Aligning Translation-Specific Understanding to General Understanding in Large Language Models

Yichong Huang[†], **Baohang Li**[†], **Xiaocheng Feng**[†‡], **Wenshuai Huo**[†‡], **Chengpeng Fu**[†‡], **Ting Liu**[†], **Bing Qin**[†‡✉]

[†]Harbin Institute of Technology [‡] Peng Cheng Laboratory

{ychuang,xcfeng,baohangli,cpfu,wshuo,tliu,qinb}@ir.hit.edu.cn

Abstract

Large Language models (LLMs) have exhibited remarkable abilities in understanding complex texts, offering a promising path towards human-like translation performance. However, this study reveals the misalignment between the translation-specific understanding and the general understanding inside LLMs. This understanding misalignment leads to LLMs mistakenly or literally translating some complicated concepts that they accurately comprehend in the general scenarios (*e.g.,* QA). To align the translation-specific understanding to the general one, we propose a novel translation process, DUAT (**D**ifficult words **U**nderstanding **A**ligned **T**ranslation), explicitly incorporating the general understanding on the complicated content incurring inconsistent understanding to guide the translation. Specifically, DUAT performs cross-lingual interpretation for the difficult-to-translate words and enhances the translation with the generated interpretations. Furthermore, we reframe the external tools to improve DUAT in detecting difficult words and generating helpful interpretations. We conduct experiments on the self-constructed benchmark Challenge-WMT[1], consisting of samples that are prone to mistranslation. Human evaluation results on high-resource and low-resource language pairs indicate that DUAT significantly facilitates the understanding alignment, which improves the translation quality (up to +3.85 COMET) and reduces the literality of the translation by - 25% $\sim$ - 51%.

1 Introduction

Recently, large language models (LLMs) have demonstrated remarkable language understanding and generation, paving the way for a higher level of performance in machine translation (Zhao et al., 2023; OpenAI, 2023; Jiang et al., 2023; Workshop, 2023). However, existing research reports that LLMs have yet to achieve as significant advances in machine translation as they have achieved in other natural language processing fields (Hendy et al., 2023; Pang et al., 2024; Zhang et al., 2023a; Jiao et al., 2023; Zhu et al., 2023b).

Question	In this Chinese sentence: "文章的前妻是马伊琍", what is the meaning of "文章"?
LLM's Answer	It refers to the **Chinese actor and singer, Wen Zhang**.

(a) LLM ***correctly understands*** the meaning of "文章" when explaining its meaning. ✓

Source Sentence	文章的前妻是马伊琍。
Reference Translation	The ex-wife of **Wen Zhang** is Ma Yili.
LLM's Translation	The ex-wife of the **article** is Ma Yili.

(b) LLM ***misunderstands*** the word "文章" as the "**creative work**" when translating the sentence. ✗

Figure 1: Illustration of the misalignment between the general understanding (Fig a) and the translation-specific language understanding (Fig b) inside the LLM (`gpt-3.5-turbo-0125`). More examples are reported in Appendix A.

In this study, we discover the misalignment between the general understanding and translation-specific understanding inside LLMs, as illustrated in Fig.1. This understanding misalignment leads to LLMs mistakenly or literally translating some complicated concepts that they accurately comprehend in general scenarios. We refer to these failures as language models' **generalization failures** on translation. Human evaluation on a total of 600 sampled sentences across six language pairs show that generalization failures account for a considerable proportion of all mistranslations (**16%-32%**), indicating serious understanding misalignment (§6.1).

To align the translation-specific understanding to the general one, we propose a novel translation process, DUAT (Difficult words Understanding Aligned Translation), explicitly incorporating the general understanding on the complicated con-

[1]The dataset is available at: ChallengeWMT

✉ means corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 5028–5041

tent incurring inconsistent understanding to guide the translation. Specifically, DUAT first detects the difficult-to-translate words in the source sentence, which could cover the generalization failures intuitively. Next, the LLM is prompted to interpret each difficult word with the target language, *i.e.,* cross-lingual interpretation, unleashing the powerful general understanding and transforming this understanding into the target language space. After that, DUAT conducts translation under the guidance of these interpretations. Unlike the CoT-based process mimicking junior translators to sequentially translate all words (Peng et al., 2023), DUAT works like senior translators to analyze the complicate words, which helps the model deep understand the source sentence and produces more nuanced translations. Furthermore, we reframe the external tool of token-level QE (Rei et al., 2023) to enhance the detection of difficult words, and design a strategy of interpretation quality control to filter hallucinated interpretations based on sentence-level QE (Rei et al., 2020).

To better analyze the understanding misalignment, we proposed the Challenge-WMT benchmark, which contains more sentences prone to mistranslation. These sentences were collected from multi-year WMT datasets and represent difficult samples that multiple state-of-the-art (SOTA) systems translate poorly. Human evaluation results indicate that DUAT significantly facilitates the understanding alignment, reducing 80%∼88% of generalization failures. Moreover, this alignment improves the translation quality, as evidenced by automatic metrics (up to +3.85 COMET), and alleviates translation literalness by -25% ∼ -51%.

2 Background

2.1 LLM-based MT

Considering the translation from source language L_s to target language L_t, LLM-based machine translation converts the source sentence x to an instruction using a translation-specific template and generates the translation by feeding the instruction to the LLM θ. To make the LLM better follow the instruction, the in-context learning (ICL) strategy (Brown et al., 2020; Dong et al., 2023) injects a few examples/demonstrations of translation into the instruction, which is shown as:

> Request: Please translate the $[L_s]$ sentence into $[L_t]$.
> *# followed by* [N Demonstrations $\mathcal{E}^{mt}$]
> Source Sentence: [Source Sentence x]

Formally, the LLM-based MT generates the translation with ICL as:

$$\hat{y} = argmax\, P_\theta(\mathcal{E}^{mt}, x), \tag{1}$$

where $\mathcal{E}^{mt} = \{(x^i, y^i)\}_{i=1}^{N}$ is the demonstrations set of translation.

