Improving Long-Context Translation via Self-Supervised Dual Learning

Shanbo Cheng[1,2*], **Shuaijie She**[1,2*†], **Yu Bao**[2], **Jianbing Zhang**[1], **Jiajun Chen**[1], **Shujian Huang**[1‡]
[1] National Key Laboratory for Novel Software Technology, Nanjing University
[2]Bytedance Seed
shesj@smail.nju.edu.cn, {chengshanbo,baoyu.001}@bytedance.com
{zjb,chenjj,huangsj}@nju.edu.cn

Abstract

Large language models (LLMs) with long context windows offer the potential to translate entire documents in a single pass, yet they frequently suffer from catastrophic information distortion, undermining the strict faithfulness required for translation. This challenge is compounded by the scarcity of document-level parallel data, which makes both supervised fine-tuning and reliable evaluation prohibitively expensive. We propose LongDu, a self-supervised post-training framework that improves long-document translation reliability via round-trip consistency. Given monolingual documents, LongDu samples multiple candidate translations, back-translates each candidate, and optimizes the model to prefer translations that best reconstruct the source. To make this signal robust for long-form generation, we design a reward that filters trivial failure modes (e.g., copying and local language drift) before applying a reconstruction and fluency score, enabling stable reinforcement learning without human annotations. We additionally introduce Long-CIRT, an automatic evaluation protocol that quantifies information distortion by measuring how much a LLM's performance degrades after a translation cycle. Across multiple base models, LongDu substantially improves information retention and translation quality, with gains that generalize beyond the training length range and to unseen target languages.

1 Introduction

The emergence of long-context large language models (LLMs) (Yang et al., 2025a; Team et al., 2025; DeepMind, 2025) makes end-to-end document-level translation feasible: a single LLM can condition on thousands of tokens and translate an entire document in one pass, potentially improving discourse-level consistency in terminology, coreference, and style. Yet, when both inputs and outputs become long, LLMs often exhibit silent but catastrophic information distortion (Liu et al., 2024a; Tu et al., 2025; Weng et al., 2020), especially omissions and long-range drift, which is unacceptable for translation.

Unlike sentence-level translation where parallel corpora abound (Cheng et al., 2025; Zheng et al., 2025; Luo et al., 2025), document-level parallel data remains scarce (Jin et al., 2023; Wang et al., 2023; Maruf et al., 2022), as manually translating and annotating lengthy documents is prohibitively costly. Meanwhile, existing evaluation and reward signals are poorly suited for long outputs: widely-used learned metrics (Sellam et al., 2020; Rei et al., 2020; Zhang et al., 2020) have limited input lengths, and long-context benchmarks (Bai et al., 2025; Hsieh et al., 2024) largely focus on "long input $\rightarrow$ short output" comprehension rather than faithful long-form translation. As a result, we lack scalable ways to both train and measure long-context translation reliability.

We propose LongDu, a self-supervised post-training framework that turns round-trip consistency into scalable supervision for long-document translation. Given a monolingual source document, LongDu samples a group of candidate translations, back-translates each candidate, and rewards translations that better reconstruct the source. The key insight is simple: if the forward translation omits or distort information, that information cannot be reliably recovered in the backward direction, yielding a measurable training signal.

To make this signal robust for long-form translation, we introduce a "Stick–Carrot" reward: we first filter and penalize long-form translation reward-hacking (e.g., source copying and local language drift) with a window-level language constraint, and then apply a reconstruction score with a fluency term, enabling stable optimization with reinforce-

*Equal contributions.
†Work done during internship at Bytedance Seed.
‡Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 27269–27280

ment learning.

To evaluate information retention, we introduce Long-CIRT, an automatic protocol that measures downstream sensitivity to translation-induced distortion. Unlike our training-time reconstruction reward, Long-CIRT fixes a solver and a question set, and compares solver's accuracy on the original document versus on the round-tripped reconstruction. By back-translating to the source language and keeping solver and questions unchanged, the resulting accuracy drop isolates information changes introduced by the translation cycle rather than confounding cross-lingual QA ability.

Empirically, LongDu yields consistent gains across model families: for Qwen3-4B-Instruct, average BLEU improves from 49.21 to 55.10, and for Qwen3-30B-A3B-Instruct from 51.99 to 58.49; it also substantially mitigates the long-document failure of a translation-specialized model (LMT-60-8B: 10.48 $\rightarrow$ 29.61 BLEU). Beyond translation quality, Long-CIRT confirms improved information retention, reducing distortion after a translation cycle (e.g., QER 2.18 $\rightarrow$ 1.03 and DER 24 $\rightarrow$ 17 with a fixed solver), and our analysis shows these gains generalize beyond the training length range up to 16k tokens and transfer to unseen target languages. Overall, LongDu and Long-CIRT offer a practical step toward reliable document-level MT at scale, enabling reference-free optimization and measurement of information preservation in long-context translation.

