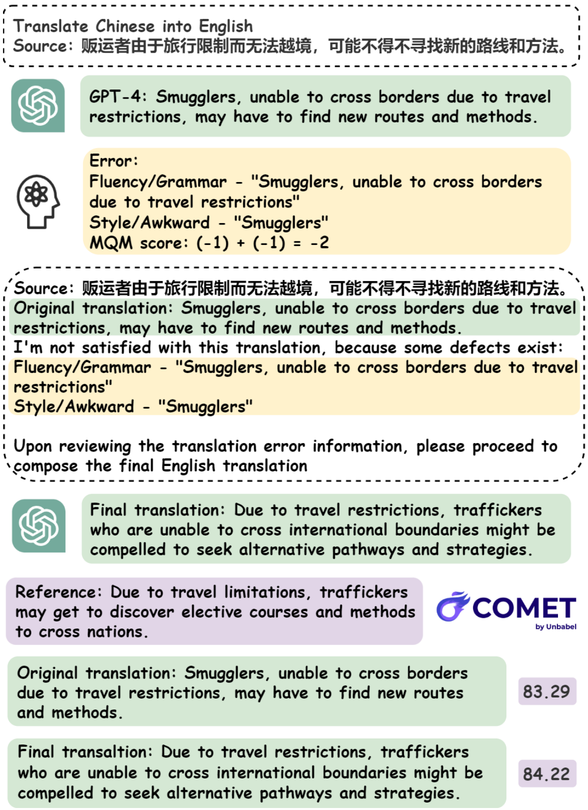
Improving LLM-based Machine Translation with Systematic Self-Correction

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have achieved impressive results in Machine Translation (MT). However, careful evaluations by human reveal that the translations produced by LLMs still contain multiple errors. Importantly, feeding back such error information into the LLMs can lead to self-correction and result in improved translation performance. Motivated by these insights, we introduce a systematic LLM-based self-correcting translation framework, named **TER**, which stands for **T**ranslate, **E**stimate, and **R**efine, marking a significant step forward in this direction. Our findings demonstrate that 1) our self-correction framework successfully assists LLMs in improving their translation quality across a wide range of languages, whether it's from high-resource languages to low-resource ones or whether it's English-centric or centered around other languages; 2) TER exhibits superior systematicity and interpretability compared to previous methods; 3) different estimation strategies yield varied impacts on AI feedback, directly affecting the effectiveness of the final corrections. We further compare different LLMs and conduct various experiments involving self-correction and cross-model correction to investigate the potential relationship between the translation and evaluation capabilities of LLMs. The code will be made available upon publication.



Translate Chinese into English
Source：贩运者由于旅行限制而无法越境，可能不得不寻找新的路线和方法。

GPT-4: Smugglers, unable to cross borders due to travel restrictions, may have to find new routes and methods.

Error:
Fluency/Grammar - "Smugglers, unable to cross borders due to travel restrictions"
Style/Awkward - "Smugglers"
MQM score: (-1) + (-1) = -2

Source：贩运者由于旅行限制而无法越境，可能不得不寻找新的路线和方法。
Original translation: Smugglers, unable to cross borders due to travel restrictions, may have to find new routes and methods.
I'm not satisfied with this translation, because some defects exist:
Fluency/Grammar - "Smugglers, unable to cross borders due to travel restrictions"
Style/Awkward - "Smugglers"

Upon reviewing the translation error information, please proceed to compose the final English translation

Final translation: Due to travel restrictions, traffickers who are unable to cross international boundaries might be compelled to seek alternative pathways and strategies.

Reference: Due to travel limitations, traffickers may get to discover elective courses and methods to cross nations.

Original translation: Smugglers, unable to cross borders due to travel restrictions, may have to find new routes and methods.

83.29

Final transaltion: Due to travel restrictions, traffickers who are unable to cross international boundaries might be compelled to seek alternative pathways and strategies.

84.22

Figure 1: The original translation is from the submission of GPT-4 (Kocmi et al., 2023) for WMT23. The MQM error label is annotated by human experts. We use OpenAI API *gpt-4* to correct the translation. The metric score increases from 83.29 to 84.22 using COMET-22 (*wmt22-comet-da*) (Rei et al., 2020) model.

1 Introduction

The results of the General Machine Translation Task [1] in WMT23 indicate that LLM-based machine translation systems (Agrawal et al., 2023; Vilar et al., 2023; Zhang et al., 2023; He et al., 2023; Gao et al., 2023; Garcia et al., 2023; Wu and Hu, 2023; Hendy et al., 2023; Peng et al., 2023a; Moslem et al., 2023; Zhu et al., 2023; Liang et al., 2023), especially GPT-4 (Achiam et al., 2023) using few-shot prompting, have taken top positions in the majority of subtasks (Kocmi et al., 2023). However, taking the whole 1976 test pairs from the WMT23 Zh-En dataset as a probing study, even with GPT-4 (the best submission in WMT23), only 332 pairs achieved a perfect score (i.e., no errors were identified upon manual inspection) according to the Multi-dimensional Quality Metrics (MQM, Freitag et al., 2021) [2]. Interestingly, when we feed back the information from the human MQM evaluations into GPT-4, asking it to correct

---

[1] WMT 2023 Shared Task: General Machine Translation

[2] https://github.com/google/wmt-mqm-human-evaluation

its initial translation based on this feedback, we observe that the errors can be corrected and there is also an improvement in the metric score, as shown in Figure 1. Such findings inspire us to leverage evaluative feedback on translations as a means to facilitate the correction and improvement of initial translations by LLMs.

Involving LLMs to automatically do the self refinement is increasingly gaining attention (Saunders et al., 2022; Welleck et al., 2022; Pan et al., 2023; Madaan et al., 2023; Shinn et al., 2023; Chen et al., 2023b; Li et al., 2023; Gou et al., 2023). In the field of Machine Translation (MT), Chen et al. (2023a) investigate the use of LLMs to rewrite translations by feeding back the previous translations, achieving changes at the lexical and structural levels while maintaining translation quality. Raunak et al. (2023) ask GPT-4 to refine translations with the inclusion of the edit proposals produced by Chain-of-Thought (CoT) (Kojima et al., 2022) strategy, indicating that GPT-4 is adept at translation post-editing. However, these efforts encounter several challenges: 1) There is a lack of clear assessment of the original translations, with evaluations either absent or merged into the correction process; 2) The feedback provided is ineffective as it lacks clear guidance or contains redundant information, leading to difficulties in comprehension by LLMs; 3) The initial translations originate from external sources, yet the capability of LLMs for self-correction remains uninvestigated; 4) The translation quality improvements are not strong, and in some scenarios, a decline has been noted.

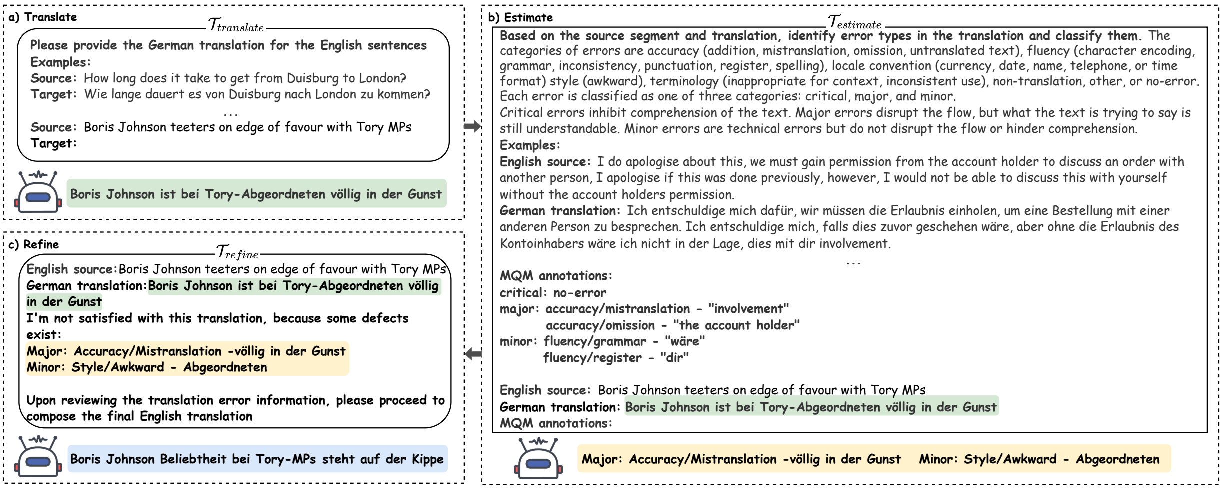
In this paper, we propose an LLM-based self-correcting translation framework, termed **TER**: **T**ranslate, **E**stimate, and **R**efine. **Translate** module utilizes an LLM for translation, ensuring that our translations are internally sourced. **Estimate** module receives the initial translation and provides systematic estimations of the translation quality. This assessment serves as feedback. **Refine** module performs translation corrections based on the information from the preceding two modules. Our experimental results indicate that our approach is more effective in improving translation quality compared to the baselines, across both high- and low-resource languages, as well as English-centric and non-English-centric direct translations, especially in semantic-related metrics (COMET (Rei et al., 2020), COMETKiwi (Rei et al., 2022), and BLEURT (Sellam et al., 2020)). We also explore the combined performance of our three modules under various prompting strategies in self-correction and cross-model correction. Additionally, we conduct a more detailed analysis of the estimation component according to types of translation errors. Our contributions are as follows:

