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

Anonymous ACL submission

Abstract

Although large language models (LLMs) have shown surprising language understanding and generation capabilities, they have yet to gain a revolutionary advancement in the field of machine translation. One potential cause of the limited performance is the misalignment between the translation-specific understanding and general understanding inside LLMs. To align the translation-specific understanding to the general one, we propose a novel translation process XIOD (**C**ross-**L**ingual **I**nterpretation **o**f **D**ifficult words), explicitly incorporating the general understanding on the content incurring inconsistent understanding to guide the translation. Specifically, XIOD performs the cross-lingual interpretation for the difficult-to-translate words and enhances the translation with the generated interpretations. Furthermore, we reframe the external tools of QE to tackle the challenges of XIOD in the detection of difficult words and the generation of helpful interpretations. We conduct experiments on the self-constructed benchmark Challenge-MT, which includes cases in which multiple SOTA translation systems consistently underperform. Experimental results show the effectiveness of our XIOD, which improves up to +3.85 COMET. Human evaluation reveals that the translation generated by XIOD accords more with the sense-for-sense translation.

1 Introduction

Recently, with the scaling of model capacities and pre-training data volume, 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; Zhang et al., 2023a; Jiao et al., 2023; Zhu et al., 2023b).

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 1: Illustration of the misalignment between the translation-specific language understanding (Fig a) and the general understanding (Fig b) in the LLM.

One potential reason leading to the limited translation performance of LLMs is the misalignment between the *general understanding* and *translation-specific understanding* inside LLMs. This claim is based on our discovery that LLMs understand many concepts accurately when explaining directly but often misunderstand these concepts in translation requests, which is illustrated in Fig 1. This paper refers to these concepts as language models' **generalization failures on translation**. Through manual statistical analysis, we found that between 3% to 15% of the translations produced by the LLM contained generalization errors across different translation directions.

In this work, we propose a novel translation process XIOD, explicitly incorporating the general understanding to guide the translation during inference. Specifically, XIOD first detects the difficult-to-translation words in the source sentence, which

✉ means corresponding author.

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 aligning this understanding into the target language space. After that, XIOD conducts translation under the guidance of these interpretations. Unlike the CoT-based process mimicking the junior translator to perform word-for-word translation (Peng et al., 2023), XIOD is more in accordance with the senior translator owning a higher performance frontier. However, XIOD faces two challenges: (1) LLMs may struggle with the accurate detection of difficult words due to their limitation in self-knowledge (Yin et al., 2023), and (2) LLMs are prone to errors and hallucinations in the interpretations (Huang et al., 2023a,b), biasing the translation from the original semantics. Therefore, 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 interpretations quality control to filter hallucinated interpretations based on sentence-level QE (Rei et al., 2020).

Towards better identifying the limitations of current MT systems, we proposed the benchmark **Challenge-MT** via collecting the difficult cases that multiple SOTA systems consistently underperform on multi-year WMT datasets. All experiments are conducted under the few-shot setting, where the demonstrations are constructed automatically in a post-explanation manner. Experimental results on six translation directions (Chinese, Estonian, and Icelandic to/from English) demonstrate the effectiveness of XIOD, which improves up to +3.85 COMET. Furthermore, human evaluation shows that xIoD accords more with the sense-for-sense translation instead of word-for-word translation.

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 QE for MT

Quality estimation (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 the translation y, the 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: XIOD

In this section, we first introduce our translation framework XIOD (§3.1). Specifically, XIOD 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 XIOD (§3.5).

3.1 Framework

The progress of XIOD is illustrated in Figure 2. Given the source sentence, XIOD first detects the difficult words or phrases in the source sentence. Once the difficult words are identified, XIOD requests the LLM to interpret each difficult word with the target language, unleashing the powerful understanding capability inside the LLM and aligning these understanding into the target language. Finally, to avoid the negative effect of incorrect and useless interpretations, XIOD removes the negative

