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

Anonymous ACL submission

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 not able to fully leverage 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 assessment results. The evaluation results across four language directions on the WMT22 benchmark reveal the effectiveness of our approach when compared to the existing methods. Our work presents a promising approach to unleash the potential of LLMs and enhance their capabilities in machine translation.

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 understanding of syntactic and semantic knowledge positions them as potent instruments for the enhancement of machine 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.

SRC	但/But 口水鸡/Poached Chicken in Chili Sauce 整盒/whole box 是/is 生肉/row meat，没办法/cannot 吃/eat，是否/whether 可以/can 退款/refund?
REF	But the Poached Chicken in Chili Sauce is just a box of inedible raw chicken, and can I get a refund for it?
Normal	But the whole box of chicken is raw meat, so you can't eat it, can you refund?
Stage 1	But ~~the whole box is meat~~, can't eat, can you refund \n [Bad]
Stage 2	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 model 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 translation errors are marked by ~~red strikethrough~~, and the highlight denote the predicted quality label.

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 supplementary information pertinent to the translation task to enhance the performance of LLMs (Ghazvininejad et al., 2023; Lu et al., 2023; He 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., 2022; Agrawal et al., 2022). 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 machine translation task of LLMs as a simple text generation task, focusing on adjusting prompts to enhance the outcomes. However, the intrinsic features of the machine translation task, such as the necessity for diverse multilingual knowledge, are often overlooked.

Some studies recommend the tuning of relatively smaller LLMs for translation, guided by a limited number of high-quality supervised instructions (Zhu et al., 2023). The adoption of instruction tuning 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 translation 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 introducing intermediate reasoning steps in the inference process significantly augments the reasoning capabilities of language models (Wei et al., 2022; Kojima et al., 2022).

In this paper, we introduce **TASTE**, a method aiming at improving the translation performance of large language models (LLMs) by instilling the ability to self-reflect on their own outputs. Specifically, we segment the translation process of LLMs 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 reflection, mirroring the common approach employed by humans to carry out tasks more effectively and impeccably. In order to establish a sufficient multitask capability for executing the entire reflective translation process, we conduct supervised fine-tuning (SFT) on LLMs using a hybrid 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 compromising tasks that are closely aligned with the TASTE process to equip LLMs with the capability to successfully 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 of LLMs. Dictionary-based approaches incorporate control hints in the prompt by bilingual or multilingual dictionaries to deal with source sentences containing rare words (Ghazvininejad et al., 2023; Lu et al., 2023). He et al. (2023) extracts translation-related knowledge, such as topics, through self-prompting and employ this information to guide the translation process. Studies in in-context learning (ICL, Brown et al., 2020) aim to provide LLMs with more relevant and high-quality translation exemplars. This approach serves to assist LLMs in retrieving bilingual knowledge, facilitating the generation of translations of the highest possible quality (Vilar et al., 2022; Agrawal et al., 2022).

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., 2021; Ouyang et al., 2022; Chung et al., 2022). 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 LLMs and calculate the additional loss. Zhang et al.

