Reinforcement Learning with Semantic Rewards Enables Low-Resource Language Expansion without Alignment Tax

Anonymous ACL submission

Abstract

Extending large language models (LLMs) to low-resource languages often incurs an "alignment tax": improvements in the target language come at the cost of catastrophic forgetting in general capabilities. We argue that this trade-off arises from the rigidity of supervised fine-tuning (SFT), which enforces token-level surface imitation on narrow and biased data distributions. To address this limitation, we propose a semantic-space alignment paradigm powered by **Group Relative Policy Optimization (GRPO)**, where the model is optimized using **embedding-level semantic rewards** rather than likelihood maximization. This objective encourages meaning preservation through flexible realizations, enabling controlled updates that reduce destructive interference with pretrained knowledge. We evaluate our approach on Tibetan–Chinese machine translation and Tibetan headline generation. Experiments show that our method acquires low-resource capabilities while markedly mitigating alignment tax, preserving general competence more effectively than SFT. Despite producing less rigid surface overlap, semantic RL yields higher semantic quality and preference in open-ended generation, and few-shot transfer results indicate that it learns more transferable and robust representations under limited supervision. Overall, our study demonstrates that reinforcement learning with semantic rewards provides a safer and more reliable pathway for inclusive low-resource language expansion.

1 Introduction

Large language models (LLMs) have achieved remarkable performance across a wide range of tasks and languages through large-scale pretraining and post-training alignment (DeepSeek-AI et al., 2025; Yang et al., 2025a; Team et al., 2025; OpenAI, 2025; Comanici et al., 2025). However, their capabilities remain highly uneven across languages:

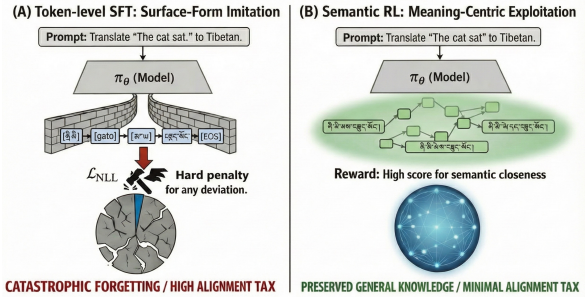

Figure 1: Token-level alignment versus semantic-space alignment in low-resource language expansion. Token-level supervised fine-tuning enforces surface-form imitation under teacher forcing, often causing catastrophic forgetting and high alignment tax. In contrast, semantic-space alignment optimizes meaning preservation with constrained reinforcement learning policy updates, allowing flexible realizations while preserving pretrained knowledge.

many languages are weakly supported, particularly for generation and reasoning-intensive tasks that require semantic abstraction rather than surface pattern matching. Improving model performance in such low-resource language settings therefore remains an important and challenging problem.

A common approach to improving low-resource language performance is further training on language-specific data, including continual pretraining and supervised instruction fine-tuning. Despite differences in data sources and training protocols, these methods share a common optimization paradigm: they rely on teacher-forced learning with token-level likelihood objectives, aligning the model to a target data distribution through surface-form imitation. Figure 1 illustrates the contrast between token-level alignment via surface-form imitation and semantic-space alignment based on meaning preservation, highlighting how different objectives lead to fundamentally different update behaviors under data scarcity. Much recent work on language expansion and adaptation follows this

paradigm, using supervised or weakly supervised finetuning to inject new linguistic capabilities into pretrained models (Csaki et al., 2024; Zhao et al., 2024). Related efforts based on continued pretraining or language-specific model construction similarly perform strong distribution-level updates on limited and domain-constrained data (Almeida et al., 2025; Bari et al., 2025).

While effective under abundant and diverse supervision, token-level distribution matching becomes problematic in low-resource language settings. Available training data are often limited in size, narrow in domain, and distributionally biased. Optimizing likelihood on such data encourages overly confident and rigid parameter updates, amplifying overfitting and interfering with representations learned during pretraining. Empirically, this interference often manifests as catastrophic forgetting: improvements in the target language are accompanied by degradation in existing high-resource language capabilities, a phenomenon that has been systematically observed in multilingual fine-tuning and low-resource representation learning settings (Liu and Niehues, 2025; Schmidt, 2025).

We argue that this phenomenon is not merely an optimization artifact but a structural outcome of the alignment objective itself. When alignment is equated with surface-form imitation on narrow distributions, representational capacity is aggressively reallocated, leading to an *alignment tax*: gains in low-resource language performance achieved at the expense of general competence. This issue persists regardless of the adaptation method (e.g., continual pretraining or instruction tuning) as long as the objective enforces token-level matching on sparse data (Yamaguchi et al., 2025).

