Intelligent Automation & Soft Computing

Tech Science Press

DOI: 10.32604/iasc.2024.052971

ARTICLE

Improving Low-Resource Machine Translation Using Reinforcement Learning from Human Feedback

Liqing Wang* **and Yiheng Xiao**

School of Information Science and Engineering, Yunnan University, Kunming, 650500, China

*Corresponding Author: Liqing Wang. Email: wlq@ynu.edu.cn

Received: 20 April 2024 Accepted: 08 July 2024 Published: 06 September 2024

ABSTRACT

Neural Machine Translation is one of the key research directions in Natural Language Processing. However, limited by the scale and quality of parallel corpus, the translation quality of low-resource Neural Machine Translation has always been unsatisfactory. When Reinforcement Learning from Human Feedback (RLHF) is applied to low-resource machine translation, commonly encountered issues of substandard preference data quality and the higher cost associated with manual feedback data. Therefore, a more cost-effective method for obtaining feedback data is proposed. At first, optimizing the quality of preference data through the prompt engineering of the Large Language Model (LLM), then combining human feedback to complete the evaluation. In this way, the reward model could acquire more semantic information and human preferences during the training phase, thereby enhancing feedback efficiency and the result's quality. Experimental results demonstrate that compared with the traditional RLHF method, our method has been proven effective on multiple datasets and exhibits a notable improvement of 1.07 in BLUE. Meanwhile, it is also more favorably received in the assessments conducted by human evaluators and GPT-4o.

KEYWORDS

Low-resource neural machine translation; RLHF; prompt engineering; LLM

1 Introduction

Neural Machine Translation (NMT) has always been one of the key research directions in Natural Language Processing and is a key technology to promote communication between different languages and cultures. With the proliferation of NMT models that depend on Large Language Models (LLMs), translation technology has achieved remarkable breakthroughs, resulting in the continuous enhancement and optimization of translation quality [1]. Nonetheless, the triumph of NMT is closely intertwined with computing resources, algorithmic models, and data resources, particularly the dependency on extensive parallel corpus data. Low-resource NMT, however, is constrained by the scale and quality of available parallel corpus data, resulting in unsatisfactory translation outcomes. Despite this, low-resource NMT continues to possess numerous practical demands and application scenarios, thereby rendering it a focal point and hotspot in the realm of machine translation.

In the past few years, LLMs based on Reinforcement Learning from Human Feedback (RLHF), represented by ChatGPT, has achieved remarkable success in various fields such as text generation and automatic summary [2]. RLHF is a technique that leverages reinforcement learning for training and directly optimizes language models using human feedback signals. By explicitly introducing human preferences into the training process of LLMs, RLHF enables the generated content to align more closely with human ethics and values. Recently, the RLHF method has also been proven effective in enhancing translation quality and better aligning with human preferences in the field of NMT [3]. This process involves human annotators ranking the model-generated content based on their preferences, establishing a reward model based on the feedback results, and then optimizing the fine-tuned language model through the Proximal Policy Optimization (PPO) algorithm during the reinforcement learning phase [4].

However, acquiring a high-quality and extensive human preference dataset necessitates significant labor costs and poses stringent demands on the linguistic proficiency of human labelers. Especially in low-resource settings, the direct acquisition of high-quality human preference datasets becomes prohibitively expensive, primarily due to the limited scale and quality of parallel corpus, as well as the rarity of bilingual talents possessing sufficient linguistic abilities. Furthermore, the inferior quality of pre-training data can lead to LLMs generating unanticipated content, encompassing fabricated facts, biased output, and text that significantly diverges from the intended sentence meanings. Overall, within the context of low-resource scenarios, the exploration of how to efficiently leverage RLHF to enhance the quality of translation models, despite limited parallel corpus and minimal labor costs, merits thorough investigation.

Addressing the aforementioned issues, this paper proposes a learning strategy that leverages LLMs and prompt learning to enhance the quality of human preference datasets. Firstly, we allowed the LLM to compare the translations generated by the fine-tuned model in a monolingual context with the labels in terms of linguistic coherence and complex language structure. After the LLM completed the analysis of the sentence information, we instructed it to refine the generated translations based on the semantic meaning of the labels, including optimizing sentence logic and supplementing missing content. After optimizing the content, human labelers rank the different versions according to their preferences, resulting in a high-quality human preference dataset that can be used to train reward models. The aforementioned improvements to the RLHF feedback stage have made this approach more suitable for low-resource machine translation tasks. Experimental results demonstrate that the model effectively improves translation quality and significantly learns human preferences. We have made our code and datasets used in the experiments publicly available.[1]

This paper's primary contributions can be summarized as follows:

(1) We explored the utilization of RLHF to enhance the quality of low-resource machine translation.
(2) We proposed a method that combines human labelers and LLM prompt learning to improve the performance of RLHF in Low-Resource Scenarios.