2 Related Work

2.1 Long Context and Machine Translation

Although position-encoding extensions (Su et al., 2023; Press et al., 2022; Peng et al., 2023; Chen et al., 2023) and continued pre-training on long-context data (Chen et al., 2024b) have enabled some progress in scaling large language models (LLMs) to longer contexts, these approaches often rely on curated long-context corpora and costly human annotations. Meanwhile, existing benchmarks such as RULER (Hsieh et al., 2024) and LongBench (Bai et al., 2025) primarily evaluate models' comprehension of long inputs (i.e., the long input $\rightarrow$ short output setting), leaving long-form generation and information retention relatively underexplored (Liu et al., 2024b). Document-level machine translation (DocMT) naturally stresses both capabilities: it requires holistic understanding of long inputs and the production of long outputs that remain faithful to the source. However, most translation datasets and LLM-based translation systems (Cheng et al., 2025; Zheng et al., 2025; Luo et al., 2025) are still predominantly sentence-level, where inputs are typically only a few hundred tokens. Prior DocMT work has largely focused on discourse-level consistency (Voita et al., 2019; Jiang et al., 2021), and investigation of long-context DocMT in the LLM era remains limited.

2.2 Dual Learning

Dual learning enhances model performance by leveraging intrinsic task symmetry, where a primal task and its complementary dual task mutually provide supervision. He et al. (2016) first introduced dual learning for machine translation, which uses bidirectional tasks (e.g., En$\rightarrow$Zh and Zh$\rightarrow$En) to generate pseudo-labels via back-translation (Sennrich et al., 2016), reducing reliance on parallel corpora—a breakthrough for low-resource language pairs. Building on this foundation, the paradigm has proven highly versatile—spanning multi-modal (Yi et al., 2017; Zhu et al., 2017) and knowledge reasoning (Dognin et al., 2020), and extending to reinforcement learning (Luo et al., 2019; Bahng et al., 2025). In modern LLMs, it further refines output quality and enforces semantic consistency (Zou et al., 2025; Chen et al., 2024a). Recently, DuPO (She et al., 2025) design complementary duality to extend dual-learning to asymmetric tasks such as math reasoning. However, extending this self-supervised paradigm to the long-context setting presents unique challenges regarding information retention and scalability that have not yet been investigated.

3 Method

In this section, we exploit the task symmetry of translation to build a round-trip framework for faithfulness of long-context translation, covering both optimization and evaluation. We first present LongDu, a training method that improves information retention without external supervision, and then introduce Long-CIRT, an automatic evaluation protocol that quantifies translation-induced information distortion.

3.1 Problem Formulation: Translation as Information Conservation

Let x be a document written in the source language L_s, and let y be its translation in the target language

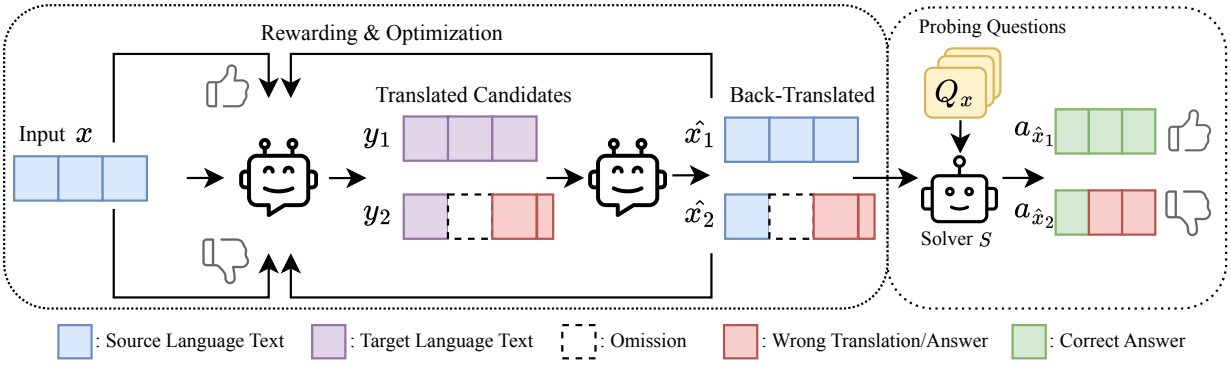

Figure 1: **Overview of LongDu training (left) and Long-CIRT evaluation (right).** *LongDu (left):* given a source-language document x, the policy π_θ samples candidate translations $\{y_i\}$ in the target language; each y_i is back-translated to a reconstruction $\hat{x}_i$ in the source language (Eq. 1). The policy is optimized (e.g., via GRPO) to prefer candidates with higher rewards based on reconstructability. *Long-CIRT (right):* each document x is paired with a probe question set $Q_x = \{q_j\}$ and gold answers. A fixed solver S answers the same questions using each reconstructed document $\hat{x}_i$; translation-induced information distortion is quantified by the solver's error increase (e.g., QER/DER) after the translation cycle.

L_t generated by a policy π_θ. Long-context translation failures are often "silent": translations remain fluent yet distort or omit entire paragraphs, entities, or numbers. Catching or correcting these omissions at scale is challenging, since document-level parallel corpora are scarce and reference-based evaluation are expensive.

We instead leverage a simple observation. If the forward translation y under-translates (omits), over-translates (adds), or mistranslates salient content in x, then even a strong back-translation system cannot reliably reconstruct the original x.

Therefore, the quality of the reconstructed document $\hat{x}$ after a translation cycle provides a natural, reference-free proxy for faithfulness. Concretely, we form a round-trip trajectory

$$x \xrightarrow{\pi_\theta(\cdot|x,L_t)} y \xrightarrow{\pi_{\text{back}}(\cdot|y,L_s)} \hat{x}, \tag{1}$$

and define a reconstruction-based reward that prefers translations that better preserve the information needed to recover the source:

$$R(x, y) \triangleq \text{Rec}(x, \hat{x}). \tag{2}$$

We could instantiate $\text{Rec}(\cdot)$ with a length-robust text reconstruction score (Sec. 3.2), enabling scalable reinforcement learning on long documents efficiently and robustly.

