvements ($p < 0.05$).

Table 6: Statistical significance of MERIT-32B improvements over GPT-4o.

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.

E Data Filtering Methodology

To improve the quality of the Chinese-centric low-resource corpora, we use a hybrid filtering pipeline that combines surface-level statistical heuristics with model-based semantic scoring. Let $\mathcal{D}_{raw} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the initial noisy LRL–Chinese parallel corpus. EPDS first removes obviously invalid sentence pairs, then assigns each remaining pair a composite quality score.

Algorithm 1: Elite Parallel Data Sampler

Input: $\mathcal{D}_{raw}$, Target Size K, LLM $\mathcal{M}$
Output: $\mathcal{D}_{clean}$

$\mathcal{D}_{valid} \leftarrow \emptyset$
foreach $(x, y) \in \mathcal{D}_{raw}$ **do**
  if $\mathbb{I}_{filter}(x, y) = 1$ **then**
    $\mathbf{f}_{stat} \leftarrow \text{ExtractFeatures}(x, y)$
    $S_{base} \leftarrow \mathbf{w}^T \cdot \Phi(\mathbf{f}_{stat})$
    $\mathcal{D}_{valid} \leftarrow \mathcal{D}_{valid} \cup \{(x, y, S_{base})\}$
foreach *batch* $B \subset \mathcal{D}_{valid}$ **do**
  $\mathbf{S}_{PPL}, \mathbf{S}_{IFD} \leftarrow \mathcal{M}.\text{Forward}(B)$
  foreach $(x, y) \in B$ **do**
    $S_{final}^{(x,y)} \leftarrow \alpha S_{base} + \beta S_{PPL}^{(x,y)} + \gamma S_{IFD}^{(x,y)}$
$\mathcal{D}_{sorted} \leftarrow \text{Sort}(\mathcal{D}_{valid}, \text{key} = S_{final})$
$\mathcal{D}_{clean} \leftarrow \{d \in \mathcal{D}_{sorted} \mid \text{rank}(d) \leq K\}$
return $\mathcal{D}_{clean}$

E.1 Statistical Feature Extraction

For each pair (x_i, y_i), we extract a feature vector $\mathbf{f}_i \in \mathbb{R}^5$ that captures surface alignment quality:

1. **Length ratio** R_{len}: $\min(|y_i|/|x_i|, |x_i|/|y_i|)$, used to penalize extreme length mismatch.

2. **Token ratio** R_{tok}: the ratio between source-side and target-side token counts.

3. **Punctuation divergence** D_{punct}: the absolute difference in punctuation ratios, computed with language-specific regular expressions.

4. **Digit divergence** D_{digit}: the absolute difference in digit proportions.

5. **Lexical diversity difference** D_{uniq}: the difference in type-token ratios.

The base statistical score is a weighted sum with normalizing functions ϕ_k:

$$S_{\text{base}}(x_i, y_i) = \sum_{k=1}^{5} w_k \phi_k(f_i^{(k)}), \qquad (25)$$

where $\phi_k(\cdot)$ normalizes each feature to a comparable range and w_k denotes its feature weight.

E.2 Model-Based Semantic Scoring

Surface statistics alone cannot reliably identify semantic mismatches. We therefore use a lightweight

instruction-tuned scoring model $\mathcal{M}_{score}$ to compute two additional signals: a conditional perplexity score and an instruction-following discrepancy score.

The conditional perplexity score rewards fluent and contextually likely target sentences:

$$S_{\mathrm{PPL}}(x_i, y_i) = \frac{1}{1 + \lambda \cdot \mathrm{PPL}(y_i \mid x_i, \mathcal{I})}, \quad (26)$$

where $\mathcal{I}$ is the translation instruction and $\lambda > 0$ is a scaling coefficient. This formulation avoids ambiguity with the sigmoid notation and ensures that lower perplexity leads to a higher score.

The instruction-following discrepancy score measures whether the target sentence relies on the source context:

$$S_{\mathrm{IFD}}(x_i, y_i) = \min\left(\frac{\mathrm{PPL}(y_i \mid \emptyset)}{\mathrm{PPL}(y_i \mid x_i, \mathcal{I})}, \tau\right) \cdot \tau^{-1}, \quad (27)$$

where τ is a normalization threshold. A higher value indicates that the target sentence is better explained when conditioned on the source, which helps filter out generic or weakly aligned targets.