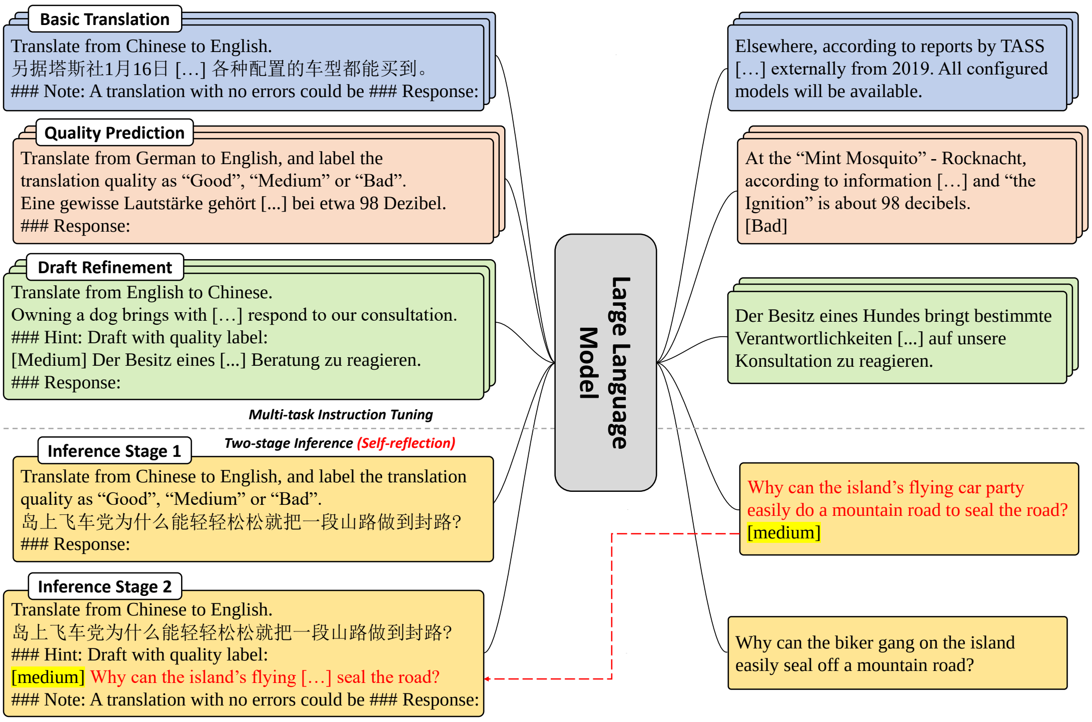

Figure 1: The framework of our proposed TASTE.

(2023b) enhances the multilingual language generation and instruction following capabilities of LLMs through interactive translation tasks. Our work represents a fusion of instruction tuning and the chain-of-thought (CoT) methodology. In our approach, we introduce a multi-step inference translation process in imitation of the self-reflection mechanism observed in humans. This capability is substantiated through the utilization of the multitask training data, comprising Basic Translation, Quality Prediction, and Draft Refinement.

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-reflect on their translation candidates, ultimately producing carefully refined outputs. This process is achieved through a two-stage inference.

In the first stage, we task the models with generating preliminary translations. Different from the conventional machine translation process, we also require the models to predict the quality of their own outputs simultaneously. These generated preliminary translations are referred to as “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:

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

$$\begin{aligned} &P(y_{1:m}, q \mid w, x; \theta) \\ =&P(q \mid y_{1:m}, w, x; \theta) P(y_{1:m} \mid w, x; \theta) \\ =&P(q \mid y_{1:m}, w, x; \theta) \prod_{t=1}^{m} P(y_i \mid y_{1:t-1}, w, x; \theta) \end{aligned} \tag{2}$$

where θ represents the parameters of the LLM, x and w denote the source sentence and the rest of the prompt (including the instruction), respectively. The preliminary translation $y_{1:m}$ is generated first, and the quality label (score) q is generated later according to $y_{1:m}$. The corresponding prompts of the first inference stage are illustrated in the "Inference Stage 1" box of 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 labels (scores), yielding the final translation candidates in a refined form. This stage of inference can be formalized into the following formula:

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

$$P(y'_{1:n} \mid y, q, w', x; \theta) = \prod_{t=1}^{n} P(y'_i \mid y'_{1:t-1}, y, q, w', x; \theta) \quad (4)$$

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

3.2 Multitask Supervised Fine-tuning

To ensure that LLMs acquire the requisite knowledge and achieve a comprehensive understanding of the task instructions, we conduct multitask supervised fine-tuning (SFT) on the models. The multitasking approach consists of three components: **Basic Translation**, **Quality Prediction** and **Draft Refinement**.

Quality Prediction We utilize translation results generated by multiple systems, paired with their evaluated quality scores, to construct fine-tuning instances. These instances are designed to teach LLMs to make quality predictions on the given inputs. Specifically, we employ the COMET score as a proxy for translation quality. The quality prediction task consists of two forms: quality estimation (QE) and text classification (TC). Please refer to Appendix A for detailed information. The ground truth of the training data would be translations with gold quality labels (either scores or categories) placed in the front. An example can be found in the corresponding block in Figure 1.

Basic Translation We utilize parallel data combined with a standardized instruction to conduct fine-tuning of LLMs for multilingual translation tasks, including German⇔ English and Chinese ⇔ 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. In order to disambiguate instructions between these two tasks and prevent LLMs from obtaining low-quality translation knowledge, we adopt the approach proposed by Zeng et al. (2023), which appends a distinguishing note, “`### Note: A translation with no errors could be.`” at the end of the Basic Translation input. This note is also incorporated into the instruction of the second inference stage to minimize errors in the models' output candidates to the greatest extent possible.

