Better Literary Translation: A Multi-Aspect Data Generation and LLM Training Approach

Zhihao Lin[1] **Ziqi Zhu**[1] **Hao Huang**[1] **Guanghui Wang**[1] **Peiyang He**[*1,2]
[1]Amazon Web Services (AWS) [2]Peking University
{lzhihao, ziqizhu, tonyhh, guanghu, peiyan}@amazon.com

Abstract

Literary translation poses unique challenges due to the scarcity of high-quality annotated data and the need to balance expression fluency with literary effect. We present a multi-aspect iterative refinement framework that generates high-quality translation references and preference data through specialized LLM translators, each targeting a distinct quality dimension. We leverage the generated data for supervised fine-tuning and reinforcement learning. Experiments show that our generated references outperform the original ground truth for SFT by 8.65 CEA100 points. For reinforcement learning, we find that DPO leads to performance degradation in this setting, while leveraging an explicit reward model for GRPO yields an additional 1.51 point improvement. We attribute this to the stability of two-stage training and GRPO's online exploration capability. Our resulting models, LitMT-8B and LitMT-14B, achieve 67.25 and 69.07 CEA100 respectively on the MetaphorTrans English-to-Chinese literary translation benchmark, competitive with Claude Sonnet 4.5 at 68.43, and demonstrate strong generalization to out-of-domain literary work (*i.e.*, O. Henry).

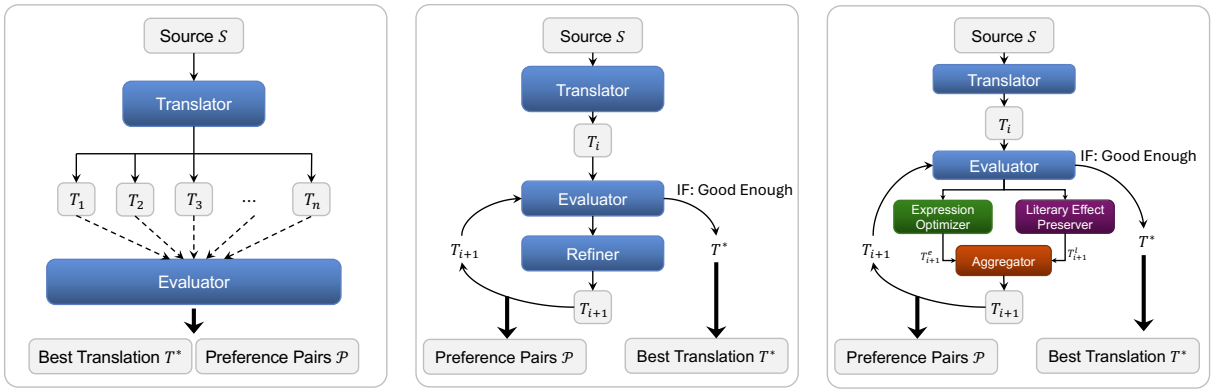
1 Introduction

Literary translation requires balancing expression fluency with literary effect. Unlike general-domain translation, literary texts demand faithful rendering of metaphors, cultural nuances, and stylistic devices (Matusov, 2019; Guerberof-Arenas and Toral, 2022), creating inherent trade-offs between naturalness and artistic preservation. High-quality literary translation enables readers across languages to experience the full artistic intent of original works—a capability of growing industrial importance, yet professional human translation remains costly and slow to scale. While large language models have shown promising results on general translation tasks (Kocmi et al., 2024, 2025), literary translation remains challenging due to both the scarcity of high-quality training data and the difficulty of capturing nuanced quality trade-offs.

Recent approaches address this by training models with long chain-of-thought (CoT) reasoning (Wang et al., 2025a, 2026, 2025b) or using massive LLMs as judges during reinforcement learning (Zheng et al., 2023; Wang et al., 2026, 2025b). While effective, these methods face critical limitations: (1) using LLM-as-a-judge as the reward signal requires expensive inference for every training sample, and this cost must be repeated for each new model, leading to prohibitive training costs; (2) long CoT reasoning increases inference latency by up to 10×, making these approaches impractical for real-time applications.

In this work, we propose a multi-aspect iterative refinement framework that generates both high-quality translation references and preference data. Unlike approaches that require costly LLM inference for every training sample, our generated references and preference pairs can be reused across multiple training runs. Our key insight is that literary translation quality can be decomposed into two dimensions—expression fluency and literary effect—which often conflict. By employing specialized LLM translators that each focus on optimizing a single dimension, we generate high-quality references through iterative refinement, while the intermediate outputs naturally form preference pairs that capture quality trade-offs (Figure 1c). This multi-aspect data provides effective targets for SFT. For preference optimization through RL training, we compare implicit reward modeling methods (*e.g.*, DPO (Rafailov et al., 2023)) with explicit reward modeling (Ouyang et al., 2022; Eisenstein et al., 2024) followed by GRPO (Shao et al., 2024), and find that training a compact reward model on our data with GRPO significantly outperforms DPO-based approaches.

*Corresponding author.



(a) Sampling + LLM Judge (b) Self-Refinement (c) Ours: Multi-Aspect Refinement

Figure 1: Comparison of data generation approaches for literary translation. (a) **Sampling + LLM Judge**: samples multiple translations and selects the best via LLM scoring. (b) **Self-Refinement**: iteratively refines translations using a single LLM. (c) **Ours: Multi-Aspect Refinement**: decomposes quality into expression fluency and literary effect with specialized modules, generating both high-quality references and preference pairs.

Since our trained models do not rely on long CoT reasoning, they remain practical for real-time applications.

Our contributions are as follows:

- We propose a **multi-aspect iterative refinement pipeline** for literary translation data generation. By decomposing translation quality into expression fluency and literary effect with specialized modules, our pipeline simultaneously produces high-quality translation references and preference pairs that capture the trade-offs between these dimensions.
- We conduct a systematic comparison between implicit reward modeling (*e.g.*, DPO) and explicit reward modeling with GRPO for literary translation LLM training. We find that combining a learned reward model with BLEU and format constraints as a composite reward for GRPO significantly outperforms DPO-based methods.
- Our resulting models, **LitMT-8B** and **LitMT-14B**, achieve 67.25 and 69.07 CEA100 respectively on the MetaphorTrans English-to-Chinese literary translation benchmark, competitive with Claude Sonnet 4.5 (68.43), and demonstrate strong generalization to out-of-domain literary work (*i.e.*, O. Henry).