A practical challenge is that round-trip training is prone to *reward hacking* in long-form translation: an LLM can inflate reconstruction reward by copying source-language spans or producing non-translation outputs, which destabilizes optimization and reinforces degeneration. LongDu addresses this by using a reward design that first enforces a strict validity filter ("Stick") and then applies the reconstruction and fluency score ("Carrot").

3.2 Optimization Framework: LongDu

We assume access to a monolingual corpus $\mathcal{D}_{\text{mono}}$ in L_s. Given $x \sim \mathcal{D}_{\text{mono}}$, our goal is to improve the translation model $\pi_\theta(y \mid x, L_t)$

Round-trip trajectory sampling. For each source document x, we sample a group of various candidate translations $\{y_i\}_{i=1}^{G}$ from the current policy model. We then back-translate each candidate to obtain its reconstruction. To reduce reward variance, we use deterministic decoding for back-translation:

$$\hat{x}_i = \arg\max_{\hat{x}} \ \pi_{\text{back}}(\hat{x} \mid y_i, L_s). \tag{3}$$

Notice that we can set $\pi_{\text{back}} = \pi_\theta$ (utilizing the same LLM prompted for the reverse direction), yielding a fully self-contained training loop. Meanwhile, as training progresses, the model's evolving capabilities naturally enhance the back-translation process, fostering a mutually reinforcing co-evolution. Compared to standard RL, LongDu only adds one deterministic back-translation per sampled translation during rollout, so the rollout cost increases by up to $\sim 2\times$ while the parameter-updating stage remains unchanged (and no extra model is needed when $\pi_{\text{back}} = \pi_\theta$).

Constraints for Language and Fluency. Long translations often fail locally (e.g., a few tokens

revert to L_s), so document-level checks are insufficient. We segment y into K overlapping windows $\{w_k(y)\}_{k=1}^{K}$ and compute a violation rate using a language identifier $\mathrm{LID}(\cdot)$:

$$v(y) \triangleq \frac{1}{K}\sum_{k=1}^{K}\mathbb{I}\left[\mathrm{LID}(w_k(y)) \neq L_t\right]. \quad (4)$$

This windowed constraint effectively filters trajectories with local language drift and source-language copying. More details in Appendix A.1.

Meanwhile, solely optimizing reconstruction can lead to "unnatural encoding", where the model generates unnatural gibberish that is nonetheless easily reconstructible. To enforce high-quality generation, we introduce a fluency reward $R_{\text{flu}}(y_i)$ by computing the geometric mean of token probabilities assigned by the reference policy π_{ref} (e.g., the initial LLM, as a standard practice in RL fine-tuning).

Reward Function. We assign each forward translation a reward as below:

$$r(x,y) = \begin{cases} -1 & \text{if } v(y) > \tau, \\ R_{\text{rec}}(x,\hat{x}) + R_{\text{flu}}(y), & \text{otherwise}, \end{cases} \quad (5)$$

We adopt BLEU (Papineni et al., 2002) as a scalable proxy for reconstruction quality. While BLEU is an n-gram overlap metric and does not fully capture semantic equivalence, it offers two practical advantages that are critical: it imposes no hard constraint on input length and is computationally lightweight, making it feasible for long-document optimization, unlike learned metrics that require aggressive truncation or incur prohibitive cost. Importantly, reconstruction is evaluated in the same source language (L_s), under which BLEU's sensitivity to n-gram mismatches serves as an effective heuristic for detecting content omission and distortion, rather than stylistic variation.

Policy Optimization. The core of our framework is to optimize LLMs using the reward $r(x, y)$, without external annotations. The objective is to maximize the expected reward based on:

$$\mathcal{J}(\theta) = \mathbb{E}_{y\sim\pi_\theta(y|x)}\left[r(x,y)\right], \quad (6)$$

where $\pi_\theta(\mathbf{y}|\mathbf{x})$ denotes the policy LLM for generating output $\mathbf{y}$. Notably, LongDu is compatible with various RL algorithms (e.g., PPO (Schulman et al., 2017)), we adopt GRPO (Shao et al., 2024) in our experiments for its simplicity and efficiency.

3.3 Evaluation Protocol: Long-CIRT

We introduce Long-CIRT, a **Long** **C**ontext **I**nformation **R**etention **T**est protocol that quantifies translation-induced information distortion without additional human annotation.

Preparation. Our evaluation can be instantiated with any long-context benchmark that provides (document, task, answer) pairs. In this work, we use long-context QA as an example: each instance contains a source-language document $x \in L_s$, a question set Q_x with gold labels $\{l_q\}_{q\in Q_x}$.

We use LongBioBench (Yang et al., 2025b) which constructs seamless contexts, mitigating shortcut cues common in NIAH-style benchmarks, an important property for our translation-based stress tests. Moreover, the data generator is controllable and length scalable, while remaining highly consistent with performance on real-world long-document tasks. More details about our settings can refer to Appendix A.2.

Round-trip translation and QA. We choose an LLM with basic QA capability as the solver S. Based on the original document x, S achieves a baseline accuracy $\mathrm{Acc}_{\text{base}}$ by answering Q_x. For a translation model π_θ under evaluation, we then obtain $(y, \hat{x})$ via the translation cycle discussed in Sec 3.1, where we set $\pi_{\text{back}} = \pi_\theta$ (the same model prompted for the reverse direction). We run the *same* solver S on the *same* questions Q_x but conditioned on the reconstructed document $\hat{x}$, yielding Acc_{rt}. If the translation process loses information, S will be less able to answer correctly from $\hat{x}$, and the resulting accuracy drop reflects information distortion.