- We propose the first LLM-based self-correcting translation framework, TER, featuring a standalone estimation module and establishing a feedback concept within self-correcting MT, weaving into an interpretable translation system.
- TER significantly improves translation quality across various language pairs with different LLMs, e.g., ChatGPT (Achiam et al., 2023), Gemini (Team et al., 2023), Claude (Anthropic, 2023), demonstrating effectiveness for both high-resource and low-resource languages as well as English-centric and non-English-centric translations.
- We conduct a comprehensive analysis to provide clear insights into how different components of TER interact and contribute to overall performance, facilitating targeted improvements and innovations in the field. We demonstrate that translation ability and quality estimation ability exhibit similar trends in specific language pairs, while error correction ability complements them in other language pairs, providing novel insights for future work.

2 TER

Figure 2 illustrates the workflow of TER. The TER framework consists of three modules: Translate, Estimate, and Refine. Given an LLM $\mathcal{M}$ with fixed parameters θ and a language pair $\mathcal{X}$-$\mathcal{Y}$ (source-target), we first utilize $\mathcal{M}$ to generate the initial translation for segment x with the prompt $\mathcal{T}_{translate}$. Got the initial translation y, we use the same LLM $\mathcal{M}$ with prompt $\mathcal{T}_{estimate}$ for estimation. Using the estimation as feedback, we leverage $\mathcal{M}$ to refine the initial translation with $\mathcal{T}_{refine}$.

Translate Using a few given exemplars as context has been proven to be effective in enhancing the LLM's translation capabilities (Agrawal et al., 2023; Vilar et al., 2023; Garcia et al., 2023). In the initial phase, we first concatenate the selected k source-target paired examples $\mathbb{E} = (x_1, y_1) \oplus (x_2, y_2) \oplus ... \oplus (x_k, y_k)$ with the testing source sentence x as the prompt $\mathcal{T}_{translate}$, and utilize $\mathcal{M}$ to



Please provide the German translation for the English sentences
Examples:
Source: How long does it take to get from Duisburg to London?
Target: Wie lange dauert es von Duisburg nach London zu kommen?
...
Source: Boris Johnson teeters on edge of favour with Tory MPs
Target:

Boris Johnson ist bei Tory-Abgeordneten völlig in der Gunst

Based on the source segment and translation, identify error types in the translation and classify them. The categories of errors are accuracy (addition, mistranslation, omission, untranslated text), fluency (character encoding, grammar, inconsistency, punctuation, register, spelling), locale convention (currency, date, name, telephone, or time format) style (awkward), terminology (inappropriate for context, inconsistent use), non-translation, other, or no-error. Each error is classified as one of three categories: critical, major, and minor.
Critical errors inhibit comprehension of the text. Major errors disrupt the flow, but what the text is trying to say is still understandable. Minor errors are technical errors but do not disrupt the flow or hinder comprehension.
Examples:
English source: I do apologise about this, we must gain permission from the account holder to discuss an order with another person, I apologise if this was done previously, however, I would not be able to discuss this with yourself without the account holders permission.
German translation: Ich entschuldige mich dafür, wir müssen die Erlaubnis einholen, um eine Bestellung mit einer anderen Person zu besprechen. Ich entschuldige mich, falls dies zuvor geschehen wäre, aber ohne die Erlaubnis des Kontoinhabers wäre ich nicht in der Lage, dies mit dir involvement.
...
MQM annotations:
critical: no-error
major: accuracy/mistranslation - "involvement"
accuracy/omission - "the account holder"
minor: fluency/grammar - "wäre"
fluency/register - "dir"

English source: Boris Johnson teeters on edge of favour with Tory MPs
German translation: Boris Johnson ist bei Tory-Abgeordneten völlig in der Gunst
MQM annotations:

Major: Accuracy/Mistranslation -völlig in der Gunst Minor: Style/Awkward - Abgeordneten

English source: Boris Johnson teeters on edge of favour with Tory MPs
German translation: Boris Johnson ist bei Tory-Abgeordneten völlig in der Gunst
I'm not satisfied with this translation, because some defects exist:
Major: Accuracy/Mistranslation -völlig in der Gunst
Minor: Style/Awkward - Abgeordneten

Upon reviewing the translation error information, please proceed to compose the final English translation

Boris Johnson Beliebtheit bei Tory-MPs steht auf der Kippe

Figure 2: TER framework includes three steps: Translate, Estimate, and Refine. All steps are executed using different prompts ($\mathcal{T}_{translate}$, $\mathcal{T}_{estimate}$, $\mathcal{T}_{refine}$). We detail our prompting strategies in Section 3.

generate the target y. This step is crucial for setting a strong foundation for subsequent refinement.

Estimate Quality Estimation (QE) in the field of MT refers to a method predicting the quality of a given translation rather than assessing how similar it is to a reference segment. Nonetheless, the learned models (Rei et al., 2022; Perrella et al., 2022; Gowda et al., 2023; Juraska et al., 2023; Guerreiro et al., 2023) focus on offering a single numerical score to gauge the quality. Even when they can predict spans of errors, they only provide information on the level of severity. This presents three main challenges: 1) The first is incorporating additional models causes inconvenience and requires a significant amount of annotated data for training; 2) The second is understanding the meaning behind the numerical representation; 3) The last is obtaining the feedback with guiding significance. Gold MQM annotation requires expensive human efforts, which runs counter to automatic QE. Recently, Kocmi and Federmann (2023) and Fernandes et al. (2023) successfully leverage LLMs to automatically annotate with MQM guideline (Freitag et al., 2021). Guerreiro et al. (2023) also showcase that using LLMs to assess translation quality can achieve performance comparable to models trained on extensive annotated datasets. Prompting with $\mathcal{T}_{estimate}$, we can make a black-and-white judgment on whether refinement is needed, rather than empirically using the learned metrics and setting a numerical threshold. Besides, when doing refinement, the result generated during the estimating process also provides important guidance.

Refine Large language models have elevated the level of machine translation to unprecedented heights. *It is relatively easy to progress from a passing level to excellence, whereas reaching the pinnacle from excellence is no easy feat.* Chen et al. (2023a) and Raunak et al. (2023) have shown the potential of leveraging LLMs to refine the translation. However, both of them lack an explicit estimation module and feedback concept, which would make this improvement process unclear. Therefore, we collect the estimation as the feedback to establish $\mathcal{T}_{refine}$ to refine the initial translation.

In a word, our TER self-correction framework requires only one LLM to accomplish all the functionalities of the aforementioned modules, providing insights into the self-evolution of LLM-based translation systems.

3 Experimental Setup

Dataset Our testset mainly comes from WMT22 and WMT23, except for Icelandic from WMT21 [3]. We evaluate 17 translation directions, including both high-resource and low-resource languages, and translations between English-centric and non-English-centric languages, totally covering 10 languages: English (En), French (Fr), German (De), Czech (Cs), Icelandic (Is), Chinese (Zh), Japanese (Ja), Russian (Ru), Ukrainian (Uk), and Hebrew (He). For each translation pair, we randomly selected 200 pairs to create our test dataset. Detailed statistics can be found in Appendix A.