2 Related Work

2.1 Fine-Tuning LLMs for Literary Translation

Recent work has fine-tuned specialized LLMs for literary translation to address both the high inference cost of commercial models and their inconsistent handling of literary devices such as metaphors and personification. DRT (Wang et al., 2025a) synthesizes long chain-of-thought data through general LLMs for SFT. DeepTrans (Wang et al., 2026) and ExTrans (Wang et al., 2025b) apply reinforcement learning with LLM-as-a-judge (Zheng et al., 2023) (*e.g.*, DeepSeek-V3) as the reward signal. While these methods achieve strong results, they require expensive LLM inference for every training sample. In contrast, our work focuses on generating high-quality training data—both references and preference pairs—that can be reused across various training paradigms without relying on massive LLMs during training.

2.2 Data Generation for LLM Training

Early approaches like Self-Instruct (Wang et al., 2023) and Alpaca (Taori et al., 2023) use LLM outputs directly as SFT targets, producing single-point outputs without quality comparisons. Rejection sampling methods such as Best-of-N (Stiennon et al., 2020; Bai et al., 2022) and RSO (Liu et al., 2024b) generate multiple outputs and select the best; however, one-shot sampling without iterative refinement often yields suboptimal results. Multi-agent approaches like MATRIX (Tang et al., 2025) and Fellowship of the LLMs (Arif et al., 2025) show promise for data generation, but have not been tailored for literary translation or addressed trade-offs between quality dimensions. Our multi-aspect approach assigns each translator a specific dimension (expression fluency vs. literary effect), generating preference pairs that capture these trade-offs.

Algorithm 1 Multi-Aspect Iterative Refinement

Require: Source S, threshold τ, max rounds K, patience N
Ensure: Best translation T^*, preference pairs $\mathcal{P}$
$T_0 \leftarrow \textsc{Translate}(S)$ {Naive translator}
$s_0, F_0 \leftarrow \textsc{Eval}(S, T_0)$ {Score and feedback}
$T^* \leftarrow T_0$; $s^* \leftarrow s_0$; $F^* \leftarrow F_0$; $n \leftarrow 0$
for $i = 1$ to K **do**
  $T_i^e \leftarrow \textsc{ExprOpt}(S, T^*, F^*)$ {Fluency-focused}
  $T_i^l \leftarrow \textsc{LitPres}(S, T^*, F^*)$ {Literary-focused}
  $T_i \leftarrow \textsc{Aggregate}(S, T_i^e, T_i^l)$ {Synthesize}
  $s_i, F_i \leftarrow \textsc{Eval}(S, T_i)$
  if $s_i > s^*$ **then**
    $T^* \leftarrow T_i$; $s^* \leftarrow s_i$; $F^* \leftarrow F_i$; $n \leftarrow 0$
  else
    $n \leftarrow n + 1$
  end if
  if $s^* \geq \tau$ **or** $n \geq N$ **then**
    break
  end if
end for
$\mathcal{P} \leftarrow \{(T_w, T_l) : s_w > s_l\}$ {All evaluated translations}
return $T^*, \mathcal{P}$

3 Approach

Our approach consists of two stages: a data generation pipeline that produces high-quality literary translations and preference pairs (§3.1), followed by a training pipeline that leverages these outputs for model optimization (§3.2).

3.1 Multi-Aspect Iterative Refinement

3.1.1 Pipeline Overview

Figure 1 (c) illustrates our pipeline. Given a source text S, we first generate an initial translation T_0 using a naive translator. We then iteratively refine through the following loop: (1) an evaluator scores the current best translation and provides feedback identifying specific issues; (2) two specialized LLM translators—Expression Optimizer and Literary Effect Preserver—each generate an improved version based on the feedback; (3) an aggregator synthesizes both versions into a single translation; (4) the evaluator scores the aggregated result and updates the best translation if the score improves. The process terminates when the score exceeds a threshold, the iteration count reaches the maximum, or the score shows no improvement for consecutive rounds.

Algorithm 1 formalizes this procedure.

3.1.2 Module Design

We now describe each module in detail. Our pipeline consists of five modules, all implemented as LLM prompts (see Appendix E for details).

Naive Translator. Generates an initial literary translation T_0 from source S without emphasis on any particular quality dimension, providing a neutral starting point for refinement.

Expression Optimizer. Focuses on target-language fluency: natural word order, idiomatic collocations, and concise expression. Given source S, current best translation T^*, and feedback F^*, it generates a fluency-optimized version T_i^e.

Literary Effect Preserver. Focuses on preserving artistic qualities: figurative language, rhetorical devices, and tone. Given the same inputs (S, T^*, F^*), it generates a literary-focused version T_i^l.

Aggregator. Synthesizes improvements from both specialized modules. Given source S and the two refined translations (T_i^e, T_i^l), it produces a final literary translation T_i that balances expression fluency with literary effect.

Evaluator. Scores literary translations based on expression fluency and literary effect, and provides textual feedback identifying specific issues to guide the refinement modules.

3.1.3 Outputs

The pipeline produces two types of training data that directly feed into our training pipeline (§3.2).

High-Quality References. For each source, we select the best translation $T^* = \arg\max_t \text{score}(t)$ across all iterations. These serve as targets for supervised fine-tuning.

Preference Pairs. From the iteration history, we construct preference pairs by enumerating all translation pairs (T_w, T_l) where $T_w \neq T_l$ and $\text{score}(T_w) > \text{score}(T_l)$. This yields multiple pairs per source, capturing fine-grained quality distinctions. These pairs support both direct preference optimization (*e.g.*, DPO (Rafailov et al., 2023), CPO (Xu et al., 2024), SimPO (Meng et al., 2024)) and reward model training (Ouyang et al., 2022).

3.2 Training Pipeline

We explore three training strategies using the generated data: supervised fine-tuning with high-quality references, and two RL approaches—implicit reward modeling (*e.g.*, DPO, SimPO, CPO) and explicit reward modeling (RM+GRPO).

3.2.1 Supervised Fine-Tuning

As a baseline, we train on the best translations y^* using cross-entropy loss:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(x,y^*)} \left[\log \pi_\theta(y^*|x)\right]. \quad (1)$$

3.2.2 Implicit Reward: DPO-Series

DPO-series methods directly optimize policy using preference pairs without training a separate reward model. Taking DPO as an example:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}\left[\log\sigma\left(\beta\log\frac{\pi_w}{\pi_w^{\text{ref}}} - \beta\log\frac{\pi_l}{\pi_l^{\text{ref}}}\right)\right], \quad (2)$$

where $\pi_w = \pi_\theta(y_w|x)$, $\pi_l = \pi_\theta(y_l|x)$, and π^{ref} denotes the reference policy. We also evaluate variants including CPO and SimPO (see Section 4.4.1).