Draft Refinement In the second stage of the reflective process, LLMs are tasked with refining drafts based on quality labels (scores) to produce final outputs. For a given source sentence, among the outputs from multiple translation systems, we designate the highest-scored output as the reference while selecting the lowest-scored one as the draft. To facilitate this process, 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 placed in front of the draft in the following format: “`### Hint: Draft with quality label: [LABEL] [Draft]`”. The complete prompt template is shown in Figure 1.

4 Experimental Setups

4.1 Data

Training Data We combined two parts of datasets to build our training set, including the WMT validation set and MTME multi-candidate dataset. Data set introduction and data size can be found in Appendix B.

Test Data To avoid possible data leakage in the training data, we evaluate the translation performance on the test sets from WMT22 competition (Kocmi et al., 2022), which covers diverse domains such as news, social, e-commerce and conversation. We mainly report the results of translations 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-7b-mt`[1] and `LLaMA-2-7b`[2] (Touvron et al., 2023) as our backbone models. The fine-tuning strategy encompasses the following approaches:

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

[1] https://huggingface.co/bigscience/bloomz-7b1-mt
[2] https://huggingface.co/meta-llama/Llama-2-7b

System	Zh⇒En		En⇒Zh		De⇒En		En⇒De	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
WMT22 Winners	33.50	81.00	54.30	86.80	33.70	85.00	38.40	87.40
NLLB-3.3b	21.07	76.92	32.52	81.56	29.54	83.42	33.98	86.23
BayLing-7b	21.54	79.45	41.96	85.15	26.80	83.96	28.23	84.26
MT-*Full*	22.81	79.25	35.49	85.01	24.05	77.61	18.84	71.31
MT-*FixEmb*	23.43	79.84	36.68	85.20	25.07	78.27	19.41	72.06
TASTE								
Full-QE	23.56	79.26	37.73	85.00	25.17	77.84	21.03	74.30
Full-TC	23.52	79.24	37.91	84.99	24.92	78.04	20.84	74.24
FixEmb-QE	24.56	80.09	39.73	85.42	26.35	78.63	21.56	75.07
FixEmb-TC	24.32	80.09	39.76	85.45	26.25	78.67	21.61	75.26
FixEmb-QE+TC	**24.62**	**80.17**	**39.97**	**85.62**	**26.60**	**79.03**	**21.89**	**75.76**

Table 2: Main results of TASTE. `BLOOMZ-7b-mt` is chosen as the backbone model. *QE* and *TC* signify that the Quality Prediction subtask takes the form of quality estimation and text classification, respectively. *QE+TC* denotes a fusion of these two approaches, combining two segments of the training data. The best results of our work are labeled using **bold font**.

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 trained on large-scale corpus during pre-training 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 can help LLMs maintain correctness and diversity in their expressions.

4.3 Baselines

The baseline models are fine-tuned on the Basic Translation data set which contains German⇔English and Chinese⇔English directions. We represent these baselines as **MT-(·)**.

Additionally, we report the results of WMT22 winners, NLLB-3.3B (Costa-jussà et al., 2022), which is a multilingual translation model trained in over 200 languages and `Bayling` (Zhang et al., 2023b), an LLM tuned for machine translation with `LLaMA-7b` as the backbone model.

5 Results

Our main results are shown in Table 2. Almost all of our methods outperform the MT baseline across both metrics, providing evidence of the effectiveness of our approach in enhancing the translation capabilities of LLMs. When employing the *QE+TC* approach, which combines the training data of both quality estimation and text classification styles, the models consistently attain the highest scores across nearly all directions. When choosing `BLOOMZ-7b-mt` as the backbone model, our approach achieves favorable results in Zh⇔ En directions, which surpasses `NLLB-3.3b` and `Bayling-7b`, approaching the performance of WMT22 winners in COMET scores (80.17 vs. 81.0 and 85.62 vs. 86.80). `LLaMA-7b` also achieves performance enhancement in Zh⇔ En directions, the details are shown in Table 3.

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.