Consequently, we propose shifting perspective to view language expansion not merely as adaptation, but as an *alignment problem under sparse supervision*. To address this, we introduce a semantic-space alignment paradigm that prioritizes meaning preservation over rigid surface-form imitation. We operationalize this framework using **Group Relative Policy Optimization (GRPO)** (Shao et al., 2024), employing **embedding-level semantic similarity** as the primary reward signal. Unlike teacher-forced training, this approach encourages the model to explore diverse linguistic realizations that maintain semantic equivalence. Crucially, by optimizing based on relative rewards within a sampled group, our method inherently incorporates the stability constraints of trust-region optimization (Schulman et al., 2017; Rafailov et al., 2023), enabling the acquisition of low-resource capabilities while strictly limiting the destructive interference typical of unconstrained likelihood maximization.

We evaluate our approach on Tibetan–Chinese machine translation (MT) and Tibetan headline generation (HG). Empirical results demonstrate that semantic-reward-driven GRPO achieves a superior trade-off between adaptation and preservation. In the MT task, our method substantially reduces alignment tax, outperforming the strong SFT baseline by **+5.15** points on the dominant-language CMRC benchmark. Similarly, in the HG task, despite lower n-gram overlap, our model is preferred by LLM-based judges with a **+16.1%** higher win rate compared to SFT. These findings suggest that semantic-space alignment offers a safer and more robust paradigm for improving low-resource language performance under data scarcity.

In summary, the main contributions of this paper are:

- We propose a **semantic-space alignment paradigm** that utilizes Group Relative Policy Optimization (GRPO) with embedding-level rewards to decouple meaning preservation from surface-form imitation.
- We demonstrate that this approach virtually eliminates the **alignment tax**, enabling significant low-resource gains while maintaining the model's general capabilities and pretrained knowledge.
- We show that our method produces **semantically superior outputs** that are preferred by LLM judges over SFT baselines, despite having lower n-gram overlap with rigid references.
- We validate that semantic RL yields **more transferable representations**, as evidenced by stronger few-shot generalization to downstream tasks compared to supervised methods.

2 Related Work

2.1 Low-Resource Language Adaptation and Expansion

Post-training language expansion andadaptation typically start from a pretrained foundation model and further train on language-specific data. Prior work studies supervised fine-tuning or instruction tuning for transferring language capabilities and scaling with data and model size (Csaki et al., 2024; Zhao et al., 2024), as well as continued pretraining

and language-specific model construction that emphasize corpus selection and composition (Almeida et al., 2025; Bari et al., 2025). Across these settings, selective adaptation on narrow distributions can induce catastrophic forgetting, degrading performance on non-target languages or tasks (Liu and Niehues, 2025; Schmidt, 2025).

Several mitigation strategies focus on constraining update magnitude or protecting previously learned capabilities, such as parameter-efficient finetuning (e.g., LoRA-style low-rank updates) and source-shielded adaptation (Yang et al., 2025b; Yamaguchi et al., 2025). While these methods can reduce interference, they generally retain teacher-forced token-level likelihood objectives. Our work is complementary: we reconsider the alignment objective itself by optimizing semantic consistency via embedding-level rewards, aiming to improve weak-language capability with lower alignment tax.

2.2 Reinforcement Learning for LLM Alignment

Reinforcement learning (RL) is commonly used in LLM alignment when optimization objectives are sequence-level or non-differentiable, enabling learning beyond token-level supervised imitation (Christiano et al., 2017; Ouyang et al., 2022). A core advantage of RL-based alignment methods is the use of constrained policy updates, such as trust-region or KL-regularized optimization, which limits drift from pretrained representations and improves stability (Schulman et al., 2015, 2017).

Recent variants retain this constrained-update principle while improving efficiency or flexibility, including direct preference optimization (Rafailov et al., 2023) and group-based policy optimization methods (Shao et al., 2024), which we adopt in this work to enable controlled alignment towards semantic-level objectives.

3 Method

3.1 Problem Formulation: Semantic-Space Alignment

We study low-resource language expansion as an alignment problem. Given a pretrained instruction-following language model π_{base}, our goal is to acquire new capabilities in a low-resource language while preserving existing competencies in dominant, high-resource languages.

Conventional supervised fine-tuning aligns models by maximizing token-level likelihood under a target data distribution. In low-resource settings, where the distribution is narrow and biased, this objective enforces surface-form imitation and often leads to overconfident updates and catastrophic forgetting. We instead frame alignment as *semantic-space alignment*: model outputs are considered correct if they preserve meaning, regardless of their specific surface realization. This formulation explicitly decouples semantic adequacy from token-level matching and allows multiple valid expressions of the same intent.

