

Source-Grounded Semantic Reinforcement Learning for Low-Resource Target-Language Generation

Zeli Su^{1,2,*} Ziyin Zhang^{3,*} Zewei Pan^{3,*} Zhou Liu^{4,*} Dingcheng Huang⁵ Dehan Li⁶

Zhankai Xu² Longfei Zheng² Xiaolu Zhang² Jun Zhou^{2,†} Wentao Zhang^{4,†}

¹ Minzu University of China ² Ant Group ³ Shanghai Jiao Tong University

⁴ Peking University ⁵ Harbin Institute of Technology ⁶ South China University of Technology

* Equal contribution † Corresponding authors

Abstract

Low-resource target-language generation is often limited by scarce parallel data, while high-resource source-language monolingual data is abundant but difficult to use with standard supervised fine-tuning. We propose **Source-Grounded Semantic Reinforcement Learning (SG-SRL)**, a resource-utilization framework that converts source-language monolingual data into cross-lingual semantic supervision. SG-SRL performs reference-free reinforcement learning (RL) on source-language data with a cross-lingual semantic reward, followed by lightweight supervised recovery. Although direct reward optimization induces verbosity-based reward hacking, recovery restores fluency, conciseness, and task format without erasing the learned semantic alignment. On Chinese-to-Thai generation, SG-SRL outperforms cold-start SFT, source-only continued pretraining, and synthetic-data SFT; human preference and training-independent metrics corroborate the gains. Recovery ablations further show that semantic retention and form restoration are jointly controlled by recovery data and intensity. An in-domain Chinese-to-Tibetan experiment shows that the reward interface can use a contrastively trained encoder when a capable multilingual reranker is unavailable, while also delimiting the framework’s dependence on reward-model discrimination.

1 Introduction

Large language models (LLMs) have achieved strong general-purpose generation abilities through large-scale pretraining and post-training (DeepSeek-AI, 2024; Guo et al., 2025; Yang et al., 2025), but their performance remains highly uneven across languages. High-resource languages

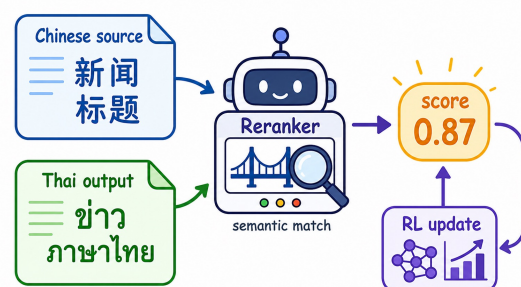


Figure 1: Reranker-based source-grounded semantic reward. A source-language input and a generated target-language output are treated as a cross-lingual query–candidate pair. The reranker estimates their semantic match and provides a scalar reward for RL without requiring a target-language reference.

benefit from abundant pretraining text, instruction data, and task-specific supervision, whereas low-resource languages often suffer from unstable generation, hallucinated content, poor factual grounding, and weak task adaptation (Üstün et al., 2024; Qin et al., 2025). Continual pretraining on target-language corpora followed by supervised fine-tuning (SFT) on downstream tasks has been employed for extending LLMs to new languages (Wang et al., 2020; Joshi et al., 2025), but it still relies on substantial target-language text or target-language task supervision, which is often unavailable for genuinely low-resource languages.

This data bottleneck becomes even more severe in cross-lingual generation. For many emerging or weakly supported target languages, high-quality parallel data is expensive to collect, and task-specific target-language annotations are even scarcer. In contrast, high-resource source-language data, such as Chinese or English news articles, is often abundant and reliable. Standard SFT can-

not directly use such source-language monolingual data, because it requires a target-language reference for each training instance. This leads to the central question of this work: *Can abundant source-language monolingual data be converted into useful supervision for low-resource target-language generation, without requiring target-language references during training?*

To this end, we need a supervision signal that can compare a source-language input with a target-language generation without a gold target-language reference. Multilingual rerankers provide a practical approximation: they score the relevance between a query and a candidate text (Nogueira and Cho, 2019), and recent LLM-based variants can distinguish semantically aligned cross-lingual pairs from unrelated ones (Sun et al., 2023; Zhang et al., 2024, 2025). As illustrated in Figure 1, we treat the source input as the query and the generated target-language output as the candidate, and use a reranker as a source-grounded semantic reward. This converts source-language monolingual data into scalable RL supervision without target-language references.

Empirically, direct optimization of this reward exposes a semantic-form trade-off. The relevance-oriented reward encourages the policy to generate longer outputs that cover more source-side concepts, leading to *verbosity-based reward hacking*: the intermediate RL checkpoint improves semantic coverage but becomes verbose, poorly formatted, and less fluent (Amodei et al., 2016). Stronger length or format constraints can reduce verbosity, but may also suppress useful source-to-target semantic learning. We therefore treat source-grounded semantic RL as *semantic mid-training* rather than final optimization.

Based on this insight, we propose **Source-Grounded Semantic Reinforcement Learning (SG-SRL)**, a train-reinforce-recover framework for low-resource target-language generation. SG-SRL first learns target-language form from a small parallel corpus, then reinforces source-grounded semantics on source-language monolingual data, and finally reuses the parallel corpus to recover fluent and concise target-language generation. On Chinese-to-Thai generation, starting from SmolLM3-3B (Bakouch et al., 2025), a model with weak Thai support, SG-SRL improves over cold-start SFT as well as source-only continued pretraining and synthetic-data SFT. Human preference and training-independent automatic metrics

agree with the LLM-based evaluation, while recovery ablations show that the semantic gains survive form restoration. We further examine an in-domain Tibetan setting where no capable LLM reranker is available. A contrastively trained Tibetan-Chinese encoder provides a different implementation of the semantic-reward interface, allowing us to separate reward-module portability from broader out-of-domain generalization.