---

[3] https://github.com/wmt-conference

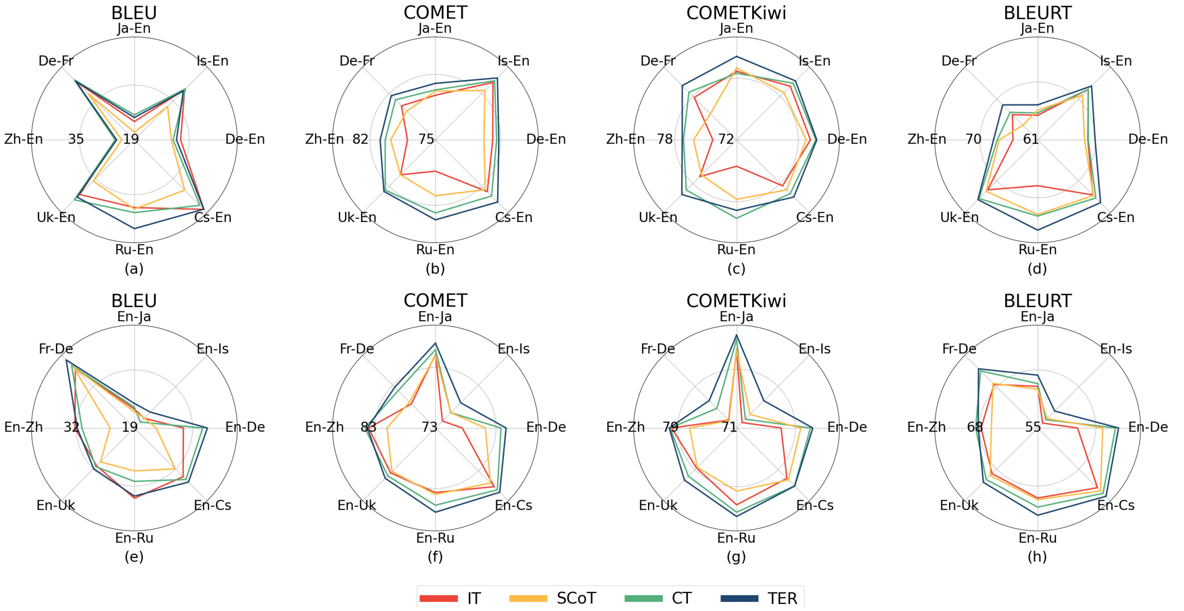


Figure 3: Results for 16 translation directions using GPT-3.5-turbo. *IT*: the initial translation using few-shot prompt; *SCoT*: Structured Chain-of-Thought (Raunak et al., 2023); *CT*: inserting the word ″bad″ to do the contrastive translation (Chen et al., 2023a).

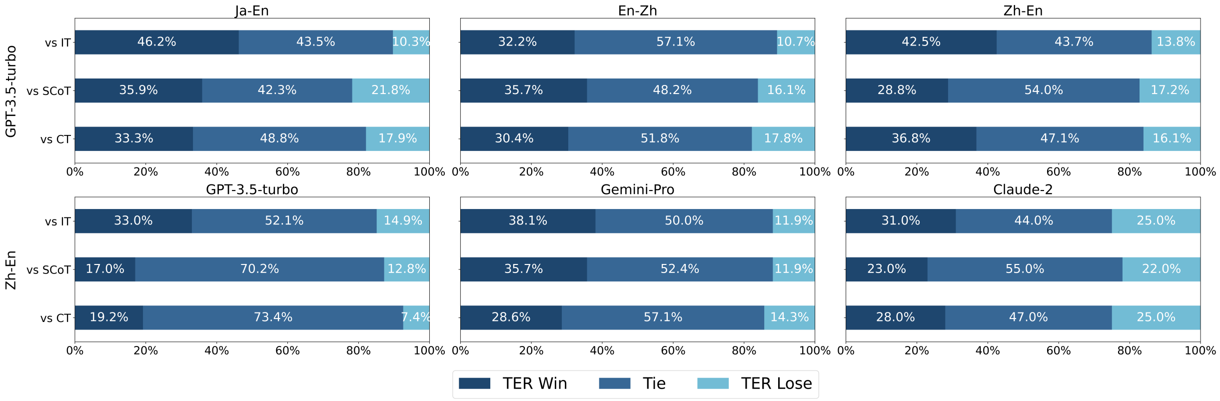


Figure 4: Results for human preference study, comparing TER with IT, SCoT, and CT. The data for the first row of subfigures comes from WMT22 tested on GPT-3.5-turbo, while the experiments for the second row of subfigures were conducted on our WMT23 Zh-En dataset using three models (GPT-3.5-turbo, Gemini-Pro, Claude-2).

LLMs In this paper, we accessed all LLMs through their APIs including GPT-3.5-turbo [4], Claude-2 [5] and Gemini-Pro [6]. Due to the limited computational resources, we did not include open-source LLMs.

Prompting Strategies For $\mathcal{T}_{translate}$, we follow the setting in Hendy et al. (2023) and Kocmi et al. (2023), using the few-shot (5-shot) high-quality translation pairs. For $\mathcal{T}_{estimate}$, we use few-shot (3-shot) multi-lingual MQM annotated examples to ask LLMs to estimate the initial target in a reference-free scenario. Kocmi and Federmann (2023) have demonstrated that this prompt method ensures the estimation can be executed across any language pairs and can rival metrics models trained on a large amount of MQM annotated data. As for $\mathcal{T}_{refine}$, $\mathcal{T}_{translate}-\alpha$ represents only using the feedback, while $\mathcal{T}_{translate}-\beta$ adheres estimation feedback to the few-shot translation examples. All prompts are detailed in Appendix G.

Evaluation methods To evaluate the quality of a translation, we utilize commonly used metrics and human preference tests.

- **Metrics of translation:** 1) a reference-based neural metric COMET-22 (Rei et al., 2020);

---

[4] We utilize gpt-3.5-turbo-0613 whose training data is up to date as of September 2021 `https://platform.openai.com/docs/models/gpt-3-5`

[5] https://www.anthropic.com/index/claude-2

[6] https://cloud.google.com/vertex-ai/docs/generative-ai/learn/models

Translate		**Estimate**		**Refine**		**BLEU**	**COMET**	**COMETKiwi**	**BLEURT**
zero-shot	**few-shot**	**zero-shot**	**few-shot**	$\mathcal{T}_{refine}-\alpha$	$\mathcal{T}_{refine}-\beta$				
✓		-	-	-	-	25.45	80.01	78.87	66.56
✓		✓		✓		24.95 (-0.50)	80.05 (+0.04)	78.95 (+0.08)	66.64 (+0.08)
✓		✓			✓	25.15 (-0.30)	80.22 (+0.21)	79.18 (+0.31)	66.91 (+0.35)
✓			✓	✓		25.56 (+0.11)	80.52 (+0.51)	79.75 (+0.88)	67.01 (+0.45)
✓			✓		✓	**25.79** (+0.34)	**80.68** (+0.67)	**79.77** (+0.90)	**67.09** (+0.53)
	✓	-	-	-	-	26.30	80.31	79.28	67.07
	✓	✓		✓		26.12 (-0.18)	80.44 (+0.13)	79.71 (+0.43)	67.24 (+0.17)
	✓	✓			✓	26.23 (-0.07)	80.54 (+0.23)	79.94 (+0.66)	67.33 (+0.26)
	✓		✓	✓		26.01 (-0.29)	80.52 (+0.21)	79.79 (+0.51)	67.66 (+0.59)
	✓		✓		✓	**26.41** (+0.11)	**80.70** (+0.39)	**80.07** (+0.79)	**67.84** (+0.77)

Table 1: Comparison of quality improvement for different variants of the TER modules in our sampled WMT23 Zh-En dataset. $\mathcal{T}_{refine}-\alpha$ uses only the feedback from the **Estimate** module, $\mathcal{T}_{refine}-\beta$ uses few-shot exemplars in the **Translate** module and the feedback. **Bold** results indicate the best in each section.

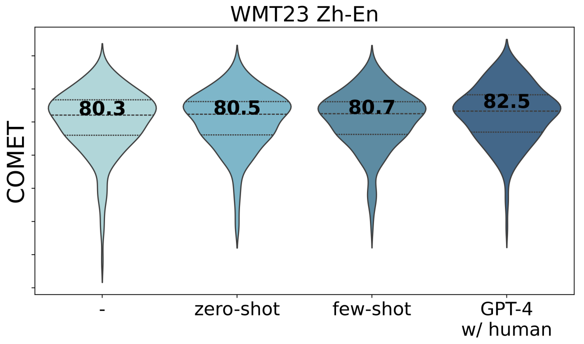


Figure 5: COMET scores for involving various feedback estimation strategies in the TER. "-" denotes the baseline where no estimation is involved, representing the initial translation (IT). *zero-shot* and *few-shot* reflect the use of different prompting methods with GPT-3.5-turbo, while *GPT-4 w/ human* indicates estimations made using GPT-4 with human assistance.

zero-shot Win	Tie	few-shot Win
35	89	76

Table 2: Pairwise human evaluation on zero-shot and few-shot estimation.