Under this perspective, alignment is defined by semantic consistency rather than distribution matching. Our objective is therefore to optimize the model to produce outputs that are semantically equivalent to reference texts, while limiting interference with representations learned during pretraining.

3.2 Two-Stage Training Paradigm

To operationalize semantic-space alignment in low-resource settings, we adopt a two-stage training paradigm.

Stage 1: Cold-start supervised fine-tuning. We first perform a lightweight supervised fine-tuning step on a small subset of low-resource data to obtain an initial policy π_{init}. Specifically, we fine-tune the base model on 5k training instances for two epochs. The goal of this stage is not to achieve strong task performance, but to bootstrap minimal output competence in the target language, such as producing text in the correct script and maintaining basic language consistency. This cold-start initialization allows the model to reliably generate non-degenerate outputs in the low-resource language, ensuring that subsequent semantic rewards are meaningful and that reinforcement learning does not collapse into uninformative exploration.

Stage 2: Reinforcement learning with semantic rewards. Starting from π_{init}, we perform reinforcement learning to align the model in semantic space. In this stage, we utilize the remaining training data to drive learning through semantic rewards rather than token-level supervision. Reinforcement learning is conducted for a single epoch, during which the model is encouraged to explore diverse surface realizations while preserving semantic equivalence to reference texts. Constrained policy optimization is applied throughout training

to control update magnitude, enabling the model to acquire low-resource language capabilities while minimizing destructive interference with pretrained representations.

3.3 Reinforcement Learning for Semantic Alignment

Optimizing semantic alignment is inherently a sequence-level problem over discrete outputs. The embedding-based semantic rewards used in our framework are not directly differentiable with respect to model parameters, making standard supervised learning objectives unsuitable. Reinforcement learning therefore provides a natural and principled framework for optimizing such non-differentiable, sequence-level objectives, allowing direct optimization of semantic consistency rather than token-level likelihood.

Beyond enabling optimization of semantic rewards, reinforcement learning also plays a critical role in preserving pretrained knowledge during low-resource adaptation. In contrast to supervised fine-tuning, which performs unconstrained likelihood maximization on narrow data distributions, constrained reinforcement learning methods explicitly limit policy updates. This controlled optimization is crucial for reducing destructive interference and mitigating catastrophic forgetting, making reinforcement learning particularly well-suited for semantic-space alignment under data scarcity.

We instantiate reinforcement learning using Group Relative Policy Optimization (GRPO, Shao et al., 2024), a value-free variant of PPO-style constrained optimization. For each input prompt x, we sample a group of candidate outputs $\{y^{(k)}\}_{k=1}^{K}$ from the current policy $\pi_\theta(\cdot \mid x)$, compute their corresponding rewards, and update the policy based on relative comparisons within the group. GRPO inherits the key stabilization mechanisms of PPO, including trust-region-style constraints that limit policy drift between updates, while avoiding the need for an explicit value function. These properties make GRPO a practical and stable choice for semantic alignment, enabling effective learning from semantic rewards while maintaining existing language capabilities.

3.4 Semantic Reward Design

A central component of our framework is a semantic reward that explicitly defines the alignment objective for low-resource language expansion. Unlike supervised fine-tuning, which implicitly aligns models through token-level likelihood, our goal is to directly guide learning toward semantic consistency. The reward therefore serves not merely as an optimization signal, but as the primary mechanism that determines what the model is encouraged to learn.

Semantic embedding model and suitability for reinforcement learning. To instantiate the semantic reward, we employ a multilingual sentence-level embedding model trained under a contrastive bilingual alignment objective. The model is adapted on parallel sentence pairs, where translations are treated as positive examples and in-batch samples provide implicit negatives. Rather than training an encoder from scratch or enforcing surface-form similarity, this adaptation strategy emphasizes meaning preservation and fine-grained semantic discriminability across different linguistic realizations.

We empirically observe that bilingual contrastive training yields stronger semantic structure than monolingual adaptation, improving both cross-lingual alignment and intra-language separability. This property is particularly important for reinforcement learning, where the reward signal must reflect graded semantic differences rather than coarse topical similarity. As a sanity check, we construct a small diagnostic set of sentence pairs spanning different degrees of semantic equivalence and find that the embedding model assigns similarity scores that consistently correlate with these graded relationships. This indicates that the resulting embedding space is suitable for use as a semantic reward for guiding alignment in reinforcement learning.