Our contributions are summarized as follows:

- We propose **SG-SRL**, a train-reinforce-recover framework that converts source-language monolingual data into semantic supervision to enable reference-free RL in the low-resource target-language generation setting.
- We instantiate SG-SRL with a multilingual reranker reward, and identify a semantic-form trade-off: direct optimization induces verbosity-based reward hacking, but the intermediate checkpoint can still learn useful cross-lingual semantic alignment.
- We show on Chinese-to-Thai generation with SmolLM3-3B that semantic RL as mid-training outperforms stronger data-use baselines under human, automatic, and LLM-based evaluation, and we quantify how recovery data and intensity trade off semantic retention against target-form restoration.
- We establish the scope of reward-module portability in an in-domain Tibetan setting: a discriminative encoder reward can replace the LLM reranker, but weak or out-of-domain reward discrimination remains a limiting condition.

2 Related Work

Low-resource language expansion. Expanding language models to low-resource languages has commonly been approached through multilingual pretraining (Conneau et al., 2020; Xue et al., 2021), continued pretraining (Wang et al., 2020; Joshi et al., 2025), or supervised fine-tuning on target-language data. However, these approaches still depend on the availability of target-language text or task-specific supervision. In genuinely low-resource settings, both pretraining corpora and high-quality downstream parallel data may be limited, motivating methods that can exploit abun-

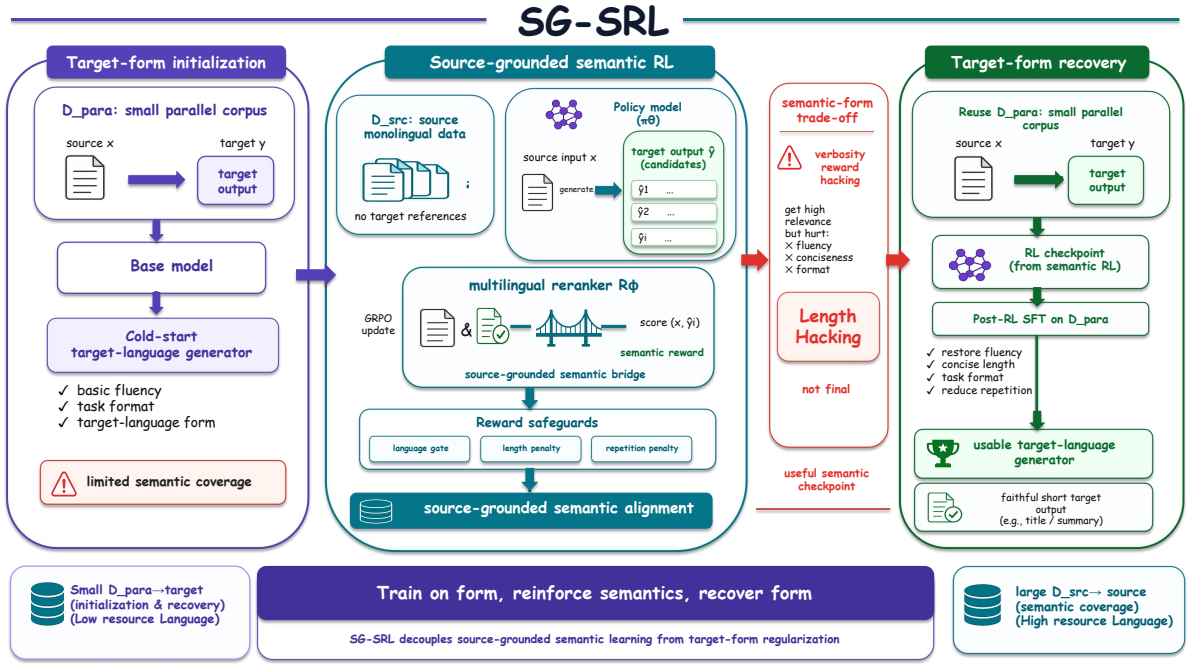


Figure 2: Overview of SG-SRL. The framework uses a small parallel corpus $\mathcal{D}_{\text{para}}$ for target-form initialization and target-form recovery, and a large source-language monolingual corpus $\mathcal{D}_{\text{src}}$ for source-grounded semantic RL. The RL stage uses a multilingual reranker as a weak semantic reward model, with language, length, and repetition safeguards. A final recovery stage reuses $\mathcal{D}_{\text{para}}$ to restore fluency, conciseness, and task format.

dant source-language data without requiring target-language references for every instance.

Semantic rewards for low-resource language learning. Recent work has explored reinforcement learning with semantic rewards as an alternative to token-level likelihood optimization for low-resource language expansion. [Su et al. \(2026\)](#) propose using embedding-level semantic rewards with GRPO to improve low-resource language capabilities while reducing alignment tax, showing that semantic-space optimization can preserve general capabilities better than conventional SFT. Our work shares the same motivation, and further extends the horizon to a reference-free, cross-lingual generation setting.

Cross-lingual semantic matching and LLM-based reranking. Dense multilingual representations and cross-lingual retrieval models provide a basis for measuring semantic correspondence across languages ([Feng et al., 2022](#); [Bonifacio et al., 2021](#)). Recent LLM-based rerankers further formulate relevance estimation as an instruction-following judgment over a query–candidate pair, often deriving a score from the model’s preference for an affirmative answer ([Sun et al., 2023](#); [Zhang et al., 2024, 2025](#)). We adopt this formulation by treating the source-language input as the query and

the generated target-language output as the candidate. Unlike standard retrieval, however, we use the reranker’s yes/no relevance probability as an online reward for policy optimization. This turns the reranker from an offline ranking module into an optimization target, which exposes relevance-oriented reward exploitation such as verbosity-based reward hacking.

Reinforcement learning and reward hacking. Reinforcement learning has become a central tool for adapting language models to human preferences and task-specific objectives ([Ouyang et al., 2022](#); [Shao et al., 2024](#)). However, optimizing imperfect proxy rewards can lead to reward hacking, where models satisfy the literal reward function while violating the intended behavior ([Amodei et al., 2016](#)). In our setting, the reranker reward is semantically informative but relevance-oriented: longer generations can cover more source-side information and thus receive higher scores, even when they degrade fluency, conciseness, or task format. Rather than treating this as a failure of RL alone, we show that the hacked intermediate policy can still acquire useful cross-lingual semantic alignment.

