DUAL-REFLECT: Enhancing Large Language Models for Reflective Translation through Dual Learning Feedback Mechanisms

Anonymous ACL submission

Abstract

Recently, large language models (LLMs) enhanced by self-reflection have achieved promising performance on machine translation. The key idea is guiding LLMs to generate translation with human-like feedback. However, existing self-reflection methods lack effective feedback information, limiting the translation performance. To address this, we introduce a DUAL-REFLECT framework, leveraging the dual learning of translation tasks to provide effective feedback, thereby enhancing the models' self-reflective abilities and improving translation performance. The application of this method across various translation tasks has proven its effectiveness in improving translation accuracy and eliminating ambiguities, especially in translation tasks with low-resource language pairs.

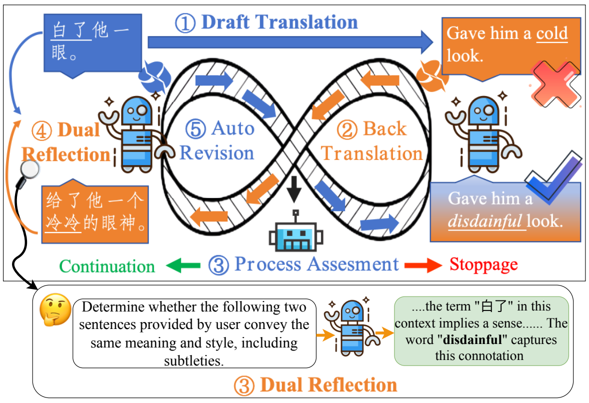

Figure 1: DUAL-REFLECT first obtains an initial translation result, then performs back-translation, and LLMs reflect on the differences between the back-translation results and the original source content to obtain feedback signals, ultimately optimizing the translation outcome.

1 Introduction

Large language models (LLMs) have recently demonstrated remarkable abilities across a variety of tasks, achieving a significant milestone in the pursuit of artificial general intelligence (AGI) (Bubeck et al., 2023a; Xu and Poo, 2023; Zhao et al., 2023). Notably, in the field of machine translation, LLMs have improved translation quality by adopting human-like methods of self-reflection (Shinn et al., 2023; Liang et al., 2023). The self-reflection process primarily relies on using LLMs to iteratively refine initial drafts through feedback loops, a method that has been widely researched and explored (Shinn et al., 2023; Park et al., 2023; Scheurer et al., 2022; Le et al., 2022; Welleck et al., 2022; Amabile and Amabile, 1983; Flower and Hayes, 1981; Simon, 1962). The lack of effective feedback limits the self-reflective capacity of Large Language Models (LLMs), thereby affecting their continuous improvement in translation (Tyen et al., 2023; Liang et al., 2023).

To address this, we introduce a framework that leverages the inherent duality property (He et al., 2016; Qin, 2020) of translation tasks to provide effective feedback to LLMs, thereby enhancing their reflective capabilities and consequently improving translation performance. This method, named DUAL-REFLECT, stands for **DUAL** learning enhanced auto-**REFLEC**tive **T**ranslation and comprises five stages: Draft Translation, Back Translation, Process Assessment, Dual-Reflection, Auto Revision. In the draft translation stage, LLMs employ their inherent translation capabilities to generate a draft translation. Subsequently, in the back translation stage, LLMs translate the draft translation back to the source language. Then, during the process assessment stage, an LLM-based agent is introduced to assess whether dual reflection is needed. If not, it outputs the final result; otherwise, the process continues to cycle through all the steps. Based on this, in the dual reflection stage, LLMs reflect on the differences between the back-translation results and the initial source input, revealing potential translation biases. LLMs further analyze the reasons for these discrepancies and propose suggestions for improvement. Finally, In the auto-revision stage, LLMs modify the initial translation by incorporating the analysis and improvement suggestions obtained through dual reflection.

We verify the effectiveness of the DUAL-REFLECT framework across four translation directions in the WMT22, covering high, medium, and low resource languages, as well as a commonsense reasoning MT Benchmark. Automatic evaluation results show that DUAL-REFLECT outperforms strong baseline methods, significantly enhancing translation performance. Notably, on low-resource translation tasks, DUAL-REFLECT achieved an average result that surpassed ChatGPT by +1.6 COMET. In addition, DUAL-REFLECT enhanced ChatGPT exceeded GPT-4 on the commonsense reasoning MT benchmark. Further human evaluation demonstrates that DUAL-REFLECT shows a better ability to resolve translation ambiguities compared to other methods. Our code will be made available at https://github.com/anonymous.