3.4.1 Embedding-Level Semantic Similarity Reward

The primary learning signal in our framework is an embedding-level semantic similarity reward. Let $f(\cdot)$ denote the sentence embedding model described above, which maps text to normalized vector representations. Given a generated output y and a reference text y^*, we compute their semantic similarity using cosine similarity:

$$s(y, y^*) = \cos\left(f(y), f(y^*)\right). \tag{1}$$

This reward directly reflects our desired learning direction: outputs are encouraged to preserve meaning, regardless of surface realization. In contrast to token-level likelihood objectives, this formulation treats semantically equivalent paraphrases as

equally valid and explicitly avoids overfitting to reference form. To stabilize optimization and focus learning on meaningful improvements beyond minimal adequacy, we apply a threshold-and-rescale shaping function:

$$R_{\text{sim}}(y, y^*) = \begin{cases} 0, & s(y, y^*) \leq \tau, \\ \frac{s(y,y^*)-\tau}{1-\tau}, & s(y, y^*) > \tau, \end{cases} \quad (2)$$

where τ corresponds to a minimal semantic adequacy level achieved after cold-start fine-tuning. This shaping ensures that reinforcement learning primarily refines semantic quality rather than amplifying noise from low-quality generations.

3.4.2 Language Consistency Reward

Because the embedding model is multilingual, optimizing semantic similarity alone may reward mixed-language or partially off-target outputs. To prevent this reward hacking behavior, we introduce a language consistency reward based on a rule-based script check using Unicode ranges and regular expressions:

$$R_{\text{lang}}(y) = \begin{cases} 0, & \text{language mixed}, \\ 1, & \text{language consistent}. \end{cases} \quad (3)$$

This acts as a hard constraint, ensuring that semantic optimization is carried out strictly within the target low-resource language space.

Final reward. The final reward combines semantic similarity and language consistency:

$$R(y, y^*) = \lambda_{\text{sim}} R_{\text{sim}}(y, y^*) + \lambda_{\text{lang}} R_{\text{lang}}(y). \quad (4)$$

Together, these components define our semantic alignment objective: the model is encouraged to improve semantic adequacy while being strictly constrained to produce linguistically consistent outputs in the low-resource language. In practice, we assign a larger weight to semantic similarity ($\lambda_{\text{sim}} = 1.5$) than to language consistency ($\lambda_{\text{lang}} = 1.0$), reflecting our design choice that semantic preservation constitutes the primary learning objective, while language consistency serves as a necessary constraint to prevent degenerate or off-language generations.

4 Experiments

We conduct a series of experiments to evaluate whether semantic-reward-driven reinforcement learning (RL) provides a better trade-off between low-resource language adaptation and preservation of existing capabilities compared to supervised fine-tuning (SFT). Our experiments are designed to answer three research questions: (1) whether RL effectively acquires low-resource language capabilities, (2) how RL and SFT differ in the trade-off between task performance and alignment tax, and (3) whether RL learns more transferable representations under data scarcity.

4.1 Experimental Setup

Base model and adaptation. All experiments are conducted on **Qwen3-4B** with parameter-efficient fine-tuning via LoRA (Hu et al., 2022). Unless otherwise specified, we apply LoRA to all linear projection layers in self-attention and MLP blocks. We use a LoRA rank of $r = 64$, scaling factor $\alpha = 128$, and dropout rate of 0.05.

Supervised fine-tuning (SFT). SFT is trained for three epochs in BF16 with a global batch size of 32, using AdamW (Loshchilov and Hutter, 2019) with learning rate 2×10^{-5} and a cosine schedule (warmup ratio 0.1).

Semantic reward model. The semantic reward described in Section 3.4 is instantiated using a bilingual sentence-embedding model built on top of CINO (Yang et al., 2022), a Tibetan-enhanced extension of XLM-R (Conneau et al., 2020). We adapt CINO into a sentence-level encoder using SENTENCETRANSFORMER, and further specialize it on Chinese–Tibetan parallel data to produce embedding-based semantic similarity scores. The resulting encoder is used as a frozen reward model during RL and is not jointly optimized with the policy model.

Reinforcement learning (GRPO). Reinforcement learning is performed with GRPO (Shao et al., 2024) starting from the SFT checkpoint, trained for one epoch in BF16 with AdamW and learning rate 5×10^{-7} (effective global batch size 32). For each prompt, we sample 8 candidates with temperature 0.8 and top-p 0.9, using max prompt/completion lengths of 256 tokens.

Controlled comparison. This unified setup ensures that observed differences primarily reflect the alignment strategy (semantic-reward-driven RL vs. SFT), rather than mismatched optimization or adaptation configurations. **Training and hyperparameter details are provided in Appendix A.**

4.1.1 Tasks and Datasets

We evaluate our approach on two representative Tibetan low-resource generation tasks: cross-lingual machine translation (MT) and monolingual headline generation (HG).