Importantly, since both QA runs are performed in the same language L_s (after back-translation), the metric is not confounded by the solver's cross-lingual QA ability. Moreover, we compare relative degradation under fixed S and fixed Q_x, which reduces sensitivity to the absolute strength of S.

Metrics. We evaluate information distortion using two error rates computed over the test set. Question Error Rate (QER) measures the average fraction of questions that are answered incorrectly, reflecting the severity of information loss. Document Error Rate (DER) measures the fraction of documents for which at least one associated question is answered incorrectly, capturing the breadth of affected content.

Formally, let the test set be $\mathcal{D} = \{x_1, \ldots, x_N\}$, where each document x_i is associated with a question set Q_{x_i} and ground-truth labels $\{l_q\}_{q \in Q_{x_i}}$. Given a fixed solver $S(\cdot, \cdot)$ and a reconstructed document $\hat{x}_i$, we define:

$$\text{QER} = \frac{1}{\sum_{i=1}^{N} |Q_{x_i}|} \sum_{i=1}^{N} \sum_{q \in Q_{x_i}} \mathbb{I}[S(\hat{x}_i, q) \neq l_q],$$

$$\text{DER} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[\exists q \in Q_{x_i},\ S(\hat{x}_i, q) \neq l_q]. \tag{7}$$

Although we instantiate LONG-CIRT with long-context QA, the same idea applies to other long-context tasks (e.g., long context summarization) with automatic verification: fix a downstream solver and measure its performance change after a translation cycle.

4 Experiment

We validate the efficacy of LongDu on long-document translation. Below, we detail the experimental setup, followed by main results.

4.1 Experiment Setup

Base Model. We evaluate LongDu on a diverse set of base models to demonstrate its effectiveness and robustness across different scales and specializations. Our primary base model is Qwen3-4B-Instruct-2507 (Yang et al., 2025a), a competitive open-source model that enables rapid and diverse experimentation. To assess scalability, we apply LongDu to a much stronger, 30B parameter MoE model, Qwen3-30B-A3B-Instruct (Yang et al., 2025a). We also evaluate QwenLong-L1.5-30B-A3B (Shen et al., 2025), a model optimized via reinforcement learning on synthetic data that achieves SOTA performance on long-context benchmarks. Furthermore, to examine its impact on models already specialized for translation, we use LMT-60-8B (Luo et al., 2025), which has been extensively trained on sentence-level translation. For comparison, we benchmark against a suite of state-of-the-art ultra large models, including DeepSeek-V3.1 (DeepSeek-AI et al., 2025), Doubao-Seed-1.6 (ByteDance Seed Team, 2025), Qwen3-235B-A22B (Yang et al., 2025a), and Kimi-K2-Thinking (Team et al., 2025), establishing a comprehensive evaluation landscape.

Dataset. We source our monolingual data from FineWeb-Edu (Penedo et al., 2024), a popular pre-training corpus for its high quality and diverse content. Based on the Qwen3 tokenizer, we sample 100k English documents between 0.5k and 2k tokens for training. We format the instruction to let LLM translate the English document to Chinese. For evaluation, our test set is stratified by length, comprising 200 documents from each bin within the 0.5k to 16k token range. All evaluation data is sourced from files disjoint from the training set. References for this test set are generated using the SOTA long-context LLM Gemini-2.5-Pro (DeepMind, 2025), exclusively used for evaluation.

Metrics. For translation quality, we report BLEU (Papineni et al., 2002) against references generated by Gemini-2.5-Pro. We exclude prominent encoder-based metrics like BLEURT (Sellam et al., 2020) and COMET (Rei et al., 2020), as their strict input length limits (e.g., 512 tokens) would necessitate severe truncation for our document-level tasks, rendering the scores unreliable. To capture semantic and discourse-level failures, we supplement this with an LLM-as-a-Judge framework. We prompt GPT-5.2-High to identify errors and apply weighted penalties based on their severity: 5 points for lexical, 10 for syntactic, and a substantial 50-point penalty for discourse-level errors (e.g., information omission or inconsistency).

For information retention, we then measure the accuracy drop when a fixed LLM as solver to answer based on the reconstructed document versus the original. A smaller accuracy drop indicates better information preservation. This metric directly targets the long-context failure mode we aim to address, content omission, which surface-level metrics like BLEU may miss. More details about training are provided in Appendix A.1.

4.2 Main Result

4.2.1 LongDu Improves Long-Document Translation without Parallel Data

Results from both automatic metrics and LLM-as-a-judge consistently show that LongDu improves the quality and faithfulness of long-document translation, indicating genuine translation gains, despite using no parallel data or human supervision.