3.2.3 Explicit Reward Modeling with GRPO

Our reward model is built by replacing the autoregressive language modeling head of an LLM with a linear layer that maps the final hidden state to a scalar reward. We train r_ϕ on the preference pairs using the Bradley-Terry loss (Bradley and Terry, 1952) following Ouyang et al. (2022) and Eisenstein et al. (2024):

$$\mathcal{L}_{\text{RM}} = -\mathbb{E}\left[\log\sigma(r_w - r_l)\right] + \lambda(r_w + r_l)^2, \quad (3)$$

where $r_w = r_\phi(x, y_w)$, $r_l = r_\phi(x, y_l)$, and the second term encourages zero-centered rewards.

We then optimize the policy with GRPO (Shao et al., 2024). For each source x, GRPO samples G literary translations $\{y_1, \ldots, y_G\}$ from the current policy and scores them using a composite reward. The rewards are group-normalized to compute advantages:

$$\hat{A}_i = \frac{r(x, y_i) - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}. \quad (4)$$

The policy is optimized using a clipped objective with KL regularization:

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}_x\Big[\frac{1}{G}\sum_{i=1}^{G}\min(\rho_i\hat{A}_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)\hat{A}_i) - \beta D_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}})\Big], \quad (5)$$

where $\rho_i = \pi_\theta(y_i|x)/\pi_{\text{old}}(y_i|x)$ is the probability ratio.

Composite Reward Design. Following Wang et al. (2026, 2025b), we adopt a composite reward for stable GRPO training. However, instead of relying on expensive LLM-as-a-judge or neural network-based quality estimation (QE) models (Rei et al., 2022), we use a lightweight yet effective combination:

$$r(x, y) = r_{\text{RM}} + 0.05 \cdot r_{\text{BLEU}} + r_{\text{fmt}}, \quad (6)$$

where r_{RM} is the learned reward model score (typically in $[-3, 3]$ after training), r_{BLEU} is SacreBLEU (Post, 2018) (0–100 scale), and r_{fmt} is a format constraint (0 if output is valid JSON with translation under the `translation` key, -5 otherwise). The BLEU component provides a stable lexical signal that complements the learned reward, while the format reward ensures structured outputs.

4 Experiments

4.1 Experimental Setup

Data. We use MetaphorTrans (Wang et al., 2025a), an English-to-Chinese literary translation benchmark containing sentences with metaphors and similes from approximately 400 English literary works in Project Gutenberg. The dataset comprises 19,264 training samples and 2,000 test samples. We apply our multi-aspect pipeline to the 19,264 training sources to generate high-quality reference translations and preference pairs. For preference data, we first split at the sample level (17,337 train / 1,927 dev), then construct pairs within each split, yielding 179,588 training pairs and 19,767 dev pairs. We also evaluate on The Essential O. Henry Collection (Wang et al., 2026) (586 test samples) for out-of-domain generalization.

Models. We train LitMT-8B and LitMT-14B from Qwen3-8B-Base and Qwen3-14B-Base (Yang et al., 2025) respectively, using our full pipeline (SFT followed by RM+GRPO). All ablation studies use Qwen3-8B-Base. We compare against three categories of baselines: (1) general-purpose LLMs including Qwen3 (Yang et al., 2025), Llama 3.3 (Grattafiori et al., 2024), DeepSeek V3.1 (Liu et al., 2024a), Kimi K2 (Team et al., 2025), Claude 4.5 (Anthropic, 2025b,a), gpt-oss (Agarwal et al., 2025), GPT-5.2 (OpenAI, 2025); (2) general translation models such as X-ALMA (Xu et al., 2025) and LMT-60 (Luo et al., 2025); and (3) specialized literary translation models trained on MetaphorTrans: DRT (Wang et al., 2025a), DeepTrans (Wang et al., 2026), and ExTrans (Wang et al., 2025b).

Metrics. We adopt LLM-based metrics with prompts designed for literary translation quality assessment (Appendix F). To ensure more accurate evaluation, we upgrade the evaluator from GPT-4o used in prior work (Wang et al., 2025a, 2026, 2025b) to Claude Opus 4.5 (Anthropic, 2025a),

Model	Params	CRF	CEA5	CEA100
X-ALMA-13B (G6)	13B	51.35	2.44	41.11
LMT-60-8B	8B	54.59	2.59	44.63
gpt-oss-20B♠	20B (3.6B)	55.97	2.59	44.70
Llama3.3-70B	70B	59.46	2.75	48.34
gpt-oss-120B♠	120B (5.1B)	62.25	2.89	51.50
Qwen3-8B	8B	62.70	2.94	52.77
MiniMax M2	230B (10B)	65.13	2.96	53.15
DRT-7B◇♠	7B	64.48	3.05	55.09
Qwen3-14B	14B	65.25	3.07	55.47
Qwen3-32B	32B	66.65	3.15	57.09
DRT-14B◇♠	14B	67.07	3.21	58.43
DeepTrans-7B◇♠	7B	70.31	3.33	61.15
ExTrans-7B◇♠	7B	71.87	3.42	62.95
Qwen3-235B-A22B	235B (22B)	74.15	3.55	65.62
LitMT-8B◇	**8B**	**73.03**	**3.61**	**67.25**
Claude Sonnet 4.5	–	77.46	3.66	68.43
GPT-5.2	–	77.45	3.68	68.68
LitMT-14B◇	**14B**	**75.20**	**3.71**	**69.07**
DeepSeek V3.1	671B (37B)	77.19	3.75	70.27
Kimi K2♠	1T (32B)	79.93	3.81	71.70
Claude Opus 4.5	–	79.95	3.90	73.30

◇trained on MetaphorTrans. ♠thinking-enabled. **Bold**/ gray : our models.

Table 1: Main results on MetaphorTrans (in-domain). Models sorted by CEA100. Params in parentheses denote activated parameters for MoE models. Qwen3-235B-A22B refers to the Instruct variant. gpt-oss models use medium reasoning effort.

