Tuning Less, Prompting More: A Cost-Effective and Flexible Training and Inference Pipeline for Natural Language Transformation

Anonymous ACL submission

Abstract

Natural language transformation (NLT) tasks, such as machine translation (MT) and text style transfer (TST), require models to generate accurate and contextually appropriate outputs. However, existing approaches face significant challenges, including the computational costs of leveraging large pre-trained models and the limited generalization ability of fine-tuned smaller models. In this paper, we propose a novel framework that combines the flexibility of prompting with the cost-effectiveness of fine-tuning. Our method enhances smaller models by integrating In-Context Examples (ICE) retrieved from training data, enabling the model to better capture contextual information and align with user preferences. We further improve performance through hierarchical contrastive learning and dynamic preference inference mechanisms. Experimental results demonstrate that our approach outperforms existing methods, such as Supervised Fine Tuning (SFT), Direct Preference Optimization (DPO), and Contrastive Preference Optimization (CPO), across both MT and TST tasks, providing a more efficient solution for resource-constrained environments.

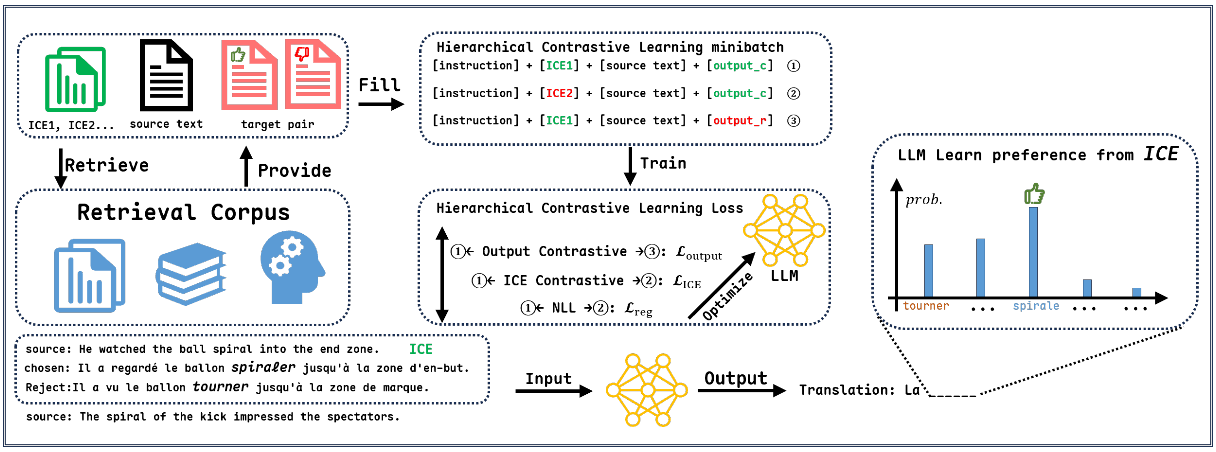
1 Introduction

Recent advancements in natural language transformation (e.g., machine translation and text style transfer) have been significantly driven by large language models (LLMs) (Achiam et al., 2023; Grattafiori et al., 2024; DeepSeek-AI et al., 2024). Depending on whether model parameters are modified, existing approaches can be categorized into two main strategies: maintaining fixed parameters to elicit capabilities from large models through prompting, and modifying parameters to optimize performance of smaller models via fine-tuning. Prompting large models leverages their inherent flexibility and preserves their generalization capabilities without compromising the model's universal applicability (Hendy et al., 2023; Jiao et al., 2023b; Zhu et al., 2023). This approach allows for effective few-shot learning and adaptability across diverse tasks. On the other hand, fine-tuning smaller models, a trend gaining traction in recent studies, offers a cost-effective alternative by tailoring models to specific translation tasks (Zeng et al., 2023; Jiao et al., 2023a; Kudugunta et al., 2024; Zan et al., 2024; Li et al., 2024; Xu et al., 2023).