2 Approach: DUAL-REFLECT

Our DUAL-REFLECT framework consists of Five key stages, described in detail as follows:

2.1 Stage-1: Draft Translation

In the draft translation stage, LLMs utilize their inherent translation capabilities to generate a draft translation from the source language L^s to the target language L^t. The instruction template for this translation task is as follows:

Translation Instruction: Translate the following text from L^s to L^t:
Input Text:

Source Sentence x

Output Text:

Target Sentence y

2.2 Stage-2: Back Translation

In this stage, the same instruction as used in the draft translation stage is adopted. The goal is to back-translate the initial translation result from the target language L^t back to the source language L^s, with the output being x'.

2.3 Stage-3: Process Assessment

We introduce an evaluation agent, denoted as PA, to supervise and control the entire translation process. This Agent has two different modes:

Judgment Mode: PA determines whether it can accurately identify the differences between x and x' within a given specific number of iterations. If $PA(x, x') = False$, the Dual Reflection stage is terminated; otherwise, the entire process continues.

Pattern Extraction: In the judgment mode, once determined to be $True$ or after exceeding the predefined number of iterations, PA is responsible for extracting the final translation result from the entire output, denoted as $PA(x, x') = final\_translation$.

2.4 Stage-4: Dual Reflection

The goal of the dual reflection stage is to reflect on the differences between the source sentences generated by back-translation and the initial source input. Then, it outputs analysis results and proposes suggestions to enhance translation performance.

Dual Reflection Instruction: Determine whether the two sentences provided by the user convey the same meaning, style, and subtleties, and based on the analysis, translate all words or phrases in the first sentence that cause the aforementioned differences into the target language:
Input Text:

Source Sentence x' and x

Output Text:

Analysis Results (AR) and Translation Suggestions (TS)

2.5 Stage-5: Auto Revision

In this stage, utilizing the output of the dual reflection and the original source sentences as input, the original source sentences are re-translated (from L^s to L^t).

Auto Revision Instruction: Translate the following text from L^s to L^t:
Input Text:

Analysis Results (AR), Translation Suggestions (TS) and x

Output Text:

Target Sentence y

Methods	En→De		En→Ja		Cs→Uk		En→Hr	
	COMET	BLEURT	COMET	BLEURT	COMET	BLEURT	COMET	BLEURT
ChatGPT	85.8	75.6	87.9	66.3	88.0	75.0	85.9	75.0
+5-shot	86.5	76.3	88.2	67.1	88.3	-	86.4	-
+Rerank	86.0	75.9	88.0	66.6	88.3	75.3	86.3	75.4
+MAPS	86.4	76.3	88.5	67.4	88.8	76.1	86.5	76.0
+Self-Reflect	86.3	76.1	88.3	66.9	88.4	76.0	86.3	75.8
+DUAL-REFLECT	**86.5**	**76.4**	**88.7**	**67.9**	**90.2**	**77.3**	**86.9**	**76.4**

Table 1: The main results from the WMT22 benchmark are presented. The meaning of the ChatGPT is to utilize the GPT-3.5-turbo API for Zero-Shot translation. The bold indicates the highest value.

3 Experiments

3.1 Experimental Setup

Test Data. To mitigate concerns of data leakage as highlighted by Bubeck et al., 2023b, Garcia et al., 2023, and Zhu et al., 2023, we leveraged the WMT22[1] (Kocmi et al., 2022) test set in our evaluation framework. Additionally, to further evaluate DUAL-REFLECT's performance in complex translation tasks, we employed the Commonsense Reasoning MT dataset (He et al., 2020), consisting of Chinese→English translation examples. See Appendix A.1 for specific details.

Comparing Systems. In our evaluation, the DUAL-REFLECT framework is compared with a range of models, including ChatGPT(Ouyang et al., 2022), GPT-4[2] (Achiam et al., 2023), ReRank, Self-Reflect(Shinn et al., 2023), MAD(Liang et al., 2023), and MAPS(He et al., 2023). See Appendix A.2 for specific details.

