LLM-based Translation Inference with Iterative Bilingual Understanding

Anonymous ACL submission

Abstract

The remarkable understanding and generation capabilities of large language models (LLMs) have greatly improved translation performance. However, incorrect understanding of the sentence to be translated can degrade translation quality. To address this issue, we proposed a novel Iterative Bilingual Understanding Translation (IBUT) method based on the cross-lingual capabilities of LLMs and the dual characteristics of translation tasks. The cross-lingual capability of LLMs enables the generation of contextual understanding for both the source and target languages separately. Furthermore, the dual characteristics allow IBUT to generate effective cross-lingual feedback, iteratively refining contextual understanding, thereby reducing errors and improving translation performance. Experimental results showed that the proposed IBUT outperforms several strong comparison methods, especially being generalized to multiple domains (e.g., news, commonsense, and cultural translation benchmarks).

1 Introduction

In the field of machine translation (MT), translations based on LLMs (LLM-MT) have become a research focus(Tyen et al., 2023; Liang et al., 2023; Guerreiro et al., 2023; Ranaldi et al., 2023; Zhang et al., 2024). Numerous studies have shown that the remarkable understanding and generation capabilities of LLMs significantly improve translation performance (Hendy et al., 2023; Jiao et al., 2023; Le Scao et al., 2023; Iyer et al., 2023; Zeng et al., 2023; Zhao et al., 2024). One type of translation paradigm can be summarized as follows: the LLM first generates a contextual understanding of the sentence to be translated, and then, based on this understanding, performs a better translation (as shown in Figure 1(a)) (He et al., 2024; Chen et al., 2024b; Liang et al., 2023; Wu et al., 2024; Chen et al., 2024a).

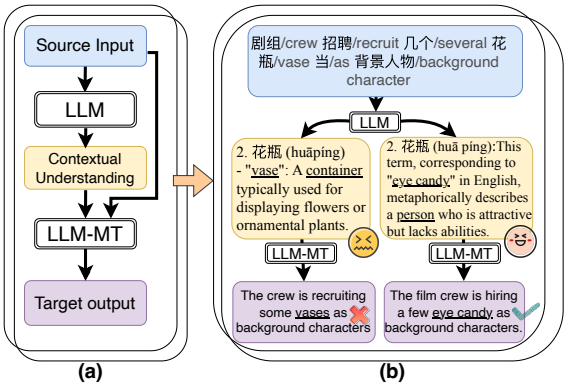

Figure 1: Illustration of the LLMs translation paradigm based on contextual understanding (Fig a). A commonsense domain example of LLM (gpt-3.5-turbo) translation from Chinese to English (Fig b).

However, our study found that when the LLM generates incorrect contextual understanding of the sentence to be translated, it negatively affects translation quality (as shown in Figure 1(b)). These generated understanding errors lead to the introduction of misleading information during the translation process, particularly when dealing with complex concepts such as commonsense and cultural domains. We refer to this issue as **Understanding Distortion** of LLMs. This study manually evaluated the Chinese-English test set in the commonsense domain. It found that Understanding Distortion makes up **40%** of translation errors, highlighting the importance of this issue (§ 6.2).

To address this issue, we propose a new method called **I**terative **B**ilingual **U**nderstanding **T**ranslation (IBUT). The IBUT method consists of four parts: Understanding Generation, Alignment Judgment, Iterative Refinement, and Understanding-Based Translation. For Understanding Generation, IBUT leverages the cross-linguistic capabilities of LLMs to generate contextual understanding for both the source and

target languages. For the alignment judgment part, IBUT uses the generated bilingual contextual understanding and employs dual learning from the translation task (He et al., 2016; Qin, 2020; Chen et al., 2024b) as supervisory signals to produce explicit verbal feedback. For the Iterative Refinement part, the verbal feedback act as a *semantic* gradient, providing LLM with a clear direction for refinement, thereby iteratively refining the bilingual contextual understanding. For the Undersderstanding-Based Translation part, LLM-MT finally translates using the source sentence and refined bilingual contextual understanding.

Empirical evaluations were conducted on closed-source and open-source LLMs including ChatGPT, GPT-4, Alpaca, and Vicuna, covering multiple domains (e.g., news, commonsense, and cultural translation benchmarks). The results show an average improvement of +1.3, +4.2, and +2.3 COMET scores compared to the baseline, confirming the effectiveness of the IBUT strategy. Additionally, analysis experiments show that IBUT positively enhances the quality of bilingual contextual understanding with each iteration, thereby improving translation performance. Our code will be made available at `https://github.com/anonymous`.

2 Related Work

Machine Translation Based on Large Language Models (LLM-MT). Large language models, such as GPT-3 (Brown et al., 2020), have demonstrated their effectiveness in machine translation across various language pairs (Hendy et al., 2023; Jiao et al., 2023; Le Scao et al., 2023; Iyer et al., 2023; Zeng et al., 2023; Puduppully et al., 2023; Karpinska and Iyyer, 2023; Moslem et al., 2023c; Wang et al., 2023; Iyer et al., 2023; Farinhas et al., 2023). Recent studies delve into the performance of LLM in machine translation, including control over formality in translation outputs (Garcia and Firat, 2022), in-context translation abilities during pre-training (Shin et al., 2022), and the impact of LLM-based machine translation on culturally sensitive texts (Yao et al., 2023).Additionally, a study has explored the bilingual capabilities of LLMs to enhance translation performance (Huang et al., 2024). For translation tasks requiring reasoning, multi-agent debates can effectively enhance the reasoning abilities of LLM-MT (Liang et al., 2023). These investigations further validate the research value of LLM-MT, offering diverse research directions for scholars.

Knowledge-based Machine Translation. Extensive research indicates that incorporating knowledge enhances translation performance. This external knowledge includes bilingual dictionaries(Arthur et al., 2016), probabilistic interpolation of dictionaries(Khandelwal et al., 2020), data augmentation through back-translation (Hu et al., 2019), phrase memory decoding (Wang et al., 2017), and entity-based denoising pre-training (Hu et al., 2021). Additionally, researchers introduced domain (Gao et al., 2023) and part-of-speech information during the inference phase and obtained multilingual translations of key terms through the NLLB translator (Lu et al., 2023), thereby enhancing the translation quality for low-resource languages. Regarding the introduction of internal knowledge into LLM-MT, LLM enhances the translation process by integrating keywords, themes, and examples similar to the given source sentences (He et al., 2023). LLM generate explanations for difficult words to enhance LLM-MT performance (Huang et al., 2024; Chen et al., 2024a).