3 Method

SG-SRL uses two data sources for different purposes: a small parallel corpus $\mathcal{D}_{\text{para}}$ for target-

form initialization and recovery, and a large source-language monolingual corpus $\mathcal{D}_{\text{src}}$ for semantic mid-training. As shown in Figure 2, SG-SRL follows a train–reinforce–recover procedure: train target-language form from $\mathcal{D}_{\text{para}}$, reinforce source-grounded semantics from $\mathcal{D}_{\text{src}}$, and recover target-language form again with $\mathcal{D}_{\text{para}}$.

3.1 Problem Formulation

We consider a high-resource source language $\mathcal{X}$ and a low-resource or weakly supported target language $\mathcal{Y}$. We are given a small parallel corpus

$$\mathcal{D}_{\text{para}} = \{(x_i, y_i)\}_{i=1}^N, \quad (1)$$

where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, and a much larger source-language monolingual corpus

$$\mathcal{D}_{\text{src}} = \{x_j\}_{j=1}^M, \quad M \gg N, \quad (2)$$

which contains no target-language references.

The goal is to train a policy model π_θ that generates a target-language output $\hat{y} \in \mathcal{Y}$ conditioned on a source-language input $x \in \mathcal{X}$. Standard supervised fine-tuning only uses $\mathcal{D}_{\text{para}}$. SG-SRL converts $\mathcal{D}_{\text{src}}$ into RL training data by using a cross-lingual semantic scorer as a weak reward model.

3.2 SG-SRL Overview

SG-SRL has three stages:

1. **Target-form initialization:** fine-tune the base model on $\mathcal{D}_{\text{para}}$ for 3 epochs to obtain a cold-start target-language generator.
2. **Source-grounded semantic RL:** optimize the initialized model on $\mathcal{D}_{\text{src}}$ for 2 epochs using a reranker-based semantic reward.
3. **Target-form recovery:** fine-tune the RL checkpoint again on $\mathcal{D}_{\text{para}}$ for 1 epoch to restore fluency, length control, and task format.

The design separates target-language form learning from source-grounded semantic learning. The small parallel corpus gives reliable but limited form supervision, while the source-language monolingual corpus provides broader semantic coverage without target-language references.

3.3 Target-form Initialization

We first perform supervised fine-tuning on $\mathcal{D}_{\text{para}}$:

$$\theta_{\text{sft}} = \arg \min_{\theta} -\mathbb{E}_{(x,y) \sim \mathcal{D}_{\text{para}}} \log \pi_{\theta}(y | x). \quad (3)$$

This stage teaches the model basic target-language form, task format, and local generation style. However, because $\mathcal{D}_{\text{para}}$ is small, the resulting cold-start model may still miss source-side entities, events, or factual relations. The next stage therefore uses $\mathcal{D}_{\text{src}}$ to inject additional semantic grounding.

3.4 Source-grounded Semantic RL

Starting from $\pi_{\theta_{\text{sft}}}$, we perform RL on $\mathcal{D}_{\text{src}}$. For each source input $x \sim \mathcal{D}_{\text{src}}$, the policy samples

$$\hat{y} \sim \pi_{\theta}(\cdot | x). \quad (4)$$

Since no target-language reference is available, we use the reranker format in Figure 1: the source input is treated as the query, and the generated target-language output is treated as the candidate document.

Reranker reward. We use Qwen3-Reranker-8B (Qwen Team, 2025) as a generative yes/no judge. Given an instruction I , source input x , and generated output $\hat{y}$, we define the semantic reward as the normalized probability of the affirmative answer:

$$\begin{aligned} z_{\phi}^{+} &= z_{\phi}(\text{yes} | I, x, \hat{y}), \\ z_{\phi}^{-} &= z_{\phi}(\text{no} | I, x, \hat{y}), \\ r_{\text{rank}}(x, \hat{y}) &= \frac{\exp z_{\phi}^{+}}{\exp z_{\phi}^{+} + \exp z_{\phi}^{-}}. \end{aligned} \quad (5)$$

This score lies in $[0, 1]$ and provides the main source-grounded semantic signal.

Reward safeguards. To reduce obvious degeneration, we combine the reranker score with the three safeguards shown in Figure 2: a hard language gate, a batch-relative length penalty, and a repetition penalty. For a candidate $\hat{y}_i$ in reward batch B , the final reward is

$$\begin{aligned} s_i &= r_i^{\text{rank}} - \lambda_{\text{len}} p_i^{\text{len}} - \lambda_{\text{rep}} p_i^{\text{rep}}, \\ r_i &= g_i \max(s_i, 0). \end{aligned} \quad (6)$$

Here, g_i is a hard target-language gate. In the Chinese–Thai setting, an output passes when at least 90% of its characters are Thai, it contains no Chinese characters, and at most 5% are Latin-script characters. If the gate fails, the reward is set to zero.

The length penalty p_i^{len} is computed with the reranker tokenizer and is relative to the current reward batch. It starts only when an output exceeds 2.5 times the batch median length and increases

linearly until 5.0 times the median length. The repetition penalty p_i^{rep} is based on repeated 4-grams and is activated when the repeated 4-gram ratio exceeds 0.15. We set $\lambda_{\text{len}} = 0.20$ and $\lambda_{\text{rep}} = 0.05$ in the main configuration. Thus, semantic alignment is the only graded positive signal, while language validity, length control, and repetition control act as safeguards.

Policy optimization. We optimize the policy with GRPO (Shao et al., 2024). For each source input, we sample a group of candidate outputs, compute their rewards, normalize rewards within the group, and update the policy according to relative advantages. The reference policy is initialized from $\pi_{\theta_{\text{sft}}}$ to reduce drift from the cold-start target-language behavior.

We use this RL stage as semantic mid-training rather than final optimization. Directly optimizing the semantic reward can improve source-grounded coverage, but may also produce verbose or poorly formatted outputs. The purpose of this stage is therefore to learn source-to-target semantic alignment from $\mathcal{D}_{\text{src}}$, not to produce the final deployable generator.