2 Related Works

2.1 Reinforcement Learning from Human Feedback

In recent years, with the tremendous success of the GPT series of works, research on applying RLHF to natural language processing tasks has significantly increased [5]. Although LLMs trained

[1] https://github.com/kachipaa/RLHF-zh-lo-NMT (accessed on 29 June 2024).

through RLHF exhibit remarkable capabilities, they may still exhibit unexpected behaviors due to the low quality of pre-training data, such as fabricating facts, generating biased or toxic text, and even content harmful to humans [6]. RLHF has been utilized by researchers to render the generated content more harmless and effective [7]. Current research efforts have been directed towards enhancing translation quality in the field of machine translation through utilizing RLHF, yielding promising results. This research demonstrates that RLHF can significantly improve the performance of translation models, but further validation in low-resource scenarios is still needed [3].

However, the use of human feedback in RLHF poses several challenges. Human preferences are often noisy and may exhibit signs of ambiguity or conflict. This uncertainty in the data can have adverse effects on the accuracy and effectiveness of the reward model [8]. Feedback collected from humans may inherently contain biases or misalignments, influenced by the human labeler's own goals or viewpoints [9]. Additionally, the process of interpreting and modeling human feedback is complex. Different evaluators may interpret the same scenario differently, leading to inconsistent feedback. This variability poses significant challenges in accurately capturing and simulating human preferences.

Moreover, the process of collecting preference datasets for RLHF is time-consuming and labor-intensive, often demanding high standards from human labelers [10]. Some researchers explored using AI to replace humans, significantly reducing training costs while achieving similar training quality to human labelers [11]. Recent research has conducted a detailed comparison between RLHF and Reinforcement Learning from AI Feedback (RLAIF) in terms of resource utilization, time efficiency, and outcomes, ultimately demonstrating that RLAIF outperforms human labelers on average [12].

Considering the reliability and difficulty in obtaining human preference data, which have been a bottleneck in the aforementioned methods. Our approach is different in that we effectively utilize LLM prompt learning to assist human labelers in providing feedback.

2.2 Low-Resource Neural Machine Translation

Addressing the challenge of training high-quality translation models with deep learning for low-resource machine translation has been a persistent research objective, attracting sustained attention. A common approach for low-resource machine translation is data augmentation. A data augmentation method based on soft context is proposed, which replaces a word with a linear combination of similar words from the current context [13]. This approach allows the generated translation to better capture contextual semantic information. In terms of modifying the model. An integrated neural machine translation model with external phrase memories is proposed in the field of external information fusion [14]. This method adds a phrase memory unit to the general model, which stores pre-summarized phrase alignment information. A study incorporated BERT into neural translation systems, enhancing the quality of low-resource translation models by incorporating external knowledge [15]. Several explorations leveraging reinforcement learning to train superior translation models have achieved optimal results across multiple datasets [16].

However, while these studies have improved translation quality, the translated content does not necessarily reflect human preferences. Our approach aims to enhance translation quality while adhering more closely to human linguistic habits.

3 Applied Methods

This paper primarily focuses on the research of low-resource neural machine translation, employing Lao-Chinese mutual translation as the primary carrier. To establish a Lao-Chinese bilingual

machine translation model that aligns with human values, we adopt the mT5 model as our baseline [17]. Training in the following steps as shown in Fig. 1:

1. Fine-tuning the baseline mT5 model using a limited Lao-Chinese parallel bilingual dataset to endow it with basic Lao-Chinese translation capabilities.
2. Incorporating data generated by the Supervised Fine-tuning (SFT) model into the LLM and optimizing it through prompt engineering. Specifically, we utilize a prompt template to allow the LLM to compare the generated content with the translation in terms of sentence meaning logic, semantic fluency, and alignment with human values. By doing so, we enhance the quality of the generated data. Subsequently, human labelers rank the data according to their preferences, ultimately leading to a high-quality human preference dataset.
3. Training a reward model using the human preference dataset. This reward model assigns higher scores to translations of higher quality. The construction of this dataset is a combination of LLM prompt learning and feedback from human labelers.
4. Leveraging the reward model, which encapsulates human preferences and rich semantic knowledge, we enhance the translation quality of the language model through the PPO algorithm.