2.2 Quality estimation (QE)

QE for machine translation, *i.e.,* reference-free MT evaluation, aims to predict the quality of the given translation only according to the source sentence, which has shown auspicious correlations with human judgments (Rei et al., 2020, 2021). Given a source sentence x and a translation y, QE score is denoted as $\psi(y \mid x)$.

Thanks to the recent advance in the interpretability of neural MT metrics (Rei et al., 2023), token-level QE is proposed to score the error degree of the given translation span by calculating the misalignment of this span against the source sentence. Given a source sentence x and the candidate translation $\widetilde{y}$, token-level QE $\phi(\cdot)$ annotates the error degree of the specific span w^t in the translation, *i.e.,* $\phi(w^t \mid \widetilde{y}, x)$ where $w^t \in \widetilde{y}$.

3 Approach: DUAT

In this section, we first introduce our translation framework DUAT (§3.1). Specifically, DUAT consists of three components: *difficult word detection* (§3.2), *cross-lingual interpretation* (§3.3), and *interpretation quality control* (§3.4). To make the LLM follow the procedure of each component as expected, we adopt the in-context learning strategy and design an automatic method for constructing demonstrations of DUAT (§3.5).

3.1 Framework

The progress of DUAT is illustrated in Fig.2. Given the source sentence, DUAT first detects the difficult words or phrases in the source sentence. Once the difficult words are identified, DUAT requests the LLM to interpret each difficult word with the target language, unleashing the powerful understanding capability inside the LLM and transforming these understandings into the target language space. Finally, to avoid the interference of incorrect and useless interpretations, DUAT removes the negative interpretations through the interpretation quality control and outputs the final translation guided by the helpful interpretations.

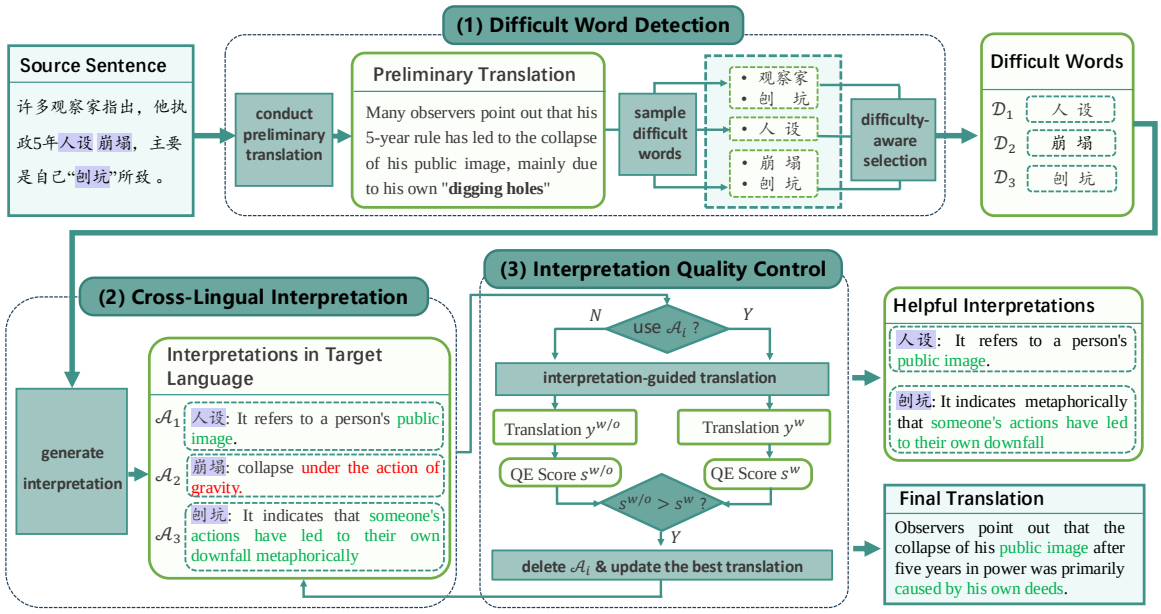

Figure 2: DUAT framework. The purple spans indicate the difficult-to-translate words, the green spans indicate the correct translation/interpretation, and the red spans indicate the incorrect ones.

3.2 Step-1: Difficult Word Detection

In practice, we found that directly inquiring LLMs to identify difficult-to-translate words is challenging. To tackle this challenge, we first conduct a preliminary translation for the given source sentence and then extract the **mistranslated words and phrases** in the source sentence **as the difficult words**. Concretely, we invent DUAT-I to do this leveraging the ***Intrinsic*** ability of LLMs at first.

DUAT-I. Given source sentence x, DUAT-I first obtains the preliminary translation $\widetilde{y}$ (also known as *draft translation*) by prompting the LLM to translate x with the in-context learning strategy, which is shown in Eq. (1). Next, the LLM is requested to output the difficult words based on the source sentence and the preliminary translation:

> Request: Given a $[L_s]$ sentence and its draft $[L_t]$ translation, output the mistranslated words and phrases in the $[L_s]$ sentence.
> *# followed by* [N Demonstrations $\mathcal{E}^{diff}$]
> Source Sentence: [Given Sentence x]
> Draft Translation: [Draft Translation $\widetilde{y}$]

DUAT-I obtains the difficult word list $\mathcal{D}$ via performing *greedy decoding* on the LLM:

$$\mathcal{D} = argmax\, P_\theta(\mathcal{E}^{diff}, x, \widetilde{y}\,), \tag{2}$$

where θ is the LLM, which is prompted with N demonstrations of difficult word detection $\mathcal{E}^{diff} = \{x^i, \widetilde{y}^i, \mathcal{D}^i\}_{i=1}^{N}$.

DUAT-E. It is frequently observed that the LLM fails to recognize the mistranslated words, due to their limitation in self-knowledge (Yin et al., 2023) Therefore, we devise DUAT-E to boost the detection with the ***External*** tool. First, DUAT-E requests the LLM with the same prompt as DUAT-I while performing *temperature sampling* for K times. Next, the union of all sampling results is taken as the candidate set of difficult words $\mathcal{D}^{cand}$:

$$\mathcal{D}^{cand} = \bigcup_{k=1}^{K} \mathcal{D}_k \sim P_\theta(\mathcal{E}^{diff}, x, \widetilde{y}, T), \tag{3}$$

where T is the hyperparameter of sampling temperature, which is set to 0.5 to capture more candidates, and K is set to 5 empirically.

