

A Preference-driven Paradigm for Enhanced Translation with Large Language Models

Dawei Zhu^{1,2*} Sony Trenous¹ Xiaoyu Shen^{1†} Dietrich Klakow²
Bill Byrne¹ Eva Hasler¹

¹Amazon AGI

²Saarland University, Saarland Informatics Campus
{daweizhu, trenous, willbyrn, ehasler}@amazon.com

Abstract

Recent research has shown that large language models (LLMs) can achieve remarkable translation performance through supervised fine-tuning (SFT) using only a small amount of parallel data. However, SFT simply instructs the model to imitate the reference translations at the token level, making it vulnerable to the noise present in the references. Hence, the assistance from SFT often reaches a plateau once the LLMs have achieved a certain level of translation capability, and further increasing the size of parallel data does not provide additional benefits. To overcome this plateau associated with imitation-based SFT, we propose a preference-based approach built upon the Plackett-Luce model. The objective is to steer LLMs towards a more nuanced understanding of translation preferences from a holistic view, while also being more resilient in the absence of gold translations. We further build a dataset named MAPLE to verify the effectiveness of our approach, which includes multiple translations of varying quality for each source sentence. Extensive experiments demonstrate the superiority of our approach in “breaking the plateau” across diverse LLMs and test settings. Our in-depth analysis underscores the pivotal role of diverse translations and accurate preference scores in the success of our approach.¹

1 Introduction

The emergence of Large Language Models (LLMs) has significantly transformed the landscape of NLP, showcasing outstanding capabilities in a spectrum of NLP tasks (Brown et al., 2020; Scao et al., 2022; Chowdhery et al., 2023; Touvron et al., 2023a). This transformation extends to machine translation (MT) (OpenAI, 2023; Jiao et al., 2023b;

Hendy et al., 2023). Through supervised fine-tuning (SFT) using a small amount of parallel data, LLMs demonstrate the capability to compete with established commercial translation services such as Google Translate, particularly in high-resource languages (Jiao et al., 2023a; Zhang et al., 2023b).

Nevertheless, SFT trains the model to imitate reference translations token by token, making it vulnerable to the noise present within the data (Ott et al., 2018; Zhou et al., 2023; Touvron et al., 2023b). The noise can stem not only from the lack of attention by annotators, but also from the inherent challenge of achieving perfect translations due to the intricate interplay of language, culture, and vocabulary. As an adept translator requires not only linguistic proficiency but also a deep understanding of cultural contexts and nuances in both the source and target, it is nearly unattainable to gather extensive parallel translations of top-notch quality (Khayrallah and Koehn, 2018; Herold et al., 2022; Maillard et al., 2023). As a result, the performance enhancement achieved through SFT often quickly reaches a plateau. Further increasing the volume of parallel translations typically yields minimal additional benefits, and may instead impair the translation capabilities of LLMs (Xu et al., 2023).

To alleviate aforementioned limitation of SFT, endeavors have been made to provide LLMs with holistic assessment of contrasting examples rather than token-level imitations. Jiao et al. (2023a); Chen et al. (2023) add a flawed translation to the reference translation in the model input, encouraging the target LLM to recognize their quality difference. Zeng et al. (2023) also use a pair of translations, but they additionally optimize the LLM to favor better translations through ranking loss. Nevertheless, these works have shared limitations. First, the flawed translations are either generated by adding artificial noise to the reference translations or by other (smaller) MT systems. These imperfections in translations can be obvious and easy for

*Work done during an internship at Amazon.

†Now with Eastern Institute of Technology, Ningbo.

¹Our dataset, code, and models are available at: <https://github.com/amazon-science/preference-driven-mt>

LLM to distinguish, weakening the learning signal. Second, they only provide the relative ranking of the two translations, without quantifying the extent of their quality differences.

In this work, we present a framework based on the Plackett-Luce model to explicitly align the generation probability of the target LLM with human preferences (Plackett, 1975). Instead of using artificial noise, we collect contrasting translations generated by our target LLM, directing our optimization efforts toward “hard negative examples” (Robinson et al., 2021). Human preferences are denoted with precise scores rather than general ranking orders to teach LLMs about the nuances in different translations. LLMs are then trained to enhance their capabilities incrementally from the learnt nuances without depending solely on the existence of “gold references”, so as to effectively break the plateau associated with SFT.

We build a dataset, which we refer to as MAPLE, to facilitate preference learning. It equips each source sentence with five translations in diverse quality, scored by professional translators. By performing preference learning on MAPLE, our final MT model outperforms other MT models based on the same foundation LLM by up to 3.96 COMET score. We further show that while the intention of creating MAPLE is to enhance our target LLM, it can be reused to improve other LLMs, helping them break the performance plateau with up to 1.4M parallel data. Finally, we analyze the key factors that make preference learning effective.

Our contributions are as follows. (1) We leverage preference learning to teach LLMs a holistic notion of translation quality. Extensive experiments show that our model consistently outperforms strong baselines on two test sets across four translation directions. (2) We revisit the underlying modelling assumptions leading to the Bradley-Terry and Plackett-Luce ranking models and discuss how preference distances can be incorporated directly into the ranking models. (3) We meticulously construct an MT-oriented preference dataset, MAPLE, employing professional human translators to obtain quality scores for multiple translations corresponding to the same source sentence. We release our dataset to facilitate future MT research. (4) Our in-depth analysis reveals that high-contrast pairs and accurate quality scores are crucial in enhancing the effectiveness of our approach, providing guidance for maximizing the benefits of preference learning.

2 Related Work

LLM-based MT. One simple and effective approach to use LLMs for translation tasks is through prompting. Research in this field involves examining the impact of model sizes, the number of examples (“shots”) used, and template choices (Zhang et al., 2023a; Bawden and Yvon, 2023; Mu et al., 2023; Zhang et al., 2024). Moreover, (Ghazvininejad et al., 2023; He et al., 2023) highlight that better translations can be achieved by adding supplementary information to prompts, or engaging LLMs in related tasks prior to translation. Alternatively, another research direction seeks to fully tailor LLMs for MT tasks. Jiao et al. (2023a); Zeng et al. (2023); Chen et al. (2023); Alves et al. (2023); Zhang et al. (2023b) further train LLMs on parallel data via (parameter-efficient) fine-tuning. Xu et al. (2023) show that increasing the size of parallel data may not further improve LLM. The diminished returns from increasing data volume are likely due to data noise. Recent analyses suggest that quality trumps quantity when it comes to data effectiveness (Zhu et al., 2023; Zhou et al., 2023). Leveraging these insights, we go beyond merely fitting the reference translations. Instead, we aim to enhance the LLM’s ability to discern translations of varying quality, encouraging the generation of more precise translations while suppressing flawed outputs.

Human preference alignment. Ouyang et al. (2022) align LLMs with human intentions and values by training a reward model for preference ranking and optimizing the LLMs through the PPO algorithm (Schulman et al., 2017). However, the online reinforcement learning nature of PPO leads to considerable computational costs and is known for its sensitivity to hyperparameters (Islam et al., 2017; Huang et al., 2022). To ease the alignment, Hu et al. (2023); Dong et al. (2023) suggest offline RL algorithms where samples are pre-generated. Further research goes a step beyond by directly employing the target LLMs as reward models. Yuan et al. (2023) use a ranking loss to steer LLMs towards generating helpful responses and avoiding harmful ones. In a similar vein, Rafailov et al. (2023); Song et al. (2023); Hejna et al. (2023) use the Plackett-Luce model (Plackett, 1975) to capture human preferences in alignment. In this work, we adopt the Plackett-Luce model to MT, teaching the model to discern nuances in different translations and to prefer accurate translations.

3 Methodology

We aim to enhance LLM in MT tasks via a two-stage optimization process. We first fine-tune the target LLM with a small set of high-quality parallel data to elicit its translation ability (Section 3.1). This mirrors the supervised fine-tuning approach used in prior work, where LLMs were tailored to follow instructions (Taori et al., 2023; Zheng et al., 2023). We then use preference learning to guide the LLM to prioritize the generation of accurate translations over flawed ones (Section 3.2).