E.3 Composite Filtering Score

The final filtering score combines statistical, fluency, and context-reliance signals:

$$S_{final} = \alpha S_{base} + \beta S_{PPL} + \gamma S_{IFD}. \quad (28)$$

Unless otherwise specified, we set $(\alpha, \beta, \gamma) = (0.3, 0.3, 0.4)$ based on development-set calibration. Sentence pairs are ranked by S_{final} in descending order, and the top-K pairs are retained for subsequent validation.

E.4 Data Integrity Validation

To ensure reliable benchmark construction, we apply Data Integrity Validation (DIV) after EPDS filtering. DIV addresses two requirements. First, it preserves the language-wise and domain-wise distribution of the filtered corpus as closely as possible. Second, it enforces exact split sizes for training, development, and test sets without allowing overlap across splits.

Let $\mathcal{D}_{clean}$ be the filtered corpus. For each sentence pair s, we use metadata to obtain its language identifier $\mathrm{lang}(s)$ and domain identifier $d(s)$. DIV is applied within each language to avoid cross-language imbalance. For language ℓ, the filtered set is divided into domain buckets $\{\mathcal{G}_k^\ell\}_{k=1}^M$, and

Algorithm 2: Data Integrity Validation

Input: $\mathcal{D}_{clean}$, Targets $\{N_{train}, N_{dev}, N_{test}\}$
Output: Splits $\{S_{train}, S_{dev}, S_{test}\}$

$\mathcal{G} \leftarrow \{(k, \{s \mid d(s) = k\})\}_{k=1}^M$
$\mathbf{P}[k] \leftarrow |\mathcal{G}[k]|/|\mathcal{D}_{clean}|, \quad \forall k$
foreach *split* $T \in \{train, dev, test\}$ **do**
  $S_T \leftarrow \emptyset$
  foreach *domain* $k \in$ *Keys*$(\mathcal{G})$ **do**
    $q_k \leftarrow \lfloor N_T \cdot \mathbf{P}[k] \rfloor$
    $B_k \leftarrow \mathrm{Sample}(\mathcal{G}[k], q_k)$
    $S_T \leftarrow S_T \cup B_k$;
    $\mathcal{G}[k] \leftarrow \mathcal{G}[k] \setminus B_k$
  $\Delta \leftarrow N_T - |S_T|$
  if $\Delta > 0$ **then**
    $U \leftarrow \bigcup_k \mathcal{G}[k]$
    $S_{supp} \leftarrow \mathrm{Resample}(U, \Delta)$
    $S_T \leftarrow S_T \cup S_{supp}$
  else if $|S_T| > N_T$ **then**
    $S_T \leftarrow S_T[1 : N_T]$
if $\exists T, |S_T| \neq N_T$ **then return Error**
return $\{S_{train}, S_{dev}, S_{test}\}$

the empirical domain distribution is:

$$p_k^\ell = \frac{|\mathcal{G}_k^\ell|}{|\mathcal{D}^\ell|}. \quad (29)$$

For a target split T with size N_T^ℓ, the initial domain quota is:

$$q_{T,k}^\ell = \left\lfloor N_T^\ell p_k^\ell \right\rfloor. \quad (30)$$

Since flooring may leave unassigned slots, DIV uses a largest-remainder compensation strategy. Remaining slots are assigned to domains with the largest fractional remainders, subject to sample availability. All sampling is performed without replacement, which prevents train-dev-test leakage. If the available pool cannot satisfy the target size, the procedure returns an error rather than silently duplicating samples.

This procedure approximately preserves domain distributions while guaranteeing exact split sizes and non-overlapping splits. In our experiments, we use fixed random seeds for all sampling operations to ensure reproducibility.

Source Lang.	Annotators	Fleiss' κ	Level
Filipino (fil)	3	0.78	Substantial
Indonesian (id)	3	0.82	Almost perfect
Vietnamese (vi)	3	0.76	Substantial
Lao (lo)	3	0.65	Substantial
Burmese (my)	3	0.61	Substantial
Macro-avg.	–	**0.72**	**Substantial**

Table 7: Inter-annotator agreement on the expert-annotated development set. Fleiss' κ is computed after discretizing 0–100 quality scores into five ordinal bins.

