

# TASTE: Teaching Large Language Models to Translate through Self-Reflection

Yutong Wang<sup>1\*</sup> Jiali Zeng<sup>2</sup> Xuebo Liu<sup>1†</sup> Fandong Meng<sup>2</sup> Jie Zhou<sup>2</sup> Min Zhang<sup>1</sup>

<sup>1</sup>Institute of Computing and Intelligence, Harbin Institute of Technology, Shenzhen, China

<sup>2</sup>Pattern Recognition Center, WeChat AI, Tencent Inc, China

wangyutong@stu.hit.edu.cn, {liuxuebo,zhangmin2021}@hit.edu.cn

{lemonzeng,fandongmeng,withtomzhou}@tencent.com

## Abstract

Large language models (LLMs) have exhibited remarkable performance in various natural language processing tasks. Techniques like instruction tuning have effectively enhanced the proficiency of LLMs in the downstream task of machine translation. However, the existing approaches fail to yield satisfactory translation outputs that match the quality of supervised neural machine translation (NMT) systems. One plausible explanation for this discrepancy is that the straightforward prompts employed in these methodologies are unable to fully exploit the acquired instruction-following capabilities. To this end, we propose the TASTE framework, which stands for translating through self-reflection. The self-reflection process includes two stages of inference. In the first stage, LLMs are instructed to generate preliminary translations and conduct self-assessments on these translations simultaneously. In the second stage, LLMs are tasked to refine these preliminary translations according to the evaluation results. The evaluation results in four language directions on the WMT22 benchmark reveal the effectiveness of our approach compared to existing methods. Our work presents a promising approach to unleash the potential of LLMs and enhance their capabilities in MT. The codes and datasets are open-sourced at <https://github.com/YutongWang1216/ReflectionLLMMT>.

## 1 Introduction

Large language models (LLMs) like GPT-4 (OpenAI, 2023) have recently demonstrated dramatic performance across a wide range of natural language processing tasks (Bubeck et al., 2023; Liang et al., 2022). Their outstanding grasp of syntactic and semantic knowledge positions them as powerful instruments for the enhancement of machine

\* Work was done when Yutong Wang was interning at Pattern Recognition Center, WeChat AI, Tencent Inc, China.

† Xuebo Liu is the corresponding author.

<b>SRC</b>	但/But 口水鸡/Poached Chicken in Chili Sauce 整盒/whole box 是/is 生肉/raw meat, 没办法/cannot 吃/eat, 是否/whether 可以/can 退款/refund?
<b>REF</b>	But the Poached Chicken in Chili Sauce is just a box of inedible raw chicken, and can I get a refund for it?
<b>Normal</b>	But the whole box of chicken is raw meat, so you can't eat it, can you refund?
<b>Stage 1</b>	But <del>the whole box is meat</del> , can't eat, can you refund \n [Bad]
<b>Stage 2</b>	But the whole box of spicy chicken is raw meat, so I can't eat it. Can I get a refund?

Table 1: An example of the TASTE approach. “Normal” denotes the output of the baseline LLM fine-tuned on a normal parallel corpus. “Stage 1” and “Stage 2” denote the outputs of the first and second inference stages of the proposed self-reflection process, respectively. The highlight denotes the quality label predicted by the LLM. Inherent translation errors generated in the first stage, such as the ~~red strikethrough~~ part, are rectified in the second inference stage.

translation, capable of producing translations of superior quality (Hendy et al., 2023; Zhang et al., 2023a; Garcia and Firat, 2022). This substantial progress represents an evolution of the paradigm in machine translation, serving as the foundation of novel translation systems characterized by enhanced quality and reliability.

Numerous studies are underway to unlock the vast potential of machine translation within LLMs. Prompt engineering aims to design effective prompt templates to guide LLMs in accomplishing specific language tasks. Some approaches attempt to integrate additional information relevant to the translation task to enhance the performance of LLMs (Ghazvininejad et al., 2023; Lu et al., 2023; He et al., 2024; Peng et al., 2023). Studies in In-Context Learning (ICL, Brown et al., 2020) seek to provide LLMs with more relevant and high-quality translation exemplars, which assists LLMs

in retrieving bilingual knowledge, facilitating the generation of translations of the highest possible quality (Vilar et al., 2023; Agrawal et al., 2023). However, assessments of LLMs reveal that, in most translation directions, their performance falls short of that exhibited by robust supervised baselines (Zhu et al., 2023). This shortfall is due to the fact that these approaches often treat the LLM machine translation task as a simple text generation task, focusing on adjusting the prompts to enhance the outcomes. However, the intrinsic features of the machine translation task, such as the need for diverse multilingual knowledge, are often overlooked.

Some studies recommend the tuning of relatively smaller LLMs for translation (Zhu et al., 2023; Xu et al., 2023). Instruction tuning of LLMs with a limited number of high-quality supervised instructions in machine translation tasks yields remarkable results in some instances (Zeng et al., 2023; Jiao et al., 2023; Zhu et al., 2023; Hendy et al., 2023). Despite these achievements, these attempts still fail to fully leverage the capacity of LLMs due to their overly straightforward inference process. Unlike supervised NMT models, LLMs generate translations through language modeling, which contains a more complicated inference process and relies more on inherent linguistic knowledge. Studies such as Chain-of-Thought (CoT) reveal that the introduction of intermediate reasoning steps in the inference process significantly increases the reasoning capabilities of language models (Wei et al., 2022b; Kojima et al., 2022).

In this paper, we introduce **TASTE**, a method that aims at improving the translation performance of LLMs by instilling the ability to self-reflect on their own outputs. Specifically, we segment the LLM translation process into two stages of inference. In the first stage, LLMs are prompted to generate preliminary translations while simultaneously making quality predictions for these translations. In the second stage, we instruct LLMs to refine these preliminary translations based on the predicted quality levels to produce final candidates. An example of the proposed process can be found in Table 1. This entire process can be regarded as a form of self-reflection, mirroring the common approach employed by humans to carry out tasks more effectively and impeccably. To establish a sufficient multitask capability for executing the entire reflective translation process, we conduct supervised fine-tuning (SFT) on LLMs using a multitask training dataset. This method demonstrates a

remarkable stimulation of the potential of LLMs, providing a novel approach to enhance the translation performance of these models.

Our contributions are summarized as follows:

- We present the **TASTE** method, which guides LLMs through a two-stage inference process, allowing them to initially generate preliminary results and subsequently refine them into improved candidates based on their self-assessment results.
- We create a multi-task training set comprising tasks that are closely aligned with the **TASTE** process to equip LLMs with the capability to execute the whole inference process.
- We find that by employing the **TASTE** method, LLMs proficiently refine their initial translation candidates, resulting in superior final outcomes, which in turn contributes to an enhancement in their translation capabilities.