Evaluation Metrics. In evaluating our translation methodology, we initially employ COMET[3] (Rei et al., 2022) and BLEURT[4] (Sellam et al., 2020) as automatic metrics, aligning with the established standards in both LLM-based translation literature (He et al., 2023; Huang et al., 2024). To further evaluate our translation method, we employ human evaluations to verify translation performance and the ability to resolve translation ambiguities. Details on human evaluations are in Appendix A.4.

3.2 Main Results

The main results are presented in Tables 1 and 2. Based on these outcomes, we derive the subsequent insights:

The effectiveness of DUAL-REFLECT has been validated across a wide range of settings. As shown in Table 1, DUAL-REFLECT achieves the best performance compared to other methods across various levels of language similarity and resource availability. Specifically, DUAL-REFLECT demonstrates an average improvement of +1.18 COMET over the baseline ChatGPT and +0.75 COMET over the Self-Reflect methods. In the low-resource Cs→Uk translation task, DUAL-REFLECT surpasses ChatGPT and MAPS by +2.2 and +1.4 COMET, respectively. These improvements indicate that DUAL-REFLECT has broad applicability across different levels of resource availability and language similarity, especially exhibiting more pronounced improvements in language pairs with low resources.

Methods	AutoMetrics	
	COMET	**BLEURT**
GPT-4	82.0	71.0
ChatGPT		
+Zero-Shot	79.7	68.2
+Rerank	80.9	68.9
+MAPS	81.9	-
+Self-Reflect	80.9	68.7
+MAD	82.0	69.4
+DUAL-REFLECT	**82.2**	**71.8**

Table 2: The main results from the Commonsense MT benchmark are presented. The bold indicates the highest value.

The effectiveness of DUAL-REFLECT in commonsense reasoning translation tasks. The results, presented in Table 2, show that in commonsense reasoning translation tasks, DUAL-REFLECT significantly outperforms other methods, achieving the best translation performance. Compared to the Self-Reflect method, it showed an improvement of +1.3 COMET, indicating more effective error correction capabilities. Moreover, DUAL-REFLECT also surpassed the MAD method, which relies on feedback from multi-agent debate, demonstrating the high quality of its feedback. Notably, in translation tasks involving logical reasoning, DUAL-REFLECT's performance even exceeded that of GPT-4, proving

[1]https://www.statmt.org/wmt22/index.html

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

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

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

its exceptional logical reasoning ability.

4 Analysis

We thoroughly analyze our approach, with results primarily reported on CommonsenseMT Zh→En unless stated otherwise.

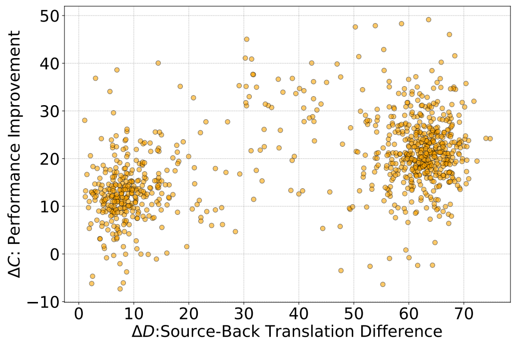

Figure 2: Effectiveness experiment of Dual Learning, each point represents a translation data from the test set.

4.1 The Effectiveness of Dual Learning

In this study, we explore the potential positive impact of a dual learning feedback mechanism on translation performance, as shown in Figure 2. The horizontal axis denotes $\Delta D = 100 - COMET(x, x')$, the disparity between the original sentence x and its back-translated version x'. The vertical axis quantifies improvement in translation performance, as a COMET metric difference (ΔC), between DUAL-REFLECT and ChatGPT. Findings show a correlation coefficient of 0.46, indicating that feedback from dual learning improves the model's reflective capabilities, thus enhancing translation accuracy. Additionally, the experimental data shows significant differences between the output x' and the original source sentence x in the initial back-translation ($\Delta D > 50$), further confirming the universality of differences obtained from the dual learning in translation tasks.

4.2 Human Evaluation

In terms of human evaluation, this study follows the method of Liang et al., 2023 to assess translation outcomes from two main dimensions: accuracy in ambiguity resolution and direct assessment of translation quality (details in Appendix 3).