F Further Evaluation and Analysis

F.1 Inter-Annotator Agreement

To quantify the reliability of our cross-review protocol, we compute Fleiss' κ on the expert-annotated development set. Since the original QE scores are defined on a 0–100 scale, we discretize them into five ordinal quality bins before computing agreement. As shown in Table 7, the macro-average agreement reaches a substantial level ($\kappa = 0.72$). Agreement on Lao and Burmese is lower than that of the other languages, likely due to script variation, tokenization instability, and orthographic inconsistency. This observation further motivates the use of SAR with a tolerance window rather than an exact-match reward.

F.2 Sensitivity of Distillation Threshold

We analyze the sensitivity of the reward-informed data distillation threshold τ during the GRPO-SAR stage. The threshold controls the trade-off between retaining diverse training examples and filtering noisy or low-utility pairs. As shown in Figure 4, a low threshold retains more data but yields weaker performance, suggesting that weakly aligned examples remain in the training pool. In contrast, overly high thresholds remove too many useful examples and reduce the Overall Score. The selected threshold $\tau = 85$ achieves the best Overall Score among the evaluated settings, balancing data quality and coverage.

F.3 Data Scale and Sampling Strategy

To verify that MERIT's gains are attributable to reward-informed data selection rather than data size alone, we conduct a controlled sampling experiment. All subsets are constrained to the same retained sizes as the corresponding MERIT stages. As shown in Figure 5, reducing the training data from 40,000 to 12,675 instances through random sampling substantially lowers the Overall Score to 49.85. EPDS-only selection mitigates this degradation but remains below the full SFT-LTP baseline. In contrast, the GRPO-SAR selected subset achieves 53.62 despite using only 31.7% of the initial training pool. A similar pattern is observed in the LLME stage, where random inclusion of LLM-generated candidates is less effective than QE-guided selection. These results indicate that the improvements are largely attributable to reward-informed selection quality rather than data volume.

Selection Method	Overall
Random selection	49.85
Moore-Lewis CED (Moore and Lewis, 2010)	50.51
Dual cross-entropy filtering (Junczys-Dowmunt, 2018)	51.13
EPDS-only heuristic filtering	51.22
LASER margin filtering (Artetxe and Schwenk, 2018)	51.49
LaBSE margin filtering (Feng et al., 2022)	51.82
COMETKiwi filtering (Rei et al., 2022)	52.17
QE scoring without GRPO	52.44
GRPO with format-only reward	52.63
GRPO-SAR (Ours)	**53.62**

Table 8: Comparison with representative data selection and QE baselines. All methods select 12,675 training instances from the same candidate pool. Overall Score is computed as the average of mean sacreBLEU and mean BERTScore across five languages.

F.4 Comparison with Data Selection Baselines

To verify that MERIT's improvements are attributable to reward-informed selection rather than generic filtering, we compare GRPO-SAR with representative data selection and quality estimation baselines. These include cross-entropy-based selection, semantic embedding-based filtering, and QE-based filtering. All methods select exactly 12,675 sentence pairs from the same candidate

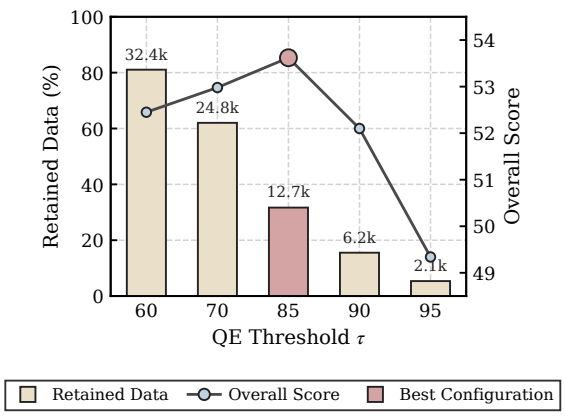

Figure 4: Sensitivity of the QE threshold τ during GRPO-SAR data distillation. Bars indicate retained data ratio relative to the 40,000-instance SFT-LTP pool, and the line indicates Overall Score.