## 2 Related Work

Efforts to enhance the translation performance of LLMs can be categorized into two research lines: prompt engineering and instruction tuning. Prompt Engineering aims to design proper prompt templates and introduce prior knowledge or supplementary information to support the inference process. Dictionary-based approaches incorporate control hints in the prompt from bilingual or multilingual dictionaries to deal with rare words in source sentences (Ghazvininejad et al., 2023; Lu et al., 2023). He et al. (2024) extracts translation-related knowledge, such as topics, by self-prompting to guide the translation process. Studies in ICL (Brown et al., 2020) aim to provide LLMs with more relevant and high-quality translation exemplars. This approach assists LLMs in retrieving bilingual knowledge, facilitating the generation of translations of the highest possible quality (Vilar et al., 2023; Agrawal et al., 2023).

Instruction tuning represents an efficient method to enhance the ability of LLMs to follow natural language instructions and yield outputs that align more closely with human preference in downstream zero-shot tasks (Wei et al., 2022a; Ouyang et al., 2022; Chung et al., 2024). Jiao et al. (2023) explore several translation instructions to improve the translation performance of LLMs. Zeng et al. (2023) employ examples in comparison to instruct

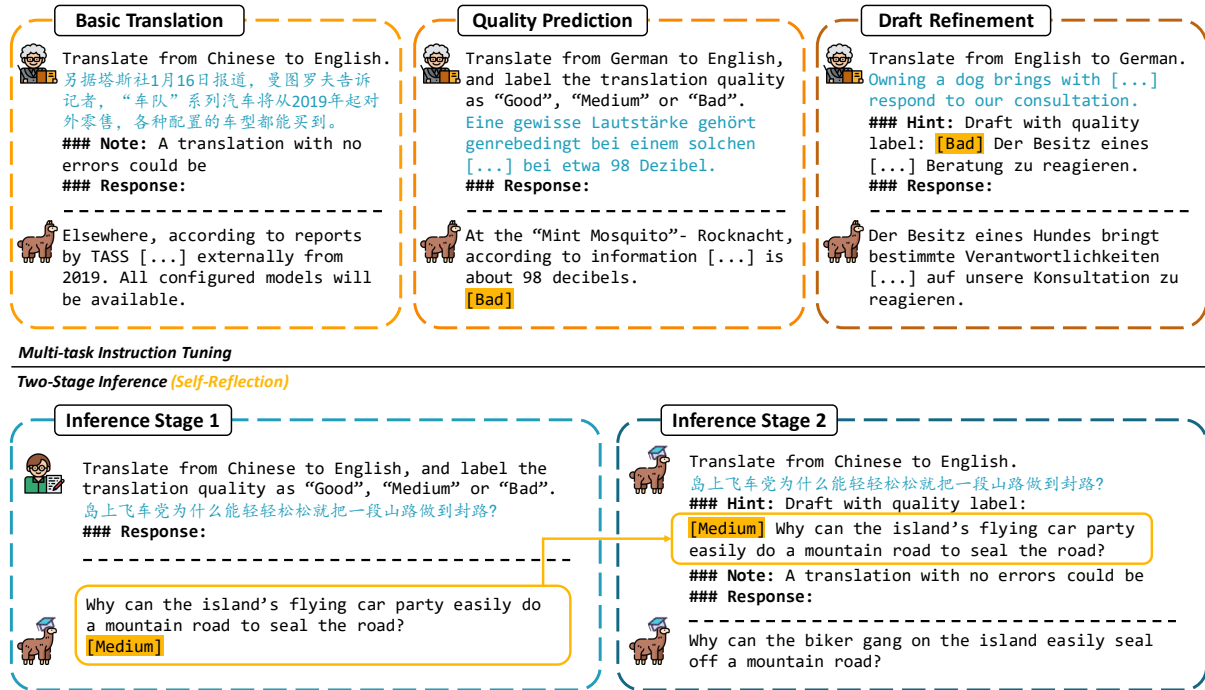


Figure 1: The framework of our proposed TASTE method.

LLMs and calculate the additional loss. Zhang et al. (2023b) enhance the multilingual language generation and instruction following capabilities of LLMs through interactive translation tasks.

Additionally, several studies proposed to facilitate a similar reflection process, utilizing confidence-guided approaches or multi-step inference, to assist the translation procedure. Lu et al. (2022) train a confidence estimation network in parallel with the backbone network to predict the confidence levels for generated translations, determining the amount of hints the model requires to produce correct translations. Xia et al. (2017) introduce a second-pass decoder to the conventional encoder-decoder structure, polishing the initial drafts and generating the final outputs. Tan et al. (2022) divide the translation process into three stages and independently apply different continuous prompts to better shift language to translation tasks. Li et al. (2023) propose a deliberate-then-generate inference framework, where LLMs are first prompted to detect error types from given candidates and then generate their final answers. Chen et al. (2023) propose to iteratively prompt LLMs to self-correct their translations. Feng et al. (2024) introduce a self-correcting inference framework for LLMs accessible via APIs, where LLMs autonomously conduct MQM self-evaluations and refine the primary candidates based on the evaluation results. Ki

and Carpuat (2024) utilize a trained fine-grained feedback model to identify defects in generated translations, subsequently directing LLMs to refine the translations based on the feedback.

Our work represents a fusion of instruction tuning and the CoT methodology. We introduce a multi-step inference translation process in imitation of the self-reflection mechanism observed in humans. The utilization of multitask training data, including Basic Translation, Quality Prediction, and Draft Refinement, substantiates not only the multi-step inference capability but also the comprehension of nuances in translation quality.

### 3 TASTE: Translate through Reflection

#### 3.1 Overall Framework

In this work, we aim to enhance the translation capabilities of LLMs by instructing them to engage in self-reflection on their translation candidates, ultimately producing carefully refined outputs. This process is achieved through a two-stage inference.

In the first stage, we ask the models to generate preliminary translations. Different from the conventional machine translation process, we also require them to predict the quality of their own outputs simultaneously. These preliminary translations are named “drafts”, and their corresponding quality predictions can take the form of either approximate labels or precise scores. This stage of inference can

be formalized into the following formula:

$$(\mathbf{y}, q) \sim P(\mathbf{y}, q \mid \mathbf{w}, \mathbf{x}; \theta) \quad (1)$$

$$\begin{aligned} & P(\mathbf{y}_{1:m}, q \mid \mathbf{w}, \mathbf{x}; \theta) \\ = & P(q \mid \mathbf{y}_{1:m}, \mathbf{w}, \mathbf{x}; \theta) P(\mathbf{y}_{1:m} \mid \mathbf{w}, \mathbf{x}; \theta) \\ = & P(q \mid \mathbf{y}_{1:m}, \mathbf{w}, \mathbf{x}; \theta) \prod_{t=1}^m P(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \mathbf{w}, \mathbf{x}; \theta) \end{aligned} \quad (2)$$

where  $\theta$  represents the parameters of the LLM,  $\mathbf{x}$  and  $\mathbf{w}$  denote the source sentence and the rest of the prompt (including the instruction), respectively. The preliminary translation  $\mathbf{y}_{1:m}$  is generated first, and the quality label (score)  $q$  is generated later according to  $\mathbf{y}_{1:m}$ . The corresponding prompts of the first inference stage are illustrated in the “Inference Stage 1” box in Figure 1.