As shown in Table 1, applying LongDu to Qwen3-4B-Instruct boosts the average BLEU score from 49.21 to 55.10 (+5.89), enabling a compact 4B model to rival and even surpass much larger LLMs, including DeepSeek-V3.1 (50.20) and Doubao-Seed-1.6 (50.29). Complementing

Model	Context Length (K)					Avg.
	0.5K–1K	1K–2K	2K–4K	4K–8K	8K–16K	
DeepSeek-V3.1	49.12	50.02	51.76	51.56	48.55	50.20
Doubao-Seed-1.6-251015	49.31	50.54	51.79	51.69	48.12	50.29
QwenLong-L1.5-30B-A3B	51.41	52.39	53.61	51.76	49.50	51.73
Qwen3-235B-A22B-Instruct-2507	55.03	55.49	57.35	58.31	58.53	56.94
Kimi-K2-thinking-251104	59.64	61.02	61.49	59.60	56.39	59.63
LMT-60-8B	33.36	15.79	3.13	0.12	0.02	10.48
w/ LongDu (ours)	51.34	50.64	38.76	6.99	0.30	29.61
Qwen3-4B-Instruct-2507	46.79	47.48	49.40	50.14	52.23	49.21
w/ LongDu (ours)	52.97	54.25	56.61	55.95	55.74	55.10
Qwen3-30B-A3B-Instruct-2507	50.63	50.47	52.06	51.72	55.07	51.99
w/ LongDu (ours)	56.29	57.56	60.06	59.03	59.51	58.49

Table 1: **Performance comparison on the Long Document Translation Benchmark across varying context lengths (0.5K to 16K tokens).** We report BLEU scores against reference from Gemini-2.5-Pro. LongDu consistently improves long-document translation across model families and length buckets, and partially mitigates catastrophic long-context degradation.

these BLEU gains, Figure 2 shows that LongDu consistently reduces translation errors and improves faithfulness across all tested lengths, indicating fewer severe discourse-level issues such as omissions and inconsistencies.

Importantly, the gains are not confined to a single backbone: LongDu improves both translation-specialized and general-purpose LLMs. For instance, it raises the average BLEU of LMT-60-8B from 10.48 to 29.61, and consistently benefits larger, stronger models such as Qwen3-30B-A3B-Instruct (51.99 → 58.49), demonstrating the robustness of our optimization across model scales.

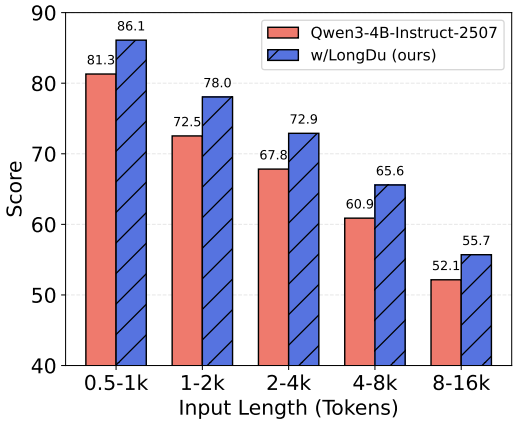

Figure 2: **LLM-as-judge scores of translation quality (0-100; higher is better) across input-length buckets.** We assess each model via GPT-5.2-High with weighted penalties for errors. LongDu yields consistently higher scores at every context length, indicating fewer severe long-document translation failures such as omissions and inconsistencies.

4.2.2 Pitfall of Current Translation LLM and Long Context Post-Training

These comparisons also reveal critical limitations of existing paradigms for long-document translation. First, models extensively trained for sentence-level MT can degrade catastrophically as input length grows. As shown in Table 1, LMT-60-8B drops from 33.36 BLEU at 0.5K–1K to 15.79 at 1K–2K, and becomes nearly unusable beyond 2K (3.13/0.12/0.02 BLEU), where we frequently observe very severely truncated outputs. This suggests that conventional MT LLM training (e.g., extensive CPT/SFT on sentence-level parallel data) substantially improves short-context translation yet fails to enforce document-level coverage and information preservation.

Conversely, generic long-context adaptation alone is insufficient. Despite being optimized for long-context usage like QA, QwenLong-L1.5-30B-A3B does not improve long-document translation over its similar-size baseline Qwen3-30B-A3B-Instruct-2507 (51.73 vs. 51.99 BLEU on average), indicating that extending or optimizing long-context capabilities without translation-specific faithfulness supervision does not reliably transfer to this setting. Together, these observations motivate LongDu, which directly targets information preservation in long-document translation via cyclic consistency without parallel data.

4.2.3 LongDu Generalizes Beyond the Training Length Range

To stress-test length generalization, we deliberately restrict LongDu training to relatively short docu-

ments (0.5k–2k tokens), and evaluate on substantially longer inputs up to 16k tokens. Table 1 shows that LongDu improves translation quality not only within the training band, but also beyond it. For Qwen3-4B, BLEU increases by +6.18 (0.5k–1k) and +6.77 (1k–2k) in-range, while remaining consistently positive out-of-distribution: +7.21 on 2k–4k, +5.81 on 4k–8k, and +3.51 on 8k–16k. Importantly, this trend is corroborated by the LLM-as-judge scores in Figure 2, which rise across every length bucket (e.g., 0.5k–1k: 81.3→86.1; 1k–2k: 72.5→78.0 ; 2k–4k: 67.8→72.9), indicating fewer severe long-document errors rather than mere lexical matching.

We further observe a length-dependent pattern: improvements persist beyond the 0.5k–2k training regime but taper at very long contexts (Table 1). A plausible explanation is that long-document translation amplifies error accumulation: small local mistakes (e.g., partial omissions or subtle drift) can propagate across sections, and maintaining discourse-level consistency becomes increasingly difficult as the context expands. This is consistent with the LLM-as-judge trends in Figure 2, which show universal gains yet smaller increments in the 8k–16k range where omission is more severe.