We can also observe inconsistencies in both the trajectory and magnitude of changes when examining BLEU and COMET scores. For instance, our approach, referred to as TASTE-*FixEmb-TC*, slightly lags behind `BayLing` in terms of BLEU scores (39.76 vs. 41.96), yet it achieves a higher

System	Zh⇒En		En⇒Zh	
	BLEU	COMET	BLEU	COMET
MT-*FixEmb*	24.30	79.02	33.33	83.62
TASTE				
FixEmb-QE	24.36	79.14	34.68	83.76
FixEmb-QE+TC	24.84	79.30	34.94	83.90

Table 3: The results of TASTE while taking `LLaMA-7b` as the backbone model. Our approach gains translation performance enhancement in both Zh⇒ En and EN⇒ Zh directions.

Model	PPL	Pred.↑	P↑	R↑	F1↑
BLOOMZ	4.2	85.3	78.7	78.2	78.1
LLaMA-2	-39.1	91.3	80.5	80.2	80.1

Table 4: Evaluation results on quality prediction task. PPL/Pred. represents Pearson's r between the perplexity values/predicted scores and the COMET scores. Precision, recall, and F1 values are calculated as weighted averages across three translation quality categories.

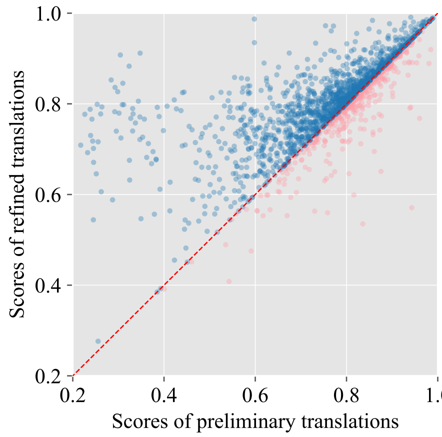

Figure 2: Comparison between the COMET scores of the preliminary and refined translations.We report the scores in Zh⇒En direction achieved by `BLOOMZ-7b-mt`.

COMET score (85.45 vs. 85.15). The limitations of BLEU have been widely discussed in recent times, primarily due to its limited correlation with human evaluation results, as highlighted by Freitag et al. (2022). It is pointed out that neural-based metrics offer a more qualified and robust means of evaluating translation quality. The observed inconsistencies in our results align with this viewpoint, emphasizing the need to prioritize the more reliable COMET scores in our assessments.

6 Analysis

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 prediction outputs. This evaluation is executed using a validation set containing all four translation directions extracted from the MTME multi-candidate data set, which does not overlap with the training data. For quality estimation, we assessed 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 gold COMET scores for comparison. For text classification, we construct gold labels for the instances according to their COMET scores following the same principle mentioned in §3.2 and we report precision, recall, and F1 values.

The results are shown in Table 4. In the quality estimation 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. And for the text classification approach, the model also exhibits a commendable level of accuracy in assigning quality labels to their translations, as evidenced by F1 values surpassing 78.1. These statistics demonstrate that our models are able to make precise quality predictions for their own generated translations, thereby providing a dependable reference for the Draft Refinement task. We can also discover from the results that `LLaMA-2` outperforms `BLOOMZ` in terms of accuracy for both quality estimation and text classification 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, respectively.

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 in which a quality improvement

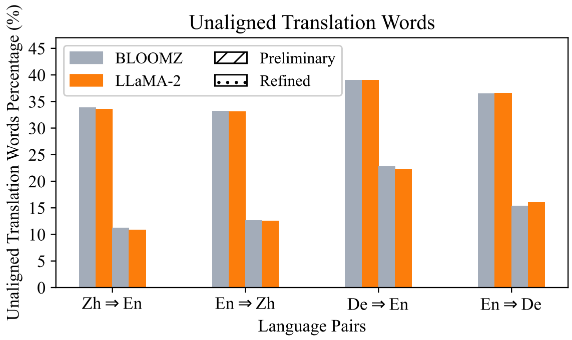

Figure 3: Comparison between the unaligned translation words percentages of the preliminary and refined translations.

Method	BLEU	COMET
MT	23.43	79.84
TASTE	24.65	80.28
w/ ConstDrafts	22.39	77.10
w/o BasicTrans	21.29	70.70
w/o QualityPred	24.29	80.06
w/o DraftRefine	22.96	76.36

Table 5: Ablation Study. We report the BLEU and COMET scores in Zh⇒En direction achieved by `BLOOMZ-7b-mt`.