The experimental results are presented in Table 3. Regarding the accuracy of ambiguity resolution, DUAL-REFLECT performs the best, indicating that dual feedback contributes to better disambiguation in translation tasks. In terms of human evaluation, DUAL-REFLECT receives the highest ratings, further demonstrating that the method achieves superior translation quality.

Methods	Human Evaluation	
	Score	ACC
GPT-4	3.9	69.8
ChatGPT		
+Zero-Shot	3.1	63.8
+Rerank	3.3	66.8
+Self-Reflect	3.4	64.9
+MAD	3.7	76.2
+DUAL-REFLECT	**4.2**	**77.4**

Table 3: The human-annotated results of the Commonsense MT benchmark.

4.3 Examine how iteration rounds affect results.

In this experimental design, we require reviewer PA to determine the final answer ($PA(x, x') = final\_translation$) in each iteration, rather than allowing adaptive termination of iterations as described in Section 2.3. Figure 3 presents the outcomes, revealing DUAL-REFLECT's superior performance over the benchmark method as iterations progress, notably achieving the highest COMET score in three iterations. This emphasizes DUAL-REFLECT's ability to provide improved translations through repeated iterations, demonstrating the effectiveness and robustness of its dual learning feedback mechanism.

5 Case Study

This section presents a case study on the DUAL-REFLECT method, assessing its effectiveness and constraints via examples (detailed in Appendix A.5). Positive instances (Figures 4 to 6) illustrate substantial enhancements in translation accuracy and semantic coherence due to its reflective and iterative processes. Conversely, negative examples (Figures 7 and 8) highlight the dependency of DUAL-REFLECT's success on Back Translation quality, suggesting limitations in its capacity for improvement. This underscores the method's reliance on the integrity of each cycle component for optimal performance.

6 Conclsuion

We introduced DUAL-REFLECT, an LLM-based machine translation method, that leverages dual learning to improve reflection and performance, excelling in resource-limited and common sense reasoning scenarios, with human evaluations confirming its effectiveness.

7 Limitations

The DUAL-REFLECT framework enhances the reflective capabilities of LLMs in translation tasks by leveraging the duality nature of translation but has several limitations. Firstly, models with stronger reflective capabilities will obtain better feedback, thereby enhancing more performance. Additionally, since our method requires multiple steps, it necessitates a significant amount of computational resources.

8 Ethics Statement

One of the core design principles of the DUAL-REFLECT framework is a strict respect for intellectual property rights. This applies to both the methods and algorithms developed within the framework as well as those cited from the literature, all adhering strictly to copyright laws. Additionally, the framework upholds this principle in the handling of translation content, ensuring its use does not infringe upon the rights of original creators.

The framework also places a strong emphasis on responsibility during the automated translation process. By integrating stages of reflection and revision, DUAL-REFLECT enhances the transparency and interpretability of the translation methodology, thereby effectively identifying and correcting potential errors in the translation process.

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*.

Teresa M Amabile and Teresa M Amabile. 1983. A theoretical framework. *The Social Psychology of Creativity*, pages 65–96.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023a. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023b. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712.

Linda Flower and John R Hayes. 1981. A cognitive process theory of writing. *College composition and communication*, 32(4):365–387.

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 *International Conference on Machine Learning*, pages 10867–10878. PMLR.

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.

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*.

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.

Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu Hong Hoi. 2022. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35:21314–21328.

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*.

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

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.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, pages 2:1–2:22. ACM.

Tao Qin. 2020. *Dual learning*. Springer.

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

Jérémy Scheurer, Jon Ander Campos, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2022. Training language models with natural language feedback. *arXiv preprint arXiv:2204.14146*, 8.

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.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning.

Herbert A Simon. 1962. The architecture of complexity. *Proceedings of the American philosophical society*, 106(6):467–482.

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*.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2022. Generating sequences by learning to self-correct. *arXiv preprint arXiv:2211.00053*.

Bo Xu and Mu-ming Poo. 2023. Large language models and brain-inspired general intelligence. *National Science Review*, 10(10):nwad267.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

A Experiment Setup

A.1 Test Data

For the WMT22 test set(Kocmi et al., 2022), the experimental analysis covers four language pairs. 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. In contrast, Cs→Uk and En→Hr are categorized as low-resource languages, being closely related and belonging to the same language family, respectively.