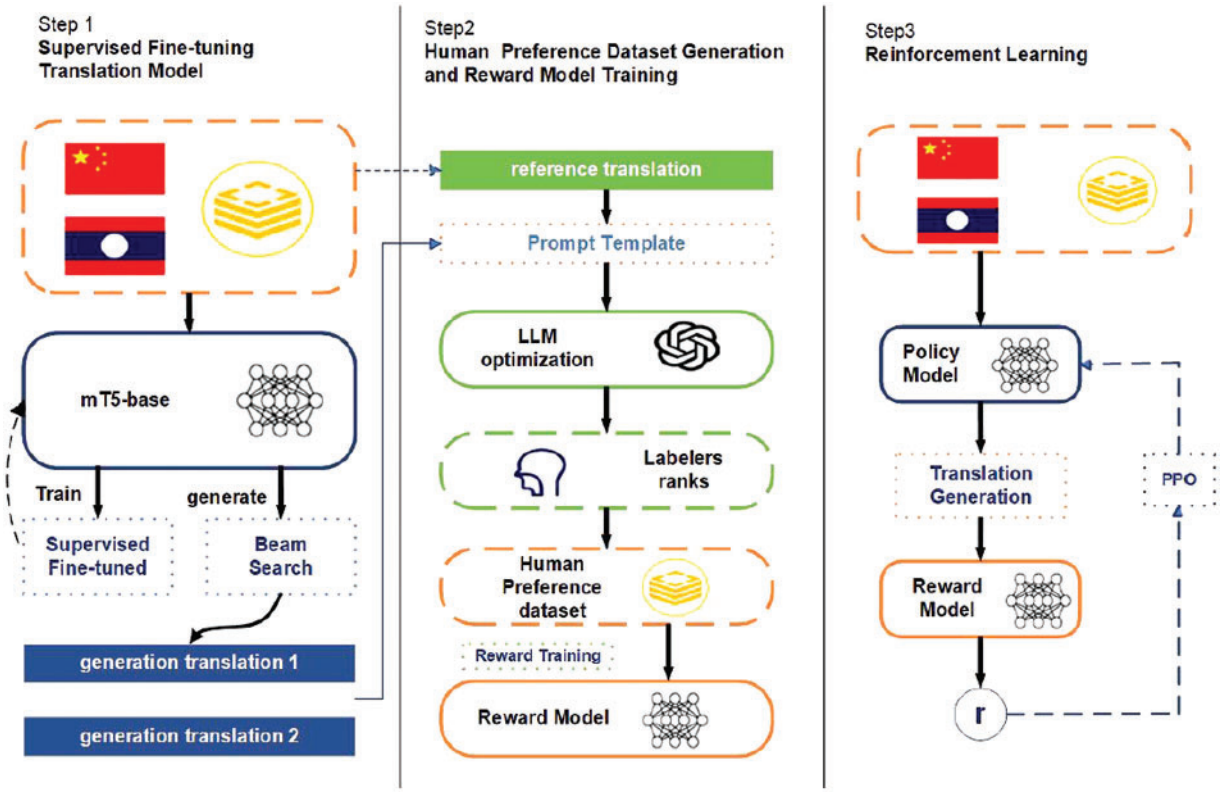

Figure 1: Overview of modeling translation preferences Using RLHF; we optimized the human preference dataset in the step 2 by comparing generated translations with labeled translations through prompting engineering to achieve cost-effective preference learning, thereby reducing the dependence on language experts

By following these steps, we aim to develop a Lao-Chinese machine translation model that not only possesses strong translation capabilities but also aligns closely with human preferences.

3.1 Supervised Fine-Tuning

In the parallel corpus dataset $D_{sft} = \{(x^{(i)},y^{(i)})\}_{i=1,\ldots,n}$, x_i denotes the source language text and y_i denotes the corresponding reference translation. We perform supervised fine-tuning on the model using the constructed Lao-Chinese parallel corpus dataset to obtain a translation language model. Utilizing the parallel corpus dataset D_{sft}, we optimize the translation model using the Cross-Entropy loss function, allowing the model to develop fundamental translation capabilities by increasing the likelihood of producing accurate reference translations. On the supervised fine-tuned model, we employ an enhanced beam search approach to generate a pair of initially diverse translations. The beam search generation function is as follows:

$$P(y|x) = \frac{1}{n^{\alpha}} \log(\Pi_{k=1}^{n} p(y_k|x, y_1, y_2, \ldots, y_{k-1})) \tag{1}$$

where n represents the sequence length, k stands for the beam size, and beam size refers to the search scope during beam search. Additionally, we have optimized the beam search by introducing length normalization to reduce the penalty for generating longer output sentences. $\alpha \in [0, 1]$ is a hyperparameter for adjusting length normalization, where $\alpha = 0$ represents no normalization, and $\alpha = 1$ represents standard length normalization. Through experimental comparison, we adjusted α to 0.7. The generated initial translations are then optimized by the LLM to obtain a human preference dataset, which is subsequently used to train a reward model.

3.2 Human Preference Dataset Generation

High-quality preference data is essential for accurately simulating human preferences. A typical method for modeling human value preferences involves prompting the model to generate two distinct outputs $(y_w, y_l) \sim \pi^{SFT}(\bullet|x)$ in response to a query, where π^{SFT} represents the translation model after supervised fine-tuning, and then allowing human labelers to choose their preference. Here, y_w and y_l represent the acceptance and rejection of the generated content. However, constructing a preference dataset for translation necessitates experts or native speakers of the specific language, which significantly increases costs. For low-resource languages, it can be impractical to find a sufficient number of qualified human labelers.