4.2.4 LongDu Reduces Information Distortion

We further assess faithfulness on Long-CIRT, which isolates translation-induced information loss via a round-trip QA protocol. Table 2 shows that applying LongDu consistently reduces both QER and DER across solvers and base translation systems. For Qwen3-4B-Instruct-2507, LongDu lowers DER from 24.0→17.0 and QER from 2.18→1.03 under Doubao-1.6-Thinking, and yields analogous gains with a weaker Qwen3-30B-A3B solver (DER: 30.5→26.5; QER: 2.74→1.94). Notably, LongDu also mitigates the severe distortion exhibited by translation-specialized LMT-60-8B on long documents, reducing QER from 68.18→62.36 and DER from 72.5→70.5 with Seed-1.6-Thinking, and from 69.03→60.72 (QER) and 75.0→70.75 (DER) with Qwen3-30B-A3B.

These consistent reductions directly support our central hypothesis: optimizing round-trip reconstructability makes omissions costly, if a forward translation discards high-entropy details (e.g., entities, dates, or relations), they cannot be reliably recovered in $\hat{x}$, leading to downstream QA failures. The similar trend also indicates that we focus on relative performance changes, implying that our approach does not strictly require a highly capable solver and thus enjoys broader applications.

System	QER (%) ↓	DER (%) ↓
Solver: Qwen3-30B-A3B-Instruct-2507		
Original Source Text	0.13	2.50
LMT-60-8B	69.03	75.00
w/ LongDu (ours)	60.72	70.75
Qwen3-4B-Instruct-2507	2.74	30.50
w/ LongDu (ours)	1.94	26.50
Solver: Doubao-1.6-Thinking		
Original Source Text	0.00	0.00
LMT-60-8B	68.18	72.50
w/ LongDu (ours)	62.36	70.50
Qwen3-4B-Instruct-2507	2.18	24.00
w/ LongDu (ours)	1.03	17.00

Table 2: Evaluation of Information Retention on Long-CIRT. We report the Error Rate (ER) on questions and document to measure faithfulness. Lower ER indicates that the translation preserves more specific factual details required to answer the questions. LongDu reduces ER, quantitatively demonstrating reduced information loss during the translation process.

4.2.5 LongDu Generalizes Across Languages

LongDu generalizes beyond the training language pair. Although LongDu training is conducted on English-Chinese, Figure 3 shows consistent BLEU improvements when evaluating on translations into (e.g., Thai, Japanese, German, and Spanish). This suggests that the learned preference signal primarily strengthens a general long-context retention behavior—maintaining recoverable information across long horizons—rather than learning language-pair-specific heuristics. Consequently, LongDu can transfer its benefits to unseen target languages, supporting scalable long-context post-training without language-specific annotation.

4.2.6 LongDu Generalizes Beyond Translation

We adopt LongGenBench (Liu et al., 2024b), a benchmark for long-form generation that requires 16K-token outputs while following structured constraints (single-point, range, and periodic instructions) across strictly sequential subtasks (e.g., diary writing, menu planning, and urban planning). Following the official protocol, we report (i) STIC-1/2, the specific-instruction completion rate computed over generated subtasks and over the full required instruction set respectively, and (ii) wAvg = CR × STIC-2 (CR, the main-task completion rate) ,

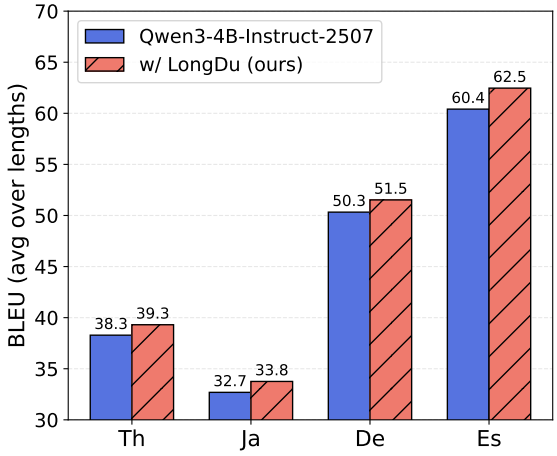

Figure 3: Cross-lingual generalization: BLEU (averaged over length buckets) for translating English long documents into four unseen target languages (Thai, Japanese, German, Spanish). Although trained only on English-Chinese, LongDu consistently improves performance across diverse languages, suggesting a language-agnostic gain in long-context information retention.

System	STIC-1	STIC-2	wAvg
Qwen3-4B-Instruct-2507	30.0	29.0	28.8
w/ LongDu (ours)	31.0	31.0	30.6

Table 3: Task Generalization on LongGen-Bench. We evaluate whether the capability learned via LongDu transfers to general long-form generation tasks beyond translation. STIC metrics measure strict instruction compliance (e.g., constraints on length or formatting). Results show that LongDu improves constraint satisfaction, suggesting the model acquires a transferable competence in controlled long-context generation.

which penalizes incomplete generation. As shown in Table 3, applying LongDu improves STIC-2 from 29.0 to 31.0 and wAvg from 28.8 to 30.6 under the 16K setting, indicating better long-horizon constraint tracking. These results demonstrate that LongDu learns a transferable competence for controlled long-form generation beyond translation.

4.2.7 LongDu Yields Stable and Progressive Improvements

Figure 4 tracks LongDu training on a translation-expert LLM (LMT-60-8B) and a general instruction model (Qwen3-4B-Instruct-2507). The language consistency remains above 99.5% for all steps, showing stable optimization and indicating that our reward design effectively prevents reward hacking via copying the source text, keeping producing genuine translation.