3.5 Target-form Recovery

After semantic RL, we obtain an intermediate policy $\pi_{\theta_{\text{rl}}}$ that has absorbed source-side semantic supervision but may have degraded surface form. We then reuse $\mathcal{D}_{\text{para}}$ for 1 epoch of lightweight supervised recovery:

$$\begin{aligned} \theta_{\text{sg-srl}} &= \arg \min_{\theta} \mathcal{L}_{\text{rec}}(\theta), \quad \theta \leftarrow \theta_{\text{rl}}, \\ \mathcal{L}_{\text{rec}}(\theta) &= -\mathbb{E}_{(x,y) \sim \mathcal{D}_{\text{para}}} \log \pi_{\theta}(y | x). \end{aligned} \quad (7)$$

Although initialization and recovery use the same parallel corpus, they serve different roles. Initialization teaches basic target-language generation from the base model, while recovery regularizes a semantically enhanced but form-degraded RL checkpoint. Reuse does not expose evaluation targets: the parallel training corpus, source-only RL corpus, and development set are mutually disjoint. In Section 4.2, we vary both nested recovery subsets and optimization intensity to separate the effect of reusable form anchors from memorization of particular examples. The final model preserves source-grounded semantic alignment while restoring fluency, conciseness, and task format.

4 Experiments

We evaluate four aspects of SG-SRL: (1) whether it improves Chinese-to-Thai generation beyond alternative uses of the same source data and target supervision, (2) how semantic alignment survives target-form recovery, (3) whether the learned alignment transfers beyond title generation, and (4) whether the semantic-reward interface can use an in-domain encoder when no capable target-language reranker is available.

4.1 Experiment 1: Effectiveness of SG-SRL

We first evaluate whether the full train–reinforce–recover pipeline improves target-language generation beyond cold-start SFT, testing the central claim that source-language monolingual data can provide useful semantic supervision when it is converted into a reference-free cross-lingual reward.

Task and data. The main task is Chinese-to-Thai news-title generation. We use CNewSum, a large-scale Chinese summarization dataset with human-annotated adequacy and deducibility levels (Wang et al., 2021), to construct a low-resource Chinese-to-Thai setting. We translate 15k Chinese titles into Thai. Among them, 10k parallel examples are used for target-form initialization and target-form recovery, and 5k examples are held out as the development set. We additionally sample 100k Chinese-only examples from CNewSum as source-language monolingual data for source-grounded semantic RL. The three splits are each deduplicated and mutually disjoint. This setting matches the target scenario of SG-SRL: a small amount of target-language supervision is available for learning output form, while substantially more source-language data can be used for semantic mid-training.

Base model and training stages. All main Chinese-to-Thai experiments use SmolLM3-3B (Bakouch et al., 2025) as the base model. The SG-SRL pipeline has three stages as described in Section 3. First, we train a cold-start supervised model on the 10k Chinese–Thai parallel examples. Second, we perform source-grounded semantic RL on the 100k Chinese-only examples. Third, we apply target-form recovery by fine-tuning the RL checkpoint for one epoch on the same 10k parallel examples. We refer to the supervised model after the first stage as COLD-START SFT, and to the final recovered model as SG-SRL.

Baselines. Beyond COLD-START SFT, we control for two simpler ways to consume additional data. CPT+SFT continues pretraining on the same 100k Chinese-only corpus before SFT on the 10k gold pairs, isolating source exposure without an online cross-lingual objective. SYNTHETIC SFT translates 50k Chinese titles with `deepseek-v4-flash` and combines them with the 10k gold pairs, yielding six times as many supervised pairs as cold-start SFT. Appendix A.7 further compares the reranker reward with an embedding-similarity RL baseline.

Evaluation protocol. We triangulate across channels that differ from the training reward. On the full 5k development set, we report ROUGE-L (Lin, 2004), COMET-DA (Rei et al., 2020), and chrF (Popović, 2015), together with a five-dimensional `deepseek-v4-flash` judge covering entity alignment, event alignment, factual consistency, fluency, and conciseness. ROUGE-L uses the SentencePiece tokenizer of CINO, an XLM-R variant, because Thai text is not whitespace segmented (Yang et al., 2022). We also sample 500 SG-SRL/COLD-START SFT pairs for preference evaluation by two annotators. The Qwen reranker used in training is reported only as a mechanism diagnostic, never as the primary effectiveness metric.

Results. Table 1 shows consistent evidence across automatic, LLM-based, and human evaluation. SG-SRL achieves the best ROUGE-L, chrF, and judge score, while its COMET-DA improvement over cold-start SFT is smaller. In the 500-pair human study, annotators prefer SG-SRL in 425 cases (85%), prefer COLD-START SFT in 15 (3%), and judge 60 (12%) as ties.

For a reference-anchored view of the judge, we also compare the two model outputs with the gold Thai title. The three-way ranking and direct pairwise comparison are reported together in Table 2.

Analysis. These controlled comparisons make two simpler explanations less likely. CPT+SFT sees the same 100k source examples but remains 0.315 below SG-SRL on the judge and 4.29 chrF points lower, so source exposure alone does not match SG-SRL in this setup. SYNTHETIC SFT has six times the supervised pairs yet is the weakest system, indicating that additional machine-translated targets do not substitute for online source-grounded optimization under the tested

configuration. Agreement between human preference, character/subword overlap metrics, COMET-DA, and both judge views also makes it unlikely that the gain only reflects the training reranker’s preferences. Absolute ROUGE-L remains low because open-ended Thai titles can preserve meaning with little single-reference subsequence overlap.

4.2 Experiment 2: Semantic Learning versus Reward Hacking

We next analyze the intermediate RL checkpoints before target-form recovery. This experiment asks why the semantic RL stage should be viewed as mid-training rather than as the final deployable generator. All variants start from the same COLD-START SFT checkpoint and optimize on the same Chinese-only source corpus, but differ in reward design or RL control.