3.1 Supervised fine-tuning

We begin with optimizing our target LLM on parallel data to specialize it for translation. Let x and y denote the source and target sentence, respectively. Following Jiao et al. (2023a) we first construct a prompt by applying an instruction template \mathcal{I} to x . The instruction template is randomly sampled from an instruction pool for each training sample. The target LLM, denoted by π_θ is optimized through the log-likelihood loss:

$$\begin{aligned} \mathcal{L}_{SFT}(\pi_\theta) &= -\log \pi_\theta(x, y) \\ &= -\sum_t \log P_{\pi_\theta}(y_t | y_1, \dots, y_{t-1}, \mathcal{I}(x)) \end{aligned} \quad (1)$$

where $\pi_\theta(x, y)$ denotes the likelihood of π_θ generating output y given input x . Note that in a standard implementation, a decoder-only LLM will also predict tokens within $\mathcal{I}(x)$, we zero-out the loss on these tokens as our main goal is to teach translation, not to model the input distribution (Touvron et al., 2023b).²

3.2 Preference learning

The goal of the preference learning stage is to explicitly optimize the target LLM to favor accurate translations over erroneous ones. Formally, consider a set of translations y^1, \dots, y^L corresponding to a source sentence x . We assume that these translations are ordered by preference: $y^i \succ_x y^j$ for $i < j$. That is, translation y^i is preferred over y^j as a translation of the source sentence x . We further assume that there is some underlying reward model r^* that reflects the quality of the translations, which we cannot access but which we can approximate. Under the Plackett-Luce ranking model (Plackett,

1975), the distribution of preferences can be formulated as follows:

$$p^*(y_{\succ_x}^{1:L} | x) = \prod_{i=1}^{L-1} \frac{\exp(r^*(x, y^i))}{\sum_{j=i}^L \exp(r^*(x, y^j))} \quad (2)$$

where $y_{\succ_x}^{1:L}$ is a shorthand for the complete preference ranking $y^1 \succ_x, \dots, \succ_x y^L$. In practice, given a training set \mathcal{D} with translations equipped with a preference ranking, a reward model r_θ can be trained via maximum likelihood estimation (Cheng et al., 2010):

$$\begin{aligned} \mathcal{L}_{PL}(r_\theta) &= -\mathbb{E}_{x, y_{\succ_x}^{1:L} \in \mathcal{D}} \sum_{i=1}^{L-1} \left[r_\theta(x, y^i) - \right. \\ &\quad \left. \log \sum_{j=i}^L \exp(r_\theta(x, y^j)) \right] \end{aligned} \quad (3)$$

Following recent work (Rafailov et al., 2023; Song et al., 2023; Hejna et al., 2023), we parameterize the reward model using the target LLM π_θ and rewrite the above objective as:

$$\mathcal{L}_{PL}(\pi_\theta) = -\mathbb{E}_{x, y_{\succ_x}^{1:L} \in \mathcal{D}} \sum_{i=1}^{L-1} \log \frac{\pi_\theta(x, y^i)}{\sum_{j=i}^L \pi_\theta(x, y^j)} \quad (4)$$

where $r_\theta := \log(\pi_\theta)$. Through optimizing Equation 4, we explicitly align the LLM generation probability with the translation quality.

A caveat when optimizing Equation 4 is that the ranking information omits any measure of absolute translation quality, which may lead to inadvertent suppression of the likelihood of good translations. Consider a case where we have a pair of translations, y^1 and y^2 , which are both acceptable translations but have different word orders that causes minor difference in preference. Optimizing Equation 4 may cause the model to raise the probability of y^1 and to suppress the probability y^2 , which may damage the model.³ To address this issue, we follow Song et al. (2023) to consider the preference

²As per Ouyang et al. (2022), we use the term ‘‘SFT’’ which is interchangeably referred to as ‘‘instruction-tuning’’ or simply ‘‘fine-tuning’’ in current literature to convey the same concept.

³Cheng and Hüllermeier (2008) show that, while the preference can be learned asymptotically solely through ranking information, incorporating additional, more detailed, preference information (e.g., distance) makes the learning process more data-efficient. Table 6 presents an ablation study.

distance in \mathcal{L}_{PLD} :

$$\mathcal{L}_{PLD}(\pi_\theta) = -\mathbb{E}_{x, y_{>x}^{1:L} \in \mathcal{D}} \sum_{i=1}^{L-1} \log \frac{\pi_\theta^{d_i^i}(x, y^i)}{\sum_{j=i}^L \pi_\theta^{d_i^j}(x, y^j)}$$

where

$$\begin{aligned} d_i^j &= r^*(x, y^i) - r^*(x, y^j), \text{ for } j > i \\ d_i^i &= \max_{j>i} (d_i^j) \end{aligned} \quad (5)$$

We obtain the ground truth preference value $r^*(x, y)$ through human annotation, which will be detailed in Section 4. Finally, we combine a SFT loss calculated on the best translation y^1 with \mathcal{L}_{PLD} , making the complete loss function:

$$\mathcal{L} = \mathcal{L}_{PLD} + \beta \mathcal{L}_{SFT} \quad (6)$$

where the hyperparameter β balances the strengths of preference learning and SFT. We use **PL** as an abbreviation of our preference learning method (i.e., optimizing Equation 6) in the subsequent text.

We now provide some justification for directly incorporating preference distances into the Plackett-Luce model by studying the original derivation of the binary case ($L = 2$) (Thurstone, 1927; Mosteller, 1951; Bradley, 1953; Hamilton et al., 2023). Denote the preferences for y^i and y^j by random variables X_i and X_j such that the probability that y^i is preferred to y^j is $\pi_{ij} = P(X_i > X_j)$. Assuming that X_i and X_j follow Gumbel distributions⁴ with locations s_i and s_j and a common scale parameter γ , the difference between the two random variables $d_{ij} = X_i - X_j$ follows a logistic distribution with location $s_i - s_j$ and scale γ :

$$d_{ij} \sim \frac{1}{4\gamma} \operatorname{sech}^2\left(\frac{d_{ij} - (s_i - s_j)}{2\gamma}\right) \quad (7)$$

By defining $\pi_i = e^{s_i}$, it follows that

$$\begin{aligned} \pi_{ij} &= P(d_{ij} > 0) \\ &= \int_0^\infty \frac{1}{4\gamma} \operatorname{sech}^2\left(\frac{d_{ij} - (s_i - s_j)}{2\gamma}\right) dd_{ij} \\ &= \frac{\pi_i^{\frac{1}{\gamma}}}{\pi_i^{\frac{1}{\gamma}} + \pi_j^{\frac{1}{\gamma}}} \end{aligned} \quad (8)$$

⁴Assuming preferences arise from a large number of i.i.d. contributions, a normal distribution results in the limit if these are averaged while the Gumbel distribution results from taking their maximum (Hamilton et al., 2023).

Usually the scale parameter γ is set to 1 which yields the Bradley-Terry model (Bradley and Terry, 1952) (and Equation 13 of Bradley (1953)).

To introduce distance information for the binary preference case, we first note that $d_1^1 = d_1^2$ for $L = 2$ (from Equation 5). We then take $\gamma = \frac{1}{d_1^1}$ and $\pi_i = \pi_\theta(x, y^i)$, which yields:

$$\pi_{12} = \frac{\pi_\theta^{d_1^1}(x, y^1)}{\pi_\theta^{d_1^1}(x, y^1) + \pi_\theta^{d_1^1}(x, y^2)} \quad (9)$$

This shows that, for the binary case, preference distances based on the ground truth preferences can be incorporated exactly into the Bradley-Terry distribution by assuming that the X_1 and X_2 have Gumbel distributions with location parameters $s_i = \log \pi_\theta(x, y^i)$ and scale parameter $\gamma = \frac{1}{r^*(x, y^1) - r^*(x, y^2)}$.

We derive and discuss the more general case of Equation 5 ($L > 2$) in Appendix A.

Connections with DPO The preference learning framework investigated here shares a common origin with DPO (Rafailov et al., 2023) in the Bradley-Terry and Plackett-Luce models over rankings (Equation 2, and Equation 18 of Rafailov et al. (2023)). Here, the target LLM π_θ serves directly as the reward function ($r_\theta = \log(\pi_\theta)$), whereas the DPO reward function also includes a reference distribution π_{ref} that arises from the KL-divergence constraint term in its RL objective function. By contrast, regularization in this work is through an external SFT term (Equation 6) distinct from the reward function. We note also that the use of distance functions based on ground truth reference values brings additional information into our ranking model beyond preference order alone.