System	Zh⇒En	En⇒Zh	De⇒En	En⇒De
Ours	79.30	83.90	83.87	83.47
ICL-7b	74.50	73.79	79.63	74.37
ICL-13b	75.21	75.32	80.10	73.55

Table 6: COMET scores gained by our approach and the In-context Learning method.

is achieved through the refinement process. As the plot demonstrates, a majority of the final candidates exhibit higher quality levels than their initial counterparts. In many cases, the candidates gain an enhancement in their COMET score of over 0.05. Furthermore, it is worth noting that the Draft Refinement process helps rectify the generation failures that may occur during the initial inference stage (instances located in the top-left region of the plot). These observations indicate the capacity of the Draft Refinement process to effectively refine the preliminary translations generated after the first inference stage and its ability to handle instances of generation failure.

Unaligned Translation Words (UTW) We measure the number of target-side words that remain unaligned in a word-to-word alignment between the source sentences and translations obtained after the first and second inference stages, respectively. 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 unaligned translation words is reduced significantly during the Draft Refinement process, with a decrease of approximately 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 Ablation Study

In order to emphasize the necessity of our multitask training set and prompt design, we conduct an ablation study. We choose `BLOOMZ-7b-mt` as the backbone model and fine-tune it using various training sets with *FixEmb-TC* method. BLEU and COMET scores evaluated in Zh⇒En direction are reported in Table 5.

Contrastive Drafts In the Draft Refinement subset of the multitask training data, we choose one low-quality candidate from the MTME multi-candidate data set as a draft to be refined. Here, we add one more candidate with the second-highest COMET score to form a pair of contrastive drafts. The task for LLMs is to generate refined translations based on the contrastive drafts with their respective quality labels. The results in the third line of Table 5 show that this approach brings no positive effects. This indicates that during the refinement stage, extra drafts are not needed by LLMs to generate higher-quality translations.

Multitask Training Set Our multitask training set contrains three parts: **Basic Translation**, **Quality Prediction** and **Draft Refinement**. Each task serves for the whole reflection process we propose. To demonstrate the rationality of this task combination, we remove a specific section of the training set separately, and the consequences are shown in the last three rows of Table 5. The performance of the model decreases when any subset of the training date is removed. This result implies that each of the sub-tasks is essential for our approach.

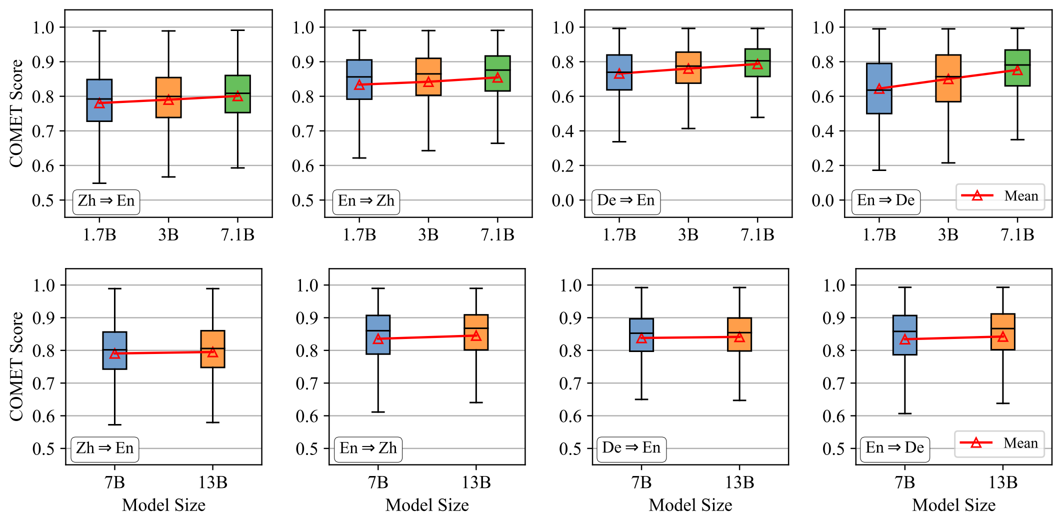

Figure 4: COMET scores obtained from `BLOOMZ` (Line 1) and `LLaMA-2` (Line 2) across different model sizes.