Machine Translation (MT). For Tibetan–Chinese machine translation, we use an internal parallel corpus collected for training a vision–language model (Wu et al., 2019). Specifically, the corpus consists of Tibetan–Chinese sentence pairs translated as part of the pretraining data construction pipeline for the VLM, rather than annotations produced by the model itself. We repurpose this parallel data as supervised training material for machine translation in our experiments. As the corpus was originally curated to support VLM pretraining, a large portion of the data is grounded in visual descriptions, resulting in a relatively narrow and domain-constrained distribution. While this dataset does not aim to be a comprehensive translation benchmark, it reflects a realistic low-resource scenario with limited domain diversity and is therefore suitable for studying alignment behavior under data scarcity.

Headline Generation (HG). For Tibetan headline generation, we use the Tibetan subset of the CMHG dataset (Xu et al., 2025). Due to the fine-grained tokenization of Tibetan in the Qwen tokenizer, raw samples often result in excessively long sequences and high memory consumption. We therefore filter out samples exceeding 1024 tokens and retain shorter instances for both training and evaluation. After filtering, the dataset contains 16,449 training samples and 621 test samples, which are used consistently across all headline generation experiments.

4.1.2 Evaluation Protocols

We adopt a multi-dimensional evaluation protocol to capture both surface-level accuracy and semantic quality. For task performance, we report standard reference-based metrics (BLEU for MT and ROUGE for HG) as well as embedding-based semantic similarity. To assess semantic quality beyond reference matching, we conduct blind pairwise evaluations using an LLM-as-a-Judge, where judgments are produced by **GPT-5.2** under a fixed evaluation prompt; the full judging prompt and evaluation policy are provided in Appendix B. To quantify alignment tax, we evaluate all models on a dominant-language benchmark (Chinese CMRC, Cui et al., 2019) before and after adaptation and report performance changes relative to the base model.

Model	BLEU-4	Similarity
Cold-start SFT	0.3953	0.5593
RL (Ours)	**0.4519**	**0.7164**

Table 1: Experiment 1 results on Tibetan–Chinese machine translation.

Model	ROUGE-L	Similarity
Cold-start SFT	0.2204	0.5774
RL (Ours)	**0.2530**	**0.6404**

Table 2: Experiment 1 results on Tibetan headline generation.

4.2 Experiment 1: Effectiveness of Semantic-Reward RL

We first evaluate whether semantic-reward-driven reinforcement learning (RL) effectively acquires low-resource language capabilities beyond minimal supervised initialization. Specifically, we compare RL against the cold-start SFT model on both Tibetan–Chinese machine translation (MT) and Tibetan headline generation (HG).

Machine Translation. Table 1 reports the results on Tibetan–Chinese MT. Starting from the same cold-start SFT checkpoint trained on 5k parallel sentence pairs, RL is further trained on approximately 90k additional samples using semantic rewards. Compared to the cold-start baseline, RL yields consistent improvements in both reference-based accuracy and semantic similarity. BLEU-4 increases from 0.3953 to 0.4519, while semantic similarity improves substantially from 0.5593 to 0.7164.

Headline Generation. We observe a similar trend on Tibetan headline generation. As shown in Table 2, the RL model trained on approximately 15k samples consistently outperforms the cold-start SFT baseline in both ROUGE-L and semantic similarity. In particular, ROUGE-L improves from 0.2204 to 0.2530, while semantic similarity increases from 0.5774 to 0.6404.

Analysis. Across both translation and generation tasks, semantic-reward-driven RL consistently improves performance over the cold-start SFT baseline. Notably, the improvements are particularly pronounced in semantic similarity, suggesting that

RL primarily refines meaning preservation rather than merely increasing surface-level overlap with references. These results confirm that embedding-level semantic rewards constitute a sufficiently informative alignment signal, enabling effective low-resource language learning beyond minimal supervised initialization.

4.3 Experiment 2: Trade-off Between Task Performance and Alignment Tax

In this experiment, we compare semantic-reward-driven RL with a Strong SFT baseline to characterize the trade-off between task performance and preservation of existing general capabilities (i.e., alignment tax). Unlike the cold-start SFT model used in Experiment 1 — which is trained only on a small 5k subset of low-resource data and serves solely as the initialization policy for RL — the **Strong SFT model is trained on the full available training data (i.e., the same combined dataset used by cold-start SFT + RL) under the same optimization and LoRA configuration**, representing the best-effort supervised adaptation outcome. Table 3 reports results on Tibetan–Chinese machine translation (MT) and Tibetan headline generation (HG), including task metrics, semantic similarity, and dominant-language performance on CMRC as a proxy for alignment tax. We additionally report LLM-based preference as a reference-free measure of semantic quality.