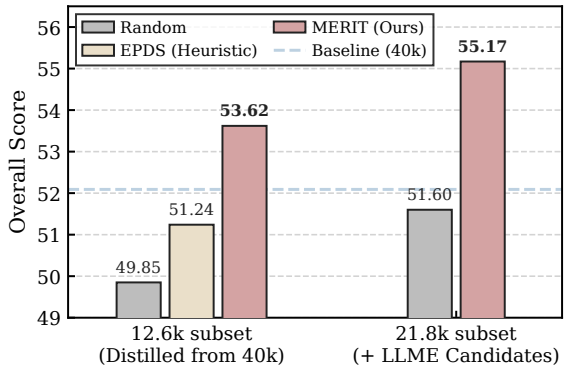

Figure 5: Controlled comparison of data scale and sampling strategy. MERIT-style reward-informed selection achieves higher Overall Score than random or heuristic selection under the same data budget.

pool, matching the GRPO-SAR data size. This controlled setting isolates the effect of selection quality from data scale. Table 8 shows that semantic and QE-based filtering methods outperform random and purely heuristic selection, confirming the importance of data quality under the same data budget. COMETKiwi and QE scoring provide stronger baselines than cross-entropy or embedding-based filtering, but they remain below GRPO-SAR. This suggests that expert-aligned reward optimization improves selection beyond static scoring alone.

G CALT Dataset Splits and Downsampling Motivation

The CALT benchmark is reconstructed from the publicly available Asian Language Treebank (ALT). Although ALT provides multilingual sentence-level alignments through shared sentence identifiers, the amount and quality of usable LRL–Chinese pairs can vary across source languages after filtering. To support stable many-to-one multilingual training and to avoid allowing relatively better-resourced or cleaner directions to dominate the optimization process, we apply balanced downsampling after the initial EPDS filtering stage. For each source language $\ell \in \{\text{fil}, \text{id}, \text{lo}, \text{my}, \text{vi}\}$, the initial training subset is upper-bounded and normalized to exactly 8,000 parallel sentences:

$$|\mathcal{D}^{\ell}_{\text{train}}| = 8{,}000.$$

This yields a balanced multilingual training pool of 40,000 sentence pairs before reward-informed distillation.

We further clarify the percentage annotations reported in Table 4. The reduction percentages attached to individual language columns are computed relative to the corresponding 8,000-sentence single-language training subset. For example, under the GRPO-SAR stage, Filipino retains 2,423 sentence pairs, giving a language-wise reduction of

$$1 - \frac{2{,}423}{8{,}000} = 69.7\%.$$

By contrast, the reduction percentages reported for the total corpus size are computed relative to the full 40,000-sentence multilingual pool. Thus, GRPO-SAR retains 12,675 sentence pairs, corresponding to a total reduction of

$$1 - \frac{12{,}675}{40{,}000} = 68.3\%,$$

while the LLME stage retains 21,793 sentence pairs, corresponding to a total reduction of

$$1 - \frac{21{,}793}{40{,}000} = 45.5\%.$$

This distinction is important because language-wise reductions and total-corpus reductions use different denominators.

H Data Leakage Prevention

To preserve the empirical integrity of evaluation, we enforce strict train–development–test separation throughout benchmark construction, candidate generation, reward-informed data selection, and final model training. The test set $\mathcal{D}_{\text{test}}$ is held out completely and is used only once for final evaluation. No test source sentences or test reference translations are used during EPDS filtering, LLME candidate generation, QE-agent training, reward rescoring, threshold selection, or model fine-tuning.

The training split $\mathcal{D}_{\text{train}}$ is the only split used for LLME candidate augmentation and final translation-model optimization. During the LLME stage, candidate Chinese translations are generated by prompting the teacher LLMs using only source-side sentences from $\mathcal{D}_{\text{train}}$. Reference translations from $\mathcal{D}_{\text{dev}}$ or $\mathcal{D}_{\text{test}}$ are never provided to the teacher LLMs, and no development or test sentence pairs are inserted into the distilled training pool.