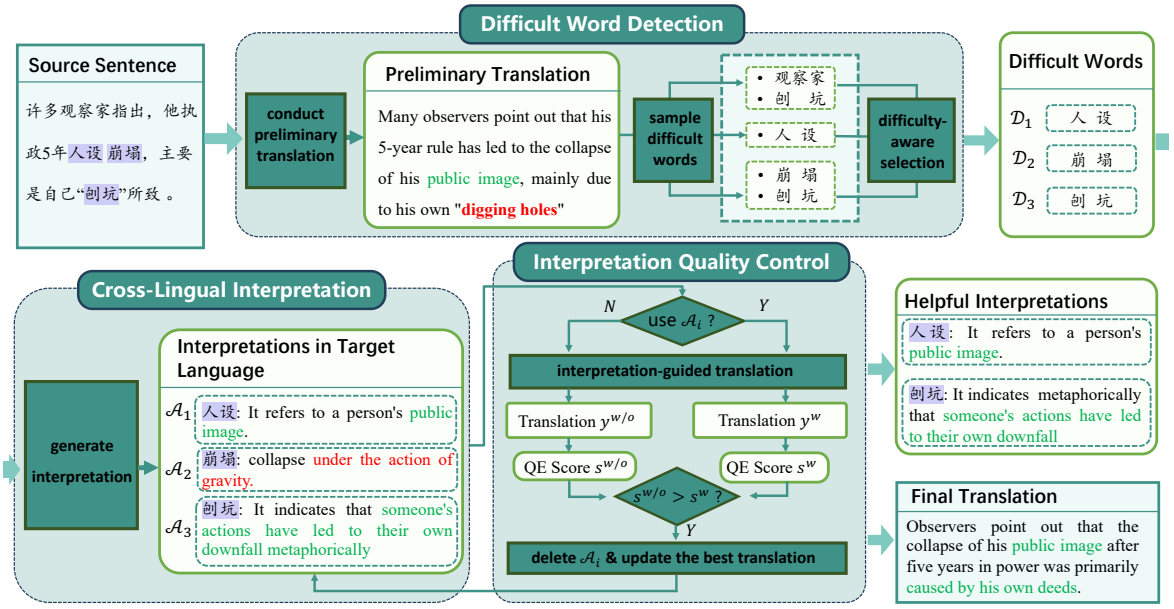

Figure 2: XIOD framework. The purple spans indicate the difficult-to-translate words, the green spans indicate the correct translation/interpretation, and the red spans indicate the incorrect ones.

interpretations through the interpretation quality control and outputs the final translation guided by the helpful interpretations.

3.2 Step-1: Difficult Word Detection

Correct understanding of the difficult words determines largely the quality of the final translation. To make the LLM comprehend these difficult words, it is necessary to recognize them at first. However, finding the translation-specific difficult words remains challenging (Lim et al., 2023; Sun, 2015). To tackle this challenge, we first conduct a preliminary translation for the given source sentence and then extract the **mistranslated words** in the source sentence **as the difficult words**. Concretely, we invent XIOD-I to do this using solely the ***Intrinsic*** knowledge of LLMs at first.

XIOD-I. Given source sentence x, XIOD-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 in the $[L_s]$ sentence.
> *# followed by* [N Demonstrations $\mathcal{E}^{diff}$]
> Source Sentence: [Given Sentence x]
> Draft Translation: [Draft Translation $\widetilde{y}$]

XIOD-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 $\mathcal{E}^{diff} = \{x^i, \widetilde{y}^i, \mathcal{D}^i\}_{i=1}^{N}$.

XIOD-E. It is frequently observed that the LLM fails to recognize the mistranslated words. Therefore, we devise XIOD-E to boost the detection with the ***external*** tool. First, XIOD-E requests the LLM with the same prompt as XIOD-I while performing the *temperature sampling* for K times ($K = 5$). 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 sampling temperature, which is set to 0.5 to capture more candidates.

Finally, XIOD-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.*,

https://platform.openai.com/docs/api-reference/chat/create#chat-create-temperature

$\phi(w^t \mid \widetilde{y}, x)$ 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:

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

Then, xIoD-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, xIoD lets the LLM generate the interpretation of each difficult word via requesting with the prompt:

> 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:

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

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. Through the generated interpretations, xIoD enhances the translation with LLMs' general understanding capability. However, LLMs may generate incorrect or 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 negative effect of the generated interpretations, xIoD removes the incorrect and useless interpretations through the interpretation quality control (**IQC**) and outputs the final translation guided by the helpful interpretations.

Concretely, given a set of interpretations $\mathcal{A}$, xIoD 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*:

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 $\hat{y}$

```
Â ← A
ŷ ← argmax P_θ(E^igt, x, ŷ, A)
ŝ ← ψ(ŷ | x)
for i ← 1 to |A| do
    ȳ ← argmax P_θ(E^igt, x, ŷ, A − {A_i}),
    s̄ ← ψ(ȳ | x)
    if s̄ > ŝ then
        A ← A − {A_i}, ŷ ← ȳ, ŝ ← s̄
    end
end
```

> 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:

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

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

Once the better translation performance is achieved by ablation, which is measured by the QE 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 detail this process in Alg. 1.