Task 1: Tibetan–Chinese Machine Translation (MT). On MT, Strong SFT achieves higher reference-based scores, improving BLEU from 0.4519 (RL) to 0.6006 and semantic similarity from 0.7164 to 0.8282, reflecting stronger surface alignment to references. However, this advantage is less pronounced under LLM-based judgment: Strong SFT is preferred in 59.2% of cases, while the RL-aligned model still wins 33.5%, indicating competitive semantic quality despite lower BLEU.

These metric gains come with a substantial alignment tax. After adaptation, SFT suffers marked degradation on CMRC (41.82 Avg / 62.99 F1), whereas RL preserves general capability significantly better (46.97 Avg / 65.79 F1). Overall, token-level imitation inflates reference-based MT metrics at the cost of forgetting, while constrained semantic alignment via RL yields safer updates with lower alignment tax.

Task 2: Tibetan Headline Generation (HG). On HG, Strong SFT again achieves higher reference-based scores (ROUGE-L 0.3095 vs. 0.2530 for RL), while the semantic similarity gap remains small (0.6499 vs. 0.6404). Both methods largely preserve dominant-language performance, with only minor differences in CMRC. However, under LLM-based judgment, RL is strongly preferred: it wins 51.2% of pairwise comparisons versus 35.1% for SFT (+16.1 points). This suggests that in open-ended generation, semantic-reward-driven RL learns generation behaviors that go beyond reference imitation, capturing alternative yet semantically appropriate realizations that are more human-preferred despite lower n-gram overlap.

Overall analysis. Across both tasks, Table 3 reveals a consistent pattern: supervised fine-tuning excels at maximizing reference-based metrics, while semantic-reward-driven RL better preserves general capabilities and improves semantic quality under preference-based evaluation. In MT, this trade-off manifests primarily as alignment tax, where SFT's metric gains coincide with substantial forgetting. In HG, where multiple valid realizations exist, RL is consistently preferred by LLM judges despite lower ROUGE, indicating that it learns generation patterns not anchored to a single reference form.

Together, these results suggest a fundamental mismatch between reference-based metrics and true semantic quality in low-resource settings. By aligning models in semantic space rather than enforcing surface imitation, constrained RL enables the acquisition of alternative, semantically valid generation paradigms that are poorly reflected by n-gram metrics but better capture human preferences.

4.4 Experiment 3: Few-Shot Transferability

Finally, we examine whether the stronger MT metrics observed for SFT in Experiment 2 translate into better *cross-task* generalization. While SFT achieves higher reference-based scores on MT, Experiment 2 also reveals a clear mismatch between such metrics and semantic quality, as reflected by alignment tax and LLM-based judgment. This raises a natural question: *do higher MT scores indicate genuinely stronger Tibetan representations, or do they primarily capture task- and reference-specific surface patterns that are unlikely to transfer?*

To answer this, we design a few-shot transfer test from MT to HG.

Concretely, we take the best MT checkpoints

Model	Task Performance		General Capability (Alignment Tax)		Semantic Quality
	Metric	Similarity	CMRC Avg	CMRC F1	LLM-Judge Win (%)
Task 1: Tibetan–Chinese Machine Translation (MT)					
Strong SFT	**0.6006**	**0.8282**	41.82	62.99	59.2
RL (Ours)	0.4519	0.7164	**46.97**	**65.79**	33.5
Gap (RL vs. SFT)	-0.1487	-0.1118	**+5.15**	**+2.80**	-25.7
Task 2: Tibetan Headline Generation (HG)					
Strong SFT	**0.3095**	0.6499	44.20	65.30	35.1
RL (Ours)	0.2530	0.6404	**45.10**	65.20	**51.2**
Gap (RL vs. SFT)	-0.0565	-0.0095	**+0.90**	-0.10	**+16.1**

Table 3: **Trade-off Analysis: Task Performance vs. Alignment Tax.** For machine translation, Strong SFT achieves higher task metrics but incurs a heavy alignment tax, reflected by a significant drop in CMRC performance. In contrast, RL preserves general language capabilities with substantially higher CMRC scores while sacrificing surface-level metrics. For headline generation, both methods exhibit comparable general capability preservation, but RL significantly outperforms SFT in semantic quality as measured by LLM-based judgment, despite lower n-gram-based scores.

Initialization	ROUGE-L	Similarity
Base Model	0.1585	0.4695
MT-SFT	**0.1935**	0.5456
MT-RL (Ours)	0.1918	**0.5690**

Table 4: Few-shot transfer from MT to HG with 1,000 HG training samples.

produced by Strong SFT and by RL, and fine-tune each of them on the HG task using only 1,000 training samples under identical training settings. This setting stresses representation reuse: with limited HG supervision, a model that learns more general and semantically grounded Tibetan features during MT should adapt more effectively than a model whose gains are dominated by task-specific imitation.