Finally, DUAT-E annotates each candidate word with its degree of misalignment with respect to the draft translation, which reflects the translation-specific difficulty. To implement this function, we adopt an external tool of token-level QE $\phi(\cdot)$. As shown in §2.2, token-level QE is originally used to score the mistranslation degree of the given translation span with respect to the source sentence, *i.e.,* $\phi(w^t \mid \widetilde{y}, x)$ where $w^t \in \widetilde{y}$. Differently, we utilize this tool in a dual manner. That is, we use $\phi(\cdot)$ to annotate the misalignment degree of the given ***source*** span with respect to the translation, *i.e.,* $\phi(w^s \mid x, \widetilde{y}\,)$ where $w^s \in x$. Formally, the misalignment score of each difficult word candidate is calculated as:

$$\phi(d) = \phi(d \mid x, \widetilde{y}\,), d \in \mathcal{D}^{cand}. \tag{4}$$

Then, DUAT-E selects candidates with misalignment score $\phi(d) > \tau$, where τ is the hyperparameter named the difficulty threshold. We refer to this procedure as the *difficulty-aware selection* in Fig.2.

3.3 Step-2: Cross-Lingual Interpretation

After the difficult words in the source sentence are detected, DUAT lets the LLM generate the interpretation of each difficult word via requesting:

> Request: Given a $[L_s]$ sentence, provide the concise interpretation for each difficult word with the $[L_t]$.
>
> *# followed by* [N Demonstrations $\mathcal{E}^{intp}$]
>
> Source Sentence: [Given Sentence x]
> Difficult Words: [Difficult Words $\mathcal{D}$]

Through access to the LLM, the interpretation set $\mathcal{A}$ is obtained:

$$\mathcal{A} = argmax\, P_\theta(\mathcal{E}^{intp}, x, \mathcal{D}), \tag{5}$$

where $\mathcal{E}^{intp} = \{x^i, \mathcal{D}^i, \mathcal{A}^i\}_{i=1}^N$, which is the demonstrations of the cross-lingual interpretation.

Prob and cons. Under the guidance of generated interpretations, DUAT can align the translation-specific understanding to the general one. However, LLMs may generate hallucinated interpretations sometimes (*e.g.,* the interpretation of "崩塌" in Fig.2), which biases the resulting translation from the original semantics. Besides, helpless interpretations that can not provide useful information also pose a risk of disturbing the translation process.

3.4 Step-3: Interpretation Quality Control

To overcome the potential interference of the generated negative interpretations, DUAT removes them through the interpretation quality control (**IQC**) and outputs the final translation guided by the helpful interpretations.

Concretely, given a set of interpretations $\mathcal{A}$, DUAT ablates each interpretation $\mathcal{A}_i$ sequentially and uses the remaining interpretations to guide the translation. The **interpretation-guided translation** is implemented in a fashion of *refinement*:

> Request: Given a $[L_s]$ sentence and its draft $[L_t]$ translation, please revise the translation according to the interpretations of the difficult words.
>
> *# followed by* [N Demonstrations $\mathcal{E}^{igt}$]
>
> Source Sentence: [Given Sentence x]
> Draft Translation: [Draft Translation $\widetilde{y}$]
> Interpretations of Difficult Words:
>
> [Interpretations $\mathcal{A}$]

Formally, the translation is obtained as:

$$\hat{y} = argmax\, P_\theta(\mathcal{E}^{igt}, x, \widetilde{y}, \mathcal{A}), \tag{6}$$

where $\mathcal{E}^{igt} = \{x^i, \widetilde{y}^i, \mathcal{A}^i, \hat{y}^i\}$, which is the demonstration set of interpretation-guide translation.

If the better translation performance is achieved by ablation, which is measured by the QE[2] tool due to the unavailable access to the reference translation, the interpretation $\mathcal{A}_i$ is removed from $\mathcal{A}$ and the current translation is taken as the best translation. We also detail this process in Alg.1.

3.5 Demonstrations Synthesis for DUAT

To make the LLM follow the procedure of DUAT as expected, we adopt the ICL strategy. Common practice constructs demonstrations manually, necessitating human translators proficient in $N \times (N-1)$ language pairs for N languages. To overcome this considerable cost, we devise a method for synthesizing high-quality demonstrations of DUAT based on parallel data. This process of synthesizing demonstrations is accomplished in a manner of post-explanation by asking the LLM to compare the baseline translation and the reference translation. We describe this process in Appendix B.2.

4 Testbed: Challenge-WMT

To better analyze the understanding misalignment problem of LLMs, we propose the benchmark **Challenge-WMT**, which contains challenge sentences that are prone to mistranslation. This benchmark is constructed by collecting samples that multiple SOTA systems translate poorly from multi-year WMT testsets of six language pairs (*Chinese* (zh), *Estonian* (et), and *Icelandic* (is) to/from *English* (en)). Additionally, we believe that this dataset could promote future research in understanding the limitations of existing MT systems.

We select three SOTA MT systems: Google Translate, ChatGPT, and NLLB (NLLB Team et al., 2024). Due to the poor performance of NLLB in the zh⇔en translation, we additionally train a zh↔en translation model based on DeltaLM (Ma et al., 2021) on the parallel corpus from OPUS[3]. Next, all of the system translations are scored with COMET metric, and the ρ of samples with the lowest score for each system are extracted as its difficult samples set. We vary the value of ρ across different language pairs to ensure an appropriate