3.5 Demonstrations Synthesis for xIoD

To make the LLM follow the procedure of xIoD as expected, we adopt the ICL strategy. Common practice constructs the demonstrations manually for ICL, 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 xIoD based on the parallel data.

Inspired by the idea of Auto-CoT (Zhang et al., 2023b), we utilize LLM to generate the difficult

We use wmt21-comet-qe-da as the QE scorer.

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 XIOD.

4 Testbed: Challenge-MT dataset

In this work, we propose a benchmark Challenge-MT, which consists of difficult translation samples, for the following reasons: 1) MT has witnessed human-like performance on many languages, which appeals to a more challenging benchmark to detect the flaws of SOTA systems, and 2) a benchmark containing more instances that models underperform helps us to analyze the issue of the understanding misalignment in LLMs effectively.

Challenge-MT is constructed by collecting the most challenge subset of the widely used WMT datasets, involving six translation directions of *Chinese* (`zh`), *Estonian* (`et`), and *Icelandic* (`is`) to/from *English* (`en`). Specifically, we first evaluate three SOTA MT systems (Google Translate, ChatGPT (`gpt-3.5-turbo`), and NLLB (NLLB Team et al., 2022)) on the multi-year WMT datasets. Due to the poor performance of NLLB on the zh⇔en translation, we additionally train a zh↔en translation model based on the DeltaLM (Ma et al., 2021) on the parallel corpus from OPUS. Next, The generated translations are scored with the COMET metric, and the ρ of instances 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 scale for each difficult sample set. Finally, the intersection of all systems' difficult sample sets is taken as the Challenge-MT benchmark. We equally split this dataset into the validation set and the test set.

As the results demonstrated in Fig 5, SOTA MT systems show extremely poor performance on the Challenge-MT benchmark, which is 10 COMET scores lower than the complete set in most translation directions. To understand the cause of the low performance on Challenge-MT, we conduct a multi-aspect comparison for the complete set and the Challenge-MT subset in Appendix A, which shows that the samples are longer and have higher perplexity in the Challenge-MT. These features indicate that **sentences that are longer and harder to understand are more difficult to translate**.

5 Experiments

5.1 Experimental Setup

Comparative Methods. We verify the effectiveness of our XIOD on the LLM `GPT-3.5-turbo` for its promising capability in following complicated instructions. Demonstrations of XIOD are gained by performing our automatic method (§3.5) on the validation set of Challenge-MT. We compare XIOD 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., 2023), 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 XIOD and other ICL-based methods, we select K=8 demonstrations (*i.e.,* 8-shot) to achieve a strong baseline performance. More details of

https://opus.nlpl.eu/

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
Existing Systems														
Google	74.85	1.87	68.21	-5.97	79.11	6.02	78.83	5.57	76.17	0.56	78.70	3.14	75.98	1.87
NLLB	68.77	-2.74	60.09	-11.27	74.20	2.40	74.35	3.57	69.37	-5.28	72.55	0.90	69.89	-2.07
GPT-4	76.15	3.00	70.77	-1.40	80.25	6.28	77.83	5.35	**77.33**	1.61	**79.39**	3.54	76.95	3.06
Baselines														
Zero-shot	74.89	1.76	71.27	-1.34	80.67	7.55	74.93	5.45	71.17	-2.73	76.22	2.70	74.86	2.23
ICL	75.47	2.31	72.22	-0.73	80.9	9.37	79.40	6.56	73.19	-1.17	77.52	3.55	76.45	3.32
+CoT	73.85	0.86	71.35	-1.00	78.03	5.24	76.78	4.77	69.72	-2.96	76.55	2.96	74.38	1.65
+Topic	75.83	2.40	72.46	-0.25	80.98	7.19	79.20	6.59	72.77	-1.58	76.49	2.80	76.29	2.86
+Keywords	73.93	1.02	71.22	-1.62	78.63	5.87	77.79	5.48	70.33	-3.40	74.55	1.48	74.41	1.47
+SimDems	75.22	2.06	72.20	-0.60	81.24	7.90	79.11	6.76	72.70	-1.52	76.78	3.01	76.21	2.94
Ours														
xIoD-I	76.92	4.07	72.94	**0.45**	82.92	9.68	79.96	7.19	76.64	2.98	78.45	4.65	77.97	4.84
xIoD-E	**77.57**	**4.61**	**73.23**	0.32	**83.07**	**10.39**	**80.01**	**7.81**	77.04	**3.03**	78.70	**4.82**	**78.27**	**5.16**

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

Methods	En⇒Is		Is⇒En	
	COMET	QE	COMET	QE
Zero-shot	77.33	1.61	79.39	3.54
ICL	80.1	3.76	81.02	4.33
xIoD-E	**81.7**	**7.00**	**81.21**	**5.75**

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

re-implementing the baselines under the few-shot setting are illustrated in Appendix B

Metrics. Following previous research of LLM-based MT (Garcia et al., 2023; Chen et al., 2023), we adopt COMET (Rei et al., 2020) as the evaluation metric as its higher correlation with human judgment than BLEU (Papineni et al., 2002). Besides, we report the QE score to alleviate the reference bias (Freitag et al., 2020) of the reference-based metrics.