3 The Proposed Method

The incorrect understanding of translated sentences generated by LLMs leads to a decline in translation quality. To address this issue, we propose a new method called **I**terative **B**ilingual **U**nderstanding **T**ranslation (IBUT). IBUT utilizes LLMs to generate bilingual contextual understanding of the source input and utilizes the dual learning of translation tasks to establish verbal feedback for iteratively refining this understanding. Finally, the iterative refinement reduces errors in bilingual contextual understanding, thereby enhancing translation performance. The IBUT consists of four parts: 1. Understanding Generation; 2. Alignment Judgment; 3. Iterative Refinement; 4.Understanding-Based Translation. We use MT to denote a translation model based on LLM, and lowercase letters s and t to represent sentences in the source language (L^s) and target language (L^t), respectively. That is, $s = (s[1], \cdots, s[n])$ and $t = (t[1], \cdots, t[m])$, where each $s[i]$ and $t[i]$ is a token.

1. Understanding Generation. For the first part of the IBUT method, as shown in

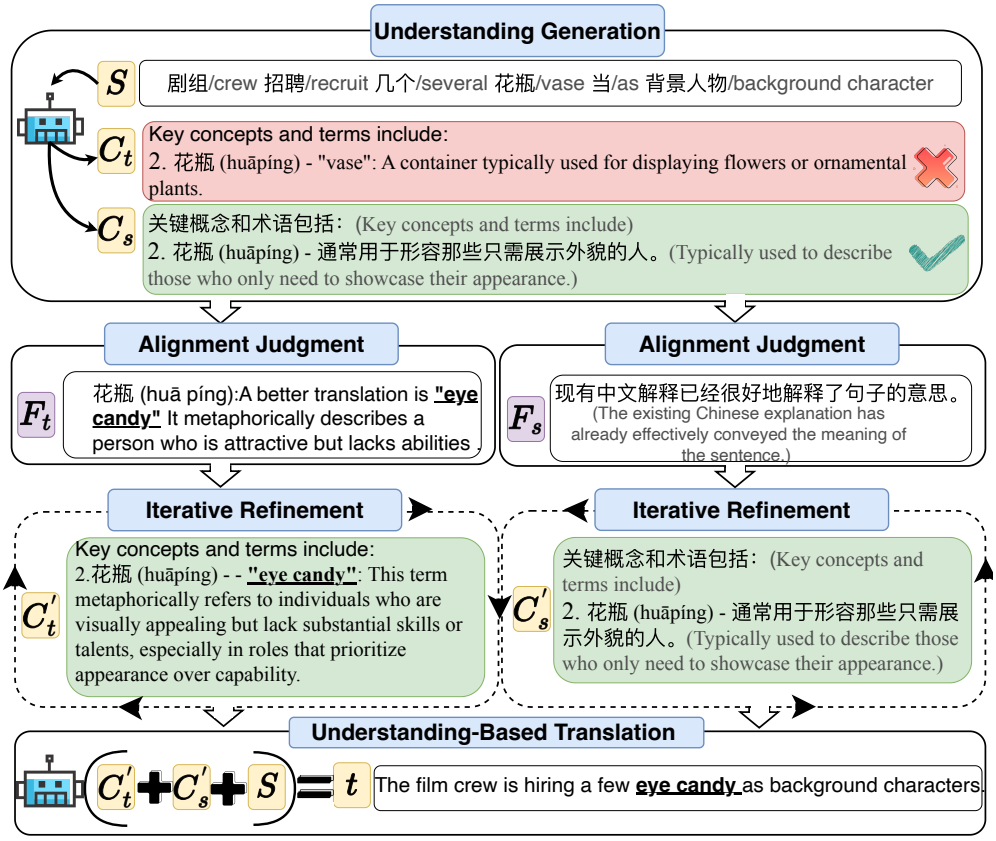

Figure 2: **IBUT translation framework.** The process involves first generating a bilingual understanding of the translation input sentence using an LLM. Next, verbal feedback is obtained via LLM, informed by the translation input and the bilingual understanding. This feedback is then used to further refine the bilingual understanding. The final step involves using LLM to perform the translation, leveraging both the bilingual understanding and the original input sentence.

Figure 2, LLMs generate contextual understanding in both the source and target languages from the source sentence, represented as C_s and C_t respectively. This understanding includes key concepts, terms, term explanations, and examples. Detailed prompts are provided in Appendix A.1.

2. Alignment Judgment. The second part of IBUT introduces an LLM-based agent, denoted as JA, which evaluates the consistency of bilingual contextual understanding and supervises the entire translation process. Based on the dual learning (He et al., 2016; Qin, 2020; Chen et al., 2024b), bilingual contextual understanding is generated from the same source sentence, and both should be consistent in form and semantics. Based on this assumption, JA initially identifies whether there are differences in the bilingual contextual understanding (C_s and C_t) generated based on the source sentence s. If $JA(C_s, C_t, s)$ = True, as shown in Figure 2, JA generates explicit verbal feedback in both the source and target languages ($F_s, F_t \leftarrow \text{JA}(C_s, C_t, s)$). The verbal feedback specifies the content of the differences between C_s and C_t and provides suggestions for refinement. If $JA(C_s, C_t, s)$ = False, the Alignment Judgment part ends, and the process transitions to the Undersderstanding-Based Translation part. Specific prompts are displayed in Appendix A.2.

3. Iterative Refinement. In the third part of IBUT, the max number of iterations (max_iter) is initially defined. As shown in Figure 2, the previously generated bilingual contextual understanding is refined based on the verbal feedback signals F_s and F_t ($C'_s \leftarrow M(s, C_s, F_s)$ and $C'_t \leftarrow M(s, C_t, F_t)$, where M is an LLM). If the number of iterations exceeds max_iter, the process will directly enter the Understanding-Based translation part. If the number of iterations does not exceed max_iter, the process will continue into the Alignment Judgment part to iteratively refine the bilingual contextual understanding. Specific prompts are displayed in Appendix A.3.