[2]We use wmt21-comet-qe-da as the QE scorer.
[3]https://opus.nlpl.eu/

Methods	En⇒Zh		Zh⇒En		En⇒Et		Et⇒En		En⇒Is		Is⇒En		Average	
	COMET	BLEURT	COMET	BLEURT	COMET	BLEURT	COMET	BLEURT	COMET	BLEURT	COMET	BLEURT	COMET	BLEURT
Existing Systems														
Google	74.85	54.95	68.21	52.65	79.11	68.71	78.83	65.46	76.17	59.67	78.70	66.54	75.98	61.33
NLLB	68.77	47.77	60.09	45.14	74.20	63.13	74.35	60.87	69.37	52.08	72.55	59.66	69.89	54.78
GPT-4	76.15	55.37	70.77	58.85	80.25	70.80	77.83	65.48	**77.33**	**61.15**	**79.39**	**68.40**	76.95	63.34
Baselines														
Zero-shot	74.89	54.25	71.27	58.24	80.67	69.10	74.93	61.75	71.17	53.12	76.22	64.05	74.86	60.09
ICL	75.47	55.79	72.22	59.56	80.9	69.93	79.40	66.63	73.19	54.87	77.52	65.81	76.45	62.10
+CoT	73.85	53.42	71.35	57.90	78.03	66.03	76.78	63.97	69.72	50.98	76.55	64.54	74.38	59.47
+Keywords	73.93	55.10	71.22	59.18	78.63	70.01	77.79	66.30	70.33	54.31	74.55	64.40	74.41	61.55
+Topic	75.83	53.60	72.46	57.98	80.98	66.78	79.20	64.64	72.77	51.98	76.49	62.06	76.29	59.51
+SimDems	75.22	55.10	72.20	59.16	81.24	70.42	79.11	66.29	72.70	54.04	76.78	64.43	76.21	61.57
Ours														
DUAT-I	76.92	56.16	72.94	59.94	82.92	72.19	79.96	**67.05**	76.64	57.71	78.45	66.80	77.97	63.31
DUAT-E	**77.57**	**56.86**	**73.25**	**60.16**	**83.07**	**72.30**	**80.01**	66.91	77.04	57.38	78.70	66.93	**78.27**	**63.42**

Table 1: Main results on Challenge-WMT. The bold indicates the highest value. '+SimDems' represents the translation strategy with demonstrations similar to the source sentence. The strategies '+Topic', '+Keywords', and '+SimDems' are proposed in MAPS. The baselines and our approaches are implemented based on `GPT-3.5-turbo`.

scale for each difficult sample set. Finally, the intersection of all systems' difficult sample sets is taken as the Challenge-WMT testbed. Challenge-WMTcomprises around 600+ sentence pairs for each language pair, which is illustrated in Tab.6. We equally split this dataset into the validation set and the test set.

We report the translation performance measured by COMET on Challenge-WMT and the complete WMT set in Fig.8, which shows that the performance decreases dramatically on Challenge-WMT (84.5⇒73.6 averagely). Next, we conduct a multi-aspect comparison for Challenge-WMT and the complete set in Appendix C, and find that the samples of Challenge-WMT have higher perplexity (214⇒252 averagely). This result indicates that **sentences in Challenge-WMT are more complex**.

5 Experiments

5.1 Experimental Setup

Comparative Methods. We verify the effectiveness of our DUAT on the LLM `GPT-3.5-turbo` for its promising capability in following complicated instructions. Demonstrations of DUAT are gained by performing our automatic method (§3.5) on the validation set of Challenge-WMT. We compare DUAT with the following methods:

- **Zero-shot**, which asks the LLM to translate the source sentence directly.
- **ICL** (In-Context Learning), enhancing the translation with K randomly selected exemplars from the validation set.
- **CoT** (Wei et al., 2022), encouraging the LLM to resolve the problem step by step. In this work, we re-implement CoT by prompting the LLM to translate the source sentence step by step.
- **MAPS** (He et al., 2024), incorporating the knowledge of *keywords*, *topic words*, and *demonstrations similar* to the given source sentence to enhance the translation process, respectively.
- Commercial and open-source systems. We also report the performance of **Google** Translate, **NLLB** (in zh⇔en translation, we replace NLLB with our trained MT model based on DeltaLM), and zero-shot translation based on **GPT4** (`GPT-4-turbo`).

For DUAT and other ICL-based methods, we select K=8 demonstrations (*i.e.,* 8-shot) to achieve a strong baseline performance. More details of re-implementing the baselines under the few-shot setting are illustrated in Appendix D.

Metrics. Following previous research of LLM-based MT (Garcia et al., 2023; Chen et al., 2023), we adopt COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020) as the evaluation metrics as their high correlations with human judgment than BLEU (Papineni et al., 2002).

5.2 Results on Challenge-WMT

The main results are illustrated in Tab.1. From the results, we have drawn the following observations:

Methods	En⇒Is		Is⇒En	
	COMET	BLEURT	COMET	BLEURT
Zero-shot	77.33	61.15	79.39	68.40
ICL	80.1	61.99	81.02	**70.20**
DUAT-E	**81.7**	**64.01**	**81.21**	69.22

Table 2: Results in En⇔Is translation based on GPT-4.

Methods	En⇒Zh		Zh⇒En	
	COMET	BLEURT	COMET	BLEURT
WMT22 Best	86.80	–	81.00	–
Zero-shot	86.91	72.51	82.55	71.12
ICL	87.10	**73.15**	82.71	71.38
DUAT-E	**87.60**	72.41	**82.75**	**71.75**

Table 3: Results on the complete WMT2022 testset. The result of WMT22 Best is reported for comparison.

(1) DUAT achieves significant improvements. On average, DUAT-E surpasses the baseline ICL by +1.82 COMET and +1.32 BLEURT, and improves Zero-shot by +3.41 COMET and +3.33 BLEURT. In the low-resource translation of En⇒Is, DUAT-E improves ICL by +3.85 COMET and Zero-shot by +5.87 COMET. These improvements show that, DUAT largely elicits the translation ability via aligning the translation-specific understanding to the general one.

(2) DUAT achieves state-of-the-art performance on Challenge-WMT. DUAT achieves the highest scores in En⇔Zh and En⇔Et translation in terms of COMET and BLEURT. In En⇔Is translation, GPT-4 achieves the best results due to its superior multilingual capabilities. To verify the effectiveness of DUAT on LLMs with stronger multilingual capabilities, we implement DUAT based on GPT-4 in En⇔Is translation, as shown in Table 2. The results suggest that DUAT can also benefit stronger multilingual LLMs by facilitating the understanding misalignment.