Table 4 reports the results. Both MT-adapted models improve substantially over the base model, confirming that MT training provides useful Tibetan signal for downstream generation. However, despite MT-SFT's strong MT performance, it does not retain a corresponding advantage in transfer. The RL-initialized model achieves a higher semantic similarity score (0.5690 vs. 0.5456), while maintaining comparable ROUGE-L (0.1918 vs. 0.1935). This indicates that the MT-SFT model's improvements are, at least in part, tied to MT-specific surface alignment and do not generalize as strongly to a different open-ended generation task, whereas semantic-reward-driven RL yields representations that transfer better under limited supervision.

Overall, this experiment supports our central claim that semantic-space alignment provides a safer and more effective adaptation paradigm in low-resource settings: while SFT can produce larger in-task metric gains, RL achieves more robust generalization across tasks, consistent with the practical needs of low-resource language expansion.

5 Conclusion

This paper argues that low-resource language expansion should be treated as an *alignment* problem, where the core objective is semantic consistency rather than token-level imitation. We propose a semantic-space alignment paradigm instantiated with reinforcement learning driven by embedding-level semantic similarity and a strict language-consistency constraint. Experiments on Tibetan–Chinese machine translation and Tibetan headline generation show that semantic-reward-driven RL acquires low-resource language capabilities while substantially reducing alignment tax, preserving dominant-language competence with near-zero forgetting. We further observe a consistent mismatch between reference-based metrics and semantic quality: despite weaker n-gram overlap, RL is often preferred by LLM-based judges in open-ended generation and yields more transferable representations under few-shot transfer.Taken together, these findings suggest that semantic-space alignment offers a scalable path for extending LLMs to weakly supported languages under data scarcity, shifting low-resource adaptation from distribution matching toward meaning-centered alignment.

Limitations

While modern LLMs nominally support many languages, identifying a language with weak model performance often implies severe data scarcity, as in the case of Tibetan. The limited and domain-narrow nature of available data (e.g., translation corpora) may cause supervised fine-tuning to achieve artificially high in-domain metrics that do not fully reflect real-world generalization, making some degree of overfitting unavoidable.

Ethical Considerations

This work promotes inclusive language modeling by extending LLMs to low-resource languages such as Tibetan. All data are publicly available or internally licensed and contain no personal or sensitive information. No human participants were involved, and all evaluations were performed automatically using LLM-based judges. While our method reduces overfitting and potential bias from narrow supervision, residual pretrained biases may persist. Future research should further assess fairness and bias in low-resource settings.

References

Thales Sales Almeida, Rodrigo Nogueira, and Hélio Pedrini. 2025. Curió-edu 7b: Examining data selection impacts in llm continued pretraining. *Preprint*, arXiv:2512.12770.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Majed Alrubaian, Ali Alammari, Zaki Alawami, and 8 others. 2025. Allam: Large language models for arabic and english. In *International Conference on Learning Representations (ICLR) 2025*. Poster.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

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.

Zoltan Csaki, Bo Li, Jonathan Lingjie Li, Qiantong Xu, Pian Pawakapan, Leon Zhang, Yun Du, Hengyu Zhao, Changran Hu, and Urmish Thakker. 2024. SambaLingo: Teaching large language models new languages. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 1–21, Miami, Florida, USA. Association for Computational Linguistics.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for Chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5886–5891, Hong Kong, China. Association for Computational Linguistics.

DeepSeek-AI and 1 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Danni Liu and Jan Niehues. 2025. Conditions for catastrophic forgetting in multilingual translation. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 347–359. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

OpenAI. 2025. Introducing gpt-4.1 in the api. OpenAI Blog. Accessed: 2025-12-25.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Fabian David Schmidt. 2025. *Robust and Scalable Cross-Lingual Transfer*. Ph.D. thesis, Bayerische

Julius-Maximilians-Universitaet Wuerzburg (Germany).

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. 2015. Trust region policy optimization. *arXiv preprint arXiv:1502.05477*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Kimi Team, Yifan Bai, and 1 others. 2025. Kimi k2: Open agentic intelligence. *Preprint*, arXiv:2507.20534.

Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, and 1 others. 2019. Large-scale datasets for going deeper in image understanding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1480–1485. IEEE.

Guixian Xu, Zeli Su, Ziyin Zhang, Jianing Liu, Xu Han, Ting Zhang, and Yushuang Dong. 2025. CMHG: A dataset and benchmark for headline generation of minority languages in China. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12350–12357, Suzhou, China. Association for Computational Linguistics.