Meanwhile, the reconstruction score increases

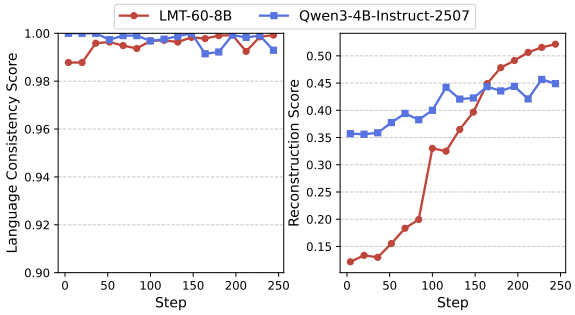

Figure 4: Training dynamics of LongDu. We report the language consistency score (left) and reconstruction score (right) throughout RL training for two base models. LongDu maintains near-perfect language consistency while steadily improving reconstruction quality.

steadily, suggesting that LongDu progressively strengthens long-context reconstruction and thus better preserves document-level semantics and facts. Notably, LMT-60-8B starts with a much lower reconstruction score, consistent with its sentence-level translation bias and weaker long-document ability, but improves rapidly during training, whereas Qwen3 starts higher and improves more smoothly. These dynamics align with the large gains in long-context translation quality reported in Table 1.

5 Conclusion

In this work, we study faithfulness degradation in end-to-end long-context translation, where human supervision is prohibitively expensive and traditional metrics fall short. We introduce LongDu, a self-supervised framework that leverages round-trip reconstruction as a training signal and optimizes with RL to better preserve information without requiring document-level parallel data. To evaluate translation-induced information loss automatically, we further propose Long-CIRT, a round-trip evaluation protocol that fixes an external solver and quantifies the accuracy drop on reconstructed documents. Across multiple backbones and a wide range of document lengths, LongDu consistently improves both translation quality and faithfulness, and transfers to held-out language pairs and length regimes beyond training. Overall, our work demonstrates a practical approach to derive self-supervised signals for both training and evaluation in the challenging long-context setting, offering a promising direction for future research.

Limitations

Although our reward design includes explicit language constraints and a fluency term, and our LLM-as-a-judge evaluation suggests improved translation quality, LongDu may nonetheless alter a model's original translation style, potentially biasing it toward more direct and concise (i.e., more literal) renderings. Carefully characterizing and controlling such stylistic drift, especially for domains where tone, register, and authorial voice matter, is left to future work. Moreover, due to computational constraints, our experiments scale post-training only up to 30B-parameter backbones and train on documents up to 2K tokens; we plan to further scale along model strength, context length, and language coverage to push the frontier of long-document translation.

Meanwhile, our method appears to rely on a basic baseline long-context competence from the base model. Take LMT-60-8B as the example, while LongDu substantially mitigates omission issues within the 0-2K regime, performance remains noticeably weaker when extending to much longer contexts (e.g., 4-16K). Developing training strategies that better bridge this gap, such as multi-round/iterative post-training curricula, or coupling with monolingual long-context corpora is an important direction for our future work.

Acknowledgement

We would like to thank the anonymous reviewers for their insightful comments and constructive feedback, which greatly enhanced the rigor and clarity of our manuscript. Correspondence should be addressed to Shujian Huang. This work is supported by the following funding sources: the National Science Foundation of China (Grant No. 62376116); the Research Project of Nanjing University-China Mobile Joint Institute (Grant No. NJ20250038); the Fundamental Research Funds for the Central Universities (Grant No. 2024300507); and the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (Grant No. JYB2025XDXM118).

References

Hyojin Bahng, Caroline Chan, Fredo Durand, and Phillip Isola. 2025. Cycle consistency as reward: Learning image-text alignment without human preferences. *arXiv preprint arXiv:2506.02095*.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3639–3664, Vienna, Austria. Association for Computational Linguistics.

ByteDance Seed Team. 2025. Seed1.5-Thinking: Advancing superb reasoning models with reinforcement learning. *Preprint*, arXiv:2504.13914.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024a. DUAL-REFLECT: Enhancing large language models for reflective translation through dual learning feedback mechanisms. *Preprint*, arXiv:2406.07232.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2024b. Longlora: Efficient fine-tuning of long-context large language models. *Preprint*, arXiv:2309.12307.

Shanbo Cheng, Yu Bao, Qian Cao, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, Wenhao Zhu, Jingwen Chen, Zhichao Huang, Tao Li, Yifu Li, Huiying Lin, Sitong Liu, Ningxin Peng, Shuaijie She, Lu Xu, Nuo Xu, Sen Yang, and 7 others. 2025. Seed-x: Building strong multilingual translation llm with 7b parameters. *Preprint*, arXiv:2507.13618.

DeepMind. 2025. Gemini 2.5. `https://deepmind.google/technologies/gemini/`. Accessed: 2025-04-18.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Pierre L. Dognin, Igor Melnyk, Inkit Padhi, Cícero Nogueira dos Santos, and Payel Das. 2020. Dualtkb: A dual learning bridge between text and knowledge base. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8605–8616. Association for Computational Linguistics.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 820–828.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*.

Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and M. Zhou. 2021. Blonde: An automatic evaluation metric for document-level machine translation. In *North American Chapter of the Association for Computational Linguistics*.