We compare four variants at a high level here, leaving the exact reward definitions and ablation rationale to Appendix A.2. GATE+BATCHLEN is the main SG-SRL reward configuration, combining a reranker semantic score with Thai-language gating, batch-relative length control, and repetition control. GATE-ONLY removes the explicit length and repetition safeguards. REFLen replaces batch-relative length control with source-conditioned absolute length control. Observing that recent analyses RL training suggest that improvements and failures can arise from the RL algorithm itself (Liu et al., 2025), we introduce another setting GRPO-CONTROL, where GRPO is replaced with Dr.GRPO (Liu et al., 2025) to examine whether verbosity and form degradation are mainly artifacts of the GRPO-style optimization, or the relevance-oriented reward.

Results. Table 3 shows a clear semantic-form trade-off. GATE+BATCHLEN achieves the best overall score and the strongest entity, event, and factual alignment, indicating that source-grounded semantic RL can inject useful cross-lingual grounding into the model. However, its average output length is 531.7 characters, far longer than the gold Thai references, whose average length is 162.1 characters. Thus, even the best intermediate RL checkpoint is not directly deployable.

Analysis. The comparison with GATE-ONLY shows why auxiliary regularization is necessary. When the reward contains only the reranker score and the language gate, the model exploits the relevance reward by generating much longer outputs (average length 866.6 characters). REFLen shows

Model	ROUGE-L	COMET-DA	chrF	Judge	Thai gate	Avg. len.	Human wins
COLD-START SFT	0.191	0.652	21.87	2.699	80.3%	136	3%
CPT+SFT	0.192	0.637	23.60	2.584	83.2%	440	–
SYNTHETIC SFT	0.175	0.597	21.51	2.387	78.4%	600	–
SG-SRL	0.223	0.658	27.89	2.899	96.9%	385	85%

Table 1: **Chinese-to-Thai title generation.** Automatic metrics and the five-dimensional judge use the full 5k development set. Human wins are pairwise win rates on 500 SG-SRL/COLD-START SFT pairs evaluated by two annotators; the remaining 12% are ties. The training reranker is excluded from effectiveness columns.

Model	Total score	Avg. score	Rank 1	Rank 2	Rank 3	Outcome	Count	Ratio
Gold	7899	1.58	3723	865	412	SG-SRL wins	3613	72.3%
SG-SRL	3189	0.64	995	2602	1403	COLD-START SFT wins	1378	27.6%
COLD-START SFT	-1088	-0.22	282	1533	3185	Tie	9	0.2%

(a) Three-way ranking with the gold reference.

(b) LLM-judge pairwise comparison.

Table 2: Reference-anchored and pairwise DeepSeek evaluation on the 5k development set. Three-way total score assigns +2, +1, and −1 to ranks 1–3. These results complement the automatic and human evidence in Table 1.

the opposite failure mode. It reduces average output length to 179.7 characters, close to the gold reference length, but its entity, event, factual, and fluency scores drop substantially, suggesting that aggressive absolute length control can suppress verbosity but may also restrict the model’s ability to explore and express source-side semantics in the target language.

The standalone GRPO-CONTROL obtains the lowest overall score while still producing long outputs. Thus, changing the RL control variant does not resolve degeneration in this experiment, which is consistent with the relevance-oriented reward structure being an important source of the failure.

These findings motivate the train–reinforce–recover design. The RL checkpoint is valuable because it learns source-grounded semantic alignment, but it should not be used as the final generator. Target-form recovery is needed to convert the semantically enhanced but form-degraded checkpoint into a usable target-language model.

Semantic retention through recovery. Table 4 follows the same policy across stages. RL raises the reranker diagnostic from 0.763 to 0.789 but inflates mean length from 136 to 532 characters. Recovery then increases the diagnostic further to 0.874 while restoring the Thai gate to 96.9% and improving the independent judge to 2.899. The semantic gain is therefore not forgotten during form restoration; recovery consolidates it into a better-formed output distribution.

Recovery must nevertheless be calibrated. With 10k recovery pairs, reducing the learning rate to 10^{-6} leaves outputs as long as the RL checkpoint, whereas a 5×10^{-6} rate or three epochs short-

ens outputs but lowers the reranker diagnostic. The selected 2×10^{-6} , one-epoch setting gives the strongest semantic score after restoring form. Nested recovery subsets show a complementary data threshold: on the common 4,524-example judge subset, 500–1k pairs are insufficient, 3k raises the judge from 2.303 to 2.777, and 5k reaches 2.823 versus 2.899 with 10k. All recovery variants are evaluated against held-out examples; their trend indicates that broader form coverage improves recovery. Appendix A.3 reports the full data-size and intensity results and distinguishes this matched subset from the full 5k evaluation in Table 4.

4.3 Experiment 3: Transfer Beyond Title Generation

We further test whether the semantic gains from SG-SRL transfer beyond the original title-generation format. This experiment provides a complementary probe of reusable source-grounded semantics under a different output format.

Task and data. The transfer task is long-form Chinese-to-Thai translation, which requires preserving more entities, events, and factual relations over a longer context than title generation. We construct a manually checked 100-example diagnostic set, using GPT-5.5 to assist translation of Chinese inputs longer than 500 characters. We also evaluate short-title translation, but report it in Appendix A.6 because the three models obtain very similar BLEU scores, making that setting less discriminative for semantic grounding.

Results. Table 5 shows that SG-SRL improves over both the base model and COLD-START SFT in BLEU. More importantly, the LLM-judge results

Checkpoint	Entity	Event	Factual	Fluency	Conciseness	Avg.	Mean Length
GATE+BATCHLEN	2.331	2.887	2.363	2.194	1.738	2.303	531.7
GATE-ONLY	1.777	2.166	1.776	1.583	1.219	1.704	866.6
REFLEN	1.401	1.638	1.393	1.194	1.595	1.444	179.7
GRPO-CONTROL	1.441	1.549	1.428	1.000	0.976	1.279	708.9

Table 3: **Semantic-form trade-off in intermediate RL checkpoints.** GATE+BATCHLEN achieves the strongest semantic alignment before recovery, but its outputs remain much longer than the gold Thai references, whose mean length is 162.1 characters. GATE-ONLY suffers from severe verbosity, while REFLEN controls length but weakens semantic learning. A qualitative case study is provided in Appendix A.5, and the five-dimensional judge prompt is provided in Table 9.