WMT22	En→De	En→Ja	Cs→Uk	Uk→Cs	En→Hr	Sah→Ru	Ru→Sah	En→Liv	Liv→En
Metrics					COMET				
ChatGPT	85.8	88.3	89.7	88.7	86.6	57.5	52.8	52.7	40.6
+5shot	86.5	88.2	88.3	89.6	86.4	58.3	53.1	55.3	42.7
+Rerank	86.2	88.0	88.3	89.7	86.3	58.6	53.8	55.5	42.9
+MAD	86.5	88.4	90.2	89.6	87.0	58.1	53.5	55.5	43.2
+MAPS	86.4	88.5	88.8	89.8	86.5	58.7	53.3	54.1	43.6
+Refine	86.0	88.6	89.8	89.3	87.0	58.3	53.8	55.5	43.9
+IBUT	**87.0**	**89.5**	**91.2**	**90.0**	**87.8**	**59.5**	**54.5**	**56.7**	**47.1**
Metrics					BLEURT				
ChatGPT	75.6	66.3	79.0	79.0	76.8	36.0	73.2	41.8	39.2
+5shot	76.3	67.1	75.6	79.1	75.6	36.0	75.4	42.1	40.9
+Rerank	75.3	66.6	75.3	79.5	75.4	36.3	75.9	42.7	41.0
+MAD	76.4	67.9	79.3	79.3	76.9	37.1	76.4	42.5	41.3
+MAPS	76.3	67.4	76.1	79.6	76.0	37.3	76.1	42.0	39.7
+Refine	75.9	67.9	79.0	79.8	76.9	37.4	76.5	42.7	40.1
+IBUT	**77.0**	**69.9**	**80.1**	**80.1**	**77.1**	**37.9**	**76.9**	**44.2**	**40.5**

Table 1: The main results from the WMT22 news benchmark are presented. ChatGPT mean to perform translation directly through Zero-Shot. The bold indicates the highest scores that are statistically significant, with p-values less than 0.01 in the paired t-test against all compared methods.

4.Understanding-Based Translation. In the final part of IBUT, the refined bilingual contextual understanding ($C_s^{'}$ and $C_t^{'}$) and the sentence to be translated are taken as inputs, and the translation is directly carried out through LLM-MT ($t = MT(s, C_s^{'}, C_t^{'})$). Specific prompts are displayed in Appendix A.4.

4 Experimental Setup

Dataset: We conduct experiments on four representative MT benchmarks, including the general news MT benchmarks (WMT22 and WMT23), commonsense MT, and cultural MT. The specific details of the dataset can be found in Appendix A.5.

Comparative Methods. In our evaluation, IBUT is compared with a range of translation methods, including Zero-shot (Wei et al., 2022), 5-shot (Brown et al., 2020; Hendy et al., 2023), Rerank (Moslem et al., 2023a), Refine (Chen et al., 2023), MAD (Liang et al., 2023), and MAPS (He et al., 2023). To validate its generalizability, we utilize three LLMs, which include closed-source models such as ChatGPT(Ouyang et al., 2022) and GPT-4(Achiam et al., 2023) [1], as well as open-source models like Alpaca-7B (Taori et al., 2023) [2] and Vicuna-7B (Chiang et al., 2023) [3]. Details on comparative methods are in Appendix A.6.

Evaluation Metrics. In evaluating our translation methodology, we initially employ COMET[4] (Rei et al., 2022a) and BLEURT[5] (Sellam et al., 2020) as automatic metrics, aligning with the established standards in LLM-based translation literature (He et al., 2023; Huang et al., 2024). For traditional translation evaluation, we use BLEU [6] (Papineni et al., 2002). To further evaluate our translation method, we employ human evaluations to verify translation performance. Details on human evaluations are in Appendix B.7.

5 Experimental Results

5.1 Main Results

The effectiveness of IBUT in general news translation tasks. In the WMT22 general news tasks, as shown in Table 1 (WMT23 results in the Appendix B.4), IBUT outperforms other methods across 13 language pairs and 3 evaluation metrics (BLEU results in the Appendix B.3). Specifically, in the news domain, the IBUT method outperforms translations directly based

[1]The ChatGPT and GPT-4 models used in this work are accessed through the gpt-3.5-turbo and gpt-4 APIs, respectively.

[2]https://huggingface.co/tatsu-lab/alpaca-7b-wdiff/tree/main

[3]https://huggingface.co/lmsys/vicuna-7b-v1.5

[4]https://huggingface.co/Unbabel/wmt22-comet-da

[5]https://github.com/lucadiliello/bleurt-pytorch

[6]https://github.com/mjpost/sacrebleu

Culture	En→Es	En→Fr	En→Hi	En→Ta	En→Te	En→Zh
Metrics	COMET/BLEURT/BLEU					
ChatGPT	83.0/69.3/35.7	77.9/58.3/31.1	73.6/61.8/18.8	67.9/57.4/11.3	69.9/52.0/13.2	83.3/64.5/35.0
+5-shot	83.2/70.3/44.0	78.0/58.5/24.0	74.3/63.7/18.7	71.8/60.2/11.2	70.6/53.6/13.3	83.2/64.9/35.3
+MAD	83.4/**70.8**/43.8	78.5/**59.0**/31.0	71.6/60.5/18.1	71.1/60.3/11.5	71.0/53.6/13.3	83.6/64.7/34.5
+MAPS	82.9/70.0/42.1	78.2/58.7/30.6	71.8/60.4/11.9	72.1/60.7/11.2	72.0/54.8/13.6	83.5/64.1/34.6
+Refine	83.0/70.1/42.1	78.0/58.3/30.4	74.3/63.2/18.8	71.8/60.9/11.7	71.7/54.6/13.7	83.0/65.1/34.7
+IBUT	**84.0**/70.7/**44.6**	**79.2**/58.9/**31.8**	**75.0/64.3/19.3**	**73.4/60.9/12.2**	**73.4/55.4/14.6**	**84.2/66.2/35.7**

Table 2: The main results from the cultural MT dataset are presented. The bold indicates the highest values that are statistically significant, with p-values less than 0.01 in the paired t-test against all compared methods.

on contextual understanding by +1.5 COMET and +1.4 BLEURT. This indicates that the IBUT method alleviates the issue of Understanding Distortion in the news domain.

The effectiveness of IBUT in low-resource tasks. We selected all low-resource tasks (Uk↔Cs, Ru↔Sah, Liv↔En, En→Hr) from WMT22. As observed in Table 1, current low-resource tasks still pose challenges to LLMs. However, compared to baseline methods, IBUT achieved an average improvement of +2.6 COMET in these low-resource tasks, with increases of +4 and +6.5 COMET for Liv↔En, respectively.

IBUT is effective across different language similarities. In WMT22, we validated the IBUT model using tasks with different language similarities. Specifically, Uk↔Cs represents closely related languages; En→De and En→Hr are from the same language family; Liv↔En, Ru↔Sah, and En→Ja are categorized as distant language families. The experimental results, as shown in Table 1, demonstrate significant improvements across different language similarities due to IBUT. Notably, for the selected distant family languages, there was an average increase of +3.4 COMT, highlighting IBUT's potential to enhance translation tasks in distant language families.

5.2 Generalizability of IBUT across multiple domains

IBUT Adapts to Cultural MT. As shown in Table 2, IBUT outperforms other methods across all 6 language pairs. For translation corpora containing cultural-specific items, the IBUT method achieved an average increase of +2.02 and +1.6 COMET compared to the ChatGPT and MAPS methods. Notably, in the En→Ta translation task, IBUT outperformed ChatGPT by +5.5 COMET. The experimental results above indicate that IBUT is suitable for translation tasks in the cultural domain.