In the second stage, we guide the models to refine their drafts based on the quality predictions. Both the drafts and quality labels/scores are formatted into the input field of the prompts for LLMs. The models proceed to make appropriate adjustments to the drafts according to the predicted label/scores, yielding the final translation candidates in a refined form. This stage of inference can be formalized into the following formula:

$$\mathbf{y}' \sim P(\mathbf{y}' \mid \mathbf{y}, q, \mathbf{w}', \mathbf{x}; \theta) \quad (3)$$

$$\begin{aligned} & P(\mathbf{y}'_{1:n} \mid \mathbf{y}, q, \mathbf{w}', \mathbf{x}; \theta) \\ = & \prod_{t=1}^n P(\mathbf{y}'_t \mid \mathbf{y}'_{1:t-1}, \mathbf{y}, q, \mathbf{w}', \mathbf{x}; \theta) \end{aligned} \quad (4)$$

where  $\mathbf{w}'$  denotes the new prompt employed in the second stage. The refined translation  $\mathbf{y}'_{1:n}$  is generated according to the preliminary translation  $\mathbf{y}$  with its predicted quality level  $q$ . The corresponding prompts of the second inference stage are shown in the “Inference Stage 2” box in Figure 1.

### 3.2 Multitask SFT

To ensure that LLMs achieve a comprehensive understanding of the task instructions, we conduct multitask SFT on the models. The multitasking approach consists of three components: **Quality Prediction**, **Basic Translation**, and **Draft Refinement**.

**Quality Prediction** In this sub-task, LLMs are tasked with generating translations and providing self-quality predictions for a given source sentence. The quality prediction task consists of two forms:

a) Text Classification (TC), entailing label predictions of “Good”, “Medium”, or “Bad”, and b) Quality Estimation (QE), involving integer score prediction ranging from 0 to 100. We utilize candidates of various qualities generated by multiple systems, along with their evaluated COMET scores, to construct fine-tuning instances. Please refer to Appendix A.1 for detailed information. The ground truth of the training data would be translations with gold quality labels/scores placed in the back.

**Basic Translation** We utilize parallel data combined with a standardized instruction to conduct fine-tuning of LLMs for multilingual translation tasks, including German  $\Leftrightarrow$  English and Chinese  $\Leftrightarrow$  English language pairs. The instruction is formulated straightforwardly as “Translate from [SRC] to [TGT]”. As shown in Figure 1, the Basic Translation instructions exhibit a high degree of similarity to their Quality Prediction counterparts, but they belong to two completely different tasks. To disambiguate instructions between these two tasks and prevent LLMs from obtaining low-quality translation knowledge, we follow Zeng et al. (2023) to append a distinguishing note “### Note: A translation with no errors could be” at the end of the Basic Translation input.

**Draft Refinement** In this sub-task, LLMs are asked to refine drafts based on quality labels/scores to produce final outputs. Given a source sentence and multiple candidates of various qualities, we designate the highest-scored output as the reference. The drafts are sampled from the remaining candidates, covering all quality levels. We incorporate a new field named “Hint” within the translation prompt. This field provides LLMs with translation drafts of the source sentence, with quality labels/scores placed in front of the drafts in the following format: “### Hint: Draft with quality label/score: [LABEL/SCORE] [Draft]”. We fill in “label” or “score” based on whether the TC or QE approach is employed. Examples of the complete prompts are shown in Table 14.

## 4 Experimental Setups

### 4.1 Data

We employ the WMT validation set to construct the training data for the Basic Translation task and utilize the MTME multi-candidate<sup>1</sup> dataset, which

<sup>1</sup><https://github.com/google-research/mt-metrics-eval>



System	Zh⇒En		En⇒Zh		De⇒En		En⇒De		Average	
	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU
WMT22 Winners	81.00	33.50	86.80	54.30	85.00	33.70	87.40	38.40	85.05	39.98
NLLB-3.3b	76.92	21.07	81.56	32.52	83.42	29.54	86.23	33.98	82.03	29.28
Backbone: LLaMA										
ParroT	75.90	20.20	80.30	30.30	82.40	27.30	81.60	26.10	80.05	25.98
Bayling	77.48	20.31	<b>84.43</b>	<b>38.19</b>	83.19	28.16	82.18	25.66	81.82	28.08
MT-Full	78.72	23.80	83.35	33.01	83.79	30.10	83.70	27.18	82.39	28.52
MT-FixEmb	79.02	24.30	83.62	33.33	84.05	30.62	83.66	27.75	82.59	29.00
TASTE										
Full-QE	79.17	24.27	83.90	34.25	83.83	30.49	83.38	27.16	82.57	29.04
Full-TC	79.31	24.23	84.00	34.51	83.92	30.17	82.95	26.74	82.55	28.91
FixEmb-QE	79.35	24.47	84.30	34.94	84.07	30.75	83.70	27.32	82.86	29.37
FixEmb-TC	<b>79.53</b>	<b>24.87</b>	84.24	34.96	<b>84.11</b>	<b>31.03</b>	<b>83.80</b>	<b>27.94</b>	<b>82.92</b>	<b>29.70</b>
Backbone: BLOOM										
ParroT	79.00	22.70	83.50	34.50	78.00	24.90	73.60	20.50	78.53	25.65
TIM	79.71	<b>24.51</b>	85.10	37.83	78.94	26.12	74.91	20.90	79.67	27.34
MT-Full	79.25	22.81	85.01	35.49	77.61	24.05	71.31	18.84	78.30	25.30
MT-FixEmb	79.84	23.43	85.20	36.68	78.27	25.07	72.06	19.41	78.84	26.15
TASTE										
Full-QE	79.36	23.15	85.05	36.84	78.42	24.87	75.41	21.18	79.56	26.51
Full-TC	79.14	23.04	84.94	36.75	78.74	24.97	75.53	21.13	79.59	26.47
FixEmb-QE	<b>80.40</b>	24.41	85.81	<b>39.31</b>	<b>79.20</b>	<b>26.28</b>	76.30	21.84	<b>80.43</b>	<b>27.96</b>
FixEmb-TC	80.28	24.20	<b>85.90</b>	39.07	78.96	26.27	<b>76.38</b>	<b>21.98</b>	80.38	27.88

Table 2: Main results of TASTE. LLaMA-2-7b and BLOOMZ-7b1-mt are chosen as the backbone model. *QE* and *TC* signify that the Quality Prediction subtask takes the form of quality estimation and text classification, respectively. The best results of each kind of backbone model are labeled using **bold font**.

contains source sentences and their candidate translations generated by multiple systems to build the training data for the Quality Prediction and Draft Refinement tasks. For Quality Prediction, candidates across various quality levels are sampled to form training instances. For Draft Refinements, the candidate with the highest COMET score is chosen as the reference, and the drafts to be refined are sampled from the other candidates covering various qualities. The data statistics and details of data building can be found in Appendix A.2.