Atsuki Yamaguchi, Terufumi Morishita, Aline Villavicencio, and Nikolaos Aletras. 2025. Mitigating catastrophic forgetting in target language adaptation of llms via source-shielded updates. *Preprint*, arXiv:2512.04844.

An Yang and 1 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Menglin Yang, Jialin Chen, Yifei Zhang, Jiahong Liu, Jiasheng Zhang, Qiyao Ma, Harshit Verma, Qianru Zhang, Min Zhou, Irwin King, and Rex Ying. 2025b. Low-rank adaptation for foundation models: A comprehensive review. *CoRR*, abs/2501.00365.

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.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *CoRR*, abs/2401.01055.

A Training and Hyperparameter Details

A.1 Base model and LoRA configuration

We adopt **Qwen3-4B-Instruct** as the base model and apply LoRA for all SFT and RL experiments. Unless otherwise specified, LoRA adapters are inserted into:

- **Self-attention projections:** `q_proj`, `k_proj`, `v_proj`, `o_proj`
- **MLP projections:** `gate_proj`, `up_proj`, `down_proj`

We use LoRA rank $r = 64$, scaling factor $\alpha = 128$, and dropout 0.05 throughout all experiments.

A.2 Supervised fine-tuning (SFT) details

- **Initialization:** Qwen3-4B-Instruct
- **Precision:** BF16
- **Training length:** 3 epochs
- **Optimizer:** AdamW
- **Learning rate:** 2×10^{-5}
- **Scheduler:** cosine decay with warmup ratio 0.1
- **Global batch size:** 32 (2 GPUs $\times$ per-device batch size 8 $\times$ gradient accumulation 2)
- **Sequence length:** determined by the data and model defaults, typically 1024–2048 tokens

A.3 Reinforcement learning (GRPO) details

We perform reinforcement learning using GRPO (Shao et al., 2024) starting from the cold-start SFT checkpoint.

- **Initialization:** SFT checkpoint (cold-start)
- **Precision:** BF16
- **Training length:** 1 epoch
- **Optimizer:** AdamW
- **Learning rate:** 5×10^{-7}
- **Effective global batch size:** 32 (per-device batch size 16, gradient accumulation 2)
- **Group size:** 8 sampled candidate generations per input prompt
- **Max prompt length:** 256 tokens
- **Max completion length:** 256 tokens
- **Sampling temperature:** 0.8
- **Nucleus sampling:** top-p = 0.9

A.4 Rationale for unified configuration

Across SFT and RL, we keep the base model, LoRA configuration, precision, and batch scale as aligned as possible. This reduces confounding factors and supports attributing performance differences to the alignment paradigm itself.

B LLM-judge Prompt

B.1 LLM-judge Prompt for Headline Generation Task

For the Tibetan headline generation task (HG), we used prompt in Figure 2 to evaluate two candidate headlines generated for a Tibetan news article. The evaluation was conducted using GPT-5.2 as the LLM-judge.

B.2 LLM-judge Prompt for Machine Translation Task

For the Tibetan-Chinese machine translation task (MT), we used the prompt in Figure 3 to evaluate two candidate translations. This evaluation was also conducted using GPT-5.2 as the LLM-judge.

In both tasks, the evaluation process was conducted blind, with no direct reference to the model outputs, ensuring an unbiased comparison of the candidate results.

LLM Judge Prompt for headline generation

You are an expert linguist specializing in Tibetan-Chinese journalism.
Your task is to evaluate two candidate headlines (Candidate 1 and Candidate 2) generated for a Tibetan news article.

Article:
[Source Text]:
{src}

[Candidate 1]:
{cand_1}

[Candidate 2]:
{cand_2}

Task:
Compare the two candidates.
- If Candidate 1 is significantly better, output: [[1]]
- If Candidate 2 is significantly better, output: [[2]]
- If both are equally good or bad, output: [[0]]

Provide a brief reason (in Chinese) before your decision.

Output Format:
Reason: <brief explanation>
Decision: [[1]] or [[2]] or [[0]]

Figure 2: Prompt for headline generation evaluation.

LLM Judge Prompt for machine translation

You are an expert linguist specializing in Tibetan-Chinese translation.
Your task is to evaluate two candidate translations (Candidate 1 and Candidate 2) for a Tibetan text.

Article:
[Source Text]:
{src}

[Candidate 1]:
{cand_1}

[Candidate 2]:
{cand_2}

Task:
Compare the two translations.
- If Candidate 1 is significantly better, output: [[1]]
- If Candidate 2 is significantly better, output: [[2]]
- If both are equally good or bad, output: [[0]]

Provide a brief reason (in English) before your decision.

Output Format:
Reason: <brief explanation>
Decision: [[1]] or [[2]] or [[0]]

Figure 3: Prompt for machine translation evaluation.