IBUT performed well in commonsense translation tasks. As shown in Table 3, IBUT significantly outperformed other methods in commonsense MT tasks, achieving the best translation performance. Compared to the MAPS method, IBUT improved by +2 in the COMET metric, demonstrating an enhanced ability to generate higher-quality contextual understanding. Moreover, IBUT surpassed the MAD method, which relies on multi-agent debate feedback, showing its outstanding feedback quality. Notably, in translation tasks involving logical reasoning, IBUT's performance was even better than GPT-4, fully showcasing its exceptional reasoning ability.

Methods	Metrics
	COMET/BLEURT/BLEU
GPT4	82.0/71.0/32.6
ChatGPT	79.7/68.2/29.8
+Rerank	80.9/69.1/29.9
+MAPS	81.9/69.4/27.2
+Refine	81.3/69.0/28.1
+MAD	82.0/70.8/29.1
+IBUT	**83.9/72.7/32.6**

Table 3: The main results from the Commonsense MT benchmark are presented. The bold indicates the highest value. The bold indicates the highest values, statistically significant with p-values less than 0.01 in the paired t-test against compared methods.

6 Analysis

6.1 Automated Evaluation of Understanding Distortion and Translation Performance

This study explored the positive impact of reducing understanding distortion issues in bilingual contextual understanding on translation performance using IBUT. We randomly selected a set of 200 Chinese→English translation sentence pairs from the Commonsense MT dataset, which provides a test subset for lexical ambiguity. Based on

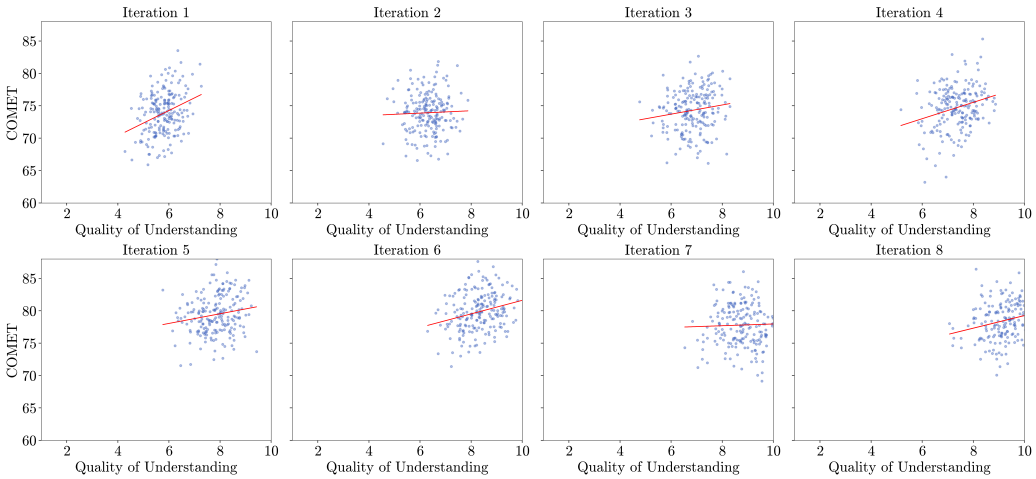

Figure 3: The experiment measures the relationship between the improvement in contextual understanding quality and translation performance during iterative refinement.

the subset, IBUT iterated 8 times ($max_iter = 8$), saving the results of bilingual contextual understanding and translation COMET scores after each iteration.

As shown in Figure 3, the vertical axis represents the translation performance, measured as the COMET score. The horizontal axis represents the scores evaluated by GPT-4 for the quality of bilingual contextual understanding affected by understanding distortion issues, with a maximum score of 10. The score for the source language is v_s and for the target language is v_t, while the overall score v is the average of the two ($v = \frac{v_s+v_t}{2}$). Details on the evaluation prompt can be found in Appendix B.2.

The experimental results, as shown in Figure 3, demonstrate a positive correlation between the quality of contextual understanding and translation performance. Additionally, as the number of iterations increases, the quality of contextual understanding progressively improves, indicating that the IBUT method effectively reduces understanding distortion issues.

6.2 Human Evaluation

Human Evaluation of Understanding Distortion Issue. In the human evaluation of understanding distortion issue, this study follows the method of Huang et al., 2024 and Chen et al., 2024b to assess translation outcomes from two main dimensions: accuracy in ambiguity resolution (commonsense domain) and the statistical results of understanding distortion issue (see Appendix B.7 for experimental setup details).

The experimental results are shown in Table 4. Understanding distortion issues accounts for a significant proportion (40%). Our method (IBUT) significantly addressed these failures, with a success rate of approximately 89%, demonstrating the effectiveness of our method. Additionally, in terms of ambiguity resolution accuracy, IBUT outperformed the baseline by 13 acc points, indicating that bilingual understanding and iterative refinement contribute to enhancing ambiguity resolution capabilities in translation tasks.

Methods	Human Evaluation	
	Nums	**ACC**
Understanding Distortion of Baseline	28 (40%)	65.9
Understanding Distortion of IBUT	**3 (-89%)**	**78.7**

Table 4: The human-annotated results of the Commonsense MT benchmark. Baseline refers to the MAPS method modified into the form shown in Figure 1(a). In the baseline method, there are 70 sentences with translation errors.

Transaltion Quality. In human evaluation of translation quality, this study adopted the method (Liang et al., 2023) to validate translation quality on both the En→Zh and Zh→En test sets of the Cultural MT and the Commonsense MT dataset (Appendix B.7 for experimental setup details).

The experimental results are displayed in Figure 4. Within the Commonsense MT Dataset, IBUT performed best in terms of ambiguity resolution

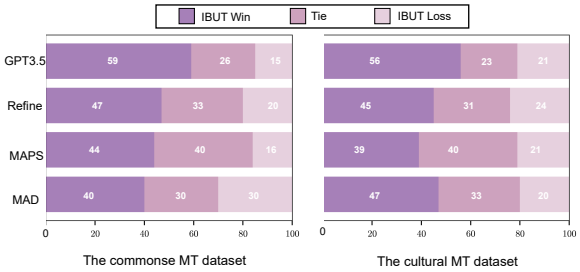

Figure 4: Human preference study comparing ChatGPT, Refine, MAPS, and MAD.

accuracy, thereby achieving higher human evaluation scores compared to other methods. In the Cultural MT Dataset, IBUT received higher human evaluation scores, indicating that its generated contextual understanding effectively enhances the performance of culturally relevant translation tasks.