Despite their advantages, both methods exhibit significant drawbacks. While prompting large models can maintain high flexibility and generalization, it incurs substantial computational costs, which not only limits their practical deployment in resource-constrained environments but also leads to unavoidable performance inefficiencies (Xu et al., 2023). Conversely, fine-tuning smaller models for natural language transformation tasks involves a trade-off, where improving performance on specific tasks often comes at the cost of reduced generalization ability. This decline in generalization is particularly noticeable when the model encounters tasks it was not explicitly trained on. This phenomenon can be explained by the "prompt shift" effect (Li et al., 2023; Li and Hoiem, 2017; Lopez-Paz and Ranzato, 2022), where even small changes in the prompt format (without any change in meaning) can lead to a significant drop in response quality. Additionally, models are typically trained on parallel corpora, which limits their inference pattern to a [source text] -> [target text] mapping. This approach struggles to handle the polysemy that arises from contextual differences in translation, as too much contextual information can disrupt the format of the mapping. Additionally, the preference alignment techniques employed during the fine-tuning of smaller models struggle to handle data from multiple distributions and diverse contexts, further restricting their applicability.

To address these challenges, we propose a novel approach that combines the flexibility of prompting

Figure 1: Framework for our Model Training and Inference. The diagram illustrates the three core components: (1) Data Augmentation through Contextual Similarity-based ICE Retrieval (top-left), (2) Hierarchical Contrastive Loss Design for fine-tuning model outputs (center), and (3) Dynamic Preference-Aware Inference for generating contextually relevant outputs (bottom-left). The arrows represent the flow of data and processes between these components.



with the cost-effectiveness of fine-tuning. Through an innovative training and inference pipeline, our method enables smaller models to extract sufficient contextual information from highly diverse In-Context Examples (ICE) within the prompting paradigm, learning fine-grained target preferences from similar contexts, thereby improving the quality of generated responses. Specifically, we retrieve similar samples from the training data, combine them into a few-shot learning format to enhance data points, and implement robustness mechanisms to handle the diversity of ICEs, further enhancing model performance through preference alignment design.

2 Related Work

In this section, we first reviewed the relevant research on Natural Language Transformation (NLT) tasks. Subsequently, we focused on two approaches based on large language models (LLMs): one involves leveraging the capabilities of general-purpose large models through the use of prompts, while the other entails fine-tuning smaller models for task-specific optimization. Finally, we analyzed the advantages, disadvantages, and challenges of both approaches in practical applications.

2.1 Natural Language Transformation

Natural Language Transformation (NLT) refers to the process of rewriting or adapting natural language texts at semantic, structural, or stylistic levels to fulfill specific task requirements or achieve desired functionalities. It encompasses a variety of tasks, including **machine translation** and **text style transfer**, which address diverse expressive objectives and application scenarios through flexible manipulation of linguistic elements. Building upon the foundational principles of Natural Language Transformation, recent advancements in large language models (LLMs) have catalyzed novel methodologies that leverage prompt-based paradigms to achieve flexible and context-aware language manipulation.

2.2 LLM-based Methods

Unlike traditional NLT approaches that often rely on task-specific architectures or explicit feature engineering, prompting-based methods exploit the intrinsic knowledge and generative capabilities of LLMs through carefully designed instructions or exemplars. This shift has enabled dynamic adaptation across diverse transformation tasks—from stylistic rewriting to domain-specific paraphrasing—by reformulating objectives as natural language prompts. Notably, techniques such as few-shot prompting, instruction tuning, and chain-of-thought reasoning have demonstrated remarkable efficacy in steering LLMs to disentangle semantic, structural, and stylistic nuances without requiring extensive fine-tuning.