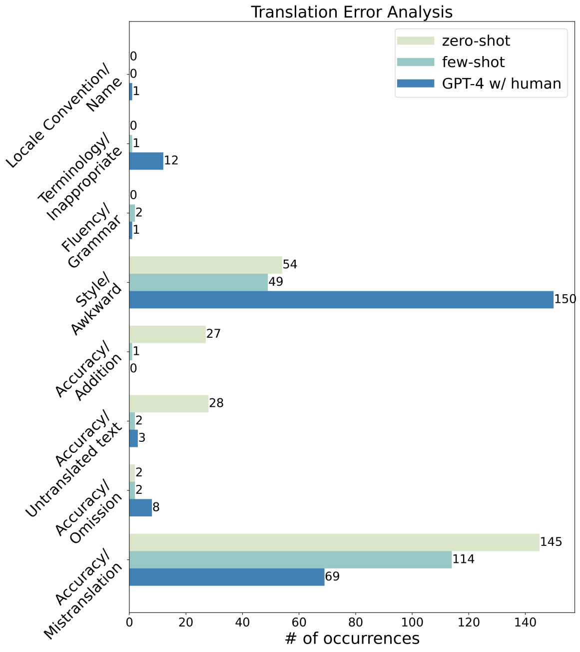


Figure 6: Error type analysis for initial translations in our WMT23 Zh-En. We quantified the errors across different estimation prompting strategies, including zero-shot, few-shot, and GPT-4 with human assistance.

2) a reference-free quality estimation model COMETKiwi (Rei et al., 2022); 3) a reference-based trained metric BLEURT-20 (Sellam et al., 2020); 4) a lexical metric SacreBLEU (Post, 2018) for completeness.

- **Human Preference:** We conduct pairwise human preference studies, engaging graduate students who are also bilingual experts. Participants were presented with a source sentence and two candidate translations, from which they were to select the superior translation.

4 Results

We compare TER with three baselines, i.e., 1) the initial translation with few-shot prompting (IT); 2) post-editing with Structured-CoT (SCoT), which uses the MQM annotation instructions (Freitag et al., 2021) to produce the intermediate CoT in the form of an MQM annotation over the source-translation pair (Raunak et al., 2023); 3) Contrastive Translate (CT), which inserts the word “bad” to hint that the previous translation is of low quality, regardless of its actual quality, to form a contrastive prompt (Chen et al., 2023a).

Figure 3 displays our test dataset results across 9 languages and 16 translation directions using GPT-3.5-turbo. Compared to self-translated IT, TER significantly outperforms them in all directions, with an average improvement of 2.48. Notably, the Ru-En pair sees a significant leap of 5.14, and

Estimate	Metric	Accuracy/ Mistranslation	Accuracy/ Omission	Accuracy/ Untranslated text	Accuracy/ Addition	Style/ Awkward	Fluency/ Grammar	Terminology/ Inappropriate	Locale Convention/ Name
zero-shot	ΔCOMET	+0.15	-0.70	+0.45	-0.18	+0.19	/	/	/
few-shot		+0.84	-1.16	+18.42	0.00	+1.09	+0.11	-0.10	/
GPT-4 w/ human		+3.34	+4.34	+31.09	/	+1.53	+4.14	+1.37	+1.98

Table 3: Relative COMET score improvements over initial translations (IT) when employing different estimation feedback strategies (zero-shot, few-shot, GPT-4 with human). We splited the testing targets based on the classification of errors by estimation strategies. "/" indicates that no testing target was segmented under this error type.

Models	BLEU	COMET	COMETKiwi	BLEURT	Rank
			En-Ru		
GPT-3.5-turbo	36.95	87.94	83.47	75.78	2
Gemini-Pro	34.25	86.34	82.02	74.51	3
Claude-2	**37.72**	**88.90**	**84.15**	**77.04**	1
			En-De		
GPT-3.5-turbo	**48.07**	**84.15**	**80.35**	**70.96**	1
Gemini-Pro	44.89	83.08	79.10	70.50	2
Claude-2	45.80	82.69	79.96	68.81	3
			He-En		
GPT-3.5-turbo	**47.83**	86.00	82.16	76.00	3
Gemini-Pro	46.46	86.02	82.05	76.13	2
Claude-2	47.37	**86.55**	**82.61**	**76.34**	1
			Zh-En		
GPT-3.5-turbo	**26.41**	**80.70**	**80.07**	**67.84**	1
Gemini-Pro	23.63	78.67	77.47	64.82	3
Claude-2	23.06	79.66	79.32	66.03	2

Table 4: The results of testing translation ability of various LLMs under few-shot setting in our sampled WMT23 datasets. **Bold** results indicate the best in each section. We rank different LLMs based on their scores from learned metrics.

the En-De pair has an impressive boost of 6.88. TER consistently outperforms several post-editing methods across all directions. In terms of BLEU scores, TER leads SCoT by 4.75 and CT by 1.0 on average. For the COMET metric, TER's performance is ahead of SCoT by 2.24 and beats CT by 0.71. With COMETKiwi, TER surpasses SCoT by 2.10 and CT by 0.68. On the Bleurt metric, it exceeds SCoT by 2.48 and CT by 1.11. To confirm the versatility of the TER framework with different models, we evaluated Gemini-Pro and Claude-2 on our WMT23 Zh-En dataset, comparing them against all baselines. TER demonstrated impressive performance in these tests as well (see Table 8).

To further compare the translation quality of our TER method, we conduct pairwise human evaluations in Ja-En, En-Zh, and Zh-En translation directions. Figure 4 demonstrates that when compared with IT, the proportion of evaluators who preferred TER translations were 46.2%, 32.2%, and 42.5% respectively, all significantly exceeding the proportion favoring IT. This illustrates the foundational strength of TER in self-correcting the deficiencies of IT. When compared with SCoT and CT, TER was also generally more preferred by human evaluators. These advantages are also clear with Genmini-Pro and Claude-2.

5 Analysis

Zero-shot vs Few-shot Prompting We looked into using both zero-shot and few-shot prompting strategies for our TER framework. The results are shown in Table 1. We noticed that using few-shot prompting strategies generally works better than using zero-shot prompting strategies. Interestingly, we also find that employing zero-shot prompting strategies for estimation can sometimes deteriorate the BLEU score. This suggests, on one hand, the limitations inherent in string-based metrics and, on the other hand, implies potential deficiencies in the quality of zero-shot estimations.

To explore the impact of estimation quality in TER, we utilized GPT-4 with human assistance as gold standard estimation. Figure 5 showcases involving feedback with better estimation quality can help improve the effect of TER in COMET scores. Table 2 shows a comparison between zero-shot and few-shot estimation methods evaluated by human experts. The results favor few-shot approaches, highlighting their superior quality. Additionally, correlation analyses with MQM scores, using GPT-4 with human assistance as gold scores, underscore this finding. We observed that few-shot estimation notably outperforms zero-shot one, as detailed in Table 11. However, the modest correlation scores for few-shot estimation and minor enhancement in TER point to the estimation process as a critical bottleneck in optimizing self-correction efficacy.

Error Analysis with Different Estimation Strategies Figure 6 present the error analysis with different estimation strategies. We observed that both few-shot GPT-3.5-turbo and GPT-4 with human assistance predominantly identify errors in the categories of *Accuracy/Mistranslation* and *Style/Awkward*. The former tends to estimate more errors in the *Accuracy/Mistranslation* category, while GPT-4 attributes more errors to the

Models	En-Ru		En-De		He-En		Zh-En	
	System(%)	Segment(%)	System(%)	Segment(%)	System(%)	Segment(%)	System(%)	Segment(%)
GPT-3.5-turbo	66.67	23.41	78.79	**34.48**	74.24	19.57	**90.48**	30.36
Gemini-Pro	62.86	19.32	**86.36**	26.64	83.33	30.98	80.95	21.36
Claude-2	**69.52**	**26.49**	83.33	30.33	**92.42**	**33.05**	88.57	**34.75**

Table 5: The system and segment level results of metrics by various LLMs using pairwise accuracy (%) and Kendall correlation (%) with human-annotated MQM scores, respectively. **Bold** results indicate the best in each section.

Module			Language Pair			
Translate	Estimate	Refine	En-Ru	En-De	He-En	Zh-En
GPT-3.5-turbo	GPT-3.5-turbo	-	83.66	81.93	83.57	78.67
		GPT-3.5-turbo	84.41 (+0.75)	**83.20** (+1.27)	83.70 (+0.13)	**79.50** (+0.83)
		Gemini-Pro	84.59 (+0.93)	82.32 (+0.39)	83.68 (+0.11)	78.19 (-0.48)
		Claude-2	**86.50** (+2.84)	80.79 (-1.14)	**84.24** (+0.67)	78.51 (-0.16)
Gemini-Pro	Gemini-Pro	-	82.02	80.53	80.22	77.87
		GPT-3.5-turbo	**86.99** (+4.97)	**83.43** (+2.90)	**85.08** (+4.86)	**80.85** (+2.98)
		Gemini-Pro	84.67 (+2.65)	81.72 (+1.19)	82.96 (+2.74)	78.47 (+0.60)
		Claude-2	86.36 (+4.34)	81.15 (+0.62)	83.59 (+3.37)	79.57 (+1.70)
Claude-2	Claude-2	-	85.71	80.84	82.07	66.47
		GPT-3.5-turbo	85.62 (-0.09)	**82.79** (+1.95)	82.33 (+0.26)	**76.08** (+9.61)
		Gemini-Pro	**86.35** (+0.64)	82.51 (+1.67)	82.27 (+0.20)	75.09 (+8.62)
		Claude-2	86.23 (+0.52)	81.77 (+0.93)	**82.56** (+0.49)	74.69 (+8.22)

Table 6: Results of the error-correction capabilities of various LLMs. We conducted cross-correction experiments on our sampled WMT23 datasets by swapping out different models in the **Refine** module. "-" denotes the baseline without refinement. **Bold** results indicate the best in each section.