Model	Params	CRF	CEA5	CEA100
X-ALMA-13B (G6)	13B	59.05	2.78	48.59
gpt-oss-20B♠	20B (3.6B)	63.68	2.91	51.63
LMT-60-8B	8B	66.78	3.15	56.88
Llama3.3-70B	70B	68.57	3.18	57.96
Qwen3-8B	8B	70.20	3.30	60.56
DRT-7B◇♠	7B	70.30	3.31	60.67
gpt-oss-120B♠	120B (5.1B)	71.35	3.34	61.49
MiniMax M2	230B (10B)	73.49	3.35	61.79
DeepTrans-7B◇♠	7B	74.58	3.46	64.17
Qwen3-14B	14B	73.38	3.49	64.32
DRT-14B◇♠	14B	73.67	3.52	65.19
Qwen3-32B	32B	75.51	3.56	65.81
ExTrans-7B◇♠	7B	75.04	3.62	67.39
LitMT-8B◇	**8B**	**76.42**	**3.77**	**70.38**
LitMT-14B◇	**14B**	**79.65**	**3.94**	**73.71**
Qwen3-235B-A22B	235B (22B)	81.61	3.94	74.01
Claude Sonnet 4.5	–	84.27	4.05	76.50
GPT-5.2	–	84.47	4.12	77.74
DeepSeek V3.1	671B (37B)	83.56	4.12	77.94
Kimi K2♠	1T (32B)	85.55	4.16	78.56
Claude Opus 4.5	–	86.22	4.26	80.65

◇trained on MetaphorTrans. ♠thinking-enabled. **Bold**/ gray : our models.

Table 2: Out-of-domain results on O. Henry Collection. Models sorted by CEA100. Qwen3-235B-A22B refers to the Instruct variant. gpt-oss models use medium reasoning effort.

one of the strongest available LLMs at the time of writing. We report three metrics: CRF (Claude Reference-Free) and CEA5/CEA100 (Claude Evaluator Agent, 5-point and 100-point scales) following prior work (Kocmi and Federmann, 2023; Wang et al., 2025a, 2026, 2025b). CEA100 serves as the primary metric due to its finer granularity. Unlike prior work (Wang et al., 2025a, 2026, 2025b) that evaluates on a sampled subset, we evaluate on the full 2,000 test samples, providing more reliable and reproducible results.

Implementation. Our refinement pipeline uses Qwen3-235B-A22B-Instruct (Yang et al., 2025; Qwen Team, 2025) as the backbone for all modules, with maximum rounds $K = 8$, early stopping patience $N = 3$, and score threshold $\tau = 4.9$ (out of 5.0). All training experiments use 8×H100 GPUs with global batch size 128. For SFT and DPO-series methods: learning rate 1e-5, warmup ratio 0.05, 3 epochs. The reward model trains for 1 epoch with $\lambda = 0.01$ from the SFT checkpoint. GRPO uses $G = 16$ generations per sample, temperature 1.0, top-p 0.9, $\beta = 0.01$, learning rate 1e-7, 3 epochs from the SFT checkpoint.

4.2 Main Results

Table 1 presents results on MetaphorTrans. Our models demonstrate strong parameter efficiency: LitMT-8B (8B) outperforms Qwen3-235B-A22B (235B, 22B activated) by 1.63 CEA100 points, achieving comparable quality using 30× fewer parameters. Notably, our training data is generated using Qwen3-235B-A22B, yet the 8B student model surpasses its teacher—demonstrating that our multi-aspect refinement pipeline genuinely improves data quality rather than merely distilling knowledge. LitMT-14B achieves 69.07 CEA100, competitive with frontier models such as Claude Sonnet 4.5 (68.43) and GPT-5.2 (68.68). Table 2 presents out-of-domain results on The Essential O. Henry Collection, which features distinct narrative styles and early 20th-century American English. Our models demonstrate strong generalization: LitMT-14B (73.71) approaches Qwen3-235B-A22B (74.01), and LitMT-8B (70.38) surpasses Qwen3-32B (65.81) by a wide margin.

4.3 Evaluation Reliability

Human evaluation of literary translation requires bilingual proficiency and literary expertise to assess stylistic nuances—such expertise is scarce and costly at scale. To nevertheless validate our LLM-

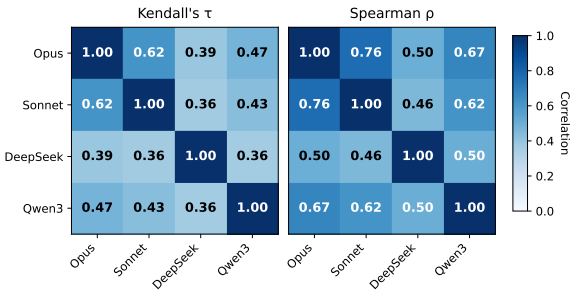


Figure 2: Sample-level correlation between four LLM judges over 12,000 instances. Left: Kendall's τ. Right: Spearman's ρ. All correlations significant at $p < 0.001$.

based evaluation, we conduct inter-judge agreement analysis following recent advances (Kocmi and Federmann, 2023; Zheng et al., 2023).

We assess agreement using four frontier LLMs: Claude Opus 4.5, Claude Sonnet 4.5, DeepSeek V3.1, and Qwen3-235B-A22B-Instruct. We select six representative models spanning different performance levels and compute pairwise system rankings, yielding $\binom{6}{2} = 15$ comparisons per judge. The judges achieve 88.9% agreement on relative ranking at the system level (see Appendix B for details). At the sample level, Figure 2 shows correlations over 12,000 instances, with Kendall's τ ranging from 0.36 to 0.62—comparable to or exceeding the τ of 0.36–0.38 reported between GPT-4 and human (Kocmi and Federmann, 2023). Cross-family judge pairs also exhibit substantial agreement (e.g., Claude–Qwen τ=0.47), confirming that our evaluation captures genuine quality distinctions.

4.4 Ablation Studies

4.4.1 Effect of Training Methods

Table 3 compares training methods using identical data. All preference optimization methods (DPO-series and RM+GRPO) are initialized from the SFT checkpoint. A notable finding is that all DPO-series methods underperform SFT, with performance drops of 2–3 CEA100 points. In contrast, RM+GRPO improves over SFT by 1.51 points, validating our hypothesis that explicit reward modeling with online exploration better leverages the preference data.

4.4.2 Effect of Multi-Aspect Refinement

Table 4 examines each component's contribution. Best-of-8 (Stiennon et al., 2020) generates 8 translations via temperature sampling and selects the best; this baseline performs worst, confirming that

Method	CRF	CEA5	CEA100
SFT	72.66	3.54	65.74
SimPO	69.98	3.41	62.62
DPO	70.50	3.44	63.39
CPO	70.69	3.45	63.67
RM+GRPO	**73.03**	**3.61**	**67.25**

Table 3: Effect of training methods. Explicit reward modeling with GRPO outperforms all implicit reward methods.