4 Human preference data collection

We build MAPLE (MACHINE translation dataset for Preference LEarning), a dataset derived from WMT20/21 test sets. It contains multiple translations per source sentence, each assigned a real-valued human preference score. MAPLE covers four translation directions: German-to-English (de→en), Chinese-to-English (zh→en), English-to-German (en→de), and English-to-Chinese (en→zh). For each direction, 1.1K source sentences are sampled from the test sets of WMT20/21. Each source sentence is associated with five translations, including one reference translation from

WMT20/21, and four translations generated by VicunaMT, our target LLM that we aim to improve through preference learning (see training details of VicunaMT in Section 5.1). Among the four translations, one is generated using beam search with a beam size of four, and three translations are obtained through nucleus sampling (Holtzman et al., 2020) with $p = 0.9$. We also build a development set containing 200 source sentences per direction sourced from News Crawl 2022. Altogether, MAPLE contains 5.2K source sentences and 26K translations with preference scores. See Appendix B.1 for more detail on the translation collecting process.

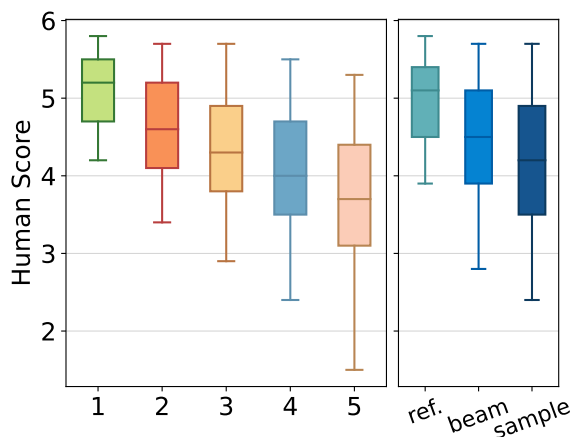


Figure 1: Human score distribution of translations by rank (left) and source (right).

Source	Zu einem großen Tuning-Treffen ist es am Samstagabend (25. Juli 2020) in Nürnberger Südstadt gekommen. (A large tuning meeting took place on Saturday evening (July 25, 2020) in Nuremberg’s Südstadt district.)
Reference translation	A large tuning meetup took place in a city south of Nürnberg this Saturday evening.
Best translation	On Saturday evening (25th July 2020) a large tuning meeting took place in Nuremberg’s south district .

Table 1: An example where the reference translation is less accurate than the best model prediction. More examples are in Appendix B.4.

Annotation guidance. We send both the source sentence and the corresponding five translations to a panel of translators for evaluation. Each example (source sentence and its translations) is assigned to two different professional translators. They observe the source and the five translations at the

same time, and assign scores between 1 (worst) and 6 (best) in increments of 0.2 using a slider. See Appendix B.2 for the full scoring rubric.

Dataset statistics. The score distribution is shown in Fig. 1. The left side shows the score distribution by rank, and we can see MAPLE contains translations that exhibit a wide range of qualities. The right side shows the score distribution by translation type, and as expected the reference is ranked highest, followed by the beam search and the nucleus samples. Nonetheless, there is considerable overlap in the score distributions, and we find that in 21% of the cases, the beam search predictions are scored higher than the reference translation. Table 1 shows an example where the reference translation contains an error.

5 Experiments

In this section, we present our MT model trained using the proposed two-stage framework and compare it with strong LLM-based MT systems.

Datasets. We train and evaluate the model on data on four translation directions: $en \leftrightarrow de$ and $en \leftrightarrow zh$. In the SFT stage, we use high-quality test sets from WMT17/18/19 for training, containing 30K parallel sentences in total across the four directions. The WMT21 test set is used for validation. In the preference learning stage, we train on MAPLE, and validation is done on the remaining data from WMT20/21 test sets which was not selected for inclusion in MAPLE. We evaluate trained models on the test sets of WMT22 (Kocmi et al., 2022) and FLORES-200 (Costa-jussà et al., 2022). Refer to Appendix C.1 for detailed data statistics.

Training. In both SFT and PL stages, we use a learning rate of $5e-6$, an effective batch size of 96, and a linear learning rate schedule with a warmup ratio of 0.1. For each training instance, one MT instruction is randomly selected from an instruction pool containing 31 MT instructions. See Appendix C.2 for the complete list of instructions.

Evaluation. At inference time, a fixed MT translation instruction is used. The maximum generation length is set to 512. We use a beam size of 4 for decoding and report BLEU (Papineni et al., 2002) and COMET (Rei et al., 2022) scores.

5.1 SFT makes good translation models

The SFT stage seeks to train a well-performing foundation MT model using parallel data. When ap-

plying SFT, we can either select a pre-trained LLM, or its instruction-tuned version. Prior research uses both types of LLMs interchangeably, leaving it unclear which is preferable in practice. To address this gap, we explore three popular families of open-access LLMs, performing SFT on both their raw (i.e., only pre-trained) and instructed-tuned versions. Specifically, we consider LLaMA-1 (Touvron et al., 2023a), Mistral (Jiang et al., 2023) and BLOOM (Scao et al., 2022); and their instruction-tuned versions, which are Vicuna (Zheng et al., 2023), Mistral-Instruct, and BLOOMZ (Muenighoff et al., 2023). The 7B parameter variants of these models are used here.

	de→en	en→de	en→zh	zh→en	Avg.
<i>WMT22</i>					
BLOOM	49.86	41.95	51.59	55.21	49.65
+SFT	77.21	69.17	84.60	78.76	77.44
BLOOMZ	74.58	62.52	83.10	78.29	74.62
+SFT	77.24	69.32	84.95	78.77	77.57
Mistral	54.18	49.08	49.10	55.47	51.96
+SFT	83.15	81.10	81.48	78.05	80.95
Mistral-Ins.	82.45	80.39	76.57	77.73	79.28
+SFT	82.68	81.23	82.49	77.73	81.03
LLaMA-1	63.29	55.29	45.80	55.17	54.89
+SFT	83.30	82.54	77.58	75.78	79.80
Vicuna	82.55	82.02	81.42	74.81	80.20
+SFT	83.55	82.79	81.27	77.39	81.25
<i>FLORES-200</i>					
BLOOM	55.03	42.36	53.82	60.25	52.86
+SFT	83.69	67.43	86.06	85.45	80.66
Mistral	42.36	32.74	33.35	42.10	37.64
+SFT	88.63	84.49	80.97	85.17	84.81
Mistral-Ins.	88.04	82.55	73.20	83.70	81.87
+SFT	88.21	83.73	82.41	84.77	84.78
LLaMA-1	58.89	52.71	42.77	49.92	51.07
+SFT	88.50	84.82	76.73	83.09	83.29
Vicuna	87.82	84.17	81.52	81.53	83.76
+SFT	88.66	86.27	80.62	84.44	85.00

Table 2: Model performance (in COMET score) before and after performing SFT on parallel data. Rows in blue indicate instruction-tuned LLMs. Best results are in **bold**. Instruction-tuned LLMs yield high COMET scores even without SFT. Raw LLMs benefit the most from SFT. Vicuna performs the best on average on both test sets. We exclude BLOOMZ on FLORES-200 as it is a part of BLOOMZ’s training data. Performance measured by BLEU score is reported in Appendix D.

Results. Table 2 presents the results before and after SFT. It can be seen that LLMs without instruction-tuning, e.g., BLOOM, perform poorly; we observe that they tend to overgenerate and re-

peat tokens in the source sentences.⁵ In contrast, instruction-tuned models work out-of-the-box and exhibit decent performance. It can be also observed that SFT dramatically boosts the performance of raw LLMs, and slightly benefits instruction-tuned LLMs. For BLOOM and Mistral, the performance gap between raw and instruction-tuned models is mostly lost after SFT. An interesting case is Vicuna, where there is a considerable improvement on en↔zh over its base model LLaMA-1. This implies that instruction-tuned LLMs may serve as a better base model for SFT. In addition, different LLMs excel in diverse translation directions and their instruction-tuned versions do not deviate from this pattern. For example, both BLOOM and BLOOMZ perform quite well on en→zh, but have a deficiency in en→de. For LLaMA-based models, the opposite holds. This could be due to the fact that German and Chinese are not included (at least, not intentionally) in BLOOM’s and LLaMA’s pre-training corpora, respectively.

The Vicuna+SFT model has the best overall performance and so we select it as our target LLM to be improved through preference learning. We call this model **VicunaMT**. The generated translations in the MAPLE dataset are produced by this model.

