A Preliminary Study of AI Agent Model in Machine Translation

Ahrii Kim
AI-Bio Convergence Research Institute
South Korea
ahriikim@gmail.com

trotacodigos/MultiAgentMT.git

Abstract

We present IR_Multi-agentMT, our submission to the WMT25 General Shared Task. The system adopts an AI-agent paradigm implemented through a multi-agent workflow, Prompt Chaining, in combination with RUBRIC-MQM, an automatic MQM-based error annotation metric. Our primary configuration follows the Translate–Postedit–Proofread paradigm, where each stage progressively enhances translation quality. We conduct a preliminary study to investigate (i) the impact of initial translation quality and (ii) the effect of enforcing explicit responses from the Postedit Agent. Our findings highlight the importance of both factors in shaping the overall performance of multi-agent translation systems.

1 Introduction

An AI agent is a computational system that operates autonomously, guided by environmental observations, and often equipped with adaptive learning capabilities (Russell and Norvig, 2010). Large Language Models (LLMs), which have become central to the development of AI agents, demonstrate advanced reasoning, contextual understanding, and flexible workflows across a wide range of tasks, including Machine Translation (MT) (Briva-Iglesias, 2025). Recent work by Briva-Iglesias (2025) explored a multi-agent system composed of four agents—a Translator, Fluency Reviewer, Adequacy Reviewer, and Editor—highlighting its potential for driving future advancements. Inspired by this line of research, we participate in this year's MT track with a multi-agent workflow. Our goal is to leverage smaller models to achieve performance competitive with, or even superior to, larger models, while simultaneously reducing computational costs.

2 Participating Task

The shared task aims to evaluate translation performance across a broad range of languages, domains, genres, and modalities. We participate in the multilingual subtask, which covers 30 target languages, with Czech, English, and Japanese serving as the source languages. Our system is categorized as unconstrained. The source data consists of 29,957 segments, corresponding to 102,060 paragraphs. Our approach performs translation inference on a segment-by-segment basis. In cases where paragraph boundaries are ignored in the original segmentation, we re-split the data into paragraphs and translate them independently.

3 IR_Multi-agentMT

3.1 Design

AI agents enable dynamic workflows through configurable architectures. We adopt the concept of Prompt Chaining, in which the output of each step serves as the input to the next, fostering systematic reasoning and iterative refinement (Briva-Iglesias, 2025). While iterative refinement may theoretically improve translation quality, cost constraints motivate us to employ a unidirectional configuration. Accordingly, we experiment with two workflow variants: Translate–Postedit (TP Workflow) and Translate–Postedit–Proofread (TPP Workflow), both of which are submitted to the competition.

3.2 Translate Agent

The Translate Agent generates target text from the given source using the official prompt provided by the organizers. We use the GPT-4o-mini model to obtain the initial translation.

3.3 Postedit Agent

The Post-edit Agent refines the translation with reference to the source text. We build upon RUBRIC-MQM (Kim, 2025), which, like its counterpart

Proceedings of the Tenth Conference on Machine Translation, pages 583–586
November 8-9, 2025

GEMBA-MQM (Kocmi and Federmann, 2023), identifies MQM-style error categories, severities, and spans. Unlike Gemba-MQM, however, Rubric-MQM reduces biases associated with the MAJOR and MISTRANSLATION labels, improves recognition of NO-ERROR cases, and increases precision in error detection.

We introduce three modifications to the original Rubric-MQM:

Transformation into a post-editor The model is instructed to propose corrected translations for each identified error span.

Severity scale adjustment The original 100-point scale is simplified to a 4-level scale, since severity is not our primary focus. While Kim (2025) stressed that the rubric scheme is essential for model effectiveness, we reduce the rubric complexity in the prompt to streamline usage.

Multilingual configuration While preserving the original in-context examples, we replace one instance from English–German with Japanese–Korean to support broader X–Y translation directions. This modification improves detection of NO-ERROR cases and prevents erroneous corrections of already error-free phrases.

Finally, the model's suggested spans are manually integrated into the sentence, and the revised output is considered the final version.

3.4 Proofread Agent

The Proofread Agent further examines and refines the translation using a Chain-of-Thought prompting strategy (Wei et al., 2022). First, the model identifies potential errors; then it is asked to propose five alternative phrasings that prioritize fluency while maintaining alignment with the source. The best alternative is selected as the final translation. This procedure is designed to resolve awkward expressions introduced during earlier manual edits and to further polish the final output.

4 Model Details

We employ prompt engineering for all agents, using `GPT-4o-mini-2024-07-18` as the primary baseline. The model is configured with a temperature of 1 and a maximum token length of 1024. Although this setup is suboptimal in terms of reproducibility, our iterative pilot studies suggest that these parameters allow the model to explore a wider error space and generate more effective modifications, thereby improving overall translation quality. Future work will focus on developing a more stable and reproducible experimental environment.

5 Experiment

We evaluate our multi-agent pipeline through experiments on the WMT24 English–Spanish dataset (Kocmi et al., 2024). We take a subset of 304 unique source segments with balanced domain distribution (literary, news, social media, and speech) and use 23 translations, summing up to 6,992 segments for analysis.

Our experiments are structured in two parts. First, we obtain initial translations from DeepL (DeepL) and MarianMT (Junczys-Dowmunt et al., 2018), cost-efficient models, and GPT-4o-mini, our baseline model, and compare the final translation quality produced by the multi-agent workflows. We use ChrF++ (Popović, 2017) and COMET (Rei et al., 2020). Second, we analyze translations generated with both the original and the refined versions of Rubric-MQM, with particular attention to cases where (i) no-error labels are produced and (ii) source phrases are erroneously marked as errors.

5.1 Analysis: Initial Translation Quality

Table 1 shows that both the initial and final translation quality are highest when using GPT-4o-mini across the two metrics. Regardless of the initial quality, the general trend in our multi-agent workflow is that translation quality decreases after the Postedit stage and increases again after Proofread, with few exceptions. For instance, the surface-level quality measured by ChrF++ drops from 78.55 to 68.18 with GPT-4o-mini, while COMET scores remain stable, indicating that the revisions primarily involve semantic edits within a similar structural framework. These findings suggest that increasing the extent of semantic-level edits can result in higher overall translation quality.

5.2 Analysis: Response of Postedit Agent

We analyze the erroneous responses produced under the original Rubric-MQM setting. As shown in Table 2, 55.38% of the model outputs are empty, indicating that the system deemed the translation perfect. In 2.8% of the cases, spans were incorrectly labeled as "no-error." Furthermore, in 36.7% of the cases, the model identified source errors, which can disrupt the agent workflow. The revised

Language	Metric	Translate	Postedit	Proofread	Final Δ
DeepL	ChrF++	45.99	38.04	40.62 ↑	-5.37
	COMET	65.72	59.96	59.71	-6.01
MarianMT	ChrF++	60.01	55.92	60.53 ↑	+0.52
	COMET	89.66	87.25	88.40 ↑	-1.26
GPT-4o-mini	ChrF++	78.55	78.93 ↑	68.18	-10.37
	COMET	94.38	91.69	96.05 ↑	+1.67

Table 1: Performance scores of the IR_Multi-agentMT system with different baselines for the initial translation (*Translate*). ↑ indicates gains over the previous stage. The gains from *Translate* to *Proofread* are reported in the *Final* column.

	NaN	No-error	Source error
#	3872	200	2567
%	55.38	2.86	36.71

Table 2: Distribution of erroneous response of Postedit Agent. Raw