To avoid possible data leakage in the training data, we evaluate the translation performance on the WMT22 test set (Kocmi et al., 2022), which covers domains such as news, social, e-commerce, and conversation. We present the translation results in German ⇔ English and Chinese ⇔ English directions. We report the BLEU scores by SacreBLEU (Post, 2018) and COMET scores by wmt22-comet-da (Rei et al., 2022).

## 4.2 Model Training

We employ BLOOMZ-7b1-mt<sup>2</sup> and LLaMA-2-7b<sup>3</sup> (Touvron et al., 2023) as our backbone models.

<sup>2</sup><https://huggingface.co/bigscience/bloomz-7b1-mt>

<sup>3</sup><https://huggingface.co/meta-llama/Llama-2-7b>

These models are all fine-tuned for 1 epoch with a batch size of 128. The learning rates are set to 2e-5, and the weight decay parameter is set to 0.0. The maximum text length is 768. We conducted the fine-tuning on eight NVIDIA A100 GPUs, using the Deep-Speed ZeRO stage3 for acceleration.

We employ two distinct training strategies, differing in the updated parameters:

**Full-Parameter Tuning (Full)** In this method, all the parameters in LLMs are involved in the training process. In comparison to methods that focus on training only a small set of parameters (such as Prefix Tuning and Low-Rank Adaption), full-parameter tuning is less susceptible to overfitting due to the larger parameter space. However, the main issue with this approach is excessive memory consumption and runtime demands.

**Tuning with Fixed Embedding Layer (FixEmb)** The embedding layer is pre-trained on large-scale corpus and reflects the general distribution of word embeddings. Further tuning, especially when the number of trainable parameters is limited or the training corpus is not abundant enough, will introduce disturbances into these distributions, leading to a decline in the model’s expressive capacity. To

overcome this problem, we freeze the embedding layers of LLMs and fine-tune the rest of the parameters. This assists LLMs in maintaining correctness and diversity in their expressions.

### 4.3 Baselines

The **MT-(·)** baseline models represent the LLMs trained exclusively with the Basic Translation dataset, as outlined in Table 11. This dataset contains the German  $\Leftrightarrow$  English and Chinese  $\Leftrightarrow$  English translation directions.

Additionally, we present the results of WMT22 winners, NLLB-3.3B (Costa-jussà et al., 2022), a multilingual translation model trained in over 200 languages, Bayling (Zhang et al., 2023b), ParroT (Jiao et al., 2023), and TIM (Zeng et al., 2023), LLMs fine-tuned for machine translation with BLOOM or LLaMA as the backbone models.

## 5 Results

Our main results are shown in Table 2. Almost all of our methods outperform the corresponding **MT-(·)** baseline across both metrics and all language pairs, providing evidence of the effectiveness of our approach in enhancing the translation capabilities of LLMs. When utilizing BLOOMZ-7b1-mt as the backbone model, our *FixEmb-(·)* approaches achieve favorable results, particularly in Zh  $\Leftrightarrow$  En directions, and outperform ParroT and TIM across all language pairs on COMET scores. While employing LLaMA-2-7b as the backbone model, our *FixEmb-(·)* approaches also gain remarkable results, particularly in De  $\Leftrightarrow$  En directions, and beat Bayling in all directions except En  $\Leftrightarrow$  Zh.

There is no significant difference in translation performance observed between two different quality prediction approaches, *(·)-QE* and *(·)-TC*. This suggests that both of these approaches effectively aid LLMs in grasping the quality differences between varying translations.

The models trained with fixed embedding layers consistently outperform their counterparts trained with full parameters across all language pairs and both evaluation metrics. We argue that this is because fixing embedding layers during fine-tuning effectively preserves the expressive capability of LLMs against word distribution biases within the training data. This facilitates the generalization of LLMs across the word domain, mitigating overfitting and thereby enhancing their capacity to produce robust and diverse translations.

Model	PPL	Pred.↑	P↑	R↑	F1↑
BLOOMZ	-37.10	76.84	70.1	68.2	67.6
LLaMA-2	0.00	80.33	70.5	70.1	69.8

Table 3: Evaluation results on quality prediction task in Zh  $\Rightarrow$  En direction. Precision, recall, and F1 values are calculated as weighted averages across three translation quality categories. PPL/Pred. represents Pearson’s  $r$  between the perplexity values/predicted scores and the COMET scores.

We also train a merged model that handles QE and TC approaches simultaneously, and conduct a comparison of the translation performance across models of different scales. Please refer to Appendix A.3 and A.4 for more details.

## 6 Analysis

Unless mentioned otherwise, the subsequent experiments are conducted in the *FixEmb-TC* setting.

### 6.1 How Good Are LLMs at Quality Prediction?

Quality Prediction constitutes an end-to-end process, where LLMs are instructed to predict quality labels or scores while generating translations. To validate the assertion that LLMs have genuinely acquired the capability to predict the quality of candidates, we evaluated the quality prediction outputs. For TC, we construct gold labels for the instances according to their COMET scores following the same principle mentioned in Appendix A.2 and report the precision, recall, and F1 values of the predicted labels. For QE, we assessed the Pearson’s correlation coefficient between the predicted quality scores and the gold COMET scores. Additionally, we present the Pearson’s correlation coefficient between the perplexity values (PPL) of the candidates and the COMET scores for comparison.

As shown in Table 3, for the TC approach, the models exhibit a commendable level of accuracy in assigning quality labels to their translations, as evidenced by F1 values surpassing 67.6. In the QE task, our models produce scores with a satisfactory correlation with COMET scores (the p-values are all smaller than 0.01), while the perplexity values demonstrate a relatively poor correlation with COMET scores. These statistics demonstrate that our models can make precise quality predictions for their own generated translations, providing a dependable reference for the Draft Refinement task.

We can also discover that LLaMA-2 outperforms

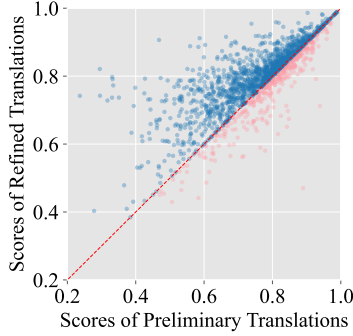


Figure 2: Comparison between the COMET scores of the preliminary and refined translations. The results are obtained by LLaMA-2-7b in Zh $\Rightarrow$ En direction.

Label	Proportion (%)	$\Delta$ COMET
Good	31.89	0.45
Medium	32.80	2.06
Bad	35.31	7.79

Table 4: Proportions of preliminary translations with different predicted quality labels and their average COMET scores increments during refinement. These results are obtained by LLaMA-2-7b in Zh  $\Rightarrow$  En direction.

BLOOMZ in terms of accuracy for both the QE and TC tasks, suggesting that LLaMA-2 possesses a more extensive bilingual knowledge base.

## 6.2 Effect of Draft Refinement

To analyze the influence of the Draft Refinement process (i.e., the second stage of inference), we perform the following two comparisons between the candidates obtained after the first and second inference stages.