In parallel to prompt-centric methodologies, recent efforts have prioritized specialized adaptation of compact language models through multi-stage fine-tuning strategies to address the unique demands of Natural Language Transformation (NLT). This paradigm typically begins with **continued pre-**

training, where models undergo domain-specific knowledge infusion via exposure to task-aligned corpora—such as stylistic parallel texts or multilingual translation pairs—to recalibrate their latent representations for transformation-centric objectives. Building upon this foundation, **instruction tuning** further optimizes models by training them on structured prompt-output pairs, enabling precise interpretation of diverse transformation intents. Finally, **preference alignment** mechanisms incorporate human or automated feedback—through techniques like reinforcement learning from human preferences or contrastive ranking—to refine outputs along critical dimensions such as stylistic consistency, lexical appropriateness, and domain-specific constraints. Collectively, this phased optimization framework allows smaller models to achieve task-aware specialization while maintaining computational tractability, thereby offering a viable alternative to large-scale LLMs in scenarios requiring strict deployment efficiency or niche-domain expertise.

2.3 Preference Alignment

Building upon the foundational stages of domain adaptation through continued pretraining and instruction tuning, preference alignment emerges as a critical mechanism to bridge the gap between model capabilities and human-centric quality requirements. While the former stages equip models with task-specific knowledge, the latter ensures that outputs adhere to nuanced desiderata: stylistic coherence, lexical precision, and context-sensitive appropriateness, which are often underspecified in textual instructions. This subsection systematically examines cutting-edge approaches to preference alignment, analyzing their technical innovations in reconciling algorithmic optimization with human judgment across diverse NLT scenarios.

Three noteworthy methods for preference alignment include Proximal Policy Optimization (PPO) (Schulman et al., 2017), Direct Preference Optimization (DPO) (Rafailov et al., 2024), and Contrastive Preference Optimization (CPO) (Xu et al., 2024). PPO sets a high-level goal for the alignment process: to maximize the "satisfaction" of a reward model trained on carefully organized preference data. While effective, this approach involves a complex pipeline, including reward modeling and iterative policy updates. In contrast, DPO and CPO bypass the need for explicit reward modeling. These methods directly utilize the general tendencies of the collected preference dataset as the alignment target, thereby simplifying the alignment process. Together, these methods provide a range of approaches for aligning model outputs with human preferences, each offering unique advantages depending on the complexity and constraints of the task at hand.

3 Methods

In this section, we present a detailed description of our framework, which comprises three core components (as illustrated in **Figure 1**): (1) Contextual Similarity-based ICE Data Augmentation, (2) Hierarchical Contrastive Loss Design that guides the model to disentangle fine-grained semantic features from augmented data, and (3) Dynamic Preference Inference Mechanism that enhances model outputs through user-customized or retrieval-augmented preference sample prompts. The subsequent sections will systematically elaborate on the design principles and implementation details of each component.

3.1 Contextual Similarity-Based ICE Augmentation via Retrieval

To address the limitations of rigid [source→target] mappings in conventional fine-tuning, we introduce a retrieval-augmented paradigm that enriches training data with explicit preference signals through in-context exemplars. The core objective is to enable smaller models to resolve contextual ambiguity by learning from semantically aligned preference pairs, thereby emulating the few-shot reasoning capabilities of large language models. Our approach constructs augmented data points by retrieving semantically congruent exemplars from the training corpus based on source text similarity. This is achieved by computing cosine distances between sentence embeddings derived from a pretrained language model, ensuring that the retrieved in-context examples (ICEs) share comparable linguistic patterns and domain characteristics with the target source.

Each retrieved ICE (ICE source, ICE chosen, ICE rejected) serves as explicit preference evidence, forming an augmented input structure: [ICE source, ICE chosen, ICE rejected; current source] $\rightarrow$ current chosen. This format allows the model to simultaneously learn from both the target instance and its contextual neighbors, reinforcing preference consistency across analogous contexts.

By integrating ICEs, the model is exposed to diverse yet semantically aligned examples, enabling it to generalize beyond the narrow [source→target] mapping and capture fine-grained preference patterns conditioned on contextual semantics.

The methodology is designed to be flexible and extensible. While our implementation utilizes in-corpus retrieval for experimental consistency, the architecture inherently supports integration with external knowledge bases through unified embedding alignment. Additionally, the approach incorporates robustness mechanisms by prioritizing high-similarity exemplars (top-k retrieval) and introducing adversarial negative samples through lexical substitution, mitigating potential noise from suboptimal retrievals. This retrieval-enhanced paradigm fundamentally repositions smaller models from passive pattern memorizers to active context interpreters, laying a critical foundation for subsequent preference alignment learning.