6.3 Impact of Iterative Refinement on Translation Performance

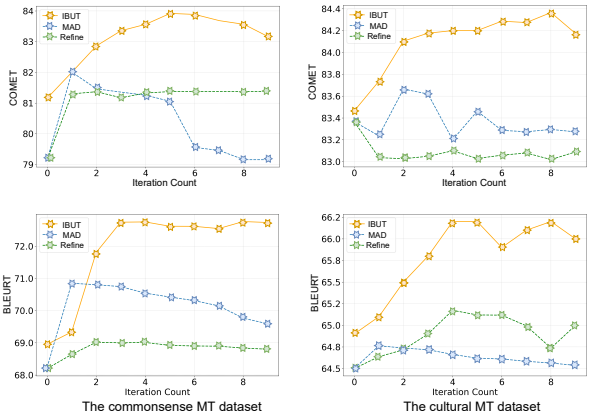

Figure 5: Analysis of the experimental setup for assessing the impact of the Iterative Refinement part on translation performance.

To further verify the impact of the Iterative Refinement part on overall translation performance, we conducted experiments on Cultural MT (En→Zh) and Commonsense MT (Zh→En), comparing methods like MAD and Refine to iteratively enhance translation quality. We set the maximum number of iterations at 9 and required that each iteration in the Iterative Refinement part obtain a new translation COMET score, rather than allowing adaptive termination in the Alignment Judgment part.

The experimental results, as shown in Figure 5, first indicate that IBUT surpasses the comparative methods in translation performance with each iteration, further proving the effectiveness of the method. Secondly, compared to the comparative methods, IBUT progressively enhances its performance in each iteration, demonstrating that the dual learning of translation can provide positive supervision signals in each iteration.

6.4 Effectiveness of Bilingual Contextual Understanding and Ablation Experiments

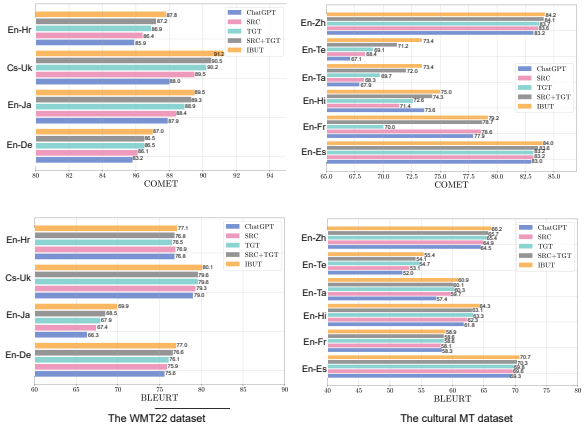

Figure 6: **Effectiveness of Bilingual Contextual Understanding and Ablation Experiments.** On the left are results for four language pairs from WMT22, and on the right are results for five language pairs from cultural MT. ChatGPT for direct translation; SRC for translation based on source language understanding; TGT for translation based on target language understanding; SRC+TGT for translation based on both source and target language understanding; IBUT as proposed method in section 3.

The IBUT introduced bilingual contextual understanding based on the source sentence to improve translation performance. To evaluate the effects of bilingual contextual understanding, we designed 5 control methods: (a) LLM-MT directly translating (ChatGPT); (b) LLM generating contextual understanding based on the source language, translated by LLM-MT (SRC); (c) LLM generating contextual understanding based on the target language, translated by LLM-MT (TGT); (d) LLM generating contextual understanding for both source and target languages, translated by LLM-MT (SRC+TGT); (e) using the IBUT method described in section 3.

The effectiveness of Bilingual Contextual Understanding. The experimental results are shown in Figure 6. On the WMT22 and cultural MT datasets, first, the translation results

based on contextual understanding surpass the baseline methods, demonstrating the validity of our research direction. Second, bilingual (SRC+TGT) contextual understanding significantly improves translation performance compared to monolingual (SRC or TGT) understanding, indicating that bilingual contextual understanding effectively enhances translation results. Additionally, experimental results show that target language understanding (TGT) contributes more to translation performance improvement than source language understanding (SRC).

Ablation Experiments on Different Parts of IBUT. As shown in Figure 6, "SRC or TGT" indicates the use of only the Understanding Generation part in IBUT, while "SRC+TGT" indicates that IBUT does not use iterative refinement. The experimental results demonstrate that the performance of the IBUT method surpasses these methods, thereby proving the rationality and effectiveness of the IBUT method design.

6.5 IBUT Demonstrates Generalizability in Model Selection

WMT22	En→De	En→Ja
Metrics	COMET/BLEURT/BLEU	
Alpaca-7B	75.5/62.2/11.3	56.6/31.4/6.6
+5shot	76.3/62.8/12.1	57.9/31.9/7.0
+MAPS	76.7/63.5/12.6	58.3/33.9/7.5
+IBUT	**78.4/64.9/13.1**	**61.3/34.8/8.2**
Vicuna-7B	79.8/67.4/15.2	82.3/58.7/9.4
+5shot	80.3/67.8/15.3	83.3/59.3/9.5
+MAPS	81.1/68.4/16.1	84.4/60.3/9.8
+IBUT	**82.0/69.1/17.3**	**85.1/61.1/11.0**

Table 5: The experimental results of IBUT on open-source models. Alpaca-7B and Vicuna-7B mean to perform translation directly through Zero-Shot. The bold indicates the highest values that are statistically significant, with p-values less than 0.01 in the paired t-test against all compared methods.

To validate the generalizability of the IBUT method on open-source models, we selected two open-source models (Alpaca and Vicuna) for experimental verification. The experimental results, as shown in Table 5, indicate that the overall performance trends of the two open-source models are consistent with those observed using the GPT3.5 model. This demonstrates the generalizability of the IBUT method in open-source models. Additionally, we further validated the effectiveness of the IBUT method in GPT-4. The results are shown in Appendix B.6.

7 Computational Resource Analysis

Since the IBUT method requires multiple iterative steps, it is necessary to discuss and analyze its resource consumption. For token consumption, we used the gpt-3.5-turbo tokenizer[7] to tokenize and then calculated the token consumption of the comparative methods requiring iteration on the commonsense dataset. The results are shown in Table 6.

Methods	Avg I/O	COMET/BLEURT/BLEU
ChatGPT	11.7/24.4	79.7/68.2/29.8
+5-shot	59.4/ 35.6	79.9/68.6/30.2
+MAPS	167.7/172.2	81.9/69.4/27.2
+MAD	202.2/224.4	82.0/70.8/29.1
+IBUT	194.6/209.3	83.9/72.7/32.6

Table 6: The statistics of methods inference cost on the commonsense dataset. The I/O represent Input/Output.