5.2 Refining through preference learning

Baselines. We continue training our VicunaMT model on MAPLE through preference learning and compare it with the following competitive systems from recent work: (1) **ParroT** (Jiao et al., 2023a) adds a “Hint” field to the model input, prompting the model to generate both correct and incorrect translations. At inference time, the “correct” version of the translations is used for evaluation. (2) **TIM** (Zeng et al., 2023) incorporates standard SFT with a ranking loss computed on a pair of correct and incorrect translations. (3) **SWIE** (Chen et al., 2023) proposes to attach an instruction adapter to enhance LLMs’ long-term attention for better translation. (4) **ALMA** (Xu et al., 2023) first continues pre-training the LLM on monolingual data, followed by performing SFT on parallel data. Furthermore, as the preference learning stage introduces additional data, a performance gain could be trivial by exposing the model with more samples. To establish a fair comparison, we design two additional

⁵Overgeneration is also noticed in (Bawden and Yvon, 2023), while it can be partially alleviated by prompt engineering and text post-processing (Srivastava et al., 2023), enhancing LLMs’ zero-shot performance is not our primary focus.

System	WMT22					FLORES-200				
	de→en	en→de	en→zh	zh→en	Avg.	de→en	en→de	en→zh	zh→en	Avg.
<i>Commercial LLMs & LLaMA-2-7B based MT systems</i>										
ChatGPT _(3.5-turbo-0613)	85.38	86.92	87.00	82.42	85.43	89.58	88.68	88.56	86.91	88.02
GPT-4 _(gpt-4-0613)	85.57	87.36	87.29	82.88	85.78	89.66	88.89	88.91	87.25	88.68
ALMA-7B _(LLaMA-2)	83.98	85.59	85.05	79.73	83.59	⊗	⊗	⊗	⊗	⊗
<i>BLOOMZ-mt-7B based LLMs</i>										
ParroT _(BLOOMZ-mt)	78.00	73.60	83.50	79.00	78.53	*	*	*	*	*
TIM _(BLOOMZ-mt)	77.65	74.16	84.89	79.50	79.05	*	*	*	*	*
SWIE _(BLOOMZ-mt)	78.80	75.17	84.53	79.15	79.41	*	*	*	*	*
<i>LLaMA-1-7B based LLMs</i>										
ParroT _(LLaMA-1)	82.40	81.60	80.30	75.90	80.05	88.40	84.60	81.20	83.40	84.40
TIM _(LLaMA-1)	82.80	82.32	80.03	75.46	80.15	88.08	85.00	80.93	83.18	84.30
SWIE _(LLaMA-1)	82.97	81.89	80.14	76.14	80.29	88.39	85.21	81.14	83.50	84.56
VicunaMT _(LLaMA-1)	83.55	82.79	81.27	77.39	81.25	88.66	86.27	80.62	84.44	85.00
+ REF	83.88	83.37	82.86	78.19	82.07	88.48	86.11	83.35	84.54	85.62
+ BEST	83.61	83.08	83.20	78.35	82.06	88.67	85.87	84.02	84.55	85.78
+ PL	84.23	84.43	84.26	79.07	83.00	88.83	86.73	84.88	84.76	86.30

Table 3: Model performance in COMET scores. Best results of LLaMA-1 based models are in **bold**. Applying preference learning (+PL) on top of our VicunaMT model consistently leads to improvements in all cases, achieving the highest average performance among all BLOOM and LLaMA-1 based MT models. Performance in BLEU scores is reported in Appendix E. ⊗: LLaMA-2 based models were not evaluated due to license constraints. WMT22 results are extracted from the original paper. *: BLOOMZ-family models use FLORES-200 for training.

baselines: (5) **REF** trains VicunaMT with the reference translations in MAPLE. (6) **BEST** trains VicunaMT with the translations that are scored highest by our annotators. See Table 1 for an example comparison of the reference and best translations. All aforementioned baselines are performed on 7B LLMs (based either on BLOOM-7B or LLaMA-7B). Finally, we also compare our model against commercial LLMs, including ChatGPT and GPT-4.

Results. We report the MT performance of various baselines in Table 3. It can be seen that our VicunaMT model performs well compared to recent MT systems. PL further increases the performance advantage. Our final model, VicunaMT+PL, achieves the highest average performance (83 on WMT22 and 86.3 on FLORES-200), consistently outperforming all LLaMA-1 based models across all directions, with the largest improvement being a 3.96 increase in COMET score. (en→zh on WMT22). Notably, LLaMA-based models are originally much weaker in directions involving Chinese. Through preference learning, VicunaMT reaches a translation performance close to BLOOM-based LLMs. This becomes practically significant when the goal is to deploy a single LLM to handle multiple translation directions. Also, the PL model scores higher than VicunaMT models

fine-tuned on the reference and best translations, indicating that the performance gain does not just come from having more data. Compared to the ALMA model, which is based on LLaMA-2 (Touvron et al., 2023b), a widely recognized superior open access LLM, our model demonstrates only a slight deficit of 0.59 COMET scores. Note that our strategy is orthogonal to ALMA’s approach, which leverages monolingual data. Combining both strategies should lead to even better performance.

We supplement our assessment with a human evaluation, contrasting VicunaMT+PL with SFT-only Vicuna variations including VicunaMT and VicunaMT+REF, as illustrated in Table 4. The human evaluation confirms the trend observed with automatic metrics, where PL substantially outperforms SFT-only variations.

6 Analysis

Reuse of preference data. MAPLE contains the translations generated by VicunaMT, which is also the target LLM we aim to improve. There would be additional value if this data could be reused to improve other LLMs. To investigate this, we train both Mistral-Instruct and BLOOMZ on MAPLE using PL. As shown in Table 5, PL improves both models, suggesting that the MAPLE is not limited for use with VicunaMT and can be reused for im-

	de→en	en→de	en→zh	zh→en
	VicunaMT+PL vs.			
VicunaMT	+3.7%	+4.4%	+5.6%	+5.7%
VicunaMT+REF	+3.7%	+2.5%	+5.0%	+3.5%

Table 4: Relative improvements of VicunaMT+PL over SFT-only models (VicunaMT and VicunaMT+REF), assessed through human evaluation on the WMT22 test set, employing the same scoring criteria as those specified in MAPLE. A two-sided t-test was conducted, with 95% confidence intervals noted as $\pm 1.7\%$. Positive values indicate the improvement achieved by VicunaMT+PL compared to the other models.

proving other LLMs.

	WMT22				Avg.
	de→en	en→de	en→zh	zh→en	
BLOOMZ [†]	77.24	69.32	84.95	78.77	77.57
+REF	77.41	68.47	84.76	79.50	77.53
+BEST	77.48	68.64	85.15	79.59	77.72
+PL	77.83	69.84	85.36	80.67	78.42
Mistral-Ins. [†]	82.68	81.23	82.49	77.73	81.03
+REF	83.06	82.63	83.39	78.07	81.79
+BEST	82.98	81.84	83.34	78.33	81.62
+PL	83.35	82.94	84.71	79.25	82.56

Table 5: Model performance in COMET scores. Best results are in **bold**. MAPLE can be reused to improve BLOOMZ and Mistral-Instruct. See results on FLORES-200 and in BLEU scores in Appendix F.1: SFT stage has already been applied to these models.

Limited gains with additional parallel data. Section 5.2 shows that the MAPLE dataset, which contains 4.4K preference examples, can be more valuable than an equivalent amount of parallel data with either the reference or the best translations. A natural follow-up question is whether adding more parallel data can close the gap. To answer this question, we collect more data by concatenating WMT20, WMT21 test data with News Commentary v16, making 1.4M parallel sentences in total.⁶ We fine-tune VicunaMT and Mistral-InstructMT (i.e., Mistral-Instruct after SFT stage) on different proportions of this data and plot the performance curve in Figure 2. In both cases, similar to observations in (Xu et al., 2023), adding more parallel data does not always improve these models and they never attain the performance level reached by using PL with MAPLE.

⁶We select News Commentary for its high-quality, domain-matching parallel data to WMT test data. WMT20/21 are included as MAPLE is built on a subset from them.