3.2 Hierarchical Contrastive Learning with Dual ICE Augmentation

Building upon the retrieval-augmented samples $[ICE, \text{prompt}, \text{chosen}, \text{reject}]$ from the previous step, we introduce a hierarchical contrastive objective that operates at both the model output and ICE levels. First, to strengthen the model's ability to distinguish between preferred and rejected outputs, we apply contrastive learning between $[ICE, \text{prompt}, \text{chosen}]$ and $[ICE, \text{prompt}, \text{reject}]$. By emphasizing the difference between the chosen and rejected outputs under otherwise identical contextual cues, this step enhances the model's ability to internalize explicit preference signals. Crucially, this contrastive signal helps the model capture more subtle features that drive user-preferred outputs, moving beyond simple pattern matching. The contrastive loss at the output level can be expressed as:

$$\mathcal{L}_{\text{output}} = -\log \sigma \left(\beta \log \frac{\pi_\theta(y_c|ICE_1, x)}{\pi_\theta(y_r|ICE_1, x)} \right) \quad (1)$$

Where, π_θ represents the current model, y_c and y_r stand for chosen output and rejected output. β is a hyperparameter, and in the experiment, we set it to 0.15.

Next, to further enhance the model's robustness to variations in the quality of ICE selections, we introduce a second set of exemplars (ICE_2) and construct new augmented input pairs: $[ICE_1, \text{prompt}, \text{chosen}]$ and $[ICE_2, \text{prompt}, \text{chosen}]$. This dual ICE augmentation enforces an additional contrastive objective, encouraging consistent representations of the same target across different in-context exemplars. By aligning semantically similar but potentially divergent examples, this approach mitigates the impact of retrieval noise, ensuring that the learned preferences remain robust despite contextual changes. The contrastive loss at the ICE level can be formulated as:

$$\mathcal{L}_{\text{ICE}} = \text{Sim}[\pi_\theta(ICE_1; x; y_c); \pi_\theta(ICE_2, x; y_c)] \quad (2)$$

where $\pi_\theta(ICE_1; x; y_c)$ represents the last token's logits after model computing, and Sim() means cosine similarity function.

Finally, to stabilize learning and maintain a coherent preference distribution, we introduce a regularization term based on the negative log-likelihood (NLL) between $[ICE_1, \text{prompt}, \text{chosen}]$ and $[ICE_2, \text{prompt}, \text{chosen}]$. This regularization term can be represented as:

$$\mathcal{L}_{\text{reg}} = \text{NLL}[\pi_\theta(y_c|ICE_1, x)] + \quad (3)$$
$$\text{NLL}[\pi_\theta(y_c|ICE_2, x)] \quad (4)$$

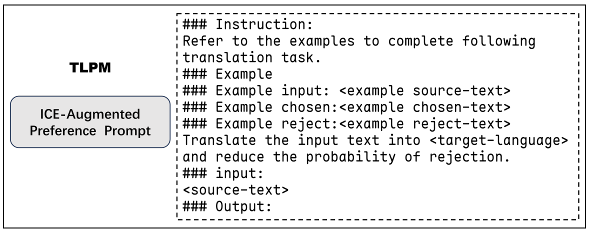
The overall training objective combines these three loss components through a weighted to balance their contributions during optimization. By integrating output-level discrimination, ICE-level consistency, and distributional regularization, the unified loss function ensures the model learns robust preference patterns while maintaining generation stability. The total loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{output}} + \min(\frac{\mathcal{L}_{\text{reg}}}{\mathcal{L}_{\text{ICE}}}, \lambda)\mathcal{L}_{\text{ICE}} + \mathcal{L}_{\text{reg}} \quad (5)$$

where λ is weighting coefficients that control the relative strength of each objective. This balanced formulation allows the model to simultaneously optimize for preference discrimination, contextual robustness, and distributional coherence. In practice, we find equal weighting ($\lambda = 1.0$) provides stable convergence, though domain-specific tuning may further enhance performance.