The experimental results show that the IBUT method increases token consumption by about 5 times compared to the 5-shot method, but it improves the COMET/BLEURT/BLEU translation metrics by +4.0/+4.1/+2.4, respectively, indicating a considerable performance gain. The IBUT method is on the same level as other strong methods (MAD and MAPS), achieving an average improvement of 2 points. For inference time analysis, refer to Appendix B.1. Additionally, we discuss this limitation in the Limitations section, noting it as a future research direction for the MT community.

8 Conclusion

In this paper, we introduce a new method called **I**terative **B**ilingual **U**nderstanding **T**ranslation (IBUT), designed to address the inherent challenges of decreased translation performance in LLM-MT due to error introduced by contextual understanding. The IBUT initially generates bilingual contextual understanding, and then constructs a supervisory signal based on the dual learning of the translation task, thereby iteratively refining bilingual contextual understanding to enhance LLM-MT performance. The method performs well in various scenarios involving general news translation tasks, commonsense MT, and cultural MT datasets, with human evaluations confirming its effectiveness.

[7] https://github.com/openai/tiktoken

9 Limitations

The IBUT method has several limitations. Firstly, models with stronger understanding and generation capabilities will obtain better contextual understanding, thereby enhancing translation performance. Additionally, since our method requires multiple steps, it necessitates a significant amount of computational resources.

References

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

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. arXiv preprint arXiv:1606.02006.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024a. Benchmarking llms for translating classical chinese poetry: Evaluating adequacy, fluency, and elegance. arXiv preprint arXiv:2408.09945.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024b. DUAL-REFLECT: Enhancing large language models for reflective translation through dual learning feedback mechanisms. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 693–704, Bangkok, Thailand. Association for Computational Linguistics.

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

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

António Farinhas, José Guilherme Camargo de Souza, and André F. T. Martins. 2023. An empirical study of translation hypothesis ensembling with large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, pages 11956–11970. Association for Computational Linguistics.

Yuan Gao, Ruili Wang, and Feng Hou. 2023. How to design translation prompts for chatgpt: An empirical study. arXiv e-prints, pages arXiv–2304.

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

Nuno Miguel Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models. CoRR, abs/2303.16104.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. Advances in neural information processing systems, 29.

Jie He, Tao Wang, Deyi Xiong, and Qun Liu. 2020. The box is in the pen: Evaluating commonsense reasoning in neural machine translation. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3662–3672, Online. Association for Computational Linguistics.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring human-like translation strategy with large language models. ArXiv, abs/2305.04118.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. Transactions of the Association for Computational Linguistics, 12:229–246.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. arXiv preprint arXiv:2302.09210.

Junjie Hu, Hiroaki Hayashi, Kyunghyun Cho, and Graham Neubig. 2021. Deep: denoising entity pre-training for neural machine translation. arXiv preprint arXiv:2111.07393.

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. arXiv preprint arXiv:1906.00376.

Yichong Huang, Xiaocheng Feng, Baohang Li, Chengpeng Fu, Wenshuai Huo, Ting Liu, and Bing Qin. 2024. Aligning translation-specific understanding to general understanding in large language models. arXiv preprint arXiv:2401.05072.

Vivek Iyer, Pinzhen Chen, and Alexandra Birch. 2023. Towards effective disambiguation for machine translation with large language models. In Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023, pages 482–495. Association for Computational Linguistics.

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

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023, pages 419–451. Association for Computational Linguistics.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Nearest neighbor machine translation. arXiv preprint arXiv:2010.00710.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In Proceedings of the Eighth Conference on Machine Translation, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In Proceedings of the Seventh Conference on Machine Translation (WMT), pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

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

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023. Chain-of-dictionary prompting elicits translation in large language models. arXiv preprint arXiv:2305.06575.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023a. Adaptive machine translation with large language models. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023, Tampere, Finland, 12-15 June 2023, pages 227–237. European Association for Machine Translation.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023b. Adaptive machine translation with large language models. arXiv preprint arXiv:2301.13294.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023c. Domain terminology integration into machine translation: Leveraging large language models. In Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023, pages 902–911. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35:27730–27744.

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.

Ratish Puduppully, Anoop Kunchukuttan, Raj Dabre, Aiti Aw, and Nancy Chen. 2023. Decomt: Decomposed prompting for machine translation between related languages using large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 4586–4602.

Tao Qin. 2020. Dual learning. Springer.

Leonardo Ranaldi, Giulia Pucci, and André Freitas. 2023. Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations. CoRR, abs/2308.14186.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In Proceedings of the Seventh Conference on Machine Translation (WMT), pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Proceedings of the Seventh Conference on Machine Translation (WMT), pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Seongjin Shin, Sang-Woo Lee, Hwijeen Ahn, Sungdong Kim, HyoungSeok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, Woomyoung Park, Jung-Woo Ha, et al. 2022. On the effect of pretraining corpora on in-context learning by a large-scale language model. arXiv preprint arXiv:2204.13509.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Gladys Tyen, Hassan Mansoor, Peter Chen, Tony Mak, and Victor Cărbune. 2023. Llms cannot find reasoning errors, but can correct them! arXiv preprint arXiv:2311.08516.

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, EMNLP 2023, Singapore, December 6-10, 2023, pages 16646–16661. Association for Computational Linguistics.

Xing Wang, Zhaopeng Tu, Deyi Xiong, and Min Zhang. 2017. Translating phrases in neural machine translation. arXiv preprint arXiv:1708.01980.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net.

Minghao Wu, Yulin Yuan, Gholamreza Haffari, and Longyue Wang. 2024. beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. arXiv preprint arXiv:2405.11804.

Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. 2023. Empowering llm-based machine translation with cultural awareness. abs/2305.14328.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023. Improving machine translation with large language models: A preliminary study with cooperative decoding. CoRR, abs/2311.02851.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Yang Xiang, and Min Zhang. 2024. Paying more attention to source context: Mitigating unfaithful translations from large language model. In Findings of the Association for Computational Linguistics ACL 2024, pages 13816–13836, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Tiejun Zhao, Muven Xu, and Antony Chen. 2024. A review of natural language processing research. Journal of Xinjiang Normal University (Philosophy and Social Sciences), pages 1–23.

A Experiment Setup

A.1 Detailed prompt for part-1

Part-1: Understanding Generation: Please fully understand the meaning of the following L^s text from your memory and describe your understanding of key concepts, definitions, examples, and explanations of specific terms related to the translation task in L^s/L^t:s

Input Text:

Source Sentence s

Output Text:

C_s or C_t

A.2 Detailed prompt for part-2

Part-2: Alignment Judgment-1: If you are a L^s and L^t linguist, determine whether provided source contextual understanding C_s and target contextual understanding C_t, based on the source sentence s, convey different key concepts, definitions, examples, and explanations of specific terms related to the translation task. If so, provide a '$True$' response; otherwise, give a '$False$' response.