**Translation Quality** We evaluate the COMET scores of the preliminary and refined translations. The results are shown in Figure 2. In the plot, each point located above the diagonal line represents an instance where a quality improvement is achieved through refinement. As the plot demonstrates, a majority of the final candidates exhibit higher quality levels than their initial counterparts.

Table 4 illustrates the proportions of preliminary translations with varying predicted quality labels and their respective average COMET score increments during the refinement process. The most significant score enhancements are observed in instances labeled as “Bad”, which constitute the largest proportion of all instances. Subsequently, “Medium” instances show a moderate improvement, while “Good” instances exhibit the least noticeable enhancement. These observations highlight the ef-

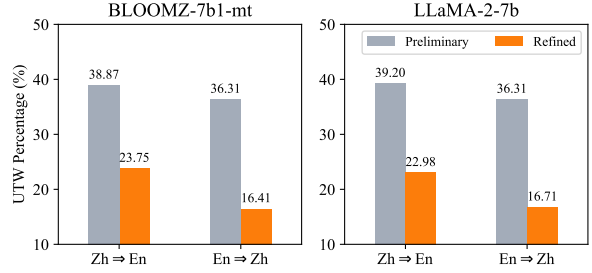


Figure 3: Comparison between the UTW percentages of the preliminary and refined translations.

Label	Edit Distance	COMET
Origin	18.98 <sup>+0.00</sup>	79.53 <sup>+0.00</sup>
Good	16.95 <sup>-2.03</sup>	79.25 <sup>-0.28</sup>
Random	18.78 <sup>-0.20</sup>	79.36 <sup>-0.17</sup>
Bad	20.20 <sup>+1.22</sup>	79.51 <sup>-0.02</sup>
Blank	18.12 <sup>-0.86</sup>	79.08 <sup>-0.45</sup>

Table 5: The edit distance between the preliminary and refined translations and the final COMET scores under different quality label configurations. “Origin” represents the configuration where predicted labels remain unmodified. “Blank” represents that quality labels are removed during refinement processes. These results are obtained by LLaMA-2-7b in Zh  $\Rightarrow$  En direction.

ficacy of the Draft Refinement process in refining the preliminary translations generated in the first inference stage as well as rectifying potential generation failures, as evidenced by instances located in the top-left region of Figure 2.

**Unaligned Translation Words (UTW)** We measure the percentages of target words that remain unaligned in a word-to-word alignment between the source sentences and translations obtained after the first and second inference stages. The alignments are extracted using the tool developed by Dou and Neubig (2021). This measurement is also used by Hendy et al. (2023) to investigate the presence of words that have no support in the source sentences. The results are shown in Figure 3. We can observe that the amount of UTW is significantly reduced during the draft refinement process, with a decrease of more than 15 percentage points. This observation suggests that the Draft Refinement process contributes to a reduction in hallucinations within the candidates, leading to a higher level of translation precision and mitigation of potential risks within the translation systems.

## 6.3 The Role of Quality Labels

To examine the impact of the predicted quality labels on the refinement process, we conduct experi-

Method	BLEU	COMET
MT-FixEmb	19.41	72.06
TASTE	21.98	76.38
w/o Basic Translation	20.00	72.12
w/o Quality Prediction	17.86	72.26
w/o Draft Refinement	19.31	72.00

Table 6: Ablation Study. We report the BLEU and COMET scores in En⇒De direction achieved by BLOOMZ-7b1-mt.

System	Zh⇒En	En⇒Zh	De⇒En	En⇒De
CoT-7b	74.50	73.79	79.63	74.37
CoT-13b	75.21	75.32	80.10	73.55
TASTE	79.53	84.24	84.11	83.80

Table 7: COMET scores gained by our approach and the CoT method.

ments by modifying these labels with the following configurations: a) All the labels are set to “Good”. b) All the labels are set to “Bad”. c) All the labels are randomly sampled among “Good”, “Medium”, and “Bad”. d) All the labels are removed from the prompts and the model is only provided with draft translations during the refinement process. Subsequently, we perform the refinement process, and calculate the average edit distances between the preliminary and refined translations as follows:

$$\begin{aligned} \bar{d} &= \frac{1}{n} \sum_{i=1}^n (1 - \text{LevRatio}_i) \\ &= \frac{1}{n} \sum_{i=1}^n \frac{\text{LevDist}_i}{\text{len}_i^1 + \text{len}_i^2} \end{aligned} \quad (5)$$

Here,  $\text{LevRatio}_i$  represents the Levenshtein distance ratio<sup>4</sup> of the  $i$ -th instance,  $\text{len}_i^1$  and  $\text{len}_i^2$  represent the lengths of two strings, respectively, and  $\text{LevDist}_i$  represents the Levenshtein distance between these strings.

We report the average edit distances and the COMET score of the refined translations in Table 5. In the cases where all the labels are set to “Good”, the edit distances between the preliminary and refined translations are relatively small. This suggests that the model tends to make fewer modifications to the preliminary translations. Conversely, when all the labels are set to “Bad”, the edit distances are relatively large, indicating that the model tends to make more modifications during refinement. Furthermore, noticeable performance decreases are observed when the labels are

<sup>4</sup><https://rapidfuzz.github.io/Levenshtein/levenshtein.html>

System	Zh⇒En	En⇒Zh	De⇒En	En⇒De
ICL-2shot	77.43	78.69	82.99	78.85
ICL-3shot	77.89	79.60	83.05	79.27
ICL-4shot	77.91	79.89	83.16	79.65
TASTE	79.53	84.24	84.11	83.80

Table 8: COMET scores gained by our approach and the ICL method.

set to “Good”, sampled randomly (i.e. Random), or removed from the prompts (i.e. Blank). These phenomena illustrate the impact of the quality labels in the refinement process, which is to assist LLMs in making reasonable adjustments based on the actual translation quality levels and generating high-quality final candidates.

## 6.4 Ablation Study

To emphasize the necessity of our multitask training set and prompt design, we conduct an ablation study. We choose BLOOMZ-7b1-mt as the backbone model and fine-tune it using various training sets with the *FixEmb-TC* method. BLEU and COMET scores in the Zh⇒En direction are reported.

Our multitask training set contains three parts: **Basic Translation**, **Quality Prediction**, and **Draft Refinement**. To demonstrate the rationality of this task combination, we remove a specific section of the training set separately, and the consequences are shown in Table 6. The performance of the model decreases when any subset of the training data is removed. This result implies that each of the sub-tasks is essential for our approach. When the Quality Prediction data is removed from the training set, the BLEU scores exhibit the most noticeable decrease. This observation suggests that the TASTE process heavily relies on the model’s ability to discern various qualities of translations.

## 6.5 Comparison with Related Methods

**TASTE vs CoT** Our approach is based on a two-stage inference, which is similar to the thought of CoT. To certify the superiority of our proposal, we perform a comparison with the CoT method. We apply the same prompts utilized in TASTE to guide a two-stage inference process with LLaMA-2-chat-7b and LLaMA-2-chat-13b, both of which undergo no fine-tuning process. The results are shown in Table 7. In many-to-English translation directions, the ICL method gains reasonable performance, yet our approach outperforms it significantly. In English-to-many directions, the



Period	De⇒En	En⇒De
Before	78.27	72.06
After	84.16	84.19

Table 9: COMET scores obtained before and after the post-editing process.