The Commonsense Reasoning MT dataset(He et al., 2020) encompasses vocabulary that requires common knowledge for resolution, along with instances of contextual/contextless grammatical ambiguity in Chinese-to-English translation data. Each translation data includes a source sentence and two contrasting translations, involving seven different types of common knowledge. Despite these elements appearing amenable to direct translation, such simplified interpretations are often misleading.

A.2 Comparative Methods

The following sections provide detailed descriptions of these comparisons.

- **Baseline**, standard zero-shot translation is performed in ChatGPT (Ouyang et al., 2022) and GPT-4 (Achiam et al., 2023) with the temperature parameter set to 0, which is the default value for our experiments.
- **Rerank** was conducted with the identical prompt as the baseline, employing a temperature of 0.3, in alignment with Moslem et al., 2023. Three random samples were generated and combined with the baseline to yield four candidates. The optimal candidate was chosen through Quality Estimation (QE).
- **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.

- **Self-Reflect** (Shinn et al., 2023), This approach requires the LLM to scrutinize and refine its translation until it deems the current output satisfactory.
- **MAD** (Liang et al., 2023), incorporating the knowledge of keywords, topic words, and demonstrations similar to the given source sentence to enhance the translation process, respectively.

A.3 Examine how iteration rounds affect results

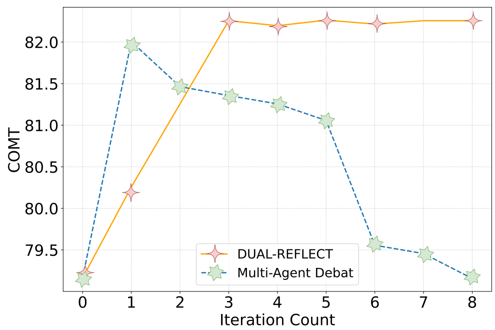

Figure 3: Translation performance varies based on the iteration of either DUAL-REFLECT or MAD.

A.4 Human Evaluations

In this section, we conduct human evaluation to measure translation quality. We assess coherence, fluency, and ambiguity resolution. Four experts were invited to participate, and 50 samples were randomly selected from translations generated by different methods. For translation quality, each sentence was rated on a scale from 1 to 5, with 3 indicating a pass, 4 showing substantial consistency with the reference, and 5 being the highest score. The final score is the average of these four ratings. Additionally, in the CommonsenseMT task, the four experts scored each sample for ambiguity resolution against the reference, awarding 1 point for resolved and 0 points for unresolved.