(3) CoT works poorly in machine translation. In Table 1, CoT incurs a dramatic performance drop over the baseline ICL. Our case studies reveal that CoT produces extremely wordy translations, which is also observed by Peng et al. (2023). We conjecture that CoT makes LLMs imitate junior translators rather than advanced translators.

(4) Difficult words are the bottleneck in translating complex sentences. The results show that incorporating the analysis of keywords and topics (He et al., 2024) has yet to gain significant improvements as DUAT. It suggests that it is the difficult words that lead to the performance bottleneck in translating intricate sentences. We also follow He et al. (2024) to experiment under the rerank setting as shown in Appendix.E, which shows the effectiveness of our method further.

(5) The external tool helps LLMs better identify difficult words. DUAT-I achieves an average improvement of +1.52 COMET, demonstrating the effective performance of LLMs in recognizing difficult words. And DUAT-E gains a further improvement of +0.3 COMET, showing the effectiveness of the external token-level QE tool in this task.

5.3 Results on the complete WMT

Experimental results on Challenge-WMT show the effectiveness of DUAT in translating complex sentences, raising the question of its impact on translating simple sentences. Therefore, we conduct additional experiments on the complete WMT2022 testset of Zh⇔En translation. As the results shown in Tab.3, our method achieves comparable results to the baseline ICL in both translation directions. These results show that DUAT has no negative impact on translating simple sentences.

6 Analysis

6.1 Human Evaluation

To quantitatively analyze the understanding misalignment problem, we employ one senior human translator for each language direction to assess generalization failures and translation literalness. Specifically, in each direction, we randomly sample 100 sentences[4] from Challenge-WMT and ask the senior translator to annotate the mistranslated words and phrases (*i.e.,* mistranslation) and score the literalness (1 to 5 score) of the translations generated by the strong baseline (`GPT-3.5-turbo` with ICL). Next, we ask the LLM to generate interpretations of the mistranslated content. These interpretations are presented to the translator, who judges whether they contain the correct understanding of the content. If the interpretation is accurate, the content is annotated as a generalization failure. Finally, the translator is asked to annotate the generalization failures of DUATand score the literalness of the translations produced by DUAT.

Analysis of generalization failures. As the results shown in Tab.4, generalization failures ac-

[4]It takes 1.4 dollars for annotating one sentence.

	En⇒Zh	Zh⇒En	En⇒Et	Et⇒En	En⇒Is	Is⇒En
#Mistranslation	19	25	26	**53**	22	34
#Generalization Failure of Baseline	5 (26%)	4 (16%)	7 (27%)	**17 (32%)**	5 (23%)	10 (29%)
#Generalization Failure of DUAT	1 (-80%)	1 (-75%)	2 (-71%)	**2 (-88%)**	1 (-80%)	2 (-80%)
Translation Literalness of Baseline	4.11	3.53	**4.75**	4.46	3.76	3.70
Translation Literalness of DUAT	2.60 (-36%)	2.63 (-25%)	**2.31 (-51%)**	2.60 (-41%)	3.12 (-17%)	2.79 (-24%)

Table 4: Human evaluation results.

Methods	**En⇒Zh**		**Zh⇒En**	
	COMET	Δ	COMET	Δ
DUAT-E	77.57	–	73.23	–
w/o. Draft	76.94	-0.63	72.68	-0.55
w/o. IQC	76.54	-1.03	72.91	-0.32
DUAT-I	76.92	–	72.94	–
w/o. Draft	76.68	-0.24	72.78	-0.16
w/o. IQC	76.45	-0.47	72.59	-0.35

Table 5: Ablation Study. Δ indicates the performance drop after removing the specific component.

count for a considerable proportion of all mistranslations (16%∼32%). Our method (DUAT-E) largely resolves these failures by 71%∼88%. We further study the unresolved failures and find that most of these unresolved failures are because they are not identified as difficult words by the LLMs in the stage of difficult word detection (Eq.3) despite actually hard to translate. It indicates that the difficult word detection remains an open question.

Analysis of translation literalness. As the results shown in Tab.4, the baseline translations are highly biased towards literal translation. DUAT-significantly reduces the bias towards literal translation, indicating that the process of interpreting the difficult words first and then translating aligns better with sense-for-sense translation.

6.2 Ablation Study

DUAT introduces the processes of (1) ***draft translation*** to precisely detect the difficult words and (2) ***IQC*** to improves the helpfulness of interpretations. To clearly elucidate the contribution of these two components, we conduct an ablation study in Table 5. Specifically, we analyze the effect of the draft translation by asking the LLM to detect difficult words directly without the draft translation. The impact of IQC is analyzed by evaluating the performance of the generated translations guided by the original noisy interpretations (*i.e.,* without the processing of IQC). The results show that removing either component leads to performance drops, and IQC plays a more important role in DUAT. Specifically, the improvement of DUAT is halved when ablating the IQC on the En⇒Zh translation.

6.3 Analysis of Difficult Word Detection

To offer an in-depth insight into the process of difficult word detection, we illustrate the relation between the number of difficult words interpreted and the resulting performance by adjusting the value of the difficulty threshold (τ), which is shown in Fig.3. Concretely, a smaller value of τ allows more difficult words to be interpreted. From the results, we have the following observations:

Increasing the number of interpretations does not necessarily lead to performance improvements, but increasing high-quality ones can. Specifically, without controlling the quality of the interpretations (*i.e.,* w/o. IQC), increasing the number of interpretations (the **green lines**) yields unpredictable performance changes (as shown by the **green bins**), as introducing either valuable information or noise. Fortunately, with IQC filtering negative interpretations, increasing the number of interpretations (the **blue lines**) leads to constant improvements (as the **blue bins** show).

Interpreting words that are more difficult brings larger improvements. Specifically, in the En⇒Zh translation, decreasing the value of τ from 0.19 to 0.17, the average number of helpful interpretations is increased from 0.23 to 0.49 (+0.26), and the performance is increased from 76.52 to 76.98 (+0.46). However, decreasing the value of τ from 0.15 to 0.10, the average number of helpful interpretations is increased from 0.91 to 1.47 (+0.56), and the performance is increased from 77.17 to 77.57 (+0.40). It should be noted that interpreting more words incurs more inference costs. Therefore, a modest value of τ (*i.e.,* 0.13 ∼ 0.15) is recommended to reach a compromise between efficiency and performance of DUAT.