5.2 Results

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

(1) xIoD achieves significant improvements. Both xIoD-I and xIoD-E outperform the baseline ICL significantly. xIoD-E achieves significant improvements of +1.72 and +3.31 COMET over the baseline ICL and the Zero-shot method in average. In the low-source En⇒Is translation, xIoD-E improves ICL by +3.45 COMET and Zero-shot by +5.47 COMET. These improvements demonstrate that xIoD elicits the translation ability largely through aligning the translation-specific understanding to the general one.

(2) xIoD achieves state-of-the-art performance. xIoD achieves the highest scores in En⇔Zh and En⇔Et translation in terms of COMET and achieves the highest scores across all language pairs in terms of QE. However, due to the stronger multilingual capabilities of GPT-4 over GPT-3.5, the results of xIoD based on GPT-3.5 are lower than the results of the zero-shot method based on GPT-4 in En⇔Is translation. To verify the effectiveness of xIoD on LLMs with stronger multilingual capabilities, we implement xIoD based on GPT-4 in En⇔Is translation, as shown in Table 2. The improvements achieved by xIoD suggest that LLMs with stronger multilingual capabilities also suffer from the issue of understanding misalignment.

(3) CoT works poorly in machine translation. As illustrated in Table 1, CoT incurs a dramatic performance drop on the baseline ICL. To investigate this phenomenon, we have conducted case analysis and observed that the translations generated by the CoT are extremely wordy and unfluent. Our finding aligns with those reported in Peng et al. (2023). This consistency suggests that the sequentially step-by-step paradigm may not be effectively applicable to the translation task.

(4) xIoD surpasses previous translation strategies. The results show that incorporating the analysis of keywords and topics has not achieved a consistent improvement as xIoD. We conjecture the

Methods	Zh		Et		Is	
	En→	→En	En→	→En	En→	→En
ICL	4.11	3.53	4.75	4.46	3.76	3.70
xIoD w/o. IQC	4.00	3.52	4.29	4.41	3.59	3.61
xIoD	**2.60**	**2.63**	**2.31**	**2.60**	**3.12**	**2.79**

Table 3: Human evaluation of translation bias towards literal translation.

Methods	En⇒Zh		Zh⇒En	
	COMET	Δ	COMET	Δ
xIoD-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
xIoD-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 4: Ablation Study. Δ indicates the performance drop after removing the specific component.

reason is that it is the difficult-to-translate words that lead to the performance bottleneck of MT due to the long-tail distribution of knowledge (Kandpal et al., 2023; Feng et al., 2023). Besides, we also follow He et al. (2023) to experiment under the rerank setting as shown in Appendix.C, which shows the effectiveness of our method further.

(5) xIoD-E achieves a further improvement over xIoD-I. Specifically, xIoD-E outperforms xIoD-I by +0.3 COMET on average. This modest improvement demonstrates that the LLM (`GPT-3.5-turbo`) has reached an acceptable level of detecting difficult words, and the external tool of token-level QE could enhance its capability further.

6 Analysis

6.1 Human Evaluation

We conduct human evaluation to measure the degree to which translation bias towards literal translation. We employ one senior translator in each translation direction to assess 100 cases in range [1,5]. As the results shown in Tab. 3, **xIoD** 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

xIoD introduces the processes of (1) ***draft translation*** to precisely detect the difficult words and (2) ***IQC*** to improves the correctness of interpretations. To clearly elucidate the contribution of these two components, we conduct an ablation study in Table 4. 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 xIoD. Specifically, the improvement of xIoD 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.,* $0.13 \sim 0.15$) is recommended to reach a compromise between efficiency and performance of xIoD.