To address this issue, we opted to utilize a LLM prompting engineering to optimize preference data, significantly reducing our reliance on language experts as shown in Fig. 2. Initially, we employed two distinct prompting templates to enable the LLM to compare generated translations with labeled translations. One template focused on comparing the differences between the two in terms of sentence meaning, syntactic logic, semantic fluency, and emotional preference, while the other template did not impose any limitations on the comparison aspects. Subsequently, we instructed the LLM to improve the generated translation to make it closer to the labeled translation in the aforementioned aspects, resulting in two optimized translations. Next, we presented the optimized translations to human labelers for preference ranking, generating a human preference dataset. The content of the translations optimized by the LLM exhibited significant improvements in translation quality and logic compared to the original content.

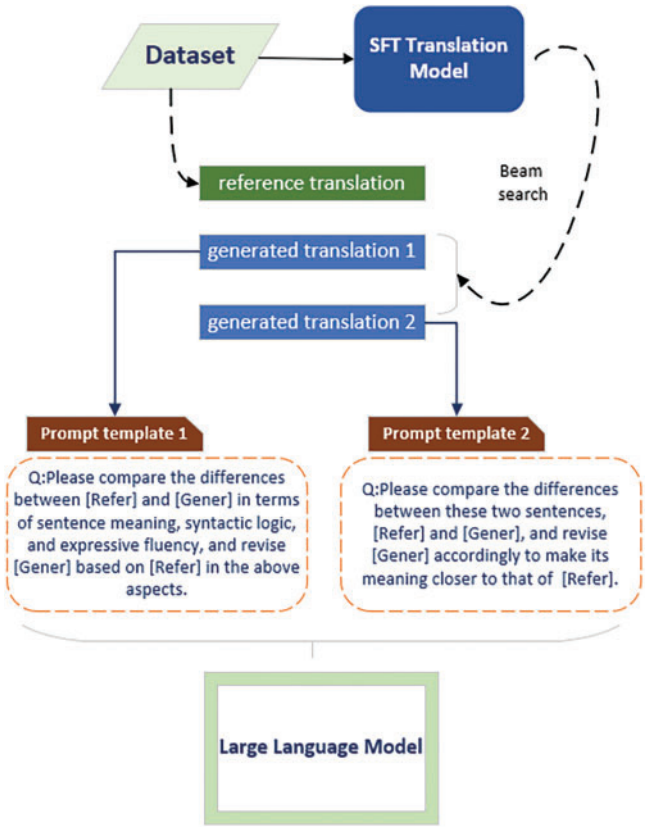

Figure 2: The process of optimizing the generated translation using an LLM involves enhancing the quality of the translation by separately improving the two versions through the application of prompting templates

3.3 Reward Model Training

Based on the optimized human preference dataset, we train a Reward Model using a BERT-based architecture, augmented with an additional linear layer that outputs a scalar value [18]. In the algorithmic research based on PPO, utilizing BERT as the baseline reward model has achieved significant results across various task domains [19–21]. Specifically, the loss function of the reward model can be expressed as:

$$\text{loss}(\theta) = -\frac{1}{\left(\frac{K}{2}\right)} E_{(x,y_w,y_l)} \sim D[\log(\sigma(r_\theta(x, y_w) - (r_\theta(x, y_l))))] \tag{2}$$

where $r_\theta(x, y)$ represents the scalar output of the reward model with parameters θ for the prompt content x and the generated content y. Among the pair of generated data, y_w is the content that more closely aligns with human preferences compared to y_l. D denotes the human preference dataset. The trained Reward Model outputs a reward score for a given prompt and its generated result, serving as a feedback signal for RL.

3.4 Optimizing translation Models through RL

During the reinforcement learning stage, we aim to employ a reward function to provide feedback to the language model. In terms of algorithm selection, we utilize the PPO algorithm to fine-tune the language model. Our objective is to maximize the following combined objective function during reinforcement learning training:

$$\text{objective}(\varnothing) = E_{(x,y)} \sim D_{\pi_{\varnothing}^{RL}}[r_\theta(x, y) - \beta \log(\pi_{\varnothing}^{RL}(y|x)/\pi^{SFT}(y|x))] \quad (3)$$

In the above equation, $\pi_{\varnothing}^{RL}$ represents the learned reinforcement learning policy and π^{SFT} denotes the model trained through supervision. The KL reward coefficient β and the pretraining loss coefficient γ regulate the intensity of the KL penalty and the pretraining gradient, respectively.