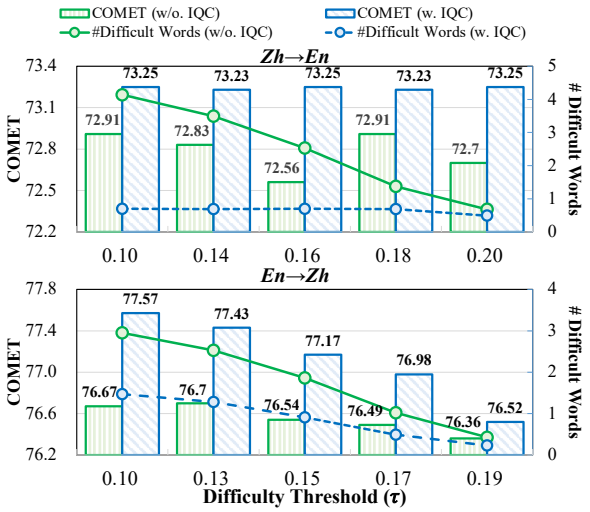

Figure 3: Effect of different values of difficulty threshold (τ) on DUAT-E.

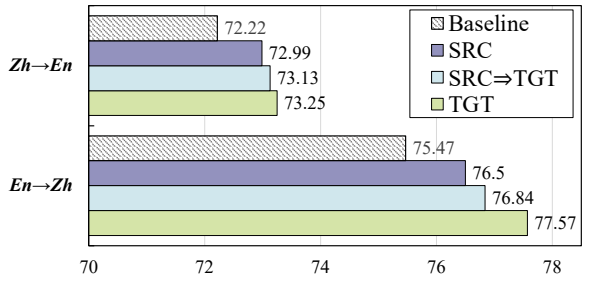

Figure 4: Effect of interpretations' language for DUAT-E.

6.4 Analysis of Interpretation Generation

Languages of interpretations. Given a difficult word, DUAT generates the corresponding interpretation with the target language (*i.e.,* cross-lingual interpretation), which implicitly comprises two stages: (1) generating the interpretation in the source language and (2) translating the interpretation into the target language. Compared with conducting these two stages explicitly, DUAT is more efficient and avoids error accumulation, which is illustrated in Fig.4. As demonstrated, interpretations in the target language (the **blue** bins) are more beneficial than the ones in the source language (the **purple** bins) owing to aligning the general understanding into the target language space, which could provide more benefits for translation. And the implicit two-stage process (the **blue** bins) is better than the explicit one (the **green** bins).

7 Related Work

Evaluation of LLMs' translation capabilities. With the remarkable progress of LLMs, researchers have assessed their translation abilities in various aspects. Zhang et al. (2023a); Vilar et al. (2023); Garcia et al. (2023); Bawden and Yvon (2023) first investigate LLM-based MT in terms of the prompt template and examples selection. Next, the evaluation is extended across more domains (Hendy et al., 2023), more languages (Zhu et al., 2023a), and document-level translation (Hendy et al., 2023; Wang et al., 2023). Other lines of work have performed in-depth assessments on the important attributes beyond accuracy, like literalness (Raunak et al., 2023) and culture awareness (Yao et al., 2023). As existing studies have shown that LLMs have achieved promising performance, our work turns out to benchmark them on hard instances towards detecting more underlying issues.

LLM-based translation strategies. Lu et al. (2023) obtain the multilingual translations of keywords in the source sentence via the translator NLLB to augment the LLM, which improves the translation of low-resource languages while hurting the performance of high-source languages. Chen et al. (2023) demonstrate that iterative refinement reduces translationese significantly. He et al. (2024) incorporate the knowledge of keywords, topics, and reference demonstrations to enhance the translation process, and use a rerank strategy to combine all candidate translations. However, there is no significant improvement to be observed when solely utilizing each single type of knowledge. Different from previous works that utilize the intrinsic knowledge of LLM, DUAT focuses on dealing with the difficult-to-translate words instead of the keywords for the reason that we argue the difficult-to-translate words lead to the performance bottleneck due to the long-tail distribution of knowledge.

LLM-based Automatic Post-Editing (APE). APE corrects the errors in the generated translation, aiming to bias the translation towards the distribution of the target language (Chen et al., 2023; Koneru et al., 2024). Differently, our work aims to leverage the powerful understanding abilities of LLMs to correct the misunderstanding of complicated concepts in the source sentence. Our work is in parallel with APE.

8 Conclusion

In this work, we propose a novel translation process, DUAT, to take the first step in resolving the misalignment between the translation-specific un-

derstanding and the general understanding. Furthermore, we utilize the token-level QE tool to enhance the detection of difficult words and the sentence-level QE tool to remove harmful interpretations. Human evaluation results on high-resource and low-resource language pairs indicate that DUAT significantly facilitates the understanding alignment, which improves the translation quality and alleviates translation literalness.

9 Limitations

Even though DUAT elicits the translation abilities of LLMs via unleashing the general understanding (intrinsic knowledge) of LLMs, they still struggle to translate concepts that require the incorporation of extrinsic knowledge, such as the translation of *neologisms*. However, Our approach lays the foundation for researching when and how to incorporate external knowledge. Besides, DUAT requires to prompt the LLM for several times, leading to an increase in latency. This latency is mainly caused by our interpretation quality control (IQC) strategy, which sequentially ablates each generated interpretation. Concretely, if $|\mathcal{D}|$ difficult words are identified, IQC needs to prompt the LLM for $|\mathcal{D}|$ times.

Acknowledgements

Bing Qin is the corresponding author of this work. We thank the anonymous reviewers for their insightful comments. This work was supported by the National Natural Science Foundation of China (NSFC) (U22B2059, grant62276078), the Key R&D Program of Heilongjiang via grant 2022ZX01A32, the International Cooperation Project of PCL, PCL2022D01 and the Fundamental Research Funds for the Central Universities (Grant No.HIT.OCEF.2023018).