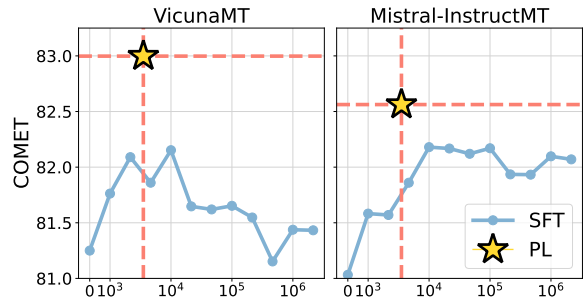


Figure 2: Performance comparison between PL using 4.4K examples from MAPLE and SFT, employing up to 1.4M parallel data. Evaluation is done on WMT22, and COMET scores are averaged across four translation directions. Performing SFT on more parallel data does not always lead to performance gain. PL consistently outperforms SFT in all cases.

Diverse translations help more. By default, we perform PL using all five translations provided by MAPLE. We now study the relation between the final model performance and the number of preference translations used. We select $K = \{2, 3, 4\}$ translations and rerun the PL algorithm on VicunaMT and Mistral-InstructMT. We explore two selection modes for selecting K translations. Given five translations sorted by human preference scores in descending order, the *forward* mode selects the first K translations (i.e., the best K), while the *reverse* mode select the first and last $K - 1$ translations. We compare both modes varying K and present the results in Figure 3. There is a clear disparity in performance with these two selection modes. The reverse mode consistently outperforms the forward mode given the same number of translations, with a larger advantage in low-resource cases, such as when $K = 2$. This is intuitive since the reverse mode always includes the highest- and lowest-scored translations and thus, PL may have a better chance to see “hard negatives” which have low human preference score but high generation probability. The general trend shows that including more preference samples is better, and using all available samples yields the best performance.

Distance information is crucial. Our framework considers the distance information in preference scores (Equation 5). We now investigate if this information can be replaced by simply using the ranking information. That is, we set $d_i^j = 1$ for all translations and rerun the PL algorithm. Table 6 shows that when the distance information is available, excluding the SFT loss does not harm

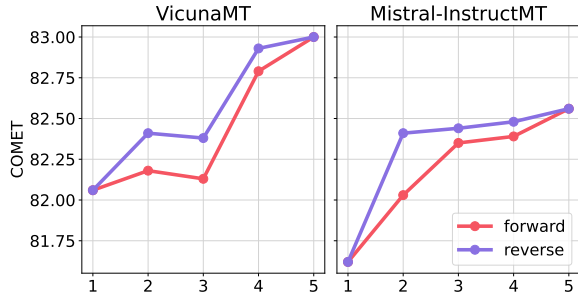


Figure 3: Model performance varying number of translations (K) per source sentence. Evaluation conducted on WMT22 and COMET scores averaged across four translation directions are reported. Reverse mode selects more diverse translations and achieves better performance, especially when fewer translations are provided.

	VicunaMT	Mistral-InstructMT
SFT stage	81.25	81.03
PL stage	83.00	82.56
w/o \mathcal{L}_{SFT}	83.00	82.54
w/o distance	82.22	81.92
w/o $\mathcal{L}_{SFT}/\text{dist.}$	74.65	60.70
\mathcal{L}_{SFT} only	82.07	81.79

Table 6: Ablation study. PL is less sensitive to \mathcal{L}_{SFT} than the distance information. Disabling both factors leads to substantial model degradation.

the performance much. In fact, we achieve the best performance when setting $\beta = 0$ for VicunaMT. However, when the distance information is withheld, we see a clear degradation in performance. We find that a larger β value is required when relying only on the ranking information, but this makes the PL algorithm closer to SFT. As a result, when only the ranking information is provided, VicunaMT performs similarly to the \mathcal{L}_{SFT} only baseline. Finally, disabling both \mathcal{L}_{SFT} and distance cause a large performance drop.

Better Model Calibration. In our preference learning framework, the model learns both translation and the ability to differentiate between different translation quality. We analyze if PL has successfully transferred human preference to the model. Using the held-out set of MAPLE, we examine the sentence-level correlation between the scores assigned by the human annotators and model generation probability. Specifically, we compute the average Pearson and Kendall’s tau correlation varying the number of preference samples (reverse mode). The results are presented in Figure 4. Com-

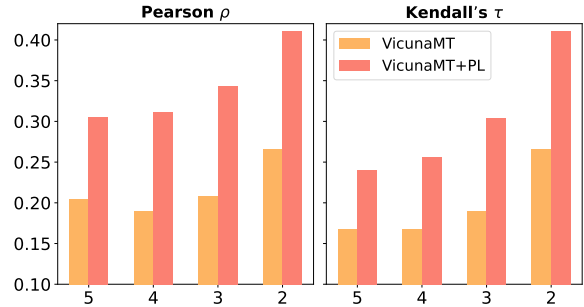


Figure 4: Sentence-level correlation between model generation probability and human preference scores varying number of translations (K). PL helps the model align better with human judgement.

pared to the SFT baseline, VicunaMT, PL substantially improves the correlation, suggesting that our final model aligns better with human preference.

7 Conclusion

We present a preference learning framework to break the performance plateau faced when performing SFT. It enhances the translation capabilities of LLMs by motivating them to differentiate the nuances in different translations. To support this framework, we have carefully curated a preference dataset, named MAPLE, featuring translations of varying quality, each scored by professional translators. Extensive experiments, including human evaluations, confirm the effectiveness of this framework. In addition, we demonstrate that MAPLE can be reused to enhance other LLMs, further bolstering its practical usability. Future research could consider extending our framework into an iterative process for continuous improvement of LLMs’ translation capabilities.

Limitations

This work demonstrates that preference learning can effectively improve LLMs’ translation capabilities. However, our study is not exhaustive and has the following limitations.

Low-resource languages. This work centers on translation directions involving high-resource languages where LLMs already exhibit proficiency. The extent to which translations for low-resource languages can leverage our framework remains uncertain. Nevertheless, it is important to emphasize that our framework is language- and model-agnostic, implying its potential applicability to low-resource languages. We leave the investigation into

this aspect to future work.

Annotation cost. Assigning preference scores for five translations per sentence can be costly, which may hinder the scaling up of the preference dataset. However, as we show in Section 6, a preference learning dataset such as MAPLE offers distinct learning signal that is not covered by massive parallel data. In addition, we highlight that the preference data can be reused to benefit other LLMs. Thus, the collected data is a valuable and reusable resource, rather than a one-time expense.

Noise in human judgement. Inevitably, human preference scores can be subjective, and annotators may not always agree. Additionally, there is a risk of annotators finding shortcuts in the annotation process (Ipeirotis et al., 2010; Hosking et al., 2023). To reduce the potential annotation mistakes, we average the scores of two translators for each sample and all translators we employ are experienced in translation assessment.

Acknowledgements

Bill Byrne holds concurrent appointments as an Amazon Scholar and as Professor of Information Engineering at the University of Cambridge. This paper describes work performed at Amazon.

We thank Felix Hieber, Lei Sun, and Tobias Domhan for their thoughtful advice, in-depth discussions, and implementation support. We would also like to thank our anonymous reviewers for their constructive feedback.