Style/Awkward category. Additionally, we find that zero-shot estimation tends to report more errors in categories like *Accuracy/Mistranslation*, *Accuracy/Addition*, and *Accuracy/Untranslated text*, which are considered *critical/major* in terms of severity. We also provide a typical case in Table 14, which demonstrates the preferences in different estimations. Considering these observations, we infer that weaker estimation strategies tend to overestimate translation errors. Table 3 further illustrates the growth in COMET scores after correcting different types of errors. When error types involve *Accuracy*, they usually pertain to higher severity levels of errors and be corrected more; whereas error types falling under the *Style* category are typically associated with lower severity errors.

LLMs Exhibit Diverse Translation and Estimation Capabilities. Table 4 and 5 present the results about how well different LLMs do in translation and estimation, based on few-shot prompting strategies. We observed that GPT-3.5-turbo performs best in translation for En-De and Zh-En, while Claude-2 excels in En-De and He-En. The average rankings for GPT-3.5-turbo, Gemini-Pro, and Claude-2 are 1.75, 2.5, and 1.75, respectively. We also find that Claude-2 achieves the highest scores in both System-level and Segment-level evaluation (see details in Appendix B) for En-Ru and He-En. As for En-De, the situation is somewhat complex, The ranking order exhibits significant differences between the system-level and segment-level evaluations. We hypothesize that the current MQM is primarily tailored for shorter sentences, potentially leading to reduced robustness when applied to longer paragraph-level tests. We consider both of these two levels in our subsequent analysis.

	En-Ru	En-De	He-En	Zh-En	Avg
System-level $\mathcal{M}$	1	-0.33	1	1	0.67
Segment-level $\mathcal{M}$	1	0.33	1	0.33	0.67

Table 7: The Kendall correlation between translation and translation evaluation capabilities. System-/Segment-level $\mathcal{M}$ means using evaluation rankings based on System-/Segment-level.

We further study the correlation between the translation and estimation capabilities of LLMs. We regard the translation rankings $\mathcal{R}_{\mathcal{M}^{x-y}_{translate}}$ from Table 4 and the estimation rankings $\mathcal{R}_{\mathcal{M}^{x-y}_{estimate}}$ from Table 5 to compute Kendall correlation. Table 7 highlights the consistency of translation and estimation capabilities in En-Ru and He-En, where the Kendall correlation scores are 1. This implies that models performing better in translation also tend to excel in evaluation. What's more, the consistency in En-De is not hypothetical, whether using system-level or segment-level evaluation metrics as a reference. This provides further evidence that using the existing MQM paradigm at paragraph level might not be robust.

Error Correction With Different LLMs Table 6 shows the error correction capabilities of different LLMs given the same translation and feedback. GPT-3.5-turbo and Claude-2, two LLMs with outstanding translation abilities as shown in Table 4, each excels in their top-ranked translation domains. When tested with the initialization settings of the other, both outperform their counterpart's self-correction in their respective top domains. Remarkably, GPT-3.5-turbo exceeds Claude-2 in all four translation directions under the settings of a third model. While Gemini-Pro trails behind the other two in its own self-correction scenarios, it doesn't always rank last under the settings of the other models. It's also noted that the capability for translation correction aligns well with translation performance in the En-De and Zh-En pairs. Intriguingly, these are the pairs where translation consistency with evaluation metrics appears weaker.

6 Related Work

LLMs for Machine Translation LLM-based MT falls into two main categories. The first focuses on developing strategies, including prompt design, in-context example selection, and more, as outlined in works (Vilar et al., 2023; Agrawal et al., 2023; Zhang et al., 2023; Garcia et al., 2023; Peng et al., 2023b). This category also encompasses the evaluation of MT in various contexts, such as low-resource, document-level, and multilingual translation, as discussed in studies (Hendy et al., 2023; Jiao et al., 2023; Karpinska and Iyyer, 2023; Zhu et al., 2023; Wang et al., 2023). For example, He et al. (2023) employ a selection mechanism based on external quality estimation to filter out unhelpful knowledge. Liang et al. (2023) introduce a Multi-Agent Debate (MAD) framework to encourage divergent thinking in LLMs to reach more accurate translations. The second focuses on training a specific translation LLM. Xu et al. (2023) propose a many-to-many LLM-based translation model, fine-tuning on monolingual data. Yang et al. (2023) construct a comprehensive parallel corpus dataset consisting of 102 languages to enhance the language capabilities of LLaMA. Wu et al. (2024) investigate the adaptation LLMs for document-level machine translation.

Besides LLM-based translation, recent studies also explore the use of LLMs as scorers to evaluate translation quality, introducing a new paradigm for automatic metrics (Kocmi and Federmann, 2023; Fernandes et al., 2023; Lu et al., 2023). Concurrent to our work, He et al. (2024) explore the use of external QE models as a reward mechanism in feedback training for machine translation. Overall, there is still no work that solely relies on an LLM to integrate evaluation and translation, using AI feedback to enhance translation quality.

Post-hoc Correction Automatic improvement is a line of research that utilizes LLMs to self-correct outputs. Self-Refine (Madaan et al., 2023) and Reflexion (Shinn et al., 2023) use LLMs to generate output, provide feedback, and refine the output based on that feedback. Recently, Gou et al. (2023) introduce CRITIC, a method designed to improve LLM outputs by integrating external feedback from diverse sources. Huang et al. (2023) examine the ability of LLMs to self-correct their reasoning responses, revealing that LLMs struggle with intrinsic self-correction. In applying post-hoc correction to MT, Chen et al. (2023a) explore using LLMs to iterate the translation, but it lacks clear guidance and the quality improvement is limited. Raunak et al. (2023) compare different post-editing strategies leveraging GPT-4, yet struggle with the lack of precise diagnostics and feedback mechanisms. In contrast, our work introduces a comprehensive self-correction translation framework that clearly separates the estimation module and introduces a structured feedback system, making the refinement process more interpretable and effective.

7 Conclusion

We introduce a self-correcting translation framework, TER, that employs large language models to perform expert-like guided revisions on translations. Experimental results demonstrate that TER surpasses existing post-editing methods in both metric scores and human preference. We conduct a comprehensive exploration of TER, showcasing the impact of different components on its performance. We delve into the estimate module and present an error-level analysis, demonstrating the key impact of estimation strategies. We also unveil potential connections between the model's translation capabilities and evaluation proficiency across various language pairs.

Limitations

Applying self-correction in the field of Machine Translation has inherent advantages, as the problem is well-defined, and the feedback information is highly directional and specific. However, identifying the optimal strategy for applying LLMs to self-correct translations remains a critical area for exploration. As we have uncovered that estimation may become a bottleneck limiting the effectiveness of TER, there is currently no universally proven estimation prompting strategies that demonstrate excellent results. What's more, due to limited computational resources, we only utilize black-box large language models in this work. It remains uncertain how the TER framework would perform on powerful open-source models.

References

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

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. Association for Computational Linguistics.

Anthropic. 2023. Model card and evaluations for claude models.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023a. Iterative translation refinement with large language models. *arXiv preprint arXiv:2306.03856*.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023b. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André FT Martins, Graham Neubig, Ankush Garg, Jonathan H Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

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. Association for Computational Linguistics.

Yuan Gao, Ruili Wang, and Feng Hou. 2023. Unleashing the power of chatgpt for translation: An empirical study. *arXiv preprint arXiv:2304.02182*.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. In *Proceedings of the 40th International Conference on Machine Learning*, pages 10867–10878. PMLR.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *ArXiv*, abs/2305.11738.

Thamme Gowda, Tom Kocmi, and Marcin Junczys-Dowmunt. 2023. Cometoid: Distilling strong reference-based machine translation metrics into Even stronger quality estimation metrics. In *Proceedings of the Eighth Conference on Machine Translation*, pages 751–755, Singapore. Association for Computational Linguistics.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection. *arXiv preprint arXiv:2310.10482*.

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*.