References

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland. European Association for Machine Translation.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023. Iterative translation refinement with large language models.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

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*, ICML'23. JMLR.org.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring Human-Like Translation Strategy with Large Language Models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine.

Sai Koneru, Miriam Exel, Matthias Huck, and Jan Niehues. 2024. Contextual refinement of translations: Large language models for sentence and document-level post-editing. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2711–2725, Mexico City, Mexico. Association for Computational Linguistics.

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023. Chain-of-dictionary prompting elicits translation in large language models.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages.

OpenAI. 2023. Gpt-4 technical report.

Jianhui Pang, Fanghua Ye, Longyue Wang, Dian Yu, Derek F. Wong, Shuming Shi, and Zhaopeng Tu. 2024. Salute the classic: Revisiting challenges of machine translation in the age of large language models.

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

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of ChatGPT for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore. Association for Computational Linguistics.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan. 2023. Do GPTs produce less literal translations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1041–1050, Toronto, Canada. Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, Marcos Treviso, Luisa Coheur, Alon Lavie, and André Martins. 2023. The inside story: Towards better understanding of machine translation neural evaluation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Toronto, Canada. Association for Computational Linguistics.

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

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

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, Toronto, Canada. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

BigScience Workshop. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. 2023. Empowering llm-based machine translation with cultural awareness.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. Association for Computational Linguistics.

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

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023b. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

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

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023b. Extrapolating large language models to non-english by aligning languages.

A Generalization Failures on Translation

In this section, we first provide the illustration of generalization failures on more LLMs, as shown in Fig.5. As we can see, all of four LLMs accurately comprehend the complex concept, which three out of them mistranslate this concept. Then, we also provide more examples of generalization failures, as shown in Fig.6 and Fig.7.

B More details of DUAT

B.1 Details of IQC

We gives a formal description of our interpretation quality control in Alg.1.

Algorithm 1: IQC

Input : source sentence x, draft translation $\widetilde{y}$, interpretations of difficult words $\mathcal{A}$, QE scorer $\psi(\cdot)$

Output : helpful interpretations $\hat{\mathcal{A}}$, final translation $\hat{y}$

$\hat{\mathcal{A}} \leftarrow \mathcal{A}$
$\hat{y} \leftarrow argmax\ P_\theta(\mathcal{E}^{igt}, x, \hat{y}, \mathcal{A})$
$\hat{s} \leftarrow \psi(\hat{y} \mid x)$
for $i \leftarrow 1\ to\ |\mathcal{A}|$ **do**
  $\overline{y} \leftarrow argmax\ P_\theta(\mathcal{E}^{igt}, x, \hat{y}, \mathcal{A} - \{\mathcal{A}_i\})$,
  $\overline{s} \leftarrow \psi(\overline{y} \mid x)$
  if $\overline{s} > \hat{s}$ **then**
    $\mathcal{A} \leftarrow \mathcal{A} - \{\mathcal{A}_i\}, \hat{y} \leftarrow \overline{y}, \hat{s} \leftarrow \overline{s}$
  end
end

B.2 Details of Demonstration Synthesis

Inspired by the idea of Auto-CoT (Zhang et al., 2023b), we utilize LLM to generate the difficult words $\mathcal{D}$ and corresponding interpretations $\mathcal{A}$ based on the given bilingual sentence pair (x, y):

> Request: Given a $[L_s]$ sentence and its $[L_t]$ translation, please output the most difficult-to-translate words in the source sentence and concisely analyze the meaning of these words.
> The input-output format is:
>
> *# the format description is omitted.*
> Source Sentence: [Source Sentence x]
> Target Translation: [Target Translation y]

Then, the response is parsed via regular expression to extract the difficult words $\mathcal{D}$ and interpretations $\mathcal{A}$. Next, we remove the noisy interpretations through a process similar to IQC (Alg. 1). The only difference is that the QE metric is replaced with the reference-based COMET (Rei et al., 2020) due to the available access to the reference translation. Finally, the generated difficult words $\mathcal{D}$ and interpretations $\mathcal{A}$ can be assembled with the source and target sentence (x, y) as demonstrations for each step of DUAT.

C Statistics of Challenge-WMT

We compare the complete WMT test set and the Challenge-WMT subset in terms of the length of source sentences, the length of target sentences, the perplexity of source sentences, average number of nouns, verbs and named entities in the source sentence. The statistics is shown in Table 6.

D Details of Experiments

We conduct experiments under the few-shot setting. To obtain the demonstrations of CoT, we ask the LLM to output the step-by-step translation process in a manner of post-explanation (*i.e.*,, given the source sentence and its translation, requesting the LLM to generate the intermediate process). To obtain the ones of MAPS, we let the LLM to perform translation with the specific strategy on the validation set, and assemble the generated intermediate process (*e.g.,* keywords) and the reference translation as demonstrations.

E Results under the Rerank setting

We follow He et al. (2024) to conduct experiments additionally under the rerank setting. For the baseline ICL, we run for 4 times with different sets of demonstrations, which are sampled randomly with seeds $\{1, 2, 3, 4\}$, and adopt QE to select the best candidate as the final translation. For MAPS, the final translation is selected from the candidates generated by the three strategies ('+topic', '+Keywords', and '+SimDems') and ICL (seed=1). For DUAT, we select the final translation from the results of DUAT and ICL (seed=1). The results are shown in Table 7.

Model	gpt-3.5-turbo-0613	gpt-3.5-turbo-0125	gpt-4-turbo-2024-04-09	gpt-4-0314
Question	In this Chinese sentence: "文章的前妻是马伊琍", what is the meaning of "文章"?			
LLM's Answer	It refers to the **Chinese actor and singer, Wen Zhang**.	In this sentence, "文章" refers to the **Chinese actor** and singer.	**Wen Zhang** (文章) is a Chinese actor.	"文章" is a person's name. It refers to a famous Chinese actor, **Wen Zhang**.
Source Sentence	文章的前妻是马伊琍。			
Reference Translation	The ex-wife of **Wen Zhang** is Ma Yili.			
LLM's Translation	The ex-wife of the **article** is Ma Yili.	The ex-wife of the **article** is Ma Yili.	The ex-wife mentioned in the **article** is Ma Yili.	The ex-wife of **Wen Zhang** is Ma Yili.