4 Experiment Result

This paper explores the feasibility of establishing a low-resource translation model for human preferences in the absence of abundant parallel corpora and human labelers. In Table 1, we illustrate how preference optimization enhances translation quality using an example. By comparing the gaps between machine translation and human translation through prompt engineering, we aim to improve the quality of preference datasets, thereby reducing the dependency on parallel datasets and high-quality human labelers. In this subsection, we first introduce the data and model sources used in our experiments, as well as the environmental details during the experimental process. We compared the translation quality of our experimental results with the baseline model through several aspects, including BLEU, ChrF and human evaluation.

Table 1: A study on the example of Lao-Chinese translation based on RLHF. The bold fonts highlight the generated results that we focus on and compare

Input	ລັດຖະມົນຕີອຸດສາຫະກຳຫຼັກແຫ່ງ NSW ໄດ້ຂໍ້ອ້າວ່າ ສະຖານທີ່ດັ່ງກ່າວຈະຖືກຮັບປະກັນ ຈົນເຖິງ 30 ວັນ ຫຼັງຈາກໄຂ້ຫວັດໃຫຍ່ຄັ້ງຫຼ້າສຸດ.
SFT (Baseline)	荷兰主要工业部长表示,该设施将隔离，到目前为止在未来 30 天内。
RLHF (ours)	**新威尔士州第一工业部长**表示，从**流感最后迹象出现之日**算起，该设施将在接下来的 30 天内被隔离。
Reference	新南威尔士州第一产业部长表示,该设施将被隔离,直到流感最后迹象出现 30 天后。
Commentary	In the SFT translation, the phrase “新南威尔士州第一产业部长” was mistranslated as “荷兰主要工业部长,” resulting in significant discrepancies in both geographical location and organizational titles compared to the original text. Additionally, the key term “流感” was omitted from the translation, leading to poor overall sentence logic and expression fluency. However, in the RLHF translation, the accuracy of the translation was improved, with the omitted information from the original text being supplemented. Furthermore, the overall sentence structure was more logical and aligned with human language habits.

4.1 Datasets

The Chinese-Lao parallel corpus used for supervised fine-tuning of the baseline model consists of 34,065 entries sourced from the Asian Language Treebank (ALT) multilingual parallel dataset. As a

contrast experiment, we also conducted training on Thai-Chinese parallel corpora. The dataset used was also ALT multilingual parallel dataset. The specific division of the dataset is shown in Table 2.

Table 2: Statistical information on Lao-Chinese and Thai-Chinese dataset

Datasets	Training set	Test set	Validation set
ALT parallel dataset (lo-zh)	18000	1000	1106
XLEnt (lo-zh)	12000	1000	959
ALT parallel dataset (th-zh)	18000	1000	1106

4.2 Model

mT5-base: Baseline model used in our experiments. The model has a parameter count of 580M and is trained on the mC4 pre-training dataset, which encompasses 101 languages. As a sequence-to-sequence model, mT5 is inherently designed for translation tasks, and it is pre-trained on a large multilingual corpus covering 101 languages, including both Chinese and Lao, all these making it a strong candidate for fine-tuning on translation-specific data.

BERT-base-Chinese: As the Init Reward Model in our experiments, this model has a parameter count of 110M and was pre-trained using a large-scale Chinese dataset. The decision to use a BERT as the reward model instead of continuing with mT5 was primarily based on the fact that a Chinese pre-trained BERT can better capture the subtle differences in Chinese semantics. Research leveraging BERT as the baseline reward model with the Proximal Policy Optimization (PPO) algorithm has achieved remarkable results in the fields of low-resource language summarization, dialogue systems, and security detection [19–21].

4.3 Implementation Details

The neural network model utilized in this experimental study was implemented using Torch 1.13, with Python 3.9 as the programming language. The experiments were conducted on two NVIDIA 2080Ti GPUs. The mT5-base model was selected as the baseline translation model. In the supervised fine-tuning process, to expedite the experimental process and reduce the demand for graphics memory. The batch size was set to 4.To mitigate the penalty imposed on long sentences, the maximum sentence length was limited to 100. A warm-up strategy was employed to adjust the learning rate, with an initial learning rate of $\eta 0 = 5e{-}4$ and a warm-up step of 4000 [22]. The Adam optimizer was used to train the model for 10 epochs, ensuring convergence. During the decoding process, to generate a pair of contrasting translations, a beam search with a width of 2 was utilized. Model performance was evaluated using both the BLEU score and human evaluation.

4.4 Ablation Study

In this section, we compare the translation quality of the SFT baseline model with the RLHF model using two metrics: BLEU and ChrF, as shown in Table 3. These three sets of experiments represent the fine-tuning of the pre-trained model, the inclusion of RLHF based on fine-tuning, and the ultimate result of our method, respectively.