The development split $\mathcal{D}_{\text{dev}}$ is used only for model selection, reward calibration, threshold tuning, and expert-aligned QE supervision. In particular, the distillation threshold τ is calibrated on development data, and the QE agent is aligned using

Model Name	Exact Version / Snapshot	API Access Date	Beam Size / Top-p
Closed-Source API baselines			
GPT-4o	`gpt-4o-2024-11-20`	March 2025	Max Tokens (Greedy)
Claude 3.5 Sonnet	`claude-3-5-sonnet-20241022`	March 2025	Max Tokens (Greedy)
Gemini 2.5 Flash	`gemini-2.5-flash-preview-04-17`	April 2025	Max Tokens (Greedy)
Open-Source LLMs			
DeepSeek-V3	`DeepSeek-v3` (Base API)	March 2025	Max Tokens (Greedy)
DeepSeek-R1 (32B/7B)	`DeepSeek-R1-Distill-Qwen-{32B,7B}`	Local Inference	Top-$p = 0.95$
Qwen-3 (32B/8B)	`Qwen/Qwen-3-{32B,8B}`	Local Inference	Max Tokens (Greedy)
Qwen-2.5 (32B/7B/3B)	`Qwen/Qwen-2.5-{32B,7B,3B}-Instruct`	Local Inference	Max Tokens (Greedy)
Traditional NMT Systems			
NLLB-200 (3.3B)	`facebook/nllb-200-3.3B`	Local Inference	Beam Size = 5
M2M-100 (1.2B)	`facebook/m2m100_1.2B`	Local Inference	Beam Size = 5

Table 9: Comprehensive architectural specifications, precise version/snapshot identifiers, API access timelines, and standard decoding configurations for all baseline systems.

expert-annotated development examples. However, development instances are not used as final SFT examples for the translation backbone, and they are not mixed with the training data during LLME augmentation or reward-informed distillation. This protocol prevents train–test contamination while allowing the development split to serve its standard role in calibration and hyperparameter selection.

I Closed-Source Baseline Alignment

We evaluate representative open-source and proprietary LLM baselines, including **Qwen2.5 and 3** (Ghosal et al., 2025), **GPT-4o** (Huang et al., 2025), **Claude-3.5** (Enis and Hopkins, 2024), **Gemini-2.5** (DeepMind Gemini Team, 2025), **DeepSeek** (Huang et al., 2025; Larionov et al., 2025). These models are selected because they represent strong contemporary multilingual LLM families and cover both open-weight and closed-source deployment settings.

Closed-source systems are evaluated under a controlled zero-shot protocol using the same test inputs, target-language instruction, and automatic evaluation pipeline as the open-source baselines. No in-context demonstrations, prompt search, test-set feedback, or reference translations are provided during evaluation. Accordingly, our comparison is not intended to claim general superiority over proprietary frontier systems, whose performance may improve under model-specific prompting or few-shot adaptation. Rather, it assesses whether a localized open-source backbone optimized with expert-aligned reward signals and reward-informed data distillation can match or exceed zero-shot API baselines on the specialized CALT LRL→Chinese benchmark.

I.1 Model Versions and API Access Timelines

Evaluating commercial closed-source models requires documenting static snapshots to prevent empirical drift caused by proprietary continuous updates. Table 9 provides a comprehensive overview of the exact model identifiers, tracking versions, designated API access dates, and their respective core decoding parameters.

I.2 Standardized Zero-Shot Prompting

To evaluate the zero-shot translation capability of both open-source and closed-source large language models uniformly without inducing prompt-overfitting, we enforce a strict, standardized cross-lingual translation prompt. The system instructions and input formats are structured to isolate the source context explicitly. The universal prompt layout applied to GPT-4o, Claude, Gemini, and the Qwen baseline families is detailed below:

Zero-Shot Translation Prompt Template

System Prompt:

```
You are a professional, highly accurate literary and technical translator. Translate the following text directly into fluent and natural Chinese (Simplified). Do not explain your reasoning, do not include any commentary, and do not alter the core semantic meaning.
```

User Prompt Template:

```
Source Language: [Language_Name]
Text to Translate: """[Source_Sentence]"""
Translation (Chinese Only):
```

For the target languages within CALT, `[Language_Name]` is dynamically mapped to *Filipino*, *Indonesian*, *Vietnamese*, *Lao*, or *Burmese* to provide the clear language discrimination boundary required for multi-task alignment.