References

- Duarte M. Alves, Nuno Miguel Guerreiro, João Alves, José Pombal, Ricardo Rei, José G. C. de Souza, Pierre Colombo, and André F. T. Martins. 2023. [Steering large language models for machine translation with finetuning and in-context learning](#). *CoRR*, abs/2310.13448.
- Rachel Bawden and François Yvon. 2023. [Investigating the translation performance of a large multilingual language model: the case of BLOOM](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023, Tampere, Finland, 12-15 June 2023*, pages 157–170. European Association for Machine Translation.
- Ralph Allan Bradley. 1953. [Some statistical methods in taste testing and quality evaluation](#). *Biometrics*, 9(1):22–38.
- Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Yijie Chen, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2023. Improving translation faithfulness of large language models via augmenting instructions. *arXiv preprint arXiv:2308.12674*.
- Weiwei Cheng, Krzysztof Dembczynski, and Eyke Hüllermeier. 2010. [Label ranking methods based on the plackett-luce model](#). In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 215–222. Omnipress.
- Weiwei Cheng and Eyke Hüllermeier. 2008. Learning similarity functions from qualitative feedback. In *LNAI 5239 Advances in Case-Based Reasoning: The 9th European Conference on Case-Based Reasoning (ECCBR-08)*, pages 129–134, Trier, Germany. Springer.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pilla, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. [Palm: Scaling language modeling with pathways](#). *Journal of Machine Learning Research*, 24(240):1–113.
- 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, Loïc Barraud, 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](#). *CoRR*, abs/2207.04672.
- Yi Dong, Zhilin Wang, Makes Narsimhan Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. 2023. [Steerlm: Attribute conditioned SFT as an \(user-steerable\) alternative to RLHF](#). *CoRR*, abs/2310.05344.
- Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. [Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. [Dictionary-based phrase-level prompting of large language models for machine translation](#). *CoRR*, abs/2302.07856.
- Ian Hamilton, Nick Tawn, and David Firth. 2023. [The many routes to the ubiquitous Bradley-Terry model](#). (arXiv:2312.13619).
- Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. [Exploring human-like translation strategy with large language models](#). *CoRR*, abs/2305.04118.
- Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W. Bradley Knox, and Dorsa Sadigh. 2023. [Contrastive preference learning: Learning from human feedback without RL](#). *CoRR*, abs/2310.13639.
- 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*.
- Christian Herold, Jan Rosendahl, Joris Vanvinckenroye, and Hermann Ney. 2022. [Detecting various types of noise for neural machine translation](#). In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2542–2551. Association for Computational Linguistics.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text degeneration](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Tom Hosking, Phil Blunsom, and Max Bartolo. 2023. [Human feedback is not gold standard](#). *CoRR*, abs/2309.16349.
- Jian Hu, Li Tao, June Yang, and Chandler Zhou. 2023. [Aligning language models with offline reinforcement learning from human feedback](#). *CoRR*, abs/2308.12050.
- Shengyi Huang, Rousslan Fernand Julien Dossa, Antonin Raffin, Anssi Kanervisto, and Weixun Wang. 2022. [The 37 implementation details of proximal policy optimization](#). In *ICLR Blog Track*. <https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/>.
- Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. 2010. [Quality management on Amazon Mechanical Turk](#). In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 64–67.
- Riashat Islam, Peter Henderson, Maziar Gomrokchi, and Doina Precup. 2017. [Reproducibility of benchmarked deep reinforcement learning tasks for continuous control](#). *CoRR*, abs/1708.04133.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *CoRR*, abs/2310.06825.
- Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023a. [Parrot: Translating during chat using large language models](#). *arXiv preprint arXiv:2304.02426*.
- Wenxiang Jiao, Wenxuan Wang, JT Huang, Xing Wang, and ZP Tu. 2023b. [Is ChatGPT a good translator? yes with GPT-4 as the engine](#). *arXiv preprint arXiv:2301.08745*.
- Huda Khayrallah and Philipp Koehn. 2018. [On the impact of various types of noise on neural machine translation](#). In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, NMT@ACL 2018, Melbourne, Australia, July 20, 2018*, pages 74–83. Association for Computational Linguistics.
- Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. [Findings of the 2022 conference on machine translation \(WMT22\)](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- R. Duncan Luce. 1959. *Individual choice behaviour*. John Wiley.

- Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. [Small data, big impact: Leveraging minimal data for effective machine translation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.
- Lucas Maystre and Matthias Grossglauser. 2015. [Fast and accurate inference of Plackett-Luce models](#). In *Advances in Neural Information Processing Systems*, volume 28.
- Frederick Mosteller. 1951. [Remarks on the method of paired comparisons: I. the least squares solution assuming equal standard deviations and equal correlations](#). *Psychometrika*, 16(1):3–9.
- Yongyu Mu, Abudurexiti Reheman, Zhiqian Cao, Yuchun Fan, Bei Li, Yinqiao Li, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2023. [Augmenting large language model translators via translation memories](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10287–10299, Toronto, Canada. Association for Computational Linguistics.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. [Crosslingual generalization through multitask finetuning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15991–16111. Association for Computational Linguistics.
- OpenAI. 2023. [GPT-4 technical report](#). *CoRR*, abs/2303.08774.
- Myle Ott, Michael Auli, David Grangier, and Marc’Aurelio Ranzato. 2018. [Analyzing uncertainty in neural machine translation](#). In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3953–3962. PMLR.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- 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.
- Robin L Plackett. 1975. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202.
- 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](#). *CoRR*, abs/2305.18290.
- Ricardo Rei, José G. C. de Souza, Duarte M. Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luísa Coheur, and André F. T. Martins. 2022. [COMET-22: Unbabel-IST 2022 submission for the metrics shared task](#). In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 578–585. Association for Computational Linguistics.
- Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. [Contrastive learning with hard negative samples](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. [BLOOM: A 176b-parameter open-access multilingual language model](#). *CoRR*, abs/2211.05100.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#). *CoRR*, abs/1707.06347.
- Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2023. [Preference ranking optimization for human alignment](#). *CoRR*, abs/2306.17492.
- Saurabh Srivastava, Chengyue Huang, Weiguo Fan, and Ziyu Yao. 2023. [Instance needs more care: Rewriting](#)

- prompts for instances yields better zero-shot performance. *CoRR*, abs/2310.02107.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.
- Louis L Thurstone. 1927. Psychophysical analysis. *The American journal of psychology*, 38(3):368–389.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. **LLaMA: Open and efficient foundation language models**. *CoRR*, abs/2302.13971.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. **Llama 2: Open foundation and fine-tuned chat models**. *CoRR*, abs/2307.09288.
- Kenneth Train. 2003. *Discrete Choice Method With Simulation*. Cambridge University Press.
- 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**. *CoRR*, abs/2309.11674.
- Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. **RRHF: rank responses to align language models with human feedback without tears**. *CoRR*, abs/2304.05302.
- Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023. Tim: Teaching large language models to translate with comparison. *arXiv preprint arXiv:2307.04408*.
- Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. **Prompting large language model for machine translation: A case study**. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.
- Miaoran Zhang, Vagrant Gautam, Mingyang Wang, Jesujoba O. Alabi, Xiaoyu Shen, Dietrich Klakow, and Marius Mosbach. 2024. **The impact of demonstrations on multilingual in-context learning: A multidimensional analysis**.
- Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023b. **Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA**. In *Proceedings of the Eighth Conference on Machine Translation*, pages 466–479, Singapore. Association for Computational Linguistics.
- 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**. *CoRR*, abs/2306.05685.
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinu Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.
- Dawei Zhu, Xiaoyu Shen, Marius Mosbach, Andreas Stephan, and Dietrich Klakow. 2023. **Weaker than you think: A critical look at weakly supervised learning**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14229–14253, Toronto, Canada. Association for Computational Linguistics.

A Incorporating multiple preferences with distance information

In Section 3.2, we demonstrated how the distance information of two preferences can be integrated into preference modeling, as illustrated in Equation 9. A similar analysis can be done for the Plackett-Luce ranking model to incorporate distance metrics across multiple preferences. Specifically, we model the probability of a particular ordering X_1, \dots, X_L as follows:

$$\begin{aligned} P(X_1 \geq X_2 \cdots \geq X_L) \\ = \prod_{i=1}^{L-1} P_i(X_i > X_j, \forall j > i) \end{aligned}$$

For each distribution P_i , let $X_j = s_j + \varepsilon_j$ for $j \geq i$, with $\varepsilon_j \sim$ standard Gumbel and independent so that (following Train (2003), Section 3)

$$P_i(X_i > X_j, \forall j > i) = \frac{e^{s_i}}{\sum_{j \geq i} e^{s_j}}$$

This ranking can be interpreted as a sequence of $L - 1$ independent choices: choose the first item, then choose the second among the remaining alternatives, etc. (Maystre and Grossglauser, 2015). It is usually assumed that each independent choice is made by the same judge whose underlying preferences do not change. If we assume $s_j = \log \pi_\theta(x, y^j)$ for this judge then Equation 4 results.

Suppose instead that, rather than a single judge, a succession of $L - 1$ different judges each make one of the sequence of independent choices. The distributions P_i should change to reflect the changing preferences of the judges. In particular, if we introduce the preference distances d_i^j for the i^{th} judge, then we obtain Equation 5 if for each P_i the location parameters are set to $s_j = d_i^j \log \pi_\theta(x, y^j)$ for $j \geq i$. We find that this modified version of the Plackett-Luce model can work well in practice although we note that these modifications may violate Luce’s Choice Axiom (Luce, 1959; Hamilton et al., 2023).