Variant	CRF	CEA5	CEA100
Best-of-8 Sampling	71.44	3.41	63.03
Naive-only	72.92	3.54	65.64
Expression Optimizer-only	72.21	3.52	65.27
Literary Preserver-only	**73.39**	3.60	66.91
All (Multi-Aspect)	73.03	**3.61**	**67.25**

Table 4: Effect of multi-aspect refinement.

targeted refinement outperforms random exploration. Naive-only uses standard iterative refinement without specialized modules. Among single-agent variants, Literary Preserver-only achieves the strongest results (66.91), suggesting that literary effect preservation is more challenging to optimize. Multi-aspect refinement combines both agents' strengths, achieving the best overall performance (67.25).

4.4.3 Effect of Data Source

Table 5 compares SFT performance across different training data sources. Our pipeline uses Qwen3-235B-A22B as the backbone for data generation. Compared to single-pass distillation from the same model, our multi-aspect refinement improves CEA100 by 4.66 points (61.08 $\rightarrow$ 65.74), demonstrating that iterative refinement genuinely enhances data quality beyond simple distillation. Our data also outperforms the original DRT ground truth by 8.65 points.

4.4.4 Effect of Base Model

Table 6 examines the effect of base model choice. Qwen3-8B-Base achieves the best performance (67.25 CEA100), outperforming Qwen3-8B by 1.48 points. However, for Qwen2.5-7B, the Instruct variant slightly outperforms the base model (64.40 vs. 63.97). The most significant factor is model family: Qwen3 models consistently outperform Qwen2.5 models by 2.8–3.3 points regardless of variant.

Data Source	CRF	CEA5	CEA100
DRT Ground Truth	66.41	3.15	57.09
Qwen3-235B-A22B Distill	70.02	3.32	61.08
DeepSeek V3.1 Distill	71.04	3.40	62.78
Ours	**72.66**	**3.54**	**65.74**

Table 5: Effect of data source (SFT only).

Base Model	CRF	CEA5	CEA100
Qwen2.5-7B	71.05	3.47	63.97
Qwen2.5-7B-Instruct	71.46	3.48	64.40
Qwen3-8B	72.12	3.55	65.77
Qwen3-8B-Base	**73.03**	**3.61**	**67.25**

Table 6: Effect of base model choice with our training approach.

4.4.5 Data Generation Statistics

Table 7 presents statistics from our multi-aspect iterative refinement pipeline. The pipeline achieves an average score improvement of 0.31 points (from 4.43 to 4.73 on a 5-point scale), with 61.6% of samples reaching the quality threshold ($\tau = 4.9$). The average best-worst score difference of 0.70 indicates meaningful quality distinctions within each sample's iteration history, providing informative preference pairs for training.

4.4.6 Reward Model Performance

Table 8 reports the pairwise classification accuracy of our 8B reward model (§3.2) on the held-out RM Dev set comprising 19,767 preference pairs. Results are stratified by the absolute score margin between chosen and rejected translations. The model attains 72.49% overall accuracy, with accuracy ranging from around 68% on ambiguous pairs (margin $<$0.5) to 96.20% on unambiguous cases (margin $\geq$3.0). This calibration validates that our preference data encodes learnable quality signals spanning the full difficulty spectrum.

5 Conclusion

We present a multi-aspect data generation and LLM training approach for literary translation. By decomposing quality into expression fluency and literary effect, our pipeline generates high-quality references and preference pairs. The generated data outperforms existing ground truth for SFT by 8.65 CEA100 points. For preference optimization, explicit reward modeling with GRPO yields an additional 1.51 points, while DPO-series methods degrade performance. Our models, LitMT-8B and LitMT-14B, achieve 67.25 and 69.07 CEA100 on MetaphorTrans, competitive with Claude Sonnet 4.5 (68.43), and demonstrate strong generalization on out-of-domain O. Henry data. Our approach provides a strong baseline for future literary translation research.

Statistic	Value
Total samples	19,264
Average refinement rounds	4.08
Average initial score	4.43
Average final score	4.73
Average best score	4.88
Average worst score	4.19
Average score improvement	0.31
Average best-worst difference	0.70
Samples reached threshold	11,863 (61.6%)

Table 7: Data generation statistics from the multi-aspect iterative refinement pipeline on MetaphorTrans training set.

Score Diff	Total	Correct	Acc. (%)
$[0, 0.25)$	10,845	7,473	68.91
$[0.25, 0.5)$	2,336	1,583	67.77
$[0.5, 1.0)$	2,287	1,695	74.11
$[1.0, 1.5)$	2,170	1,761	81.15
$[1.5, 2.0)$	745	587	78.79
$[2.0, 2.5)$	946	813	85.94
$[2.5, 3.0)$	201	189	94.03
≥ 3.0	237	228	96.20
Overall	**19,767**	**14,329**	**72.49**

Table 8: Reward model (8B) accuracy on RM Dev set by score difference (0–5 scale) between chosen and rejected translations.

Limitations

Our work has several limitations. First, while we empirically show that explicit reward modeling with GRPO outperforms DPO-series methods, a theoretical understanding of this phenomenon remains an open question. Second, whether incorporating long chain-of-thought reasoning (Wang et al., 2025a, 2026, 2025b) into our framework yields further gains remains to be explored. Third, our experiments focus on English-to-Chinese; extending to other language pairs is left for future work.

Ethical Considerations

We discuss the main ethical considerations of our work as follows: (1) **Copyright.** We use the MetaphorTrans benchmark (Wang et al., 2025a), which mines literary sentences from approximately 400 English books provided by the Project Gutenberg public-domain book repository.[1] These books are typically more than fifty years old and their copyrights have expired. The O. Henry Collection used for out-of-domain evaluation is also in the public domain. (2) **Potential Biases.** The backbone models used in our data generation pipeline (Qwen3-235B-A22B-Instruct) and training (Qwen3-8B-Base, Qwen3-14B-Base) may exhibit biases present in their pretraining data. Our models, LitMT-8B and LitMT-14B, might inherit similar biases.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Amazon Web Services. 2025. Amazon bedrock pricing. https://aws.amazon.com/bedrock/pricing/.

Anthropic. 2025a. Claude opus 4.5. https://www.anthropic.com/claude/opus.

Anthropic. 2025b. Claude sonnet 4.5. https://www.anthropic.com/claude/sonnet.