Zhiwei He, Xing Wang, Wenxiang Jiao, Zhuosheng Zhang, Rui Wang, Shuming Shi, and Zhaopeng Tu. 2024. Improving machine translation with human feedback: An exploration of quality estimation as a reward model. *arXiv preprint arXiv:2401.12873*.

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*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

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

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. *arXiv preprint arXiv:2304.03245*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 2023. Findings of the 2023 conference on machine translation (wmt23): Llms are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42.

Tom Kocmi and Christian Federmann. 2023. Gemba-mqm: Detecting translation quality error spans with gpt-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494.

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.

Miaoran Li, Baolin Peng, and Zhu Zhang. 2023. Self-checker: Plug-and-play modules for fact-checking with large language models. *arXiv preprint arXiv:2305.14623*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. *arXiv preprint*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023a. 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.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023b. Towards making the most of chatgpt for machine translation. *arXiv preprint arXiv:2303.13780*.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. 2022. MaTESe: Machine translation evaluation as a sequence tagging problem. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 569–577, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Vikas Raunak, Amr Sharaf, Hany Hassan Awadallah, and Arul Menezes. 2023. Leveraging gpt-4 for automatic translation post-editing. *arXiv preprint arXiv:2305.14878*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC De Souza, Taisiya Glushkova, Duarte Alves, Luísa Coheur, et al. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2022. Generating sequences by learning to self-correct. In *The Eleventh International Conference on Learning Representations*.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *arXiv preprint arXiv:2401.06468*.

Yangjian Wu and Gang Hu. 2023. Exploring prompt engineering with GPT language models for document-level machine translation: Insights and findings. In *Proceedings of the Eighth Conference on Machine Translation*, pages 166–169, Singapore. Association for Computational Linguistics.

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.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages. *arXiv preprint arXiv:2305.18098*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23.

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*.

A Dataset Statistics

Table 9 and 10 present details of our testset.

B Meta Evaluation of MQM

When comparing translation evaluation capabilities of different LLMs, we consider two dimensions used in WMT Shared Metrics task: system-level and segment-level. For the system-level, we utilize pairwise accuracy of system-ranking (Kocmi et al., 2021),

$$Acc = \frac{[sign(metric\Delta) == sign(human\Delta)]}{|all\ system\ pairs|}$$

where Δ represents the difference between the scores of the two systems. For the segment-level, we follow Freitag et al. (2022) to adopt the average of three types of Kendall correlation across all translation pairs.

C Primary Exploration for Iterative Self-Correction

In our primary experiments, we also explored whether continuously iterating the self-correcting process could enhance performance on the WMT23 Zh-En. We found that this might lead to a decrease in performance, with a single round of self-correcting yielding the best results (see Figure 7). We speculate that low-quality estimation from GPT-3.5 can produce hallucinations, thus leading to poor correction. Additionally, for errors that are easy to identify and correct, one round of self-correction is sufficient. However, for more challenging cases, one round or multiple rounds make a minor difference. We also consider this issue as a direction for future work.

D Exploring TER Components from Another Perspective

Table 12 shows the metrics against different parts of initial translations. Zero-shot prompting translations, in comparison to few-shot, have lower scores, indicating poorer quality. Under the same estimation and refinement strategies, the scores for zero-shot IT typically show greater enhancement. We also find that the average score of cases requiring refinement is lower than the all sampled dataset (see Table 1). Additionally, the relative score difference for few-shot is higher than for zero-shot, indicating that a stronger Estimate module (*few-shot*), compared to a weaker one (*zero-shot*), can

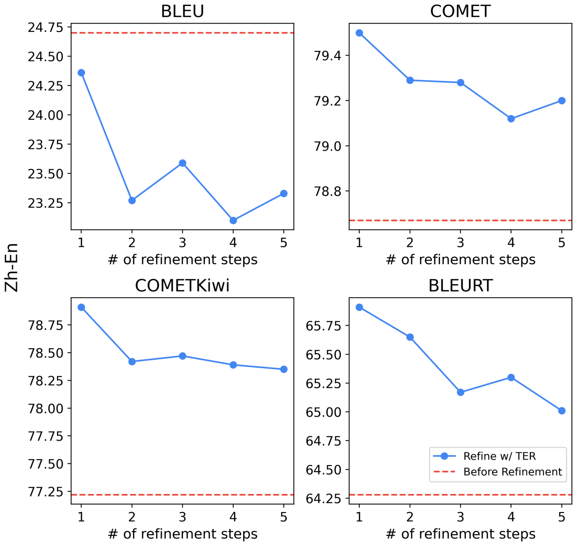


Figure 7: Metric scores for TER as a function of the number of refinement steps on Zh-En. Dashed red line denotes the initial translation baseline.

better discriminate and identify cases. Moreover, we observed using feedback alone is sufficient to improve scores, demonstrating the effectiveness of AI feedback in correcting translations. Compared to the feedback-only approach, the $\mathcal{T}_{refine} - \beta$ strategy further enhances the improvement effect.

E Different LLMs Leading Different Correction Execution Rate

Table 13 demonstrates that Gemini-Pro tends to tag fewer targets that need refinement, but always successfully execute correction. Figure 8 indicates that LLMs typically succeed in performing corrections when initialized with Gemini-Pro (i.e., using Gemini-Pro to translate and estimate). In Table 6, we observe that models generally achieve relatively high improvement scores when initialized with Gemini-Pro. This trend is not consistently present in GPT-3.5-turbo and Claude-2.

F Case Study

Table 14 showcases different estimation strategies working on the same initial translation. The weaker model and prompting method make it easy to produce hallucinations and overestimate the translation errors, and also lead to worse downstream refinement.

G Prompt Templates

Table 15, 16, 17, 18, 19, and 20 show the prompts we used in this work.

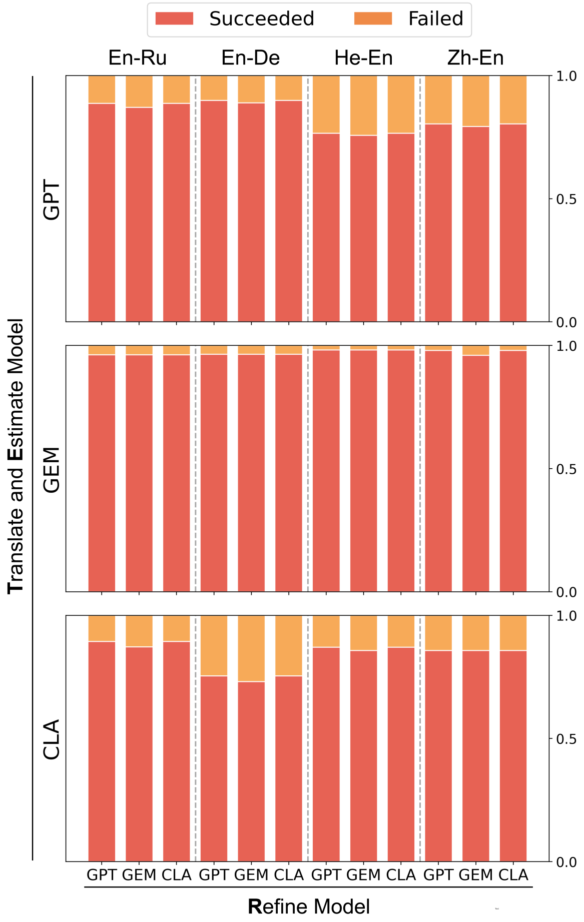


Figure 8: Correction execution rate under different initial settings. *GPT*: GPT-3.5-turbo; *GEM*: Gemini-Pro; *CLA*: Claude-2. We first use the same model for translation and estimation. Then, we do self-correction and cross-correction. *Succeeded* denotes the translation is modified after refinement. *Failed* represents the refined translation is the same as the initial translation.

	WMT23 Zh-En			
Method	**BLEU**	**COMET**	**COMETKiwi**	**BLEURT**
	GPT-3.5-turbo			
IT	24.70	78.67	83.57	78.67
SCoT	**25.10**	79.17	78.46	65.26
CT	23.96	77.98	77.10	64.34
TER	24.36	**79.50**	**78.91**	**65.91**
	Gemini-Pro			
IT	**22.17**	77.87	75.64	63.04
SCoT	19.26	76.96	75.55	63.36
CT	18.77	78.27	76.60	**64.93**
TER	19.84	**78.47**	**76.89**	64.19
	Claude-2			
IT	18.72	66.47	66.28	47.86
SCoT	16.04	69.23	69.37	52.38
CT	18.68	74.40	**76.00**	**57.39**
TER	**19.37**	**74.69**	**76.00**	57.21

Table 8: Results of the WMT23 Zh-En refined translations. The best results are in **bold**. TER is compared with three baseline methods (IT, SCoT, and CT) using GPT-3.5-turbo, Gemini-Pro, and Claude-2.