Through this layered contrastive framework, the model develops a more comprehensive and robust

Figure 2: Inference Prompt in Machine Translation



understanding of user-driven preference patterns, laying a strong foundation for subsequent dynamic preference inference tasks. The modular design of this approach allows for easy adaptation and expansion to other domains, making it both flexible and extensible for future research and applications.

3.3 Dynamic Preference-Aware Inference with Retrieval-Augmented Prompts

After completing the hierarchical contrastive learning process described earlier, the model enters the inference phase for dynamic preference-driven tasks. At this stage, the model leverages the ICEs and contrastive learning patterns acquired during training to generate outputs that align with user preferences.

During inference, the model takes source text as input and dynamically customizes ICEs by retrieving relevant examples based on specific scenarios or user preferences. These are then combined with the source text to form an enhanced input in the format "[ICEs, source text]." This approach ensures the model's reasoning is guided by preference-aligned examples, producing outputs that maintain linguistic accuracy while adapting to different requirements.

The dynamic preference-aware inference mechanism enables the model to flexibly handle diverse scenarios, delivering outputs that are accurate, personalized, and contextually relevant—making it suitable for real-world applications where preferences and context play a decisive role.

4 Experiments

To comprehensively evaluate the effectiveness of our proposed framework, we conduct extensive experiments across two dimensions: (1) comparing performance against state-of-the-art baselines, and (2) analyzing the robustness of our hierarchical contrastive learning mechanism through ablation studies. All experiments are designed to answer the key research question: How does our retrieval-augmented training inference paradigm improve preference alignment compared to conventional fine-tuning?

4.1 Data

We conducted experiments on the MT and TST tasks. In this section, we will provide a detailed explanation of how we obtained, processed, and assembled our data.

4.1.1 Preference Data

For MT, we conducted the experiments on the **ALMA-R-Preference** dataset which (Xu et al., 2024) released, and selected the chosen and rejected translations for the target language based on the average quality of each data item. We also performed supplementary experiments on translation tasks involving other low-resource languages using the Flores-200 dataset (Team et al., 2022). We converted the FLORES-200 dataset into a pairwise preference dataset by using GPT-4o-mini to generate candidate translations, and applied the Feedback from Inductive Biases method (Jiang et al., 2024) to construct preference directions. For the TST task, we directly used an open-source preference dataset[1]. This dataset is designed to convert modern language into the writing style of specified literary works, encompassing idiomatic vocabulary, syntactic structures, and rhetorical devices. The target literary works include China's Four Great Classical Novels.

4.1.2 Data preparation

In all experiments, we split a preference dataset into training and testing sets with an 4:1 ratio. In the implementation of the retrieval component, we use **SentenceTransformer** with the "**xlm-r-bert-base-nli-stsb-mean-tokens**" model (Reimers and Gurevych, 2019) to generate dense vector representations of input queries. The model is configured with a maximum sequence length of 512. The generated embedding vectors serve as the foundation for retrieval, effectively capturing semantic relationships within the source text. To enable efficient similarity retrieval, we construct an IVF-Flat clustering index based on **Faiss** (Douze et al., 2024). The embedding space is partitioned into 50 clusters to facilitate scalable retrieval. In our experiments, we use the original training set as the retrieval space. For each sample point, we identify the two most

[1] https://github.com/stylellm/stylellm_models

Table 1: The main result in Translating to English (xx→en). Our methods significantly outperform all comparable methods. The dark blue boxes indicates a significant improvement compared to their original versions , while light blue boxes represents only a small but noticeable enhancement. All red colors indicate a slight decrease in performance.