Samee Arif, Sualeha Farid, Abdul Hameed Azeemi, Awais Athar, and Agha Ali Raza. 2025. The fellowship of the llms: Multi-model workflows for synthetic preference optimization dataset generation. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 30–45.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alexander Nicholas D'Amour, Krishnamurthy Dj Dvijotham, Adam Fisch, Katherine A Heller, Stephen Robert Pfohl, Deepak Ramachandran, Peter Shaw, and Jonathan Berant. 2024. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. In *First Conference on Language Modeling*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ana Guerberof-Arenas and Antonio Toral. 2022. Creativity in translation: Machine translation as a constraint for literary texts. *Translation Spaces*, 11(2):184–212.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, and 1 others. 2024. Findings of the wmt24 general machine translation shared task: The llm era is here but mt is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46.

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

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. 2024b. Statistical rejection sampling improves preference optimization. In *The Twelfth International Conference on Learning Representations*.

Yingfeng Luo, Ziqiang Xu, Yuxuan Ouyang, Murun Yang, Dingyang Lin, Kaiyan Chang, Tong Zheng, Bei Li, Peinan Feng, Quan Du, and 1 others. 2025. Beyond english: Toward inclusive and scalable multilingual machine translation with llms. *arXiv preprint arXiv:2511.07003*.

[1] https://www.gutenberg.org/

Evgeny Matusov. 2019. The challenges of using neural machine translation for literature. In *Proceedings of the qualities of literary machine translation*, pages 10–19.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235.

OpenAI. 2025. Introducing gpt-5.2. https://openai.com/index/introducing-gpt-5-2/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3-235b-a22b-instruct-2507. https://huggingface.co/Qwen/Qwen3-235B-A22B-Instruct-2507.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC De Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, and 1 others. 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.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021.

Shuo Tang, Xianghe Pang, Zexi Liu, Bohan Tang, Rui Ye, Tian Jin, Xiaowen Dong, Yanfeng Wang, and Siheng Chen. 2025. Synthesizing post-training data for llms through multi-agent simulation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23306–23335.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Jiaan Wang, Fandong Meng, Yunlong Liang, and Jie Zhou. 2025a. Drt: Deep reasoning translation via long chain-of-thought. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6770–6782.

Jiaan Wang, Fandong Meng, and Jie Zhou. 2025b. Extrans: Multilingual deep reasoning translation via exemplar-enhanced reinforcement learning. *arXiv preprint arXiv:2505.12996*.

Jiaan Wang, Fandong Meng, and Jie Zhou. 2026. Deeptrans: Deep reasoning translation via reinforcement learning. *Transactions of the Association for Computational Linguistics*, 14:47–63.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 13484–13508.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2025. X-ALMA: Plug & play modules and adaptive rejection for quality translation at scale. In *The Thirteenth International Conference on Learning Representations*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: pushing the boundaries of llm performance in machine translation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Additional Experiments

A.1 Effect of Composite Reward Design

Table 9 presents an ablation study on the reward weighting coefficients in our composite reward formulation $r = w_{\text{RM}} \cdot r_{\text{RM}} + w_{\text{BLEU}} \cdot r_{\text{BLEU}} + r_{\text{fmt}}$ (§3.2). Employing the learned reward in isolation ($w_{\text{BLEU}} = 0$) yields 64.48 CEA100. Incorporating BLEU as a complementary lexical signal progressively improves performance, with optimal results at $w_{\text{BLEU}} = 0.05$ (67.25). However, increasing the BLEU coefficient to 0.1 degrades performance to 66.80, indicating that excessive reliance on surface-level lexical matching is detrimental. Conversely, using BLEU alone ($w_{\text{RM}} = 0$) achieves 65.49, demonstrating that the learned reward model captures quality dimensions beyond n-gram overlap. These results suggest that the optimal reward design balances learned quality judgments with lexical grounding.

w_{RM}	w_{BLEU}	CRF	CEA5	CEA100
1	0	69.71	3.50	64.48
1	0.01	71.86	3.58	66.10
1	**0.05**	73.11	3.60	**67.25**
1	0.1	73.40	3.59	66.80
0	1	72.75	3.53	65.49

Table 9: Effect of reward weighting in composite reward design. Performance peaks at $w_{\text{BLEU}} = 0.05$; higher weights degrade quality.

B Evaluation Reliability Analysis

This section provides detailed analysis supporting the evaluation reliability results presented in §4.3.

Human evaluation of literary translation poses distinctive methodological challenges: annotators must demonstrate bilingual proficiency alongside sufficient literary expertise to assess stylistic nuances, figurative language preservation, and aesthetic quality. Such specialized expertise is scarce and prohibitively expensive to obtain at scale. Following recent advances in LLM-based evaluation (Kocmi and Federmann, 2023; Zheng et al., 2023), we adopt automated assessment and conduct a comprehensive reliability analysis to validate our evaluation protocol.

Model-Level Consistency. We evaluate six representative systems on the MetaphorTrans test set using four frontier LLMs as judges: Claude Opus 4.5 (Anthropic, 2025a), Claude Sonnet 4.5 (Anthropic, 2025b), DeepSeek V3.1 (Liu et al., 2024a), and Qwen3-235B-A22B-Instruct (Yang et al., 2025; Qwen Team, 2025). Table 10 reports CEA100 scores with corresponding rankings from each judge. Despite variation in absolute score magnitudes, all four judges consistently rank LitMT-8B and LitMT-14B above the baseline systems (DRT-7B, DeepTrans-7B, Qwen3-8B). Across all 15 pairwise system comparisons, the judges exhibit 88.9% agreement on relative ordering, providing robust evidence that our principal findings are judge-invariant.

Model	Opus	Sonnet	DeepSeek	Qwen3
Kimi K2	71.70 (1)	78.06 (1)	87.19 (4)	82.01 (2)
LitMT-14B	69.07 (2)	74.50 (2)	88.51 (1)	82.29 (1)
LitMT-8B	67.25 (3)	71.94 (3)	88.09 (2)	81.70 (3)
DeepTrans-7B	61.15 (4)	64.41 (4)	88.07 (3)	79.30 (4)
DRT-7B	55.09 (5)	62.07 (5)	81.22 (5)	67.99 (5)
Qwen3-8B	52.77 (6)	59.96 (6)	77.17 (6)	64.22 (6)

Table 10: CEA100 scores (rank) from four LLM judges on MetaphorTrans test set. Opus = Claude Opus 4.5, Sonnet = Claude Sonnet 4.5, DeepSeek = DeepSeek V3.1, Qwen3 = Qwen3-235B-A22B-Instruct. Models sorted by Opus score.