Dataset	Language Pair	Domain types	Total Segments	Sampled Segments	#tokens per segment
WMT22	En-Ru	News, E-commerce	2016	200	16.38
WMT23	En-De	Social, News, Meeting notes, E-commerce	557	200	59.46
WMT23	He-En	Social, News	1910	200	21.66
WMT23	Zh-En	Manuals, News, E-commerce	1976	200	22.85

Table 9: Statistics of our testset. *#tokens per segment* indicates the average number of tokens of the translation pairs, calculated based on the sampled text from the English portion of either the source or reference. *Systems*: translation systems that are annotated in the WMT for the given year.

Dataset	Language Pair	Segments	Systems	Total Segments	Systems Selected
WMT22	En-Ru	59	15	885	HuaweiTSC, JDExploreAcademy, Lan-Bridge, M2M100_1.2B-B4, Online-A, Online-B, Online-G, Online-W, Online-Y, PROMT, QUARTZ_TuneReranking, SRPOL, bleu_bestmbr, comet_bestmbr, eTranslation
WMT23	En-De	80	12	960	AIRC, GPT4-5shot, Lan-BridgeMT, NLLB_Greedy, NLLB_MBR_BLEU, ONLINE-A, ONLINE-B, ONLINE-G, ONLINE-M, ONLINE-W, ONLINE-Y, ZengHuiMT
WMT23	He-En	80	12	960	GTCOM_Peter, GPT4-5shot, Lan-BridgeMT, NLLB_Greedy, NLLB_MBR_BLEU, ONLINE-A, ONLINE-B, ONLINE-G, ONLINE-Y, ZengHuiMT, Samsung_Research_Philippines, UvA-LTL
WMT23	Zh-En	80	15	1200	ANVITA, GPT4-5shot, Lan-BridgeMT, NLLB_Greedy, NLLB_MBR_BLEU, ONLINE-A, ONLINE-B, ONLINE-G, ONLINE-M, ONLINE-W, ONLINE-Y, ZengHuiMT, HW-TSC, IOL_Research, Yishu

Table 10: For MQM annotated datasets, we exclude data with missing annotations and sample 80 translation pairs form the former 200 sampled translation pairs, except for En-Ru, where the count is 59. The selected testing systems vary across different language pairs.

	Kendall (%)	Pearson (%)
zero-shot	11.93	8.91
few-shot	20.22	18.72

Table 11: The Kendall and Pearson correlation between zero/few-shot estimation scores (MQM typology) using GPT-3.5-turbo and gold GPT-4 with human assistance score.

Translate		Estimate		Refine		# of refinements	BLEU	COMET	COMETKiwi	BLEURT
0-shot	**5-shot**	**0-shot**	**3-shot**	$\mathcal{T}_{refine} - \alpha$	$\mathcal{T}_{refine} - \beta$		**Score (Δ)**	**Score (Δ)**	**Score (Δ)**	**Score (Δ)**
✓		✓		-	-	123	25.51	79.78	78.85	65.84
✓		✓		✓			24.29 (-1.22)	79.85 (+0.07)	78.98 (+0.13)	65.98 (+0.14)
✓		✓			✓		24.54 (-0.97)	80.12 (+0.34)	79.35 (+0.50)	66.42 (+0.58)
✓			✓	-	-	99	24.24	78.97	76.93	65.86
✓			✓	✓			24.15 (-0.09)	80.00 (+1.03)	78.71 (+1.78)	66.77 (+0.91)
✓			✓		✓		24.60 (+0.36)	80.32 (+1.35)	78.75 (+1.82)	66.94 (+1.08)
	✓	✓		-	-	160	27.06	80.07	79.17	66.75
	✓	✓		✓			26.59 (-0.47)	80.23 (+0.16)	79.72 (+0.55)	66.96 (+0.21)
	✓	✓			✓		26.71 (-0.35)	80.35 (+0.28)	80.01 (+0.84)	67.07 (+0.32)
	✓		✓	-	-	94	24.70	78.67	77.22	64.28
	✓		✓	✓			23.66 (-1.04)	79.12 (+0.45)	78.33 (+1.11)	65.53 (+1.25)
	✓		✓		✓		24.36 (-0.34)	79.50 (+0.83)	78.91 (+1.69)	65.91 (+1.63)

Table 12: Comparison of quality improvement for different variants of the TER modules on our WMT23 Zh-En refined translations. *# of refinements* refers to how many cases are judged to need refinement by the **Estimate** module. $\mathcal{T}_{refine} - \alpha$ uses only the feedback from the **Estimate** module, $\mathcal{T}_{refine} - \beta$ uses few-shot exemplars in the **Translate** module and the feedback. Δ indicates the relative quality against the results without **Refine** module.

Models	En-Ru		En-De		He-En		Zh-En	
	CN	CU	CN	CU	CN	CU	CN	CU
GPT-3.5-turbo	62	7	99	10	130	37	94	17
Gemini-Pro	52	2	45	1	56	2	100	4
Claude-2	94	10	130	32	70	8	42	7

Table 13: The statistics of cases that require refinement after estimation, "CN" represents the cases that need to be corrected after estimation, "CU" represents the cases in "CN" but remaining unchanged after refinement.

Case	
Source	我惊愕了，音质竟然是环绕3D立体！
Reference	I was astounded to discover that the sound quality was surround 3D!
IT	I was amazed that the sound quality was actually 3D surround!
Estimate (reference-free)	
0-shot	Based on the source segment and machine translation, the errors in the translation can be identified and classified as follows: {"critical": accuracy/mistranslation - ’was actually’, "major": " "minor":" "}. The critical error is the mistranslated text. This error inhibits comprehension of the text.
3-shot	"critical": no-error, "major": accuracy/mistranslation - ’actually’ for ’竟然’", "minor": style/awkward - ’I was amazed that’"
GPT-4 w/ human	"critical": no-error, "major": no-error, "minor": style/awkward - ’I was amazed that’"
Refine (with feedback from estimation)	
Fdb (0-shot)	I was amazed that the sound quality was actually 3D surround!
Fdb (3-shot)	I was astonished that the sound quality was actually 3D surround!
Fdb (GPT-4 w/ human)	I was astonished that the sound quality was actually 3D surround!

Table 14: Case study of different estimation feedback. Top: translation defects annotated by human expert (in red); Middle: translation defects estimated successfully by models (in red), overestimated errors (under lines); Bottem: unsuccessfully corrected (in red); successfully corrected (in blue). *Fdb*: source of the estimation feedback.

Zero-shot Translate
Please provide the {tgt_lan} translation for the {src_lan} sentences: Source: {origin} Target:

Table 15: **Zero-shot Translate Prompt**. {tgt_lan}: target language; {src_lan}: source language; {origin}: the source test sentence.

Few-shot Translate
Please provide the {tgt_lan} translation for the {src_lan} sentences: Example: Source: {src_example_1} Target: {tgt_example_1} Source: {src_example_2} Target: {tgt_example_2} Source: {src_example_3} Target: {tgt_example_3} Source: {src_example_4} Target: {tgt_example_4} Source: {src_example_5} Target: {tgt_example_5} Source: {origin} Target:

Table 16: **Few-shot Translate Prompt**. {tgt_lan}: target language; {src_lan}: source language; {src_example_i}: the source sentence of example i; {tgt_example_i}: the target sentence of example i; {origin}: the source test sentence.

Zero-shot Estimate
Please identify errors and assess the quality of the translation. The categories of errors are accuracy (addition, mistranslation, omission, untranslated text), fluency (character encoding, grammar, inconsistency, punctuation, register, spelling), locale convention (currency, date, name, telephone, or time format), style (awkward), terminology (inappropriate for context, inconsistent use), non-translation, other, or no-error. Each error is classified as one of three categories: critical, major, and minor. Critical errors inhibit comprehension of the text. Major errors disrupt the flow, but what the text is trying to say is still understandable. Minor errors are technical errors but do not disrupt the flow or hinder comprehension. {src_lan} source: {origin} {tgt_lan} translation: {init_trans} MQM annotations:

Table 17: **Zero-shot Estimate Prompt**. {src_lan}: source language; {origin}: the source test sentence; {tgt_lan}: target language; {init_trans}: the initial translation of the source test sentence.