A.5 Case Study

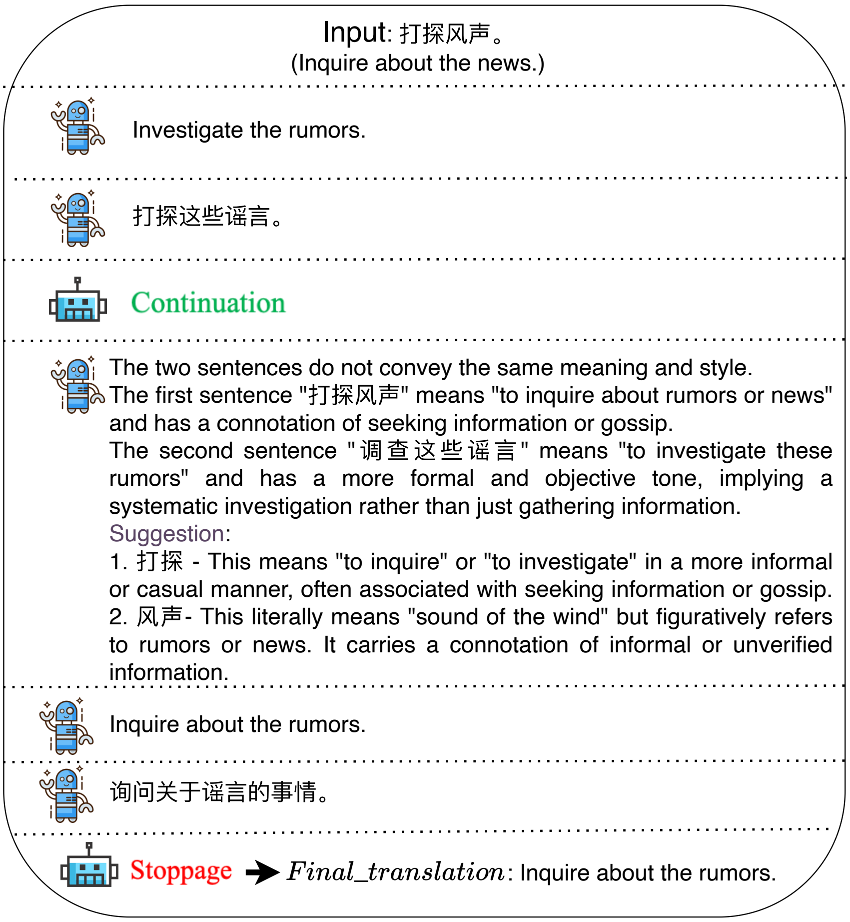

Figure 4: The DUAL-REFLECT methodology for translating positive examples within Chinese sentences.

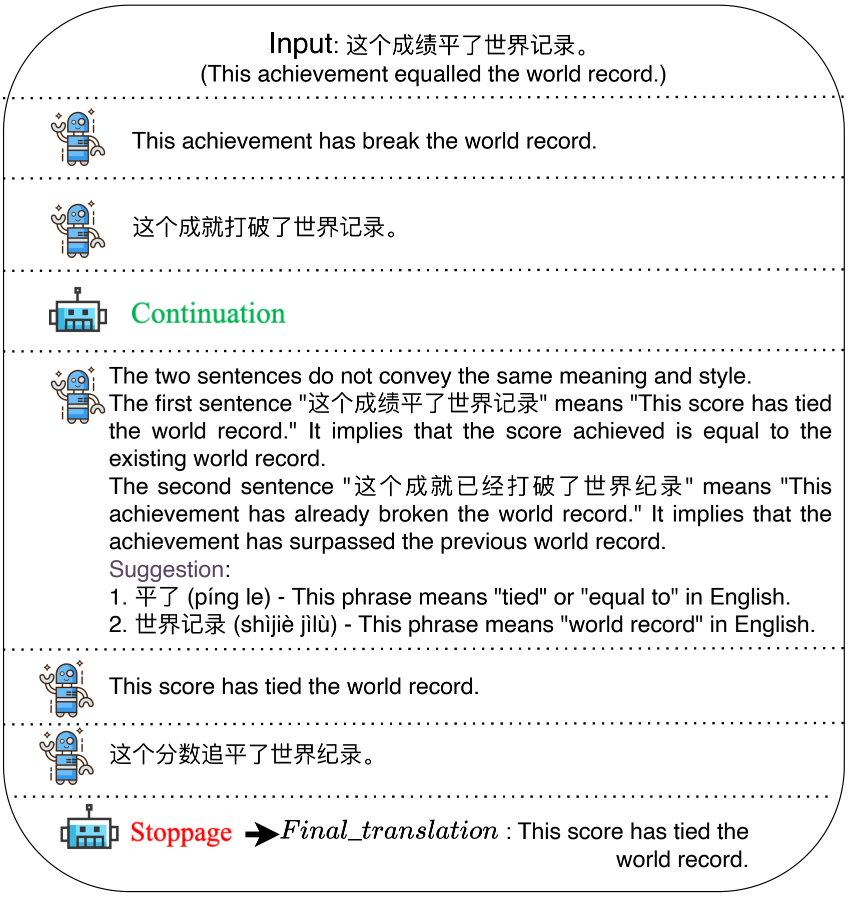

Figure 5: The DUAL-REFLECT methodology for translating positive examples within Chinese sentences.