Sample-Level Correlation. We compute pairwise correlations between all judge combinations over 12,000 evaluation instances (2,000 test samples $\times$ 6 systems). Figure 2 (in §4.3) visualizes Kendall's τ and Spearman's ρ coefficients. All correlations achieve statistical significance ($p < 0.001$). The inter-judge Kendall's τ ranges from 0.36 to 0.62, comparable to or exceeding the τ of 0.36–0.38 reported between GPT-4 and human MQM annotations (Kocmi and Federmann, 2023). This indicates that our LLM judges exhibit mutual agreement at levels commensurate with human–LLM concordance. Notably, cross-family judge pairs also demonstrate substantial correlation (e.g., Claude–Qwen τ=0.47, Claude–DeepSeek τ=0.39), confirming that our evaluation captures genuine quality distinctions rather than model-specific scoring artifacts.

Evaluation Stability. To assess reproducibility, we execute CEA100 evaluation with Claude Opus 4.5 on LitMT-8B across three independent runs. The resulting scores are 67.25, 67.25, and 67.23, yielding a standard deviation of 0.01. This negligible variance demonstrates that our LLM-based evaluation protocol produces highly stable and reproducible measurements.

Source	This it was, more than any thing else, that roused such a **tempest** in my **soul**.
Qwen3-8B	正是这一点，比任何其他事情都更激起了我内心的巨大风暴。
DRT-7B	这比任何事情都更深深激起了我内心中的这场风暴。
DeepTrans-7B	正是这一点，胜过一切其他因素，在我心中激起了如此狂风巨浪。
LitMT-8B	正是这一点，胜过其他任何事，激起了我灵魂中的狂风骤雨。
LitMT-14B	正是这一点，胜过世间一切，于我心魂深处掀起滔天巨浪。
Source	O, then the earth shook to see the **heavens on fire**, And not in fear of your **nativity**.
Qwen3-8B	哦，大地震动，目睹苍穹燃起战火，却并非因你生辰的恐惧。
DRT-7B	哦，于是大地因天火熊熊而震颤，非因惧您降临尘世。
DeepTrans-7B	哦，于是大地震颤，目睹苍穹燃起熊熊烈焰，却非因你降世而生畏惧。
LitMT-8B	啊，大地因苍天燃起烈火而震颤，却并非畏惧你降生的威仪。
LitMT-14B	啊，大地震颤，只因天穹烈焰腾起，并非因惧你降生而惶然惊悸。
Source	He said he'd **cowhide** me till I was **black and blue** if I didn't raise some money for him.
Qwen3-8B	他说如果我不为他筹些钱，他就会把我打得青一块紫一块。
DRT-7B	他说，若我不帮他弄到些钱，他就会用牛皮鞭子狠狠地抽我，直到我遍体鳞伤。
DeepTrans-7B	他扬言要重重抽我一顿，直到我浑身青一块紫一块，除非我给他筹来些钱。
LitMT-8B	他说，如果我不为他弄到钱，就用牛皮抽我，直到我青一块紫一块。
LitMT-14B	他说，如果我不替他弄到钱，就拿牛皮鞭把我抽得青一块紫一块。

Table 11: Case studies of literary translation. Green indicates good word choices, while red indicates errors. Example 1: "soul" should be 灵魂/心魂 not 内心/心中; "tempest" rendered as 狂风骤雨/滔天巨浪 is more vivid. Example 2: "heavens on fire" should not add "war" (战火); "nativity" means birth (降生), not birthday (生辰). Example 3: "cowhide" specifically means whipping with leather (牛皮/牛皮鞭); "black and blue" means bruises (青一块紫一块), not wounds (遍体鳞伤).

C Translation Case Studies

Table 11 presents representative examples comparing translations from different models. Green indicates good word choices, while red indicates problematic translations.

D Cost Analysis

We compare the computational costs of our approach against methods that use LLM-as-a-Judge for online reward signals during GRPO training. All experiments use the MetaphorTrans training set (19,264 samples).

D.1 LLM API Costs

Table 12 presents the token consumption and associated costs. We estimate token usage as follows:

Our Method. The multi-aspect iterative refinement pipeline (§3.1) processes each sample through naive translation, expression optimization, literary effect preservation, aggregation, and evaluation, with an average of 4.08 refinement rounds (Table 7). This is a **one-time** data generation cost —the resulting SFT data and preference pairs can be reused across multiple training runs.

DeepTrans/ExTrans. These methods require LLM evaluation during GRPO training. We estimate costs for a single training run (2 epochs, 8 rollouts per sample). DeepTrans (Wang et al., 2026) uses DeepSeek-V3 for scoring; ExTrans (Wang et al., 2025b) uses DeepSeek-R1 to generate reference translations and DeepSeek-V3 to compare model outputs against these references.

Method	Model	Input	Output	Cost
Ours	Qwen3-235B	192.8M	143.2M	**$168**
DeepTrans	DeepSeek-V3	122.2M	146.8M	$318
ExTrans	DeepSeek-R1	3.8M	175.9M	$955
	DeepSeek-V3	116.7M	9.2M	$83
	Total	–	–	$1,038

Table 12: LLM API costs. Our method is one-time data generation; DeepTrans/ExTrans costs are per GRPO run (2 epochs, 8 rollouts). Prices from AWS Bedrock (Amazon Web Services, 2025): Qwen3-235B ($0.22/$0.88 per 1M input/output tokens), DeepSeek-V3.1 ($0.58/$1.68, as proxy for V3), DeepSeek-R1 ($1.35/$5.40).

Since DeepTrans and ExTrans compute reward signals online, they require fresh API calls for every training run. For a typical development cycle with 5–10 hyperparameter iterations, our cost remains $168 while DeepTrans scales to $1,590–$3,180 and ExTrans to $5,190–$10,380.

D.2 GPU Training Costs

Training LitMT-8B requires 80 H100 GPU hours: 2 hours for SFT, 8 hours for reward model training,

and 70 hours for GRPO. The reward model enables training without any LLM API calls. Adopting online LLM-as-a-Judge as in DeepTrans or ExTrans would increase GPU costs by several times or even an order of magnitude, as GPUs idle while waiting for API responses.