Linghao Jin, Jacqueline He, Jonathan May, and Xuezhe Ma. 2023. Challenges in context-aware neural machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15246–15263, Singapore. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguistics*, 12:157–173.

Xiang Liu, Peijie Dong, Xuming Hu, and Xiaowen Chu. 2024b. Longgenbench: Long-context generation benchmark. *arXiv preprint arXiv:2410.04199*.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5116–5122. ijcai.org.

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

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2022. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2):45:1–45:36.

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

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Preprint*, arXiv:2406.17557.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*.

Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. *Preprint*, arXiv:2108.12409.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for mt evaluation. *Preprint*, arXiv:2009.09025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

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

Shuaijie She, Yu Bao, Yu Lu, Lu Xu, Tao Li, Wenhao Zhu, Shujian Huang, Shanbo Cheng, Lu Lu, and Yuxuan Wang. 2025. Dupo: Enabling reliable llm self-verification via dual preference optimization. *Preprint*, arXiv:2508.14460.

Weizhou Shen, Ziyi Yang, Chenliang Li, Zhiyuan Lu, Miao Peng, Huashan Sun, Yingcheng Shi, Shengyi Liao, Shaopeng Lai, Bo Zhang, Dayiheng Liu, Fei Huang, Jingren Zhou, and Ming Yan. 2025. Qwenlong-l1.5: Post-training recipe for long-context reasoning and memory management. *Preprint*, arXiv:2512.12967.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. Roformer: Enhanced transformer with rotary position embedding. *Preprint*, arXiv:2104.09864.

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*.

Lifu Tu, Rui Meng, Shafiq Joty, Yingbo Zhou, and Semih Yavuz. 2025. Investigating factuality in long-form text generation: The roles of self-known and self-unknownan. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*, pages 322–336, Suzhou, China. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Rongxiang Weng, Heng Yu, Xiangpeng Wei, and Weihua Luo. 2020. Towards enhancing faithfulness for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2675–2684, Online. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yijun Yang, Zeyu Huang, Wenhao Zhu, Zihan Qiu, Fei Yuan, Jeff Z Pan, and Ivan Titov. 2025b. A controllable examination for long-context language models. *arXiv preprint arXiv:2506.02921*.

Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. 2017. Dualgan: Unsupervised dual learning for image-to-image translation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2868–2876. IEEE Computer Society.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025. Hunyuan-mt technical report. *Preprint*, arXiv:2509.05209.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

Wei Zou, Sen Yang, Yu Bao, Shujian Huang, Jiajun Chen, and Shanbo Cheng. 2025. TRANS-ZERO: Self-play incentivizes large language models for multilingual translation without parallel data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12337–12347, Vienna, Austria. Association for Computational Linguistics.

A Appendix

A.1 Experimental Setup Details

Reinforcement Learning Configuration We implement our reinforcement learning (RL) pipeline using the veRL (Sheng et al., 2024). For each optimization step, we employ a global batch size of 128 prompts. We perform $T = 1$ rollout per iteration with $n = 32$ responses sampled for each prompt. The maximum input prompt length is capped at the maximum length of the corresponding dataset, while the maximum generation length is dynamically set to $1.5\times$ the input length.

We utilize the Adam (Kingma and Ba, 2017) optimizer with a learning rate of 1×10^{-6} and a weight decay of 0.1. Notably, the KL divergence coefficient is set to 0 in our primary experiments to allow the policy maximum flexibility within the constrained action space. For the PPO clipping mechanism, we adopt a specialized configuration with $\epsilon_{low} = 0.2$, $\epsilon_{high} = 0.28$, and $\epsilon_c = 10$ as default setting in veRL.

Evaluation Configuration For models with predefined prompts (e.g., LMT-60-8B), we follow their settings. For general LLMs, we use the default prompt A.1; for reasoning models, we remove the reasoning traces and extract only the translation for evaluation.

To ensure reproducibility, we employ greedy decoding with a temperature of $T = 0$. The maximum output token limit is extended to the model's architectural maximum to prevent truncation.

Language Identification We segment each generated translation using fixed-size windows of three words/characters. For each segment, we apply an off-the-shelf language identification tool and compute the fraction of segments whose detected language does not match the target language. Translations with a violation rate above 5% are treated as invalid, effectively filtering local language drift and source-language copying.

Prompt template for translation

```
Please translate the following text
directly into {tgt_lang} without
any explanation.

{text}
```

A.2 LongBioBench Setup Details

We place 30 probe questions at uniformly spaced positions ranging from the top 10% to 80% of each document. Each question is multiple-choice with seven options to reduce the chance of random guessing. We construct four document-length bins (1, 2, 4, and 8). For each length bin, we sample 100 documents and generate 3,000 questions in total. For evaluation, we only extract the final chosen option from the model output and verify its correctness.

A.3 Information About Use Of AI Assistants

AI Assistants is only used during the writing process, primarily to check spelling and grammatical errors; all outputs and suggestions from the AI Assistants are manually reviewed.

A.4 Used Scientific Artifacts

Below are the scientific artifacts used in our work. For the sake of ethics, our use of these artifacts is consistent with their intended use.

- *Transformers (Apache-2.0 license)*, a framework to facilitate downloading and training state-of-the-art pretrained models.
- *FineWeb-Edu ((ODC-By) v1.0 license)*, A pretraining dataset sourced from internet text, used for LLM training.
- *veRL (Apache-2.0 license)*, A reinforcement learning code framework for training LLMs.