Input Text:

Source Sentence s and source/target contextual understanding C_s/C_t

Output Text:

$True$ or $False$

Part-2: Alignment Judgment-2: If you are a linguist proficient in both L^s and L^t, based on the core meaning of the source sentence s, analyze the source contextual understanding C_s / the target contextual understanding C_t. Generate verbal feedback in the language of C_s/C_t to correct any current errors in C_s/C_t.

Input Text:

Source Sentence s, source/target language understanding C_s/C_t

Output Text:

F_s or F_t

A.3 Detailed prompt for part-3

Part-3: Iterative Refinement: If you are a linguist proficient in both L^s and L^t, based on the core meaning of the source sentence s and the opinions from F_s/F_t, further modify the current C_t/C_s.

Input Text:

Source Sentence s, source/target contextual understanding C_s/C_t and source/target verbal feedback F_s/F_t

Output Text:

C_s or C_t

A.4 Detailed prompt for part-4

Part-4:Understanding-Based Translation: Based on C_t and C_s, translate the following text from L^s to L^t.

Input Text:

Source Sentence s, source/target contextual understanding C_s/C_t

Output Text:

Target Sentence t

A.5 Dataset Detail

For the WMT22 test set (Kocmi et al., 2022), the experimental analysis covers 9 language pairs. We used the full test dataset. Among these languages, Sah↔Ru, Uk↔Cs, En→Hr and En↔Liv are classified as low-resource languages, respectively.

For the WMT23 test set (Kocmi et al., 2023), the experimental analysis covers 4 language pairs. We used the full test dataset. Among them, En→De and En→Ja are identified as high and medium-resource languages, with the former belonging to the same language family and the latter exhibiting significant differences.

The Commonsense MT dataset (He et al., 2020) encompasses vocabulary that requires common knowledge for resolution, along with instances

of ambiguity in Zh→En translation data. Each translation data includes a source sentence and two contrasting translations, involving seven different types of common knowledge. Although these sentences appear suitable for direct translation, they often lead to misleading interpretations.

The cultural MT dataset (Yao et al., 2023) introduces a culturally relevant parallel corpus, enriched with annotations of cultural-specific items. This dataset encompasses 6 language pairs: En→Es, En→Fr, En→Hr, En→Ta, En→Te, and En→Zh. It also encompasses over 7,000 cultural-specific items from 18 concept categories across more than 140 countries and regions.

A.6 Comparative Methods

The following content will provide detailed descriptions of these comparative methods:

- **Baseline** is standard zero-shot translation performed in ChatGPT (Ouyang et al., 2022) and GPT-4 (Achiam et al., 2023). The temperature parameter set to 0, which is the default value for our experiments.

- **5-Shot** (Hendy et al., 2023) involves prepending five high-quality labelled examples from the training data to the test input.

- **Rerank** (Moslem et al., 2023a) was conducted with the identical prompt as the baseline, employing a temperature of 0.3 (Moslem et al., 2023b). Three random samples were generated and combined with the baseline to yield four candidates. The best candidate was chosen through GPT-4.

- **Refine (Chen et al., 2023)** first requests a translation from ChatGPT, then provides the source text and translation results, and obtains a refined translation through multiple rounds of modifications.

- **MAPS (He et al., 2023)** incorporate the knowledge of keywords, topic words, and demonstrations similar to the given source sentence to enhance the translation process.

- **MAD (Liang et al., 2023)** enhance the capabilities of large language models (LLMs) by encouraging divergent thinking. In this method, multiple agents engage in a debate, while an agent oversees the process to derive a final solution.

- **IBUT** is proposed method in Sec.3. The method uses only ChatGPT with a max number of iterations set to 8 ($max_iter = 8$).

B Experiment Results

B.1 Computational Costs

We illustrate with our method based on Vicuna-7B, using a single A100 GPU with 80G. Our proposed IBUT method has an inference speed of 6.712s/sample with a batch size of 2 and memory usage of 17657MiB. If using Vicuna-7B for zero-shot inference, under the same batch size settings, the inference speed is 4.72s/sample with memory usage of 149658MiB.

B.2 The Experiment Setting of Error Reduction and Translation Enhancement

For the Commonsense MT lexical ambiguity subset, first manually annotate the correct understanding of ambiguous words. The annotated data includes the source language Chinese and the target language English. Details of the scoring prompt for GPT-4, focusing on the reduction of error in bilingual contextual understanding after iterative refinement, are as follows:

Prompt for GPT-4 Evaluation Please evaluate the source input s, contextual understanding C_s/C_t, and the manually annotated meanings of lexical ambiguities to assess if the contextual understanding includes error content to the translation.

Scoring Guide:

1-2 points: The contextual understanding completely deviates from the source input, leading to generated content that is severely incorrect or irrelevant.

3-4 points: The contextual understanding partially deviates from the source input, resulting in partially relevant content with evident issues.

5-6 points: Although the contextual understanding does not completely deviate, there are errors in the interpretation of the source input, leading to content that is partially correct but flawed.

7-8 points: The contextual understanding is fundamentally accurate, correctly handles the source input, and the generated content is largely correct with only minor errors.

9-10 points: The contextual understanding is completely accurate, perfectly handles the source input and lexical ambiguities, and the generated content fully meets the requirements, successfully avoiding irrelevant content.

Based on these guidelines, score the model response from 0 to 10. Provide only the total score (just a number), without scores or explanations for each aspect. The score is __.

Input Text:

Source Sentence s, source/target context understanding C_s/C_t

Output Text:

The score is __

B.3 WMT22 Results on BLEU

The translation results based on WMT22 under the BLEU evaluation metric are shown in Table 7. The experimental results of BLEU show the same trend as other evaluation metrics.

B.4 Results on WMT23

To further validate the generalizability of the method, we conducted experiments on the WMT23 test set. The experimental results are shown in Table 8.

B.5 Results on Reference-free metric

To further clarify the robustness of our evaluation, we incorporated COMET-KIWI[8] (Rei et al., 2022b), a reference-free metric in the COMET series. The experimental results are shown in Table 9.

These results demonstrate that our method still outperforms comparison methods in terms of COMET-KIWI scores, thereby further confirming the robustness of our evaluation.

B.6 General Performance

To demonstrate the generalizability of the method, we conducted experiments in Section 6.5, verifying that IBUT is effective not only on closed-source models but also on open-source models. Finally, since GPT-4 is an updated model of GPT-3.5, our method's effectiveness on GPT-3.5 theoretically implies effectiveness on GPT-4. To further illustrate this point, we conducted experiments on GPT-4 for commonsense MT. The experimental results are shown in Table 10.