ICL method failed to generate stable outcomes by the inference chain, primarily due to a severe off-target issue that kept the models from producing translations in correct target languages.

**TASTE vs ICL** We also conduct a comparative analysis between TASTE and ICL methodologies. We employ LLaMA-2-chat-7b as the backbone model and incorporate source-target pairs randomly sampled from the Base Translation training set as examples within the prompts. The ICL experiment encompasses settings ranging from 2-shot to 4-shot scenarios. 2-shot to 4-shot settings are involved in the experiment. The results, showcased in Table 8, reveal a significant performance margin between the ICL methods and our TASTE approach.

## 6.6 TASTE as an APE Tool

In the proposed TASTE framework, the fine-tuned LLMs are employed for the evaluation and refinement of their **own** draft translations. This naturally leads to the question: *Are the fine-tuned TASTE LLMs able to evaluate base translations generated by arbitrary systems and refine them as an Automatic Post-Editing (APE) tool?*

To answer this question, we conducted an experiment utilizing TASTE as an automatic post-editing tool. Initially, we select BLOOMZ-7b in the *MT-FixEmb* baseline setting to generate base translations. Subsequently, we employ LLaMA-2-7b in the *FixEmb-TC* setting as the APE model. We concatenate the base translation behind the prompt for the first inference stage and input it into the APE model to generate the quality label. Finally, we format the base translation and quality label into the prompt for the second inference stage to obtain the refined translation.

The results of this experiment, as indicated by the COMET scores before and after APE, are detailed in Table 9. Notable quality enhancements through the APE process can be observed, and the results even outperform the TASTE LLaMA-2-7b model due to the multi-system voting mechanism. This indicates that TASTE can not only serve as

an effective inference framework for a single LLM but also as an APE tool to enhance translations generated by other translation systems.

## 7 Conclusion

We introduce TASTE, a novel approach that enables LLMs to translate through the self-reflection process. Our approach allows LLMs to initially generate a preliminary translation and autonomously assess its quality. Subsequently, the translation is refined based on the evaluation results, resulting in the final candidate. Our experiments and analyses provide evidence of the effectiveness of TASTE, as it successfully enhances the translation quality through the refinement process, consistently producing high-quality candidates across various translation directions. Furthermore, our findings underscore that LLMs possess significant potential for the translation quality prediction task. The translation process can leverage this capacity to discern different qualities among translations, leading to the generation of high-quality outcomes.

## Limitations

The performance enhancement introduced by our approach exhibits inconsistency across different translation directions. We assume that this phenomenon is caused by the inherent uneven multilingual knowledge within the model, and a more in-depth exploration of the underlying principles is warranted. Additionally, considering the two inference stages in the TASTE process, the computation cost is twice that of the conventional translation generation process. However, it’s worth noting that this extra time consumption can be mitigated through acceleration methods, such as quantification and speculative decoding.

## Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62206076), Guangdong Basic and Applied Basic Research Foundation (Grant No. 2024A1515011491), Shenzhen Science and Technology Program (Grant Nos. ZDSYS20230626091203008 and KJZD20231023094700001). Xuebo Liu was sponsored by CCF-Tencent Rhino-Bird Open Research Fund. We would like to thank the anonymous reviewers and meta-reviewer for their insightful suggestions.

## References

- Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. [In-context examples selection for machine translation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. [Sparks of artificial general intelligence: Early experiments with gpt-4](#). *ArXiv preprint*, abs/2303.12712.
- Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023. [Iterative translation refinement with large language models](#). *ArXiv preprint*, abs/2306.03856.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. [Scaling instruction-finetuned language models](#). *Journal of Machine Learning Research*, 25(70):1–53.
- Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. [No language left behind: Scaling human-centered machine translation](#). *ArXiv preprint*, abs/2207.04672.
- Zi-Yi Dou and Graham Neubig. 2021. [Word alignment by fine-tuning embeddings on parallel corpora](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.
- Zhaopeng Feng, Yan Zhang, Hao Li, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. [Improving llm-based machine translation with systematic self-correction](#). *ArXiv preprint*, abs/2402.16379.
- Xavier Garcia and Orhan Firat. 2022. [Using natural language prompts for machine translation](#). *ArXiv preprint*, abs/2202.11822.
- Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. [Dictionary-based phrase-level prompting of large language models for machine translation](#). *ArXiv preprint*, abs/2302.07856.
- Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. [Exploring human-like translation strategy with large language models](#). *Transactions of the Association for Computational Linguistics*, 12:229–246.
- 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*, abs/2302.09210.
- Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. [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.
- Dayeon Ki and Marine Carpuat. 2024. [Guiding large language models to post-edit machine translation with error annotations](#). *ArXiv preprint*, abs/2404.07851.
- Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, et al. 2022. [Findings of the 2022 conference on machine translation \(wmt22\)](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). *Advances in neural information processing systems*, 35:22199–22213.
- Bei Li, Rui Wang, Junliang Guo, Kaitao Song, Xu Tan, Hany Hassan, Arul Menezes, Tong Xiao, Jiang Bian, and JingBo Zhu. 2023. [Deliberate then generate: Enhanced prompting framework for text generation](#). *ArXiv preprint*, abs/2305.19835.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. [Holistic evaluation of language models](#). *ArXiv preprint*, abs/2211.09110.
- Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023. [Chain-of-dictionary prompting elicits translation in large language models](#). *ArXiv preprint*, abs/2305.06575.
- Yu Lu, Jiali Zeng, Jiajun Zhang, Shuangzhi Wu, and Mu Li. 2022. [Learning confidence for transformer-based neural machine translation](#). In *Proceedings*

- of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2353–2364, Dublin, Ireland. Association for Computational Linguistics.
- OpenAI. 2023. [Gpt-4 technical report](#).
- 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.
- Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. [Towards making the most of ChatGPT for machine translation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633, Singapore. Association for Computational Linguistics.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa 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)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Zhixing Tan, Xiangwen Zhang, Shuo Wang, and Yang Liu. 2022. [MSP: Multi-stage prompting for making pre-trained language models better translators](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6131–6142, Dublin, Ireland. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *ArXiv preprint*, abs/2307.09288.
- David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. [Prompting palm for translation: Assessing strategies and performance](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427.
- Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. [Finetuned language models are zero-shot learners](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. [Chain-of-thought prompting elicits reasoning in large language models](#). *Advances in neural information processing systems*, 35:24824–24837.
- Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2017. [Deliberation networks: Sequence generation beyond one-pass decoding](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 1784–1794.
- 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](#). *ArXiv preprint*, abs/2309.11674.
- Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023. [Tim: Teaching large language models to translate with comparison](#). *ArXiv preprint*, abs/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*, pages 41092–41110. PMLR.
- Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangdong Gui, Yunji Chen, Xilin Chen, et al. 2023b. [Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models](#). *ArXiv preprint*, abs/2306.10968.
- 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*, abs/2304.04675.