Table 3: Experimental results for Lao-Chinese and Thai-Chinese translation. Compared to the baseline, RLHF only with human ranking and our proposed method, evaluating the translation quality from two dimensions: BLEU and ChrF

Datasets	Methods	BLEU	Δ	ChrF	Δ
ALT parallel dataset (lo-zh)	SFT (baseline)	9.64	–	32.68	–
	RLHF (only ranks)	11.26	+1.62	49.21	+16.53
	RLHF (ours)	12.33	+2.69	55.84	+23.16
XLEnt	SFT (baseline)	15.24	–	45.71	–
	RLHF (only ranks)	16.52	+1.28	65.41	+19.7
	RLHF (ours)	17.91	+2.67	68.5	+22.79
ALT parallel dataset (th-zh)	SFT (baseline)	15.79	–	43.1	–
	RLHF (only ranks)	17.11	+1.32	58.66	+15.56
	RLHF (ours)	17.63	+1.84	62.04	+18.94

According to the experimental data, on the ALT parallel dataset and XLEnt dataset, our method achieved 2.69-point and 2.67-point improvement in BLEU compared to the baseline,1.07-point and 1.39-point improvement compared to RLHF with only ranking as shown in Figs. 3 and 4. Additionally, our method also demonstrated a significant improvement in ChrF scores. As a comparison, in the Thai-Chinese task, our method achieved a 1.84 improvement in BLEU score compared to the baseline, and a 0.52 improvement in BLEU score compared to RLHF with only ranking.

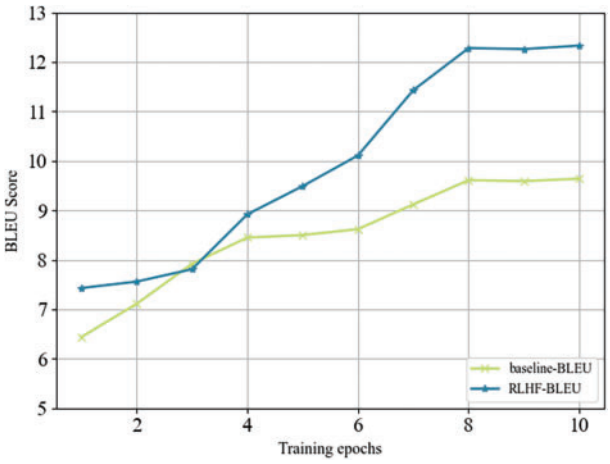

Figure 3: The comparison of our method with the SFT model on the ALT dataset in terms of BLEU score

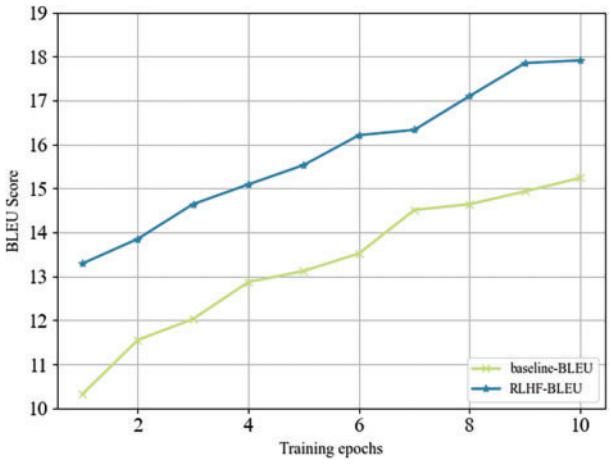

Figure 4: The comparison of our method with the SFT model on the XLEnt dataset in terms of BLEU score

The experimental results indicate that the model acquires basic translation capabilities after SFT, but there is still a significant gap compared to the optimal level. The introduction of RLHF into low-resource machine translation represents an effective attempt, as models trained with RLHF achieve better performance across all datasets. Meanwhile, optimizing the feedback phase of RLHF

can effectively enhance its performance in low-resource machine translation tasks. This not only demonstrates the effectiveness of our approach but also illustrates the sensitivity of the reward model to semantics, indicating that higher-quality data can maximize the effectiveness of the reward model. Notably, our approach yields better results for language pairs with scarcer parallel corpora. Compared to Lao-Chinese corpora, Thai-Chinese corpora are more abundant and have a higher proportion in the mT5 pre-training data. Consequently, the translation quality after fine-tuning alone is higher for Thai-Chinese. However, the improvement brought by our method in the Thai-Chinese language pair is somewhat reduced compared to Lao-Chinese.

4.5 Evaluation

We recruited human evaluators and GPT-4o to compare different models in terms of accuracy, fluency, and harmlessness. Accuracy refers to whether there is any discrepancy in the transmission of key information between the translated text and the original text; Fluency describes whether the generated translation is smooth and readable in the context of Chinese; Harmlessness assesses whether there is any potential unfriendly information hidden in the generated translation. We distributed a questionnaire to volunteers and provided GPT-4o with the same information. The volunteers we recruited for our study are native Chinese speakers found both on campus and on the internet. The survey questionnaire consists of 100 data entries, and each entry needs to be evaluated from three dimensions. Each data entry consists of three pairs of translations generated by two different models. When evaluating each dimension, volunteers need to select the sentence that better meets the requirements based on their own preferences. Meanwhile, we feed the same data to GPT-4o for evaluation using the same methodology. The results are presented as percentages in Table 4.