Stage	Reranker	Judge	Thai gate	Length
Cold-start SFT	0.763	2.699	80.3%	136
RL checkpoint	0.789	2.321	99.9%	532
Recovery (5k)	0.873	2.823	91.7%	449
Recovery (10k)	0.874	2.899	96.9%	385

Table 4: **Recovery trajectory.** The training reranker is shown only to diagnose retention; the judge and form statistics characterize final quality.

show a clearer advantage across semantic dimensions. SG-SRL achieves the highest entity alignment, event alignment, factual consistency, and completeness while maintaining Thai fluency.

Analysis. The transfer results clarify the type of improvement learned by SG-SRL. The benefit is not a uniform gain on short translation instances; rather, it becomes more visible when the input requires longer-context semantic preservation and robust cross-lingual grounding. On this diagnostic set, the result is consistent with reusable semantic alignment beyond the title-generation format.

4.4 Experiment 4: In-Domain Reward Substitution for Tibetan

Finally, we test whether the semantic-reward interface can be implemented without an LLM reranker. Tibetan lacks a capable multilingual reranker, but it is covered by CINO, a Chinese minority-language encoder (Yang et al., 2022). We therefore train a CINO-based Tibetan-Chinese embedding scorer and use its similarity as the semantic reward. Accordingly, this experiment isolates in-domain reward-module substitution under parallel supervision.

Task and data. The task is Chinese-to-Tibetan generation with an embedding-based semantic reward. The data comes from the VLM portion of FTibSuite, a resource suite for Tibetan vision-language modeling (Xu et al., 2026). All data are drawn from a 100k-example bilingual caption corpus: 10k examples are used for cold-start SFT, 10k for development, and 80k for RL. The

scorer is trained from parallel pairs in this domain; consequently, embedding similarity is a reward-aligned diagnostic, while BLEU provides the independent comparison. The Chinese-reference RL variant uses the Chinese side as its semantic anchor, whereas the Tibetan-reference variant also uses the paired target.

The embedding model is trained on the Tibetan-Chinese caption pairs with a contrastive objective, mapping matched pairs closer in representation space and pushing mismatched pairs apart. Unlike a cross-encoder reranker, this model returns a similarity score in a shared embedding space. We compare two RL reward settings: one uses Tibetan text as the semantic reference, and the other uses Chinese text as the semantic reference. The comparison tests whether the learned embedding space can provide a usable cross-lingual reward when the semantic anchor is placed on either side of the bilingual pair.

Results. Both encoder-reward variants improve over cold-start SFT. The source-anchored Chinese-reference reward raises BLEU from 0.3953 to 0.4523 and reward-aligned embedding similarity from 0.5593 to 0.7011; its BLEU also matches the Tibetan-reference variant. This provides evidence that an in-domain encoder can guide target-language generation from a source-side anchor.

Analysis. The Tibetan experiment should be interpreted together with the embedding-reward comparison summarized in Appendix A.7. That appendix replaces the reranker reward in the Thai setting with Qwen3-Embedding-4B similarity and shows that an arbitrary semantic reward is not sufficient: the generic embedding model provides less separation between good and partial candidates and yields weaker post-recovery performance than the reranker. Reward discriminability, rather than model family alone, determines whether the interface supplies a useful RL signal.

The CINO reward has sufficient discrimination

Model	BLEU	Entity	Event	Factual	Thai Fluency	Completeness	Avg.
Base	0.11	2.600	2.800	2.720	2.420	2.680	2.644
COLD-START SFT	0.14	2.700	2.700	2.680	3.240	2.180	2.700
SG-SRL	0.17	3.420	3.540	3.520	3.280	3.260	3.404

Table 5: **Transfer to long-form Chinese-to-Thai translation.** BLEU (Papineni et al., 2002) provides a surface-level automatic metric, while the LLM-judge dimensions evaluate semantic preservation and Thai quality. SG-SRL improves over both the base model and COLD-START SFT, with the largest gains on entity alignment, event alignment, factual consistency, and completeness. The long-text translation judge prompt is provided in Table 10.

Setting	BLEU	Embedding Similarity
Cold-start SFT	0.3953	0.5593
RL w/ Tibetan ref.	0.4519	0.7164
RL w/ Chinese ref.	0.4523	0.7011

Table 6: **In-domain Tibetan reward substitution.** Both encoder-reward variants improve over the pre-RL model; the source-anchored Chinese-reference variant matches the target-anchored variant on BLEU.

in the same caption domain used for contrastive training and RL. This shared domain defines the evidence boundary: Table 6 supports portability of the semantic-reward interface under a reliable in-domain scorer, but does not test transfer across domains or establish that arbitrary embedding rewards will generalize beyond their reward-model coverage.

5 Conclusion

We presented SG-SRL, a train–reinforce–recover framework that converts source-only data into reference-free cross-lingual semantic supervision. Its central design choice is to use semantic RL as mid-training: the RL checkpoint expands source-grounded coverage, and a controlled recovery pass restores target-language form. Stronger source-data and synthetic-supervision baselines, human preference, and training-independent metrics all support the resulting gains, while recovery ablations characterize how data coverage and optimization intensity control semantic retention and form restoration. A long-form diagnostic provides initial evidence that the learned alignment extends beyond title format. The Tibetan study establishes a narrower but useful form of portability: given a sufficiently discriminative in-domain cross-lingual scorer, the semantic-reward interface need not be an LLM reranker. Together, these results characterize SG-SRL as a way to amortize limited target-language supervision over abundant source-language data, with reward quality and recovery strength as explicit operating conditions.

Limitations

First, a capable multilingual reranker may be unavailable for genuinely low-resource target languages. Our Chinese-to-Thai instantiation therefore uses Qwen3-Reranker-8B, while the Tibetan experiment uses an encoder-based reward as a practical alternative. Building stronger reward models for genuinely low-resource languages remains an important direction. Second, the Tibetan reward depends on an in-domain scorer trained from bilingual pairs in the same caption domain used for RL. Its results therefore do not establish broad out-of-domain portability; reward discrimination may change when the language pair or domain changes.