Consider the case of $L = 3$. The Choice Axiom requires the odds of choosing X_2 over X_3 are independent of the presence of X_1 as an option, i.e. that the odds should not depend on whether this is a choice for the first or the second position

$$\frac{P_1(X_2 > X_j, j = 1, 3)}{P_1(X_3 > X_j, j = 1, 2)} = \frac{P_2(X_2 > X_3)}{P_2(X_3 > X_2)}$$

With the location parameters from above, the Choice Axiom requires

$$\frac{\pi_\theta(x, y^2)^{d_1^2}}{\pi_\theta(x, y^3)^{d_1^3}} = \frac{\pi_\theta(x, y^2)^{d_2^2}}{\pi_\theta(x, y^3)^{d_2^3}}$$

or that $\pi_\theta(x, y^2)^{(d_1^2 - d_2^2)} = \pi_\theta(x, y^3)^{(d_1^3 - d_2^3)}$. This holds for the default setting, $d_i^j = 1$, leading to Equation 4, but appears not to hold in general.

We find that the ground truth preference values can be introduced as preference distances in the binary comparison case, but that doing so in the more general case, while useful, may not satisfy the Axiom of Choice.

B More details on MAPLE

B.1 Data Construction

The source sentences in the training data of MAPLE are sampled from the test sets of WMT20 and WMT21. As mentioned in Section 4, four of the five translations are produced by VicunaMT. Considering that VicunaMT is already a strong MT system, often providing accurate translations free of mistakes, randomly selecting source sentences from WMT data could predominantly yield translations that are trivial for VicunaMT to translate, resulting in the collection of many uninformative samples with high human preference scores. To mitigate this, we prioritize source sentences that present difficulties for VicunaMT. Specifically, we use reference translations as a proxy to assess the quality of the model translations through COMET scores. We give priority to samples where the beam search output falls within a COMET score range of [75,85] and where there is a significant standard deviation in COMET scores among the four translations. Following these criteria, we select 1.1K samples for each translation direction. For the development set in MAPLE, we use monolingual data from News Crawl 2022. The sampling and selection process are the same as that of the training set, except that we do not have reference translations, instead, we use a strong commercial MT system to generate pseudo “reference” translations.

B.2 Scoring Rubric

The annotators are asked to judge the translation on a scale of 1 to 6, following the guidelines outlined in the following scoring rubric. They can assign scores in increments of 0.2, allowing for more detailed assessments, such as 1.2, 1.4, and so on.

- **Score it a 1** when the translation has nothing to do with the source; or when the translation has many unknown words; or when the translation looks like word salad.
- **Score it a 2** when you can understand why some of the words in the translation are there, but when the meaning of the source sentence is lost.
- **Score it a 3** when you understand why all or almost all the words in the translation are there and when some of the meaning of the source sentence are adequately transferred into the target language, but when the main meaning of the source sentence is lost.
- **Score it a 4** when the meaning of the source sentence is generally preserved, but when the translation is mechanical and possibly has vocabulary, grammatical, or date / numbering errors.
- **Score it a 5** when the meaning of the source sentence is fully preserved and the translation has no grammatical errors, but when the translation does not sound like the translation a native target language speaker would produce given the style and register of the source sentence.
- **Score it a 6** when the translation is perfect in every sense of the word – something a professional translator/interpreter would come up with when she understands well the context in which the source sentence was produced.

B.3 Annotation UI

The UI shows the different translations in a blind and randomized order. All translations are scored simultaneously. A screenshot of the UI is shown in Figure 5.

B.4 More Examples

Table 7 shows two additional examples in which the model’s translation scores higher than the reference translation. This once again highlights the presence of noise in parallel datasets.

C More implementation details

C.1 Dataset statistics

The data statistics are presented in Table 8. We use different validation sets in different training stages

Source	Other MPs criticised Twitter for allowing the tweets to remain visible .
Reference translation	其他议员也批评 Twitter 未能及时删贴 。 (Other MPs have <i>also</i> criticized Twitter for <i>failing to promptly delete tweets in time</i> .)
Best of five translation	其他议员批评了推特允许这些推文 仍然可见 。 (Other MPs have criticized Twitter for allowing these tweets to <i>remain visible</i> .)
Source	When he refused , the officials tipped his cart over, destroying all the eggs, the boy alleged.
Reference translation	男孩说，他 拒绝交出100卢比 后，那些官员就把他的小车掀翻，把所有鸡蛋砸碎。 (The boy said that after he <i>refused to hand over 100 rupees</i> , the officials overturned his car and smashed all the eggs.)
Best of five translation	当他拒绝时 ，官员将他的车子推倒，破坏了所有的蛋，男孩称。 (<i>When he refused</i> , officials pushed his car over and broke all the eggs, the boy said.)

Table 7: Two additional examples showing the reference translations can be less accurate than the best model prediction.

because MAPLE contains a subset of the parallel data in WMT20/21.

C.2 Prompt format

For each source sentence, we attach a MT instruction asking the LLM to generate the translation. The MT instructions come from a instruction pool based on the list of MT instructions released by (Jiao et al., 2023a)⁷. We list all 31 instructions in our instruction pool in Table 9. During training (in both SFT and PL stages), an instruction is randomly sampled from the instruction pool. During evaluation, the first instruction from Table 9 is always used. In addition to instructions, instruction-tuned models like Vicuna requires specific prompt formats. Table 10 presents a depiction of the conversion process from raw data points to the final model input.

C.3 Hyper-parameter search

Hyper-parameter search is done for $\beta \in [0.0, 0.05, 0.1]$, and best values are selected according to the validation loss.

⁷<https://github.com/wxjiao/Parrot>

Figure 5: User interface of translation assessment.

	Training stage	Data source	Number of samples			
			de→en	en→de	en→zh	zh→en
Training	SFT stage	WMT17	3004	3004	2001	2001
		WMT18	2998	2998	3981	3981
		WMT19	2000	1997	1997	2000
	PL stage	MAPLE	1100	1100	1100	1100
Validation	SFT stage	WMT21	1000	1002	1002	1948
	PL stage	WMT20 & 21*	500	500	500	500
Test	-	WMT22	1984	2037	2037	1875
		FLORES-200	1012	1012	1012	1012
Preference testing	-	MAPLE-dev	217	195	208	180

Table 8: Datasets used for training, validation and testing. *: a subset WMT20 and WMT21 is used.

C.4 Evaluation packages

We use the `Unbabel/wmt22-comet-da` model⁸ to compute the COMET scores and `sacreBLEU`⁹ for computing BLEU scores. The signature of the `sacreBLEU` package is `nrefs:1, case:mixed, eff:no, tok:13a, smooth:exp, version:2.0.0` for all translation directions but `en→zh`, in which we use `tok:zh`.

C.5 Hardware specifications and runtime

All experiments are either run on a host with eight NVIDIA A100-40GB GPUs or with eight H100-80GB GPUs. Mixed precision with `bfloat16` is used in both SFT and PL. `DeepSpeed`¹⁰ zero-stage

⁸<https://github.com/Unbabel/COMET>

⁹<https://github.com/mjpost/sacrebleu>

¹⁰<https://github.com/microsoft/DeepSpeed>

3 is used when running PL with five preference samples. Each experiment runs no longer than 15 minutes on H100 GPUs.

D SFT results in BLEU score

We present model performance after SFT stage measured by BLEU score in Table 11. While the general trend remains consistent in comparison to the performance evaluated by COMET, there are some exceptions. For example, although `VicunaMT` still achieves the top average score on `FLORES-200`, it is outperformed by `MistralMT` (i.e., `Mistral` + SFT) on `WMT22`.