Methods	de		zh		ru		cs		ind	
	BLEU	XCOMET	BLEU	XCOMET	BLEU	XCOMET	BLEU	XCOMET	BLEU	XCOMET
SFT	33.12	93.67	25.13	90.45	39.12	90.42	41.22	86.54	31.05	91.86
DPO	31.99	93.24	25.17	89.94	39.11	89.16	42.15	86.70	31.58	91.92
CPO	32.74	94.72	26.32	91.73	38.26	91.85	43.13	89.91	30.27	92.60
Ours	34.21	96.31	26.22	92.86	39.13	93.44	41.47	91.37	37.47	94.71

Table 2: The main result in Translating from English (en→xx).

Methods	de		zh		ru		cs		ind	
	BLEU	XCOMET	BLEU	XCOMET	BLEU	XCOMET	BLEU	XCOMET	BLEU	XCOMET
SFT	30.25	88.81	27.94	87.94	27.12	88.37	26.22	87.62	25.93	89.48
DPO	29.50	90.03	27.33	88.23	26.32	87.03	26.25	88.68	25.41	89.43
CPO	30.54	90.16	24.87	89.85	25.14	89.63	27.13	88.73	27.21	90.04
Ours	31.22	91.74	26.33	90.51	27.22	90.47	29.47	89.53	26.01	90.13

similar samples based on its source text and designate them as ICE_1 and ICE_2.

4.2 Experiments Setting

In this section, we present the baselines used in our comparative experiments, along with detailed experimental configurations and the hardware setup. Following this, we introduce the baselines chosen for our experiments and explain the rationale behind these selections.

4.2.1 Baseline

SFT Using Supervised Fine-Tuning to adapt large language models to specific downstream tasks is a fundamental approach. Its effectiveness has been validated through extensive practical experiments. Therefore, SIT on prefer data serves as the first baseline in our experiments.

DPO Direct Preference Optimization is a method designed to directly optimize models for preference alignment, focusing on aligning model outputs with user preferences rather than traditional loss functions. It has gained widespread use in the field of preference learning, especially in scenarios where the goal is to predict or rank items based on user preferences. DPO has become a popular choice in many preference alignment applications. Before applying the DPO method, we conducted preliminary training with the selected data to simulate the typical pipeline for preference alignment using DPO.

CPO We also compared the commonly used preference alignment methods in the machine translation field. Contrastive Preference Optimization is derived from the same optimization goal but reflect different training objectives with DPO. Therefore, these two methods serve as the primary comparative methods for the evaluation of preference alignment.

4.2.2 Training Details

Our experiment primarily focuses on comparing fine-tuning methods rather than specific base models. We conducted our main experiments on widely used open-source large language models (Touvron et al., 2023; Dubey et al., 2024). The experiments employed earlier versions, namely LLaMA2-13B and LLaMA3-8B, with the core research findings presented in the paper. Specifically, the MT experiments utilized LLaMA2, while the TST experiments were performed using LLaMA3.

Training with LoRA For all models, we employ the AdamW optimizer with a learning rate of 2e-5. We fine-tuned the models using a batch size of 8, across 3 epochs, and set the maximum sequence length to 666 tokens to ensure efficient handling of longer input sequences. The LoRA technique is utilized with a rank of 32. Regarding the model architecture, the following target modules are updated during training: q_proj, k_proj, v_proj, and o_proj. To prevent overfitting and enhance generalization, we apply a dropout rate of 0.05 during the training phase. The combination of these hyperparameters ensures efficient adaptation of the model to the specific task while minimizing the computational overhead added by the LoRA modi-

Table 3: The ablation result in Translating to English (xx→en).

Methods	de		zh		ru		cs		ind	
	BLEU	XCOMET	BLEU	XCOMET	BLEU	XCOMET	BLEU	XCOMET	BLEU	XCOMET
Ours	34.21	96.31	26.22	92.86	39.13	93.44	41.47	91.37	37.47	94.71
- $\mathcal{L}_{\text{ICE}}$	35.33	95.96	25.89	93.13	37.55	92.35	43.34	90.67	30.11	93.67
- $\mathcal{L}_{\text{output}}$	31.19	95.43	25.78	91.76	35.54	91.97	40.13	88.72	35.02	94.39

Table 4: The ablation result in Translating from English (en→xx).