Table 4: Human evaluators compared different models in terms of accuracy, fluency, and harmlessness. Their preferences were expressed as percentages

Evaluator	Comparison	Accuracy	Fluency	Harmless
Human volunteers	Only-rank RLHF *vs*. SFT	86%	80%	60%
	RLHF (ours) *vs*. SFT	94%	92%	56%
	RLHF (ours) *vs*. Only-rank RLHF	68%	74%	52%
GPT-4o	Only-rank RLHF *vs*. SFT	73%	69%	56%
	RLHF (ours) *vs*. SFT	77%	64%	58%
	RLHF (ours) *vs*. Only-rank RLHF	62%	54%	49%

According to human evaluators, the translations produced by the RLHF model exhibited significant improvements in both accuracy and fluency compared to the baseline. This phenomenon can be primarily attributed to the fact that we only used approximately 20,000 Chinese-Lao parallel corpora to fine-tune the translation model, resulting in numerous issues such as mistranslation and omitted content in the initial translations. When comparing the RLHF model with only ranking capabilities to our proposed method, it is evident that human evaluators preferred the latter. From the overall results, GPT-4o also prefers the model optimized by our method. However, compared to human evaluations, GPT-4o does not demonstrate a strong preference, especially when comparing our method with the RLHF-only method, where there is no significant preference in terms of fluency and harmlessness. We believe this is due to the differences in perception and understanding between LLMs and human evaluators, which also indicates that the translations generated by our method are more recognized by

humans. This indicates that our approach of optimizing the human preference dataset through prompt engineering results in translations that better align with human preferences and values.

4.6 The Impact of Prompts on the Generation of LLMs

In the stage of optimizing the translation output with LLM, to achieve higher-quality generation results, we experimented with various prompt templates, some of which are exemplified in the Table A1. We observed that, for our language optimization task, decomposing the task into multiple subtasks frequently led to better outcomes, rather than directly instructing the LLM to optimize the input's generated output based on the reference translation. In our case, we divided the task into understanding and analyzing the reference translation, analyzing the translation to be revised, comparing the differences in information between the two, and modifying and optimizing the translation accordingly. By leveraging LLM's sensitivity and accuracy for simple subtasks, we were able to consistently achieve higher-quality translations that better aligned with the experimental requirements, consistently with minimal revisions [23].

5 Conclusions

This paper explores the application of RLHF in Lao-Chinese low-resource machine translation to enhance translation quality. We propose a learning strategy suitable for low-resource settings that incorporates prompt engineering in the preference data feedback stage, leveraging LLM to optimize the generated preference data. This approach improves the reward model's understanding of Lao-Chinese semantics, enhancing translation quality while reducing the reliance on high-quality human labelers. Experimental results demonstrate that RLHF is highly effective for low-resource neural machine translation, significantly enhancing the accuracy and fluency of translations, validating that the reward model's linguistic capabilities and the quality of the data affect the effectiveness of RL. Additionally, our method reduces the dependence on high-quality human labelers and extensive parallel corpora during the training process, and the incorporation of prompt engineering in the feedback stage can effectively enhance training efficiency.

Acknowledgement: I would like to give my heartfelt thanks to my institution "School of Information Science and Engineering Yunnan University, Kunming, Yunnan, China" for the support and facilities provided for the research. I am also extremely grateful to all those who devote much time to reading this paper and give me much advice, which will benefit me in my later study.

Funding Statement: This research was supported by the National Natural Science Foundation of China under Grant No. 61862064.

Author Contributions: Study conception and design: Liqing Wang, Yiheng Xiao; data collection: Yiheng Xiao; analysis and interpretation of results: Liqing Wang, Yiheng Xiao; draft manuscript preparation: Yiheng Xiao. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] T. Kocmi and C. Federmann, "Large language models are state-of-the-art evaluators of translation quality," in *Proc. Euro. Assoc. Mach. Trans.*, Tampere, Finland, 2023, pp. 193–203.

[2] D. M. Ziegler *et al.*, "Fine-tuning language models from human preferences," arXiv preprint arXiv:1909.08593, 2019.

[3] N. Xu, J. Zhao, C. Zu, T. Gui, Q. Zhang and X. Huang, "Advancing translation preference modeling with RLHF: A step towards cost-effective solution," arXiv preprint arXiv:2402.11525, 2024.

[4] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2017.

[5] A. Glaese *et al.*, "Improving alignment of dialogue agents via targeted human judgments," arXiv preprint arXiv:2209.14375, 2022.

[6] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg and D. Amodei, "Deep reinforcement learning from human preferences," in *Proc. Adv. Neural Inform. Process. Syst. 30*, Long Beach, CA, USA, 2017, pp. 4299–4307.