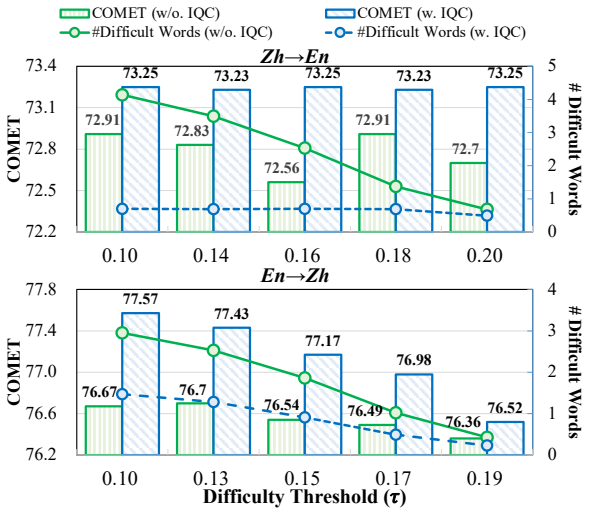

Figure 3: Effect of different values of difficulty threshold (τ) on xIoD-E.

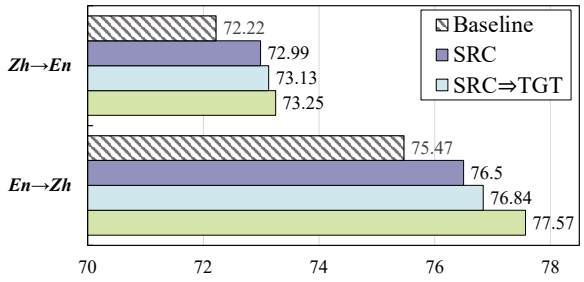

Figure 4: Effect of interpretations' language for xIoD-E.

6.4 Analysis of Interpretation Generation

Languages of interpretations. Given a difficult word, xIoD 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, xIoD 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. (2023) 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, xIoD 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.

8 Conclusion

In this work, we propose a novel translation process xIoD to take the first step in resolving the misalignment between the translation-specific understanding and the general understanding. Furthermore, we utilize the token-level QE to enhance the detection of difficult words and the sentence-level QE to remove harmful interpretations. Experimental results on the proposed Challenge-MT benchmark illustrate the effectiveness of our method.

9 Limitations

Even though xIoD 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.

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.

Zhangyin Feng, Weitao Ma, Weijiang Yu, Lei Huang, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting liu. 2023. Trends in integration of knowledge and large language models: A survey and taxonomy of methods, benchmarks, and applications.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

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. 2023. Exploring human-like translation strategy with large language models.

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.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2023b. The factual inconsistency problem in abstractive text summarization: A survey.

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.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Zheng Wei Lim, Trevor Cohn, Charles Kemp, and Ekaterina Vylomova. 2023. Predicting human translation difficulty using automatic word alignment. In *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada. 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. 2022. No language left behind: Scaling human-centered machine translation.

OpenAI. 2023. Gpt-4 technical report.

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.

Sanjun Sun. 2015. Measuring translation difficulty: Theoretical and methodological considerations. *Across Languages and Cultures*, 16(1):29 – 54.

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 Fine-grained Statistics of Challenge-MT

We compare the complete WMT test set and the Challenge-MT 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 5.

B 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.

C Results under the Rerank setting

We follow He et al. (2023) to conduct experiment 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 XIOD, we select the final translation from the results of XIOD and ICL (seed=1). The results are shown in Table 6.

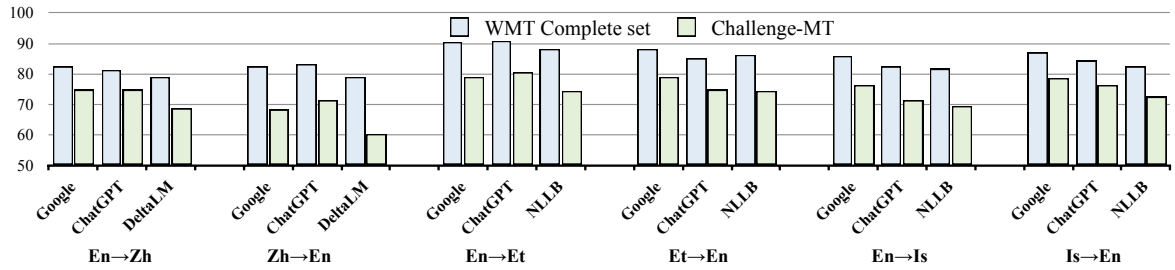

Figure 5: Translation performance on the complete WMT test set and the Challenge-MT test set.

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	140.9	**164.7**	40.1	**79.3**	127.7	**156.2**	823.4	**924.7**	111.4	**146.8**	39.9	39.9	213.9	**251.9**
#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 5: Fine-grained comparison of the complete WMT test set (*Comp.*) and the Challenge-MT subset (*Chal.*). 'NE' is the abbreviation of "Named Entities".

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														
xIoD-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
xIoD-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 6: Experimental results under the rerank setting.