Methods	de		zh		ru		cs		ind	
	BLEU	XCOMET	BLEU	XCOMET	BLEU	XCOMET	BLEU	XCOMET	BLEU	XCOMET
Ours	31.22	91.74	26.33	90.51	27.22	90.47	29.47	89.53	26.01	90.13
- $\mathcal{L}_{\text{ICE}}$	31.41	90.66	25.89	90.23	26.20	90.11	27.88	89.27	23.73	89.26
- $\mathcal{L}_{\text{output}}$	29.96	89.91	25.78	88.37	27.93	89.13	28.25	87.28	25.37	89.21

fications. We adhere to the default β value of 0.1 as suggested by Rafailov et al. (2024)

4.3 Result

We present the primary result of MT in Table 1 and Table 2. Our evaluation metrics include both statistical and neural metrics, but we place a primary emphasis on neural metrics, using statistical metrics only as a reference with a limited level of confidence. For neural metrics, we adopted the XCOMET (Rei et al., 2020) series models [2], and for statistical metrics, we used BLEU (Papineni et al., 2002). For all metrics, we calculate result with Chosen term of test data point. In translation tasks in five languages, including German, Chinese, Russian, Czech, and Indonesian, Our methods achieved an average score of 92.10 of XCOMET, CPO averaged 90.92, DPO 89.43, and SIT 89.51. The experimental results demonstrate that our method outperforms the baseline approaches across multiple evaluation metrics. A more in-depth analysis will be presented in the section 5.

For TST, we present the primary results in Figure 3. For this evaluation task, we adopted the LLM-as-Judge method[3] as the main evaluation metric, leveraging the capabilities of large language models to assess the quality, fluency, and stylistic consistency of the generated text. The LLM-as-Judge approach provides a more nuanced and context-aware evaluation, making it particularly suitable for capturing the subtle nuances in style transfer tasks (Zheng et al., 2023). We conducted a chain comparison of SFT, CPO, and our method, and additionally compared the labeled preference data in the test set using the same judge method

Figure 3: The experimental results of the TST task, evaluated through LLM-as-Judge, show that our method achieves better transfer consistency with the target style.

to verify that the evaluation approach can accurately identify the preferences. The comparison results demonstrate that our method outperforms both SFT and CPO under consistent preference evaluation. The evaluation results are presented in terms of win rate, which is calculated based on pairwise comparisons conducted by LLM-as-Judge. To ensure the reliability and robustness of the evaluation, each pairwise comparison was performed twice independently. Only when the two independent assessments yielded consistent results was a winner determined; otherwise, the outcome was considered a tie. This rigorous approach minimizes potential biases and enhances the credibility of the evaluation process.

5 Analysis

In this section, we conduct ablation studies and perturbation analysis to validate the effectiveness of our approach, along with generalization validation on more test sets and the introduction of new comprehensive metrics.

Specifically, the ablation experiments demonstrate the contribution of our hierarchical con-

[2]we use XCOMET-XL

[3]We use deepseek-chat as the judge model

trastive design. By perturbing the ICE (a critical inference component) during inference, we provide a theoretical explanation for the superiority of our method. Furthermore, we perform additional evaluations on the WMT24 test set and employ MetricX-24(Juraska et al., 2024) as an evaluation metric, further verifying the robustness and generalization capability of our approach.

Ablation Study To systematically validate the effectiveness of our proposed hierarchical contrastive learning framework, we conducted ablation experiments by progressively removing the loss terms defined in Section 3.2. The experiments were carried out on the machine translation (MT) task, with a comprehensive quality evaluation using the XCOMET metric. As shown in Table 3 and Table 4, removing the ICE-level contrastive loss leads to a performance drop, which demonstrates the importance of maintaining consistency across different contextual examples. Further removal of the output-level contrastive loss also results in a performance decline, indicating its crucial role in distinguishing subtle differences between selecting and rejecting responses.