E Data Generation Prompts

The following prompts are used in our multi-aspect iterative refinement pipeline (§3.1):

- Naive Translator (Appendix E.1)
- Expression Optimizer (Appendix E.2)
- Literary Effect Preserver (Appendix E.3)
- Aggregator (Appendix E.4)
- Evaluator (Appendix E.5)

F Evaluation Prompts

The following prompts are used for the evaluation metrics reported in our experiments (§4):

- CRF (Appendix F.2)
- CEA5: uses the same prompt as the Evaluator (Appendix E.5)
- CEA100 (Appendix F.1)

E.1 Naive Translator

```
Please translate the following text from English to {target_language}. Output your
translation in XML format.

Input text:
{source}

Output format:
<translation>
your translation here
</translation>
```

E.2 Expression Optimizer

```
You are an expression optimization expert. Your role is to ensure the translation reads
naturally and fluently in {target_language}, eliminating translationese and awkward
phrasing.

<source_text>
{source}
</source_text>

<current_translation>
{translation}
</current_translation>

<evaluator_feedback>
{evaluator_feedback}
</evaluator_feedback>

Analyze the translation for expression quality issues:

1. Word Order: Does it follow natural {target_language} syntax?
   - Subject-verb-object arrangement
   - Modifier placement (adjectives, adverbs)
   - Clause ordering and sentence flow

2. Collocation: Are word combinations natural and idiomatic?
   - Verb-noun, adjective-noun pairings
   - Common {target_language} expressions vs. literal translations
   - Avoiding awkward or non-native combinations

3. Redundancy: Is the expression concise without unnecessary repetition?
   - Removing verbose constructions
   - Eliminating translationese patterns
   - Maintaining clarity while being concise

Provide specific improvements to enhance naturalness and fluency.

Output in XML format:
<result>
<issues>
Specific expression issues: word order, collocation, or redundancy problems
</issues>
<improved_translation>
Improved translation (return original if no improvement needed)
</improved_translation>
</result>
```

E.3 Literary Effect Preserver

```
You are a literary effect expert. Your role is to preserve and recreate the literary and
aesthetic qualities of the source text, including metaphors, rhetorical devices, tone,
and emotional impact.

<source_text>
{source}
</source_text>

<current_translation>
{translation}
</current_translation>

<evaluator_feedback>
{evaluator_feedback}
</evaluator_feedback>

Analyze the translation for literary quality:

1. Metaphors and Imagery: Are figurative expressions appropriately handled?
   - Identify metaphors, similes, and imagery in the source
   - Decide: preserve literally, adapt culturally, or explicitate?
   - Example: "drowning in work" can be preserved literally or adapted idiomatically
   - Ensure metaphors are understandable in target culture

2. Rhetorical Devices: Are stylistic features recreated?
   - Parallelism, repetition, alliteration, rhythm
   - Wordplay, puns, double meanings
   - Can these be recreated or compensated in {target_language}?

3. Tone and Emotional Impact: Does it evoke similar feelings?
   - Formal/informal, serious/humorous, poetic/prosaic
   - Emotional atmosphere and mood
   - Cultural connotations and associations

Provide specific improvements to enhance literary and aesthetic quality.

Output in XML format:
<result>
<issues>
Specific literary issues: metaphors, rhetorical devices, or tone/emotion problems
</issues>
<improved_translation>
Improved translation (return original if no improvement needed)
</improved_translation>
</result>
```

E.4 Aggregator

```
You are a translation integration expert. You have translation results from two experts.
Please synthesize this information and provide an optimal translation.

<source_text>
{source}
</source_text>

<expression_optimization_expert>
{expression_translation}
</expression_optimization_expert>

<literary_effect_expert>
{literary_translation}
</literary_effect_expert>

Please synthesize the two experts' translations and provide an optimal translation that
balances expression fluency and literary effect.

Output in XML format:
<result>
<reasoning>
Integration reasoning explanation
</reasoning>
<final_translation>
Final translation
</final_translation>
</result>
```

E.5 Evaluator

```
[System]
Please evaluate the following {target_language} translation of an English text. Rate the
translation on a scale of 0.00 to 5.00 (with two decimal places), where:
- 1: Poor translation; the text is somewhat understandable but contains significant
errors and awkward phrasing that greatly hinder comprehension for a {target_language}
reader.
- 2: Fair translation; the text conveys the basic meaning but lacks fluency and contains
several awkward phrases or inaccuracies, making it challenging for a {target_language}
reader to fully grasp the intended message.
- 3: Good translation; the text is mostly fluent and conveys the original meaning well,
but may have minor awkwardness or slight inaccuracies that could confuse a {
target_language} reader.
- 4: Very good translation; the text is smooth and natural, effectively conveying the
intended meaning, but may still have minor issues that could slightly affect
understanding for a {target_language} reader.
- 5: Excellent translation; the text is fluent and natural, conveying the original
meaning clearly and effectively, with no significant issues that would hinder
understanding for a {target_language} reader.

Please provide the reason first, followed by a score. Format your evaluation in the XML
structure below:
<evaluation>
<reason>
reason for the score
</reason>
<score>
float (0.00 to 5.00, two decimal places)
</score>
</evaluation>

[User]
<text>
{source}
</text>
<translation>
{translation}
</translation>
```

F.1 CEA100 (Claude Evaluator Agent, 100-point scale)

The CEA100 metric uses the same prompt structure as CEA5, with the scoring scale adjusted to 0–100:

```
[System]
Please evaluate the following {target_language} translation of an English text. Rate the
translation on a scale of 0 to 100, where:
- 10 points: Poor translation; the text is somewhat understandable but contains
significant errors and awkward phrasing that greatly hinder comprehension for a {
target_language} reader.
- 30 points: Fair translation; the text conveys the basic meaning but lacks fluency and
contains several awkward phrases or inaccuracies, making it challenging for a {
target_language} reader to fully grasp the intended message.
- 50 points: Good translation; the text is mostly fluent and conveys the original meaning
 well, but may have minor awkwardness or slight inaccuracies that could confuse a {
target_language} reader.
- 70 points: Very good translation; the text is smooth and natural, effectively conveying
 the intended meaning, but may still have minor issues that could slightly affect
understanding for a {target_language} reader.
- 90 points: Excellent translation; the text is fluent and natural, conveying the
original meaning clearly and effectively, with no significant issues that would hinder
understanding for a {target_language} reader.

Please provide the reason first, followed by a score. Format your evaluation in the XML
structure below:
<evaluation>
<reason>
reason for the score
</reason>
<score>
int
</score>
</evaluation>

[User]
<text>
{source}
</text>
<translation>
{translation}
</translation>
```

F.2 CRF (Claude Reference-Free)

```
[System]
Score the following translation from English to {target_language} on a continuous scale
from 0 to 100, where score of zero means "no meaning preserved" and score of one hundred
means "perfect preservation of meaning, with faithfulness, expressiveness, and elegance".

Format your score in the XML structure below:
<score>
int
</score>

[User]
English source: {source}
{target_language} translation: {translation}
```