Few-shot Estimate

Please identify errors and assess the quality of the translation.
The categories of errors are accuracy (addition, mistranslation, omission, untranslated text), fluency (character encoding, grammar, inconsistency, punctuation, register, spelling),
locale convention (currency, date, name, telephone, or time format) style (awkward), terminology (inappropriate for context, inconsistent use), non-translation, other, or no-error.\n
Each error is classified as one of three categories: critical, major, and minor. Critical errors inhibit comprehension of the text. Major errors disrupt the flow, but what the text is trying to say is still understandable. Minor errors are technical errors but do not disrupt the flow or hinder comprehension.
Example1:
Chinese source: 大众点评乌鲁木齐家居商场频道为您提供居然之家地址，电话，营业时间等最新商户信息， 找装修公司，就上大众点评
English translation: Urumqi Home Furnishing Store Channel provides you with the latest business information such as the address, telephone number, business hours, etc., of high-speed rail, and find a decoration company, and go to the reviews.
MQM annotations:
critical: accuracy/addition - "of high-speed rail"
major: accuracy/mistranslation - "go to the reviews"
minor: style/awkward - "etc.,"
Example2:
English source: I do apologise about this, we must gain permission from the account holder to discuss an order with another person, I apologise if this was done previously, however, I would not be able to discuss this with yourself without the account holders permission.
German translation: Ich entschuldige mich dafür, wir müssen die Erlaubnis einholen, um eine Bestellung mit einer anderen Person zu besprechen. Ich entschuldige mich, falls dies zuvor geschehen wäre, aber ohne die Erlaubnis des Kontoinhabers wäre ich nicht in der Lage, dies mit dir involvement.
MQM annotations:
critical: no-error
major: accuracy/mistranslation - "involvement"
  accuracy/omission - "the account holder"
minor: fluency/grammar - "wäre"
  fluency/register - "dir"
Example3:
English source: Talks have resumed in Vienna to try to revive the nuclear pact, with both sides trying to gauge the prospects of success after the latest exchanges in the stop-start negotiations.
Czech transation: Ve Vídni se ve Vídni obnovily rozhovory o oživení jaderného paktu, přičemže obě partaje se snaží posoudit vyhlídky na úspěch po posledních výměnách v jednáních.
MQM annotations:
critical: no-error
major: accuracy/addition - "ve Vídni"
  accuracy/omission - "the stop-start"
minor: terminology/inappropriate for context - "partake"

{src_lan} source: {origin}
{tgt_lan} translation: {init_trans}
MQM annotations:

Table 18: **Few-shot Estimate Prompt**. {src_lan}: source language; {origin}: the source test sentence; {tgt_lan}: target language; {init_trans}: the initial translation of the source test sentence.

$\mathcal{T}_{refine} - \alpha$

Please provide the {tgt_lan} translation for the {src_lan} sentences.
Source: {raw_src}
Target: {raw_mt}
I'm not satisfied with this target, because some defects exist: {estimate_fdb}
Critical errors inhibit comprehension of the text. Major errors disrupt the flow, but what the text is trying to say is still understandable. Minor errors are technical errors but do not disrupt the flow or hinder comprehension.
Upon reviewing the translation examples and error information, please proceed to compose the final {tgt_lan} translation to the sentence: {raw_src}. First, based on the defects information locate the error span in the target segment, comprehend its nature, and rectify it. Then, imagine yourself as a native {tgt_lan} speaker, ensuring that the rectified target segment is not only precise but also faithful to the source segment.

Table 19: **Refine Prompt** $\mathcal{T}_{refine} - \alpha$. {tgt_lan}: target language; {src_lan}: source language; {raw_src}: the source test sentence; {raw_mt}: the initial translation of the source test sentence; {estimate_fdb}: the estimation feedback.

$\mathcal{T}_{refine} - \beta$

Please provide the {tgt_lan} translation for the {src_lan} sentences.
Example:
Source: {src_example_1} Target: {tgt_example_1}
Source: {src_example_2} Target: {tgt_example_2}
Source: {src_example_3} Target: {tgt_example_3}
Source: {src_example_4} Target: {tgt_example_4}
Source: {src_example_5} Target: {tgt_example_5}
Now, let's focus on the following {src_lan}-{tgt_lan} translation pair.
Source: {raw_src}
Target: {raw_mt}
I'm not satisfied with this target, because some defects exist: {estimate_fdb}
Critical errors inhibit comprehension of the text. Major errors disrupt the flow, but what the text is trying to say is still understandable. Minor errors are technical errors but do not disrupt the flow or hinder comprehension.
Upon reviewing the translation examples and error information, please proceed to compose the final {tgt_lan} translation to the sentence: {raw_src}. First, based on the defects information locate the error span in the target segment, comprehend its nature, and rectify it. Then, imagine yourself as a native {tgt_lan} speaker, ensuring that the rectified target segment is not only precise but also faithful to the source segment.

Table 20: **Refine Prompt** $\mathcal{T}_{refine} - \beta$. {tgt_lan}: target language; {src_lan}: source language; {src_example_i}: the source sentence of example i; {tgt_example_i}: the target sentence of example i; {raw_src}: the source test sentence; {raw_mt}: the initial translation of the source test sentence; {estimate_fdb}: the estimation feedback.

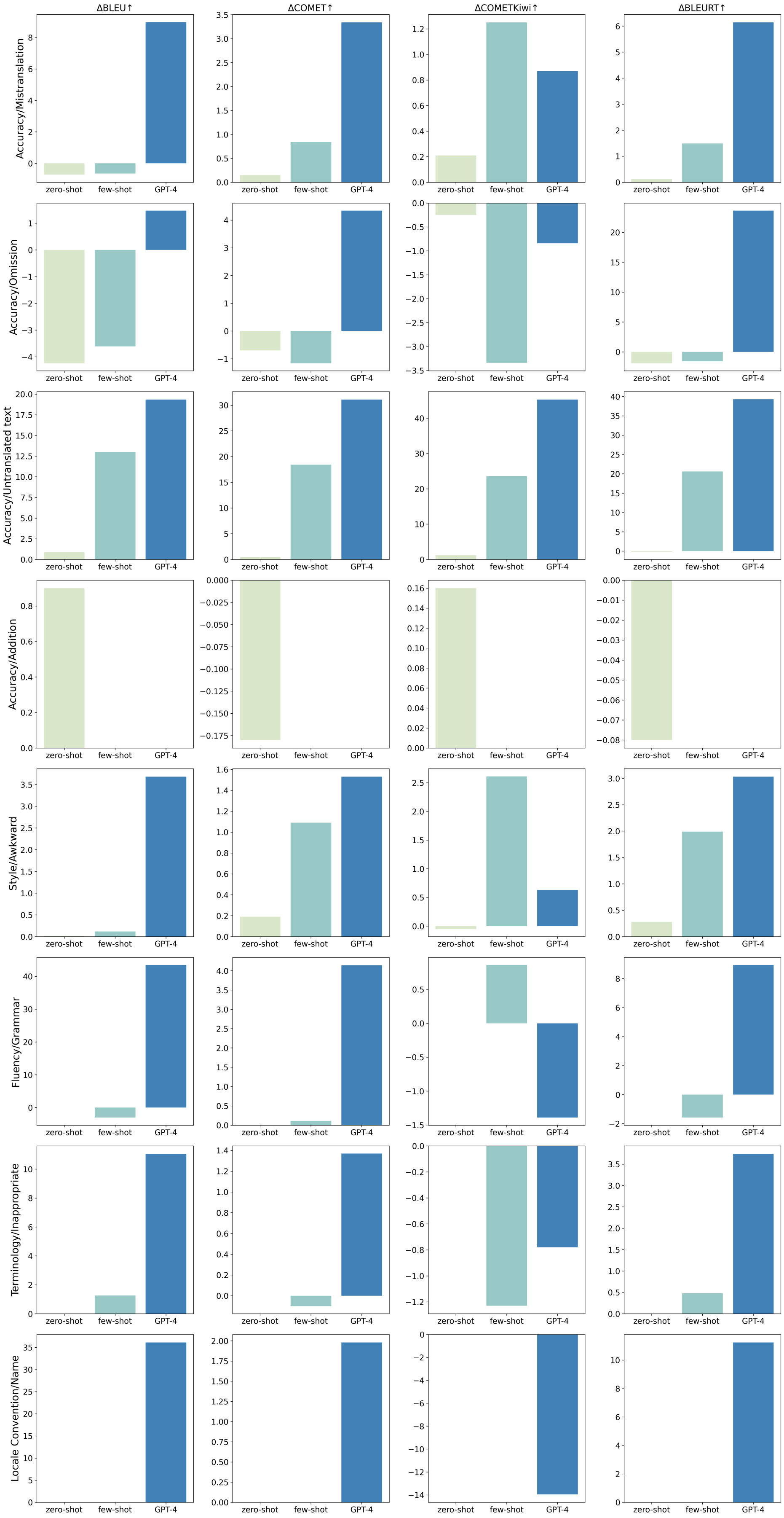


Figure 9: Improvement performance against the initial translation based on error type using different **Estimate** strategies. *GPT-4* denotes using GPT-4 with human estimation.