Perturbation Analysis We also conducted perturbation experiments on the similarity ranking of the ICE used during retrieval in the test phase to investigate the impact of ICE quality on the results. Specifically, we experimented with four types of ICE: rank1 (the most similar to the source text), rank2, rank3 (less similar), and a fixed ICE that bypassed the retrieval process entirely. We performed ablation experiments on the MT (machine translation) task, and the specific results are shown in Table 5. From the results, we observed a clear trend: the performance of the model improved significantly when the ICE was more similar to the source text. Specifically, rank1 ICE, which exhibited the highest similarity to the source text, led to the most substantial performance gains, while the fixed ICE, which lacked similarity-based retrieval, resulted in the least improvement. This demonstrates that the quality and relevance of the ICE, particularly its similarity to the source text, play a critical role in enhancing model performance.

Generalization Validation We evaluated all our original baseline models on WMT24(de, cs, zh, ru) and also included a few-shot prompting on base model as an additional baseline for comparison with our proposed method. For the final evaluation, we adopted both XCOMET-XL and MetricX-24-Hybrid-Large-v2p6 as our neural metrics to ensure a more comprehensive evaluation.

Table 5: Performance trend for different rank of example

Model	ICE1	ICE2	ICE3
XCOMET	95.25	94.67	94.59
	Constant ICE	**SIT**	**CPO**
XCOMET	93.60	92.13	93.62

As shown in Table 6, our additional experiments confirm that our method maintains its advantage even when using a limited amount of training data and under degraded retrieval conditions, which aligns with the trends observed in Table 5.

Table 6: Performance for different metrics and baseline on WMT24

Metric	SFT	CPO
XCOMET	83.16	85.08
MetricX	4.88	4.06
Metric	**few-shot**	Ours
XCOMET	81.43	85.42
MetricX	5.22	3.96

6 Conclusion

This paper proposes a novel approach that combines prompt learning and fine-tuning for machine translation and text style transfer tasks. By retrieving similar contextual examples, the method enables lightweight models to better capture user preferences. Experiments demonstrate its superiority over conventional techniques in aligning with user needs. The framework can be extended to other NLP tasks, with further improvements achievable by optimizing the retrieval mechanism and contrastive learning strategy.

Limitation

While our proposed framework successfully integrates prompting and fine-tuning for efficient context-aware language transformation, it exhibits several limitations that the method depends on high-quality retrieval, faces scalability challenges with large embeddings, and struggles with balancing the diversity and relevance of retrieved examples for effective model performance.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *Preprint*, arXiv:2401.08281.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten

Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Ruili Jiang, Kehai Chen, Xuefeng Bai, Zhixuan He, Juntao Li, Muyun Yang, Tiejun Zhao, Liqiang Nie, and Min Zhang. 2024. A survey on human preference learning for large language models. *Preprint*, arXiv:2406.11191.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023a. ParroT: Translating during chat using large language models tuned with human translation and feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15009–15020, Singapore. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023b. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*, 1(10).

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. Metricx-24: The google submission to the wmt 2024 metrics shared task. *Preprint*, arXiv:2410.03983.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2024. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *Transactions of the Association for Computational Linguistics*, 12:576–592.

Moxin Li, Wenjie Wang, Fuli Feng, Yixin Cao, Jizhi Zhang, and Tat-Seng Chua. 2023. Robust prompt optimization for large language models against distribution shifts. *arXiv preprint arXiv:2305.13954*.

Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. *Preprint*, arXiv:1606.09282.

David Lopez-Paz and Marc'Aurelio Ranzato. 2022. Gradient episodic memory for continual learning. *Preprint*, arXiv:1706.08840.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

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 (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *Preprint*, arXiv:2309.11674.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *Preprint*, arXiv:2401.08417.

Changtong Zan, Liang Ding, Li Shen, Yibing Zhen, Weifeng Liu, and Dacheng Tao. 2024. Building accurate translation-tailored llms with language aware instruction tuning. *arXiv preprint arXiv:2403.14399*.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023. Tim: Teaching large language models to translate with comparison. *arXiv preprint arXiv:2307.04408*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.