The experimental results demonstrate that our method achieves significant improvements when applied to GPT-4, thereby indicating the generalizability of our approach.

B.7 Human Evaluations

Human Evaluation of Understanding Distortion Issue. In this section, we conduct a human evaluation to measure translation quality. We assess understanding distortion issues and ambiguity resolution. We invited one annotator to participate (a professional translator). The annotator first identifies and counts the sentences with ambiguity errors in the Baseline translation. Then, among these erroneous sentences, the annotator further filters and counts those where the errors are caused by contextual understanding. Finally, the annotator identifies and counts the sentences where the Baseline translation is incorrect but the IBUT translation is correct, and where the contextual understanding in the IBUT translation generates the correct sentence. Additionally, in the CommonsenseMT task, the five experts scored each sample for ambiguity resolution against the reference, awarding 1 point for resolved and 0 points for unresolved.

Human Evaluation of Translation Quality. We conducted a human preference study on both the English-Chinese and Chinese-English test sets of the Cultural MT Datasets and the Commonsense MT Dataset. We invited one annotator to participate (a professional translator), and we randomly selected 100 translation results of the same source sentences generated by methods such as ChatGPT, Refine, MAPS, MAD, and IBUT. In terms of translation quality, the annotators compared the translation results of IBUT against other comparative methods. For the same source

[8] https://github.com/Unbabel/COMET

WMT22	En→De	En→Ja	Cs→Uk	Uk→Cs	En→Hr	Sah→Ru	Ru→Sah	En→Liv	Liv→En
Metrics	BLEU								
ChatGPT	32.3	17.3	29.9	30.6	26.9	5.9	1.9	2.4	8.5
+5shot	32.9	17.9	29.3	31.2	25.8	6.4	2.3	2.7	8.8
+Rerank	33.6	21.2	29.5	31.9	26.9	6.5	2.6	2.9	8.9
+MAD	32.9	19.7	31.6	31.6	26.5	6.7	2.6	3.1	9.7
+MAPS	33.1	21.2	29.5	31.4	27.0	6.7	2.2	2.9	9.7
+Refine	33.8	23.4	30.3	32.8	27.5	6.7	2.5	3.3	9.5
+IBUT	**34.5**	**24.3**	**31.9**	**34.3**	**28.5**	**6.9**	**4.9**	**4.7**	**10.1**

Table 7: The BLEU results from the WMT22 news benchmark are presented. The bold indicates the highest values that are statistically significant, with p-values less than 0.01 in the paired t-test against all compared methods.

WMT23	En→De	En→Ja	En→He	Cs→Uk
Metrics	COMET/BLEURT/BLEU			
ChatGPT	83.5/69.1/39.7	87.3/60.2/9.7	82.1/69.3/22.3	86.7/74.1/27.2
+5shot	83.7/69.4/40.1	87.8/61.5/10.1	82.5/69.8/22.5	87.3/74.5/27.5
+MAD	83.9/70.3/41.6	88.0/63.1/9.4	82.9/70.0/24.0	87.5/74.9/28.5
+MAPS	83.6/69.9/42.1	87.9/62.6/9.8	82.5/69.3/23.1	87.8/74.6/28.0
+Refine	83.5/68.9/41.8	87.6/62.4/10.8	82.3/68.8/23.7	87.3/74.1/28.3
+IBUT	**84.3/71.8/42.6**	**88.5/63.8/14.0**	**83.1/72.1/24.9**	**88.1/77.9/30.4**

Table 8: The main results from WMT23 are shown. The highest values are in bold, with p-values less than 0.01.

Methods	En-De	En-Ja	Cs-Uk	En-Hr
ChatGPT				
+Rerank	82.1	84.4	83.6	83.4
+MAPS	82.4	84.2	83.0	83.4
+MAD	82.0	83.7	83.6	83.3
+IBUT	**83.6**	**84.7**	**84.2**	**83.8**

Table 9: WMT22 evaluation results on COMET-KIWI metric.

Methods	COMET/BLEURT/BLEU
GPT-4	82.0/71.0/32.6
+5 shot	82.3/71.5/32.9
+Rerank	82.9/72.0/32.9
+IBUT	84.3/73.6/32.8

Table 10: General Performance of general performance on commonsense MT

sentences, if IBUT's translation quality is superior, it is marked as IBUT Win; if the translation qualities are comparable, it is marked as Tie; if the translation quality of other methods is better, it is marked as IBUT Loss. We conducted three rounds of revisions on all evaluation results to increase the fairness of the assessments as much as possible. For the content with Chinese ambiguity in the commonsense MT dataset, we ensured the correctness of the source side understanding by confirming it with classmates whose native language is Chinese.

B.8 IBUT Demonstrates Generalizability on Low-Resource Languages

To further explore whether the IBUT method can be effective in low-resource translation tasks using open-source models, we conducted experiments on the low-resource directions of WMT23[9]. The experimental results are shown in Table 11, demonstrating that our method significantly improves the performance of open-source models in low-resource translation, thereby further validating the generalizability of IBUT.

B.9 Introduce the Full Names of Languages.

To better understand the experimental setup, we present the language codes and their corresponding full language names in Table 12.

[9]https://www.statmt.org/wmt22/translation-task.html

WMT22	Cs→Uk	En→Hr
Metrics	COMET/BLEURT/BLEU	
Alpaca-7B	74.1/52.4/8.31	65.9/53.2/8.1
+5shot	75.9/53.1/8.3	67.9/53.6/8.3
+MAPS	76.3/53.7/9.2	68.1/54.2/8.9
+IBUT	**77.9/54.3/9.5**	**69.2/55.1/9.0**
Vicuna-7B	74.9/57.8/10.5	69.3/57.7/9.9
+5shot	76.3/58.3/10.9	70.2/58.1/10.7
+MAPS	77.2/59.6/11.1	71.1/58.8/11.6
+IBUT	**78.3/60.7/11.5**	**72.9/60.4/13.1**

Table 11: The experimental low-resource results of IBUT on open-source models. Alpaca-7B and Vicuna-7B mean to perform translation directly through Zero-Shot. The bold indicates the highest values that are statistically significant, with p-values less than 0.01 in the paired t-test against all compared methods.

Language Codes	Full Name of Language Code
En	English
JA	Japanese
Cs	Czech
uk	Ukrainian
De	German
Hr	German
Ru	Russian
Hi	Hindi
Ta	Tamil
Te	Telugu
Fr	French
Es	Spain
Sah	Yakut
Liv	Livonian

Table 12: The language names corresponding to the language codes.