[7] Y. Bai *et al.*, "Training a helpful and harmless assistant with reinforcement learning from human feedback," arXiv preprint arXiv:2204.05862, 2022.

[8] R. Zheng *et al.*, "Secrets of RLHF in large language models part I: PPO," arXiv preprint arXiv:2307.04964, 2023.

[9] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," in *Proc. Adv. Neural Inform. Process. Syst.*, New Orleans, LA, USA, 2022, pp. 27730–27744.

[10] Z. Yao *et al.*, "Deepspeed-chat: Easy, fast and affordable RLHF training of chatGPT-like models at all scales," arXiv preprint arXiv:2308.01320, 2023.

[11] H. Lee *et al.*, "RLAIF: Scaling reinforcement learning from human feedback with ai feedback," arXiv preprint arXiv:2309.00267, 2023.

[12] S. Höglund and J. Khedri, "*Comparison between RLHF and RLAIF in Fine-Tuning a Large Language Model,*" 2023. Accessed: Jun. 29, 2024. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-331926.

[13] L. Gao, J. Schulman, and J. Hilton, "Scaling laws for reward model overoptimization," in *Proc. 40th Int. Conf. Mach. Learn.*, Honolulu, HI, USA, 2023, pp. 10835–10866.

[14] Y. Tang, F. Meng, Z. Lu, H. Li, and P. L. H. Yu, "Neural machine translation with external phrase memory," arXiv preprint arXiv:1606.01792, 2016.

[15] J. Zhu *et al.*, "Incorporating bert into neural machine translation," arXiv preprint arXiv:2002.06823, 2020.

[16] X. Wu, S. Lv, L. Zang, J. Han, and S. Hu, "Conditional BERT contextual augmentation," in *Proc. Comput. Sci.-19th Int. Conf. (ICCS 2019)*, Faro, Portugal, 2019, pp. 84–95.

[17] L. Xue *et al.*, "mT5: A massively multilingual pre-trained text-to-text transformer," arXiv preprint arXiv:2010.11934, 2020.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[19] V. N. M. Abadi and F. Ghasemian, "*Enhancing Persian Text Summarization using the mT5 Transformer Model: A Threephased Fine-Tuning Approach and Reinforcement Learning*," 2023. doi: 10.21203/rs.3.rs-3682780/v1

[20] P. Jha, J. Scott, J. S. Ganeshna, M. Singh, and V. Ganesh, "BertRLFuzzer: A BERT and reinforcement learning based fuzzer (Student Abstract)," in *Proc. AAAI Conf. Artif. Intell.*, Vancouver, BC, Canada, 2024, pp. 23521–23522.

[21] H. Wang, H. Wang, Z. Wang, and K. Wong, "Integrating pretrained language model for dialogue policy evaluation," in *Proc. IEEE Int. Conf. Acoust. Speech Sig. Process. (ICASSP)*, Toronto, ON, Canada, 2021, pp. 6692–6696.

[22] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 1–67, 2020.

[23] X. García and O. Firat, "Using natural language prompts for machine translation," arXiv preprint arXiv:2202.11822, 2022.

Appendix A. Example of Prompt

Table A1: Some prompt templates explored during the optimization feedback phase of utilizing LLM

Prompt templates
"You are an expert in translation, and I require you to optimize and modify the translation based on the provided reference sentences in a fair and objective manner. While making minimal changes to the translation, please modify it from the following perspectives: Translation Accuracy: The translation must accurately reflect the content and meaning of the reference. No information should be introduced or omitted. Translation Fluency: The translation should not be strictly limited to the structure of the reference. It should express the meaning in a natural and fluent manner. Translation Logic: The logical flow of the translated sentences should align with the reference. Optimization Suggestions: Please review the current translation and provide suggestions to improve its quality based on the evaluation criteria. Consider the following aspects: Are there any stiff or awkward phrases that can be rephrased for a more natural flow? Are there any inaccurate or missing details that need to be corrected or added? Is there any syntactic logic in the translation, such as active-passive voice, that contradicts the reference? ***Input*** [reference] [translation] ***output***
You are a language expert, I require you to objectively optimize and modify the translation based on the provided reference sentences. Firstly, you need to understand and analyze the key information contained in the reference translation. Based on the analyzed information, you must determine if the information in the translation aligns with the reference. If there are differences, omissions, or additions, you should make minimal changes to ensure the translation's information is consistent with the reference. Finally, you need to ensure the translation flows smoothly with minimal revisions. Please review the current content and provide suggestions to improve its quality based on the evaluation criteria. Consider the following aspects: Are there any stiff or awkward phrases that could be rephrased for a more natural flow? Are there any inaccurate or missing details that require correction or addition? Is there any syntactic logic in the translation, such as active or passive voice, that contradicts the reference?" ***Input*** [reference] [translation] ***output***