Broader Impact

The framework may help extend generation systems to languages for which task-specific target data are costly to collect, but performance remains bounded by the coverage and biases of the multilingual reward model and the small parallel corpus. In practical use, reward errors could reinforce mistranslations or omit culturally specific content while still receiving high semantic scores. We therefore view automatic reward optimization as a complement to, rather than a replacement for, evaluation by target-language speakers, particularly for high-impact or community-facing applications.

Acknowledgments

This work is supported by Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM113), National Natural Science Foundation of China (92470121, 62402016), National Key R&D Program of China (2024YFA1014003), Zhongguancun Academy (C20250204, C20250602), Beijing Major Science and Technology Project (Z251100008125043, Z251100008425023), and High-performance Computing Platform of Peking University.

References

- Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. [Concrete problems in AI safety](#). *arXiv preprint arXiv:1606.06565*.
- Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Noumane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, and 4 others. 2025. SmolLM3: smol, multilingual, long-context reasoner. <https://huggingface.co/blog/smollm3>.
- Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2021. [mMARCO: A multilingual version of the MS MARCO passage ranking dataset](#). *arXiv preprint arXiv:2108.13897*.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- DeepSeek-AI. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Ariavazhagan, and Wei Wang. 2022. [Language-agnostic BERT sentence embedding](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, and 1 others. 2025. [DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning](#). *Nature*, 645(881):633–638.
- Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjan Wartikar, and Eileen Long. 2025. [Adapting multilingual LLMs to low-resource languages using continued pre-training and synthetic corpus: A case study for Hindi LLMs](#). In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 50–57, Abu Dhabi. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. [Understanding R1-Zero-like training: A critical perspective](#). In *Proceedings of the 2nd Conference on Language Modeling*.
- Rodrigo Nogueira and Kyunghyun Cho. 2019. [Passage re-ranking with BERT](#). *arXiv preprint arXiv:1901.04085*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.
- Maja Popović. 2015. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395. Association for Computational Linguistics.
- Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. [A survey of multilingual large language models](#). *Patterns*, 6(1):101118.
- Qwen Team. 2025. Qwen3-Reranker-8B. <https://huggingface.co/Qwen/Qwen3-Reranker-8B>. Accessed: 2026-05-24.
- Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2685–2702. Association for Computational Linguistics.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [DeepSeekMath: Pushing the limits of mathematical reasoning in open language models](#). *arXiv preprint arXiv:2402.03300*.
- Zeli Su, Ziyin Zhang, Zhou Liu, Xuexian Song, Zhankai Xu, Longfei Zheng, Xiaolu Zhang, Rong Fu, Guixian Xu, and Wentao Zhang. 2026. [Reinforcement learning with semantic rewards enables low-resource language expansion without alignment tax](#). *arXiv preprint arXiv:2605.14366*.
- Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking

- agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.
- Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.
- Danqing Wang, Jiaze Chen, Xianze Wu, Hao Zhou, and Lei Li. 2021. Cnewsun: A large-scale summarization dataset with human-annotated adequacy and deducibility level. In *Natural Language Processing and Chinese Computing*, pages 389–400, Cham. Springer International Publishing.
- Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. [Extending multilingual BERT to low-resource languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.
- Guixian Xu, Yide Liang, Zeli Su, Xuexian Song, Ziyin Zhang, Yushuang Dong, Ting Zhang, and Xu Han. 2026. [FTibSuite: A comprehensive resource suite for tibetan vision-language modeling](#). ACL Rolling Review, January 2026. Submission 6640.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, 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 2 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. 2022. [CINO: A Chinese minority pre-trained language model](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3937–3949, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Additional Details

A.1 Metric Sensitivity

Semantic-reward RL should not be evaluated only with the same reward model used for training. The Qwen reranker supplies a reference-free training signal and a diagnostic of whether recovery retains the optimized alignment, but it is excluded from the effectiveness columns in Table 1. Our main evidence instead combines human preference, a separate DeepSeek judge with a different objective, and training-independent ROUGE-L, chrF, and COMET-DA. The metrics have complementary sensitivities: character and subword overlap distinguish the strongest title-generation systems, while COMET-DA is comparatively compressed at the top (0.658 for SG-SRL versus 0.652 for cold-start SFT). Absolute ROUGE-L is low because a single Thai title reference under-covers valid meaning-preserving rephrasings, a known mismatch for semantic-reward optimization (Su et al., 2026).

For human evaluation, we sample 500 SG-SRL/COLD-START SFT pairs. Two annotators independently compare semantic adequacy, factual fidelity, and Thai output quality; their decisions are aggregated, with unresolved preferences recorded as ties. GPT-5.5 and Google Translate are available only as auxiliary cross-checks. SG-SRL is preferred in 425 cases, cold-start SFT in 15, and 60 are ties. Neither the training reranker score nor the DeepSeek judge decision is used as the human label.

A.2 Reward Details for RL Diagnostics

The intermediate-checkpoint comparison in Section 4.2 uses four reward/control variants. GATE+BATCHLEN is the main SG-SRL configuration: it combines the reranker relevance score with a Thai-language gate, a batch-relative length penalty, and a repeated 4-gram penalty. GATE-ONLY keeps the reranker score and Thai-language gate but removes explicit length and repetition penalties. REFLen replaces batch-relative length control with a source-conditioned absolute length constraint. GRPO-CONTROL keeps the main reward form but changes the RL control variant to test whether degeneration is mainly caused by the optimization recipe rather than the relevance-oriented reward.

A.3 Recovery Ablations

Amount of recovery data. Starting from one shared RL checkpoint, we recover on nested subsets of the 10k parallel training corpus. Automatic metrics and the Thai gate use the full disjoint 5k development set; judge scores use the common 4,524-example subset shared by all recovery variants. Table 7 shows a threshold rather than a benefit from repeatedly seeing a tiny subset: 500–1k examples do not improve the judge over the RL checkpoint, 3k produces a large form-quality transition, and 5k–10k gives the strongest reference-based metrics. The non-monotonic Thai gate at 3k also shows why language validity alone is not a sufficient recovery criterion.