## A Appendix

### A.1 Quality Prediction Task Designs

The quality prediction task is designed in two forms: text classification (TC) and quality estimation (QE).

**Text Classification (TC)** We instruct LLMs to categorize translations into three classes by the instruction “Translate from [SRC] to [TGT], and label the translation quality as “Good”, “Medium” or “Bad”.” For the candidates with the top 10% COMET scores, the gold

Task	Good	Medium	Bad
Quality Prediction	30.0k	30.0k	30.0k
Draft refinement	8.0k	8.0k	4.0k

Table 10: Numbers of instances of all three quality categories in the training set for the Quality Prediction and Draft Refinement sub-task.

Task	Size	Source
Basic Translation	45.4k	WMT Dev
Draft Refinement	20.0k	MTME
Quality Prediction	90.0k	MTME

Table 11: Data sizes and sources of the training sets.

labels are assigned as “Good”, while those with the bottom 50% of COMET scores are labeled as “Bad”. Candidates falling within the remaining range are designated as “Medium”.

**Quality Estimation (QE)** We request LLMs to simultaneously predict integer quality scores ranging from 0 to 100 while generating translations by the following instruction: “Translate from [SRC] to [TGT], and score the translation quality from 0 to 100.” Here, the placeholders “[SRC]” and “[TGT]” denote the source and target language, respectively. We amplify the COMET scores by a factor of one hundred and round them to use as gold scores.

The QE task can be regarded as a more precise version of the TC task, which is perceived as more challenging for generative language models. The methodologies employed during the training and test phase will remain consistent.

## A.2 Data Details

**WMT Development Data** We use human-written validation data from previous WMT competitions as the basic MT training data to align LLMs on the MT task. Specifically, we choose the newstest2017-2021 of German  $\Leftrightarrow$  English and Chinese  $\Leftrightarrow$  English as our MT training set.

**MTME Multi-Candidate Data** This is a dataset containing source sentences and translation candidates of multiple MT systems on the WMT Metrics Shared Tasks built by Google Research. We use the candidates of newstest2019-2021 in German  $\Leftrightarrow$  English and Chinese  $\Leftrightarrow$  English directions to build training data for the Quality Prediction and Draft

Refinement task. For Quality Prediction, the inputs for the LLMs are the instructions and the source sentences, and the text generation labels are sampled candidates with their corresponding quality labels/scores attached at the end. For Draft Refinement, we choose the candidate with the highest COMET score among all candidates of one source sentence as the label for the LLMs, and the draft translation is sampled from the rest of them. The inputs for the LLMs are the instructions, the source sentences, and the drafts with their corresponding quality labels/scores attached in the front.

To enable the LLMs to have a good understanding of the translation quality, we carefully designed the proportion of the candidates with different quality levels. We classified the candidates into three categories by the COMET scores evaluated by `wmt-22-comet-da`. Candidates with the top 10% COMET scores are classified as “Good”, while those with the bottom 50% of COMET scores are classified as “Bad”. Candidates falling within the remaining range are designated as “Medium”. For the Quality Prediction and Draft Refinement training set, the numbers of instances constructed by candidates of all three quality categories are shown in Table 10.

The sizes and sources of the training data for the three tasks are represented in Table 11. Examples of the complete prompts and labels for these tasks are shown in Table 14.

## A.3 Merged Model

We also train a model that merges two types of Quality Prediction approaches, Text Classification (TC) and Quality Estimation (QE), to facilitate the TASTE self-reflection process and generate both preliminary and refined translations. Users have the flexibility to specify the approach by instructing the model in the first inference stage. If the instruction is “Translate from [SRC] to [TGT], and label the translation quality as “Good”, “Medium” or “Bad””, then the TC approach is adopted, and the model predicts quality labels for the preliminary translation. Otherwise, if the instruction is “Translate from [SRC] to [TGT], and score the translation quality from 1 to 100”, the model employs the QE approach and predicts quality scores. For training the merged model, we utilized 45.4k instances of Basic Translations, 45k instances for each of the two Quality Prediction approaches (TC and QE), and 20k instances for each of the two Draft Refinement styles



Model Size	Zh⇒En		En⇒Zh		De⇒En		En⇒De		Average	
	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU
MT-FixEmb	79.84	23.43	85.20	36.68	78.27	25.07	72.06	19.41	78.84	26.15
TASTE										
FixEmb-QE	80.40	24.41	85.81	39.31	79.20	26.28	76.30	21.84	80.43	27.96
FixEmb-TC	80.28	24.20	85.90	39.07	78.96	26.27	76.38	21.98	80.38	27.88
FixEmb-Mix-QE	79.97	24.26	85.65	38.87	78.63	26.29	75.19	21.15	79.86	27.64
FixEmb-Mix-TC	80.11	24.19	85.60	38.73	78.48	25.90	75.04	21.02	79.81	27.46

Table 12: COMET and BLEU scores achieved by the merged model. *FixEmb-Mix-QE* and *FixEmb-Mix-TC* represent the results obtained by the merged model employing QE and TC approaches during the inference process, respectively.

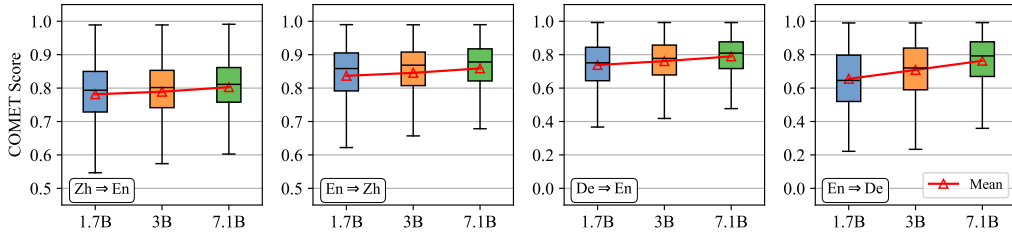


Figure 4: COMET scores obtained from BLOOMZ across different model sizes.

Model Size	Zh⇒En		En⇒Zh		De⇒En		En⇒De		Average	
	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU
1.7B	78.15	20.76	83.67	34.96	73.82	21.80	65.53	17.21	75.29	23.68
3B	78.91	22.54	84.56	36.43	76.14	23.88	70.90	19.12	77.63	25.49
7.1B	80.28	24.20	85.90	39.07	78.96	26.27	76.38	21.98	80.38	27.88

Table 13: COMET and BLEU scores achieved by BLOOMZ across different model sizes.

(TC and QE). BLOOMZ-7b1-mt is employed as the backbone model.

The results are shown in Table 12. We observe that although there is a marginal decrease in translation performance, the merged model demonstrates the capability to handle two types of quality expression approaches simultaneously and successfully conducts the normal inference process as the non-merged models.

#### A.4 Effect of Model Size

We report the COMET and BLEU scores yielded by BLOOMZ of various model sizes in Figure 4 and Table 13.