E Model comparison in BLEU score

We present model performance measured by BLEU score in Table 12. In this case, there is no clear

Instruction pool
Translate the following text from [SRC] to [TGT]:
Please provide the [TGT] translation for the following text
Convert the subsequent sentences from [SRC] into [TGT]:
Render the listed sentences in [TGT] from their original [SRC] form:
Transform the upcoming sentences from [SRC] language to [TGT] language:
Translate the given text from [SRC] to [TGT]:
Turn the following sentences from their [SRC] version to the [TGT] version:
Adapt the upcoming text from [SRC] to [TGT]:
Transpose the next sentences from the [SRC] format to the [TGT] format.
Reinterpret the ensuing text from [SRC] to [TGT] language.
Modify the forthcoming sentences, converting them from [SRC] to [TGT].
What is the meaning of these sentences when translated to [TGT]?
In the context of [TGT], what do the upcoming text signify? The text is:
How would you express the meaning of the following sentences in [TGT]?
What is the significance of the mentioned sentences in [TGT]?
In [TGT], what do the following text convey?
When translated to [TGT], what message do these sentences carry?
What is the intended meaning of the ensuing sentences in [TGT]?
How should the following sentences be comprehended in [TGT]?
In terms of [TGT], what do the next sentences imply?
Kindly furnish the [TGT] translation of the subsequent sentences.
Could you supply the [TGT] translation for the upcoming sentences?
Please offer the [TGT] rendition for the following statements.
I'd appreciate it if you could present the [TGT] translation for the following text:
Can you deliver the [TGT] translation for the mentioned sentences?
Please share the [TGT] version of the given sentences.
It would be helpful if you could provide the [TGT] translation of the ensuing sentences.
Kindly submit the [TGT] interpretation for the next sentences.
Please make available the [TGT] translation for the listed sentences.
Can you reveal the [TGT] translation of the forthcoming sentences?
Translate from [SRC] to [TGT]:

Table 9: An instruction pool containing 31 MT prompts. An instruction is randomly sampled from this pool to form a training sample. At inference time, the first instruction is always used. [SRC] and [TGT] will be replaced by the source and target sentence in the dataset, respectively.

winner. Interestingly, VicunaMT+PL attains lower BLEU scores than VicunaMT on en→de and zh→en when evaluated on WMT22. However, both COMET score and our human evaluation in Table 4 show the opposite, highlighting that BLEU scores may be less correlated to human judgement, as also noticed in (Freitag et al., 2022).

F Data reuse in BLEU score and Results on FLORES-200

We reuse MAPLE to enhance BLOOMZMT and MistralInstructMT (i.e., BLOOMZ and MistralInstruct after the SFT stage) and report model performance on WMT22 in BLEU score in Table 13. In addition, we evaluate MistralInstructMT on FLORES and present the results in Table 14.

Model	Instruction template		
Vicuna	USER:	[MT Instruction]	\nASSISTANT:\n
Mistral-Instruct	[INST]	[MT Instruction]	\n[\INST]
BLOOMZ	USER:	[MT Instruction]	\nASSISTANT:\n

(a)

Example

USER: Translate the following text from English to German: Hello, world.
 \nASSISTANT:\n Hallo, Welt.

(b)

Table 10: (a) Instruction template used for Vicuna, Mistral-Instruct, and BLOOMZ. Raw template is marked in red. BLOOMZ shares the same template as Vicuna at the SFT and PL stage. When performing BLOOMZ on zero-shot tasks, we directly use the first instruction from Table 9 without any instruction template. (b) An example that converts the raw input (marked in green) to the final input.

	de→en	en→de	en→zh	zh→en	Avg.
<i>WMT22</i>					
BLOOM	1.51	0.53	1.74	5.43	2.30
+SFT	23.73	16.15	35.15	21.64	24.17
BLOOMZ	21.59	6.79	28.72	18.54	18.91
+SFT	23.89	16.79	35.41	21.01	24.28
Mistral	4.32	2.65	4.93	7.01	4.73
+SFT	29.39	24.60	31.51	22.09	26.90
Mistral-Ins.	28.04	21.27	21.85	17.77	22.23
+SFT	28.26	24.61	31.90	20.60	26.35
LLaMA-1	6.30	4.00	0.88	3.01	3.55
+SFT	28.28	19.09	25.31	20.27	23.24
Vicuna	26.16	22.11	26.26	13.91	22.11
+SFT	29.26	25.70	29.98	20.61	26.39
<i>FLORES-200</i>					
BLOOM	3.88	1.48	7.00	3.75	4.03
+SFT	31.85	16.26	34.66	23.78	26.64
Mistral	3.58	1.37	0.16	1.06	1.54
+SFT	40.48	29.18	29.43	24.67	30.94
Mistral-Ins.	36.81	25.64	19.81	19.25	25.38
+SFT	39.16	27.79	29.77	23.10	29.96
LLaMA-1	4.08	2.80	1.73	1.60	2.55
+SFT	40.70	29.95	20.21	20.66	27.88
Vicuna	35.07	26.86	26.09	17.53	26.39
+SFT	41.90	30.63	28.52	23.34	31.10

Table 11: Model performance (in BLEU score) before and after performing SFT on parallel data. Rows in blue indicate instruction-tuned LLMs. Best results are in **bold**. Instruction-tuned LLMs perform well even without SFT. Raw LLMs benefits the most from SFT. We exclude BLOOMZ on FLORES-200 as it is a part of BLOOMZ’s training data.

System	WMT22					FLORES-200				
	de→en	en→de	en→zh	zh→en	Avg.	de→en	en→de	en→zh	zh→en	Avg.
<i>Commercial & LLaMA-2-7B based MT systems</i>										
ChatGPT _(3.5-turbo-0613)	33.13	33.56	44.59	25.63	31.62	43.06	40.07	45.69	25.57	36.55
GPT-4 _(gpt-4-0613)	33.72	34.84	42.75	26.33	34.41	43.79	41.81	46.10	27.39	39.77
ALMA-7B _(LLaMA-2)	29.49	30.31	36.48	23.52	29.95	⊗	⊗	⊗	⊗	⊗
<i>BLOOMZ-7B based LLMs</i>										
ParroT _(BLOOMZ-mt)	24.90	20.50	34.50	22.70	25.65	-*	-*	-*	-*	-*
TIM _(BLOOMZ-mt)	24.31	20.63	37.20	23.42	26.39	-*	-*	-*	-*	-*
SWIE _(BLOOMZ-mt)	25.95	21.83	36.88	23.33	27.00	-*	-*	-*	-*	-*
<i>LLaMA-1-7B based LLMs</i>										
ParroT _(LLaMA-1)	27.30	26.10	30.30	20.20	25.98	39.40	30.70	29.10	21.30	32.38
TIM _(LLaMA-1)	27.91	25.02	30.07	19.33	25.58	39.15	29.31	28.43	22.30	29.80
SWIE _(LLaMA-1)	30.48	27.10	31.08	21.19	27.47	40.20	31.41	29.07	21.59	30.57
VicunaMT _(LLaMA-1)	29.26	25.70	29.98	20.61	26.39	41.90	30.63	28.52	23.34	31.10
+ REF	31.12	24.72	30.07	20.38	26.58	39.03	29.36	28.87	22.84	30.03
+ BEST	29.44	24.93	30.91	20.39	26.16	41.29	29.34	30.07	23.48	31.05
+ PL	30.63	24.63	31.52	20.44	26.81	40.07	29.33	30.50	21.99	30.47

Table 12: Model performance in BLEU scores. Best results with LLaMA-1 based models are in **bold**. ⊗: LLaMA-2 based models were not evaluated due to license constraints. WMT22 results are extracted from the original paper. *: BLOOMZ-family models use FLORES-200 for training.

	WMT22				
	de→en	en→de	en→zh	zh→en	Avg.
BLOOMZ [†]	23.89	16.79	35.41	21.01	24.28
+REF	24.51	15.26	33.43	21.80	23.75
+BEST	23.80	16.33	34.99	21.49	24.15
+PL	24.84	16.81	36.48	23.15	25.32
Mistral-Ins. [†]	28.26	24.61	31.90	20.60	26.35
+REF	30.94	25.62	31.66	21.52	27.44
+BEST	29.76	24.30	31.12	20.83	26.50
+PL	29.32	24.78	33.00	21.76	27.47

Table 13: Model performance on WMT22 in BLEU scores. Best results are in **bold**. †: SFT stage has already been applied to these models.

	FLORES-200				
	de→en	en→de	en→zh	zh→en	Avg.
<i>COMET</i>					
Mistral-Ins. [†]	88.21	83.73	82.41	84.77	84.78
+REF	88.10	85.04	83.59	84.74	85.37
+BEST	88.41	84.55	83.46	84.94	85.34
+PL	88.56	84.98	83.86	85.34	85.67
<i>BLEU</i>					
Mistral-Ins. [†]	39.16	27.79	29.77	23.10	29.96
+REF	38.10	28.39	31.24	23.09	30.21
+BEST	39.35	28.33	30.46	22.98	30.28
+PL	39.80	27.97	31.00	23.44	30.55

Table 14: Model performance on FLORES-200 in COMET and BLEU scores. Best results are in **bold**. †: SFT stage has already been applied to these models.