Recovery intensity. Table 8 varies the learning rate and number of epochs while fixing all 10k recovery pairs. Weak recovery at 10^{-6} leaves mean length (534) essentially unchanged from the RL checkpoint (532). Stronger recovery shortens output and improves COMET-DA, but lowers the reranker diagnostic to 0.829–0.831. The selected 2×10^{-6} one-epoch configuration instead yields the highest reranker, ROUGE-L, and judge combination. Recovery therefore has a controllable semantic-retention/form-restoration frontier rather than acting as an unconditional cleanup pass.

Initialization and recovery deliberately reuse the limited parallel corpus but begin from different policies and serve different objectives. All recovery variants are evaluated on held-out development examples, and the nested-subset trend shows that broader form anchors improve recovery while preserving the alignment diagnostic.

A.4 Judge Prompts

For the main title-generation evaluation, the judge receives the Chinese source, the gold Thai title, and model outputs in randomized order. It ranks outputs by semantic adequacy, factual faithfulness, Thai fluency, and title-style conciseness. For pairwise evaluation, the judge compares SG-SRL and COLD-START SFT directly and may return a tie only when the two outputs are indistinguishable in quality.

A.5 Qualitative RL Diagnostic Case

The qualitative cases follow the same pattern as the aggregate results in Table 3. GATE-ONLY outputs often mention more source-side content but become excessively long and sometimes drift

Recovery data	Reranker	ROUGE-L	chrF	Judge	Fluency	Conciseness	Thai gate
0 (RL checkpoint)	0.789	—	—	2.303	2.194	1.738	99.9%
500	0.818	0.194	22.42	2.210	2.000	1.610	99.4%
1k	0.813	0.193	22.70	2.243	2.060	1.660	99.6%
3k	0.869	0.199	22.17	2.777	2.650	2.330	83.5%
5k	0.873	0.208	25.34	2.823	2.810	2.530	91.7%
10k	0.874	0.223	27.89	2.899	3.090	2.810	96.9%

Table 7: Recovery-data-size ablation. Judge scores use the common 4,524-example subset; automatic metrics and Thai gate use the full 5k development set. Reranker is a retention diagnostic; all other measures are independent of the training reward.

Setting	Rerank	R-L	COMET	Judge	Len.
10^{-6} , 1 epoch	0.847	0.198	0.594	2.585	534
2×10^{-6} , 1 epoch	0.874	0.223	0.658	2.899	385
5×10^{-6} , 1 epoch	0.829	0.216	0.677	2.891	266
2×10^{-6} , 3 epochs	0.831	0.206	0.674	2.923	270

Table 8: Recovery-intensity ablation with 10k parallel pairs.

away from title format. REFLen better controls length but frequently drops entities or event details. GATE+BatchLen provides the best compromise among the intermediate RL checkpoints, although it still requires target-form recovery before deployment.

A.6 Short-Title Translation

We also evaluated a short-title translation setting. The task is less discriminative than long-form transfer because the source inputs contain fewer entities and event relations, and the compared models obtain similar surface-overlap scores. We therefore report the more informative long-form transfer results in the main text, where semantic preservation over longer contexts better reveals the effect of source-grounded semantic RL.

A.7 Embedding Reward Comparison

As a same-category control, we replace the Thai reranker reward with Qwen3-Embedding-4B cosine similarity and apply the same recovery stage. On the 4,829 predictions shared by both scorers, the embedding scorer’s prediction-to-gold gap is about 0.10, compared with about 0.24 for the reranker. A 200-example FLORES control likewise gives only a 0.13 gold-to-partial gap with substantial overlap. This weak within-group contrast makes it harder for GRPO to distinguish useful candidates. The equal-recovery comparison below is a separate control run and is interpreted within-table rather than against the full-dev main run in Table 1.

ROUGE-L is effectively tied, but the reranker leads on chrF, COMET-DA, the judge, and pair-

wise preference. Because these metrics are independent of either training reward, the result indicates that an arbitrary semantic similarity objective is insufficient; the scorer must discriminate fine-grained cross-lingual adequacy in the task domain. This motivates the Tibetan experiment in Section 4.4, where the encoder is contrastively trained in-domain rather than assumed to transfer generically.

Field	Prompt instruction
Input	Given a Chinese news input and a Thai candidate title, evaluate whether the Thai title preserves the source-side entities, events, and factual relations.
Entity	Score whether named entities, quantities, locations, organizations, and other salient participants are correctly preserved.
Event	Score whether the main action, event, or state described in the source is correctly expressed.
Factual	Score whether the output avoids unsupported claims, contradictions, and hallucinated details.
Fluency	Score whether the Thai output is grammatical, natural, and readable.
Conciseness	Score whether the output follows title style and avoids unnecessary verbosity or formatting artifacts.
Output	Return integer scores for the five dimensions and a brief justification.

Table 9: Five-dimensional judge prompt used for intermediate RL-checkpoint diagnostics.

Field	Prompt instruction
Input	Given a long Chinese passage and a Thai translation, evaluate translation quality without relying only on word overlap.
Entity	Score whether important entities and numerical information are preserved.
Event	Score whether the major events, actions, and relations are translated correctly.
Factual	Score whether the translation is faithful to the source and avoids hallucination.
Thai fluency	Score whether the Thai text is fluent, grammatical, and coherent.
Completeness	Score whether the translation covers the important information in the source passage.
Output	Return dimension scores and an overall assessment.

Table 10: Judge prompt used for long-form Chinese-to-Thai translation transfer evaluation.

Reward	Judge	R-L	COMET	chrF
Embedding similarity	2.848	0.205	0.672	24.5
Reranker	2.922	0.200	0.712	29.0

Table 11: Thai semantic-reward substitution after equal recovery. The reranker model also wins 62.7% of direct pairwise comparisons.