We can observe that with the increase in the number of model parameters, both the median and mean scores are consistently rising. This indicates that our proposed method is robust in terms of model parameter scaling. As mentioned in §5, LLMs depend on large amounts of parameters to memorize task-specific knowledge to perform multi-tasking. In addition, the instructions we designed for differ-

ent tasks are highly similar, which makes it more challenging but essential for LLMs to grasp different types of knowledge.

Another observation is that the distribution of scores achieved by larger models tends to be more concentrated than that obtained by smaller ones. This indicates that as the number of model parameters increases, the performance of LLMs is not only enhanced but also stabilized, which means bad cases occur less frequently, guaranteeing the lower bound of the capacity.

#### A.5 Case Study

Several cases of the translation process of TASTE are shown in Table 15. The quality labels are predicted as “Bad” while some severe lexical or syntactic problems occur in the preliminary translations. In Case 1, an ambiguous Chinese character “扫” is inaccurately translated into “sweep”, and the term “二维码” is literally translated as “two-dimensional code” instead of “QR code”. In Case 2, the preliminary translation is incomplete, omit-

Task	Prompt
Basic Translation	<p>Write a response that appropriately completes the request.\n\n###Request:\n Translate from Chinese to English.\n 一辆 1948 年的福特水星汽车穿过佐治亚州门罗小镇的一群围观者，朝着小小的摩尔滩桥隆隆奔行。 \n\n### Note: A translation with no errors could be\n\n### Response: A 1948 Ford Mercury passed through a group of onlookers in rural Monroe, Georgia, and rumbled toward the small Moore’s Ford Bridge.</p>
Quality Prediction (TC)	<p>Write a response that appropriately completes the request.\n\n###Request:\n Translate from English to German, and label the translation quality as “Good”, “Medium” or “Bad”\n 北京大兴国际机场首航开启了北京“双机场”时代。 \n\n### Response: The first flight of Beijing Daxing International Airport ushered in the era of Beijing’s “double airport.”\n[Good]</p>
Quality Prediction (QE)	<p>Write a response that appropriately completes the request.\n\n###Request:\n Translate from Chinese to English, and score the translation quality from 0 to 100.\n 7月26日在上海拍摄的公共卫生防疫专业委员会成立仪式现场。 \n\n### Response: The scene of the inauguration ceremony of the Public Health Epidemic Prevention Professional Committee taken in Shanghai on July 26.\n[83]</p>
Draft Refinement	<p>Write a response that appropriately completes the request.\n\n###Request:\n Translate from Chinese to English.\n 虽然朱雨玲连追3分，但丁宁还是利用发球以11：9拿下首局。 \n\n### Hint:\nDraft with quality label:\n[Bad] Although he had only three points, he took the ball to 11:9.\n\n### Note: A translation with no errors could be\n\n### Response: Although Zhu Yuling chased three points in a row, but Ding Ning used his serve to take the first set 11-9.</p>

Table 14: Examples of the prompts and labels for the LLMs. We follow [Jiao et al. \(2023\)](#) to surround the inputs with “Write a response that appropriately completes the request.\n\n### Request:\n” and “### Response:” to guide the LLMs to complete specific tasks. The contents behind “### Response:” are the labels for the text generation fine-tuning of the LLMs.

ting the latter part of the source sentence. In Case 3, the word order in the preliminary translation is notably awkward. All these issues are effectively addressed during the second inference stage, resulting in refined translations of high quality. Even when the predicted quality labels are designated as “Medium” or “Good”, as seen in Case 4 and Case 5, the second stage inference continues to perform fine-tuning on the preliminary translations based on the actual context and linguistic nuances.

<b>Case 1</b>	
<b>SRC</b>	扫二维码下载贴吧客户端
<b>REF</b>	Scan QR code to download Tieba app
<b>Stage 1</b>	Sweep the two-dimensional code to download the client [Bad]
<b>Stage 2</b>	Scan the QR code to download the client
<b>Case 2</b>	
<b>SRC</b>	“部级领导干部历史文化讲座”20周年精华珍藏版 王春瑜、张大可、熊召政、葛剑雄、邓小南等 14位史学泰斗、文化名家、知名学者献给部级领导干部的公开课
<b>REF</b>	“History and Culture Lectures for Ministerial Cadres” 20th Anniversary Essence Collector’s Edition, an open class dedicated to ministerial cadres by 14 leading historians, cultural masters and well-known scholars including Wang Chunyu, Zhang Dake, Xiong Zhaozheng, Ge Jianxiong, Deng Xiaonan
<b>Stage 1</b>	"20th Anniversary of the History and Culture Lecture of Leading Cadres at the Ministerial Level" [Bad]
<b>Stage 2</b>	The 20th Anniversary of the "Historical and Cultural Lectures for Leading Cadres at the Ministerial Level" - 14 Famous Historians, Cultural Celebrities, and Renowned Scholars Including Wang Chunyu, Zhang Dake, Xiong Zhengxiong, and Deng Xiaonan Present Public Lectures for Leading Cadres at the Ministerial Level
<b>Case 3</b>	
<b>SRC</b>	当时刚创立的台积电，在英特尔这个巨无霸面前就是小喽罗。
<b>REF</b>	The newly established Taiwan Semiconductor Manufacturing Company was only a small player in front of Intel the giant.
<b>Stage 1</b>	At that time, the newly established Taiwan accumulator, in the face of the giant Intel, is a small squeak. [Bad]
<b>Stage 2</b>	At that time, the newly established Taiwan Semiconductor was just a small fry in the face of Intel, the giant.
<b>Case 4</b>	
<b>SRC</b>	比如专业、年龄、想学习的语言，甚至在哪个城市，这些都会和你学习或者想要从事的行业有关系的。
<b>REF</b>	Such as your major, age, the language you want to learn, and even which city you are in; all of these are related to what you want to learn or what industry you want to go into.
<b>Stage 1</b>	For example, professional, age, want to learn the language, even in which city, these will be related to your learning or want to engage in the industry. [Medium]
<b>Stage 2</b>	For example, your major, your age, the language you want to learn, and even the city you want to study in, are all related to the industry you want to work in.
<b>Case 5</b>	
<b>SRC</b>	箱中袋装水，其包装盒采用100%可回收利用，可减少包装66%，减少碳排放97%。
<b>REF</b>	The packaging of the bagged water in the box is 100% recyclable, which can reduce packaging by 66% and carbon emission by 97%.
<b>Stage 1</b>	The packaging box uses 100% recyclable packaging, which can reduce packaging by 66% and reduce carbon dioxide emissions by 97%. [Good]
<b>Stage 2</b>	The bagged water in the box is packed in a 100% recyclable packaging box, which can reduce packaging by 66% and carbon dioxide emissions by 97%.

Table 15: Cases of translation process of TASTE in Chinese  $\Rightarrow$  English direction. The backbone model is LLaMA-2-7b trained with its embedding layer fixed. “Stage 1” represents the preliminary translation generated during the first inference process, and “Stage 2” represents the refined translation generated during the second inference process. The predicted quality labels for the drafts are marked using **highlights**.