6.4 Comparison with In-context Learning

Our approach is based on a two-stage inference, which is similar to the thought of ICL (In-context Learning). To certify the superiority of our proposal, we perform a comparison with the ICL method. We apply the same two-stage inference procedures used in our approach to `LLaMA-2-chat-7b` and `LLaMA-2-chat-13b`, both of which undergo no training process. The results are shown in Table 6. In many-to-English translation directions, the ICL method gains reasonable performance, yet our approach outperforms it significantly. And in English-to-many directions, substantial performance gaps are observed between the ICL method and our approach. The ICL method failed to generate stable outcomes by the inference chain, primarily due to a severe off-target issue which keeps the models from producing translations in correct target languages.

6.5 Effect of Model Size

We report COMET scores yielded by LLMs of various sizes, with `BLOOMZ` and `LLaMA-2` trained by *FixEmb-QE* method as backbone models.

As shown in Figure 4, 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 different tasks are highly similar, which makes it more challenging but essential for LLMs to grasp different type 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. Regarding `LLaMA-2`, the observed improvement is more substantial in many-to-English directions. However, the underlying reasons for this phenomenon remain unexplored and will be focused on in future works.

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 enhance the translation quality through the refinement process, consistently producing high-quality candidates across various translation directions. Moreover, our findings demonstrate that performance improves with model scaling, suggesting that our methodology can be extended to larger LLMs, potentially yielding even more promising results and providing a valuable approach for machine translation using large language models.

Limitations

The performance enhancement introduced by our approach exhibits inconsistency across different translation directions. The improvement in certain directions is more substantial than in others, and this observation persists even when employing model scaling. We assume that this phenomenon is caused by the inherent uneven multilingual knowledge within the model, which is strongly influenced by the data distribution during the pretraining process of LLMs. A more in-depth exploration of the underlying principles of this phenomenon is essential, and further experiments involving additional language pairs are warranted.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. arXiv preprint arXiv:2212.02437.

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.

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 arXiv:2303.12712.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416.

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 arXiv:2207.04672.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In Conference of the European Chapter of the Association for Computational Linguistics (EACL).

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.

Xavier Garcia and Orhan Firat. 2022. Using natural language prompts for machine translation. arXiv preprint arXiv:2202.11822.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. arXiv preprint arXiv:2302.07856.

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. arXiv preprint arXiv:2305.04118.

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.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Parrot: Translating during chat using large language models. arXiv preprint arXiv:2304.02426.

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.

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 arXiv: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 arXiv:2305.06575.

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.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting palm for translation: Assessing strategies and performance. arXiv preprint arXiv:2211.09102.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35:24824–24837.

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. arXiv preprint arXiv:2301.07069.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong 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 arXiv: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 arXiv:2304.04675.

Task	Size	Source
Basic Translation	45.4k	WMT Dev
Draft Refinement	18.6k	MTME
Quality Prediction	45.0k	MTME

Table 7: Data sizes and sources of the training sets.

A Quality Prediction Task Designs

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

Quality Estimation (QE) We request LLMs to simultaneously predict quality scores on a scale 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 it to use as gold scores.

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".`" Translations with COMET scores greater than 0.85 are expected to be classified as *Good*, those less than 0.65 as *Bad*, and the remainder as *Medium*.

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

B 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 machine translation task. Specifically, we choose the newstest2017-2021 of German ⇔ English and Chinese ⇔ English as our MT training set. Source and target sentences in this training set are formed into the **MT Prompt**.

MTME Multi-Candidate Data This is a data set containing source sentences and outputs of multiple MT systems on the WMT metrics shared tasks built by Google Research[3]. We use the outputs on newstest2019-2021 MT task of German ⇔ English and Chinese ⇔ English to build training data for the Translation Classification and Draft Refinement task. We decide the quality labels of each output by calculating the COMET score with the wmt-22-comet-da model. Candidates with scores above 0.85 are labeled as **[*Good*]**, while those with scores below 0.6 are labeled as **[*Bad*]**, and the rest of them are labeled as **[*Medium*]**. The Translation Classification and Draft Refinement data are formed into the **Classification Prompt** and **Refinement Prompt**, respectively.

The sizes and sources of the training data for the three tasks are represented in Table 7.

[3] https://github.com/google-research/mt-metrics-eval