Figure 5: Illustration of understanding misalignment in more LLMs.

Source Sentence	许多观察家指出，他执政5年人设崩塌，主要是自己“刨坑”所致 。
Reference Translation	Observers point out that the collapse of his reputation after five years in power was primarily caused by **his own deeds.**
LLM's Translation	Many observers point out that his 5-year rule has led to the collapse of his public image, mainly due to his own "**digging holes**"

(a) LLM ***misunderstands*** the word "刨坑" as a physical activity during translating the source sentence . ✗

Question	In this Chinese sentence "许多观察家指出，他执政5年人设崩塌，主要是自己“刨坑”所致 ", what is the meaning of "刨坑"?
LLM's Response	It is used **metaphorically to indicate that someone's actions have led to their own downfall**.

(b) LLM correctly ***understands*** the metaphorical meaning of "刨坑" during explaining its meaning. ✓

Figure 6: Illustration that understanding misalignment leads to LLMs literally translating some complicated concepts.

Source Sentence	Judge Nigel Daly told Clifton, who admitted ABH, that the alcohol had "played a significant part" in the night's violence.
Reference Translation	法官奈杰尔·戴利告诉对造成实际身体伤害（简称 ABH）供认不讳的克利夫顿，酒精在当晚的施暴中“发挥了重要作用”。
LLM's Translation	承认犯有**轻伤罪**的克利夫顿被法官奈杰尔·戴利告知，酒精在当晚的暴力事件中“起到了重要作用”。

(a) LLM ***misunderstands*** the term "ABH". ✗

Question	在这句英文中：Judge Nigel Daly told Clifton, who admitted ABH, that the alcohol had "played a significant part" in the night's violence. "ABH"是什么意思?
LLM's Response	ABH是Actual Bodily Harm的缩写，指的是**实际身体伤害罪**。

(b) LLM correctly ***understands*** the term of "ABH" during explaining its meaning. ✓

Figure 7: Illustration that understanding misalignment leads to LLMs mistranslating some terminology.

Language pair	**En⇒Zh**		**Zh⇒En**		**En⇒Et**		**Et⇒En**		**En⇒Is**		**Is⇒En**		**Average**	
Dataset	*Comp.*	*Chal.*	*Comp.*	*Chal.*	*Comp.*	*Chal.*	*Comp.*	*Chal.*	*Comp.*	*Chal.*	*Comp.*	*Chal.*	*Comp.*	*Chal.*
#Samples	6215	**675**	7207	**615**	4000	**644**	4000	**602**	3004	**641**	3004	**694**	4572	**645**
SRC-Len	22.4	**24.2**	47.4	**52.0**	19.6	**20.3**	14.9	**15.1**	21.4	**24.6**	18.9	**20.8**	24.1	**26.2**
TGT-Len	42.5	**50.9**	28.9	**34.1**	14.9	**15.5**	19.6	**20.8**	20.6	**25.1**	20.4	**23.2**	24.5	**28.3**
SRC-PPL	141	**165**	40	**79**	128	**156**	823	**925**	111	**147**	40	**40**	214	**252**
#Noun	4.2	**4.9**	3.4	**4.1**	4.6	**4.7**	**1.7**	1.6	5.8	**7.1**	**2.1**	2.0	3.6	**4.1**
#Verb	5.2	**5.9**	3.1	**3.7**	3.0	**3.2**	2.5	2.5	3.8	**4.4**	2.6	**2.8**	3.4	**3.8**
#NE	0.8	**1.1**	2.4	2.4	**0.9**	0.8	1.6	1.6	0.9	**1.2**	**1.9**	1.8	1.4	**1.5**

Table 6: Fine-grained comparison of the complete WMT test set (*Comp.*) and the Challenge-WMT subset (*Chal.*). 'NE' is the abbreviation of "Named Entities".

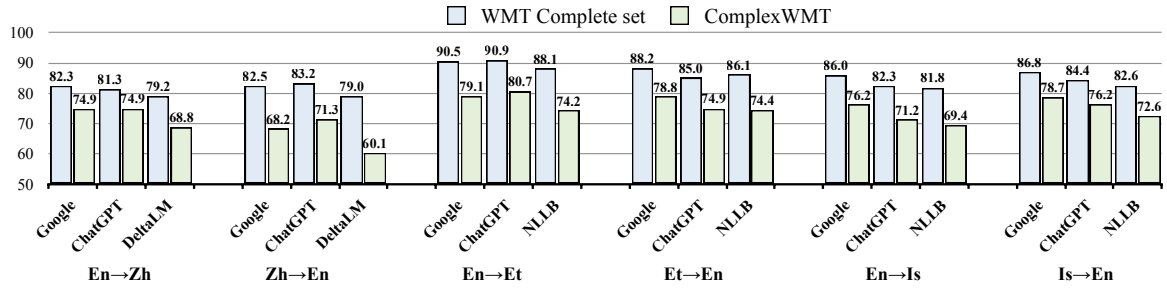

Figure 8: Translation performance on the complete WMT test set and the Challenge-WMT test set.

Methods	En⇒Zh		Zh⇒En		En⇒Et		Et⇒En		En⇒Is		Is⇒En		Average	
	COMET	QE	COMET	QE	COMET	QE	COMET	QE	COMET	QE	COMET	QE	COMET	QE
Baselines														
ICL	76.79	3.94	72.67	0.43	82.10	9.37	79.98	7.44	73.42	-1.21	78.88	4.80	77.31	4.13
MAPS	77.24	4.56	73.17	**1.70**	83.05	10.57	80.12	**8.28**	75.67	2.61	78.47	5.13	77.95	5.48
Ours														
DUAT-I	77.36	4.37	73.30	1.08	83.06	10.39	**80.22**	7.96	76.88	3.12	78.93	5.29	78.29	5.37
DUAT-E	**77.78**	**5.04**	**73.36**	0.88	**83.21**	**10.93**	80.10	8.06	**77.39**	**3.97**	**79.22**	**5.31**	**78.51**	**5.70**

Table 7: Experimental results under the rerank setting.