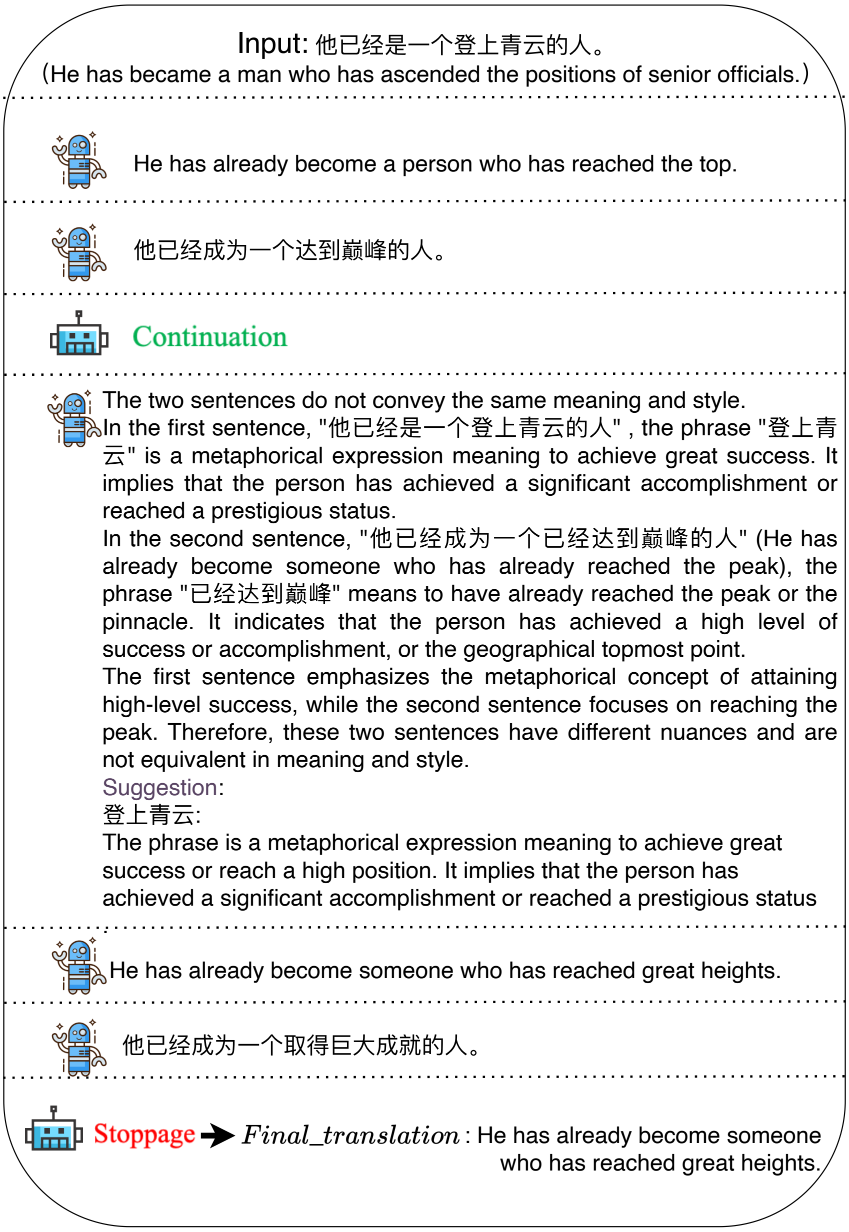

Figure 6: The DUAL-REFLECT methodology for translating positive examples within Chinese sentences.

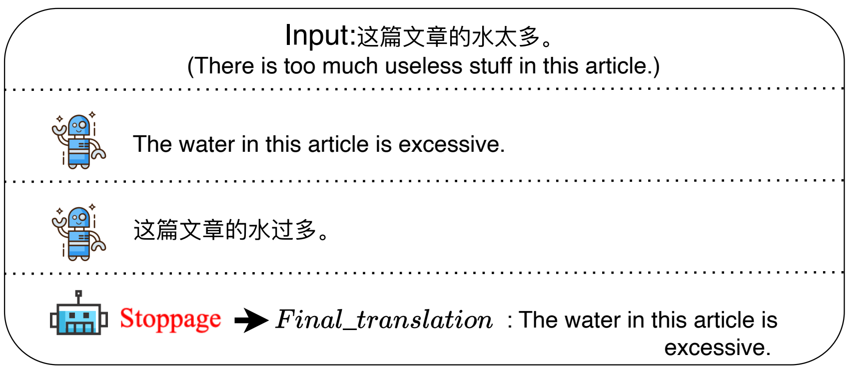

Figure 7: The DUAL-REFLECT methodology for translating negative examples within Chinese sentences.

Figure 8: The DUAL-REFLECT methodology for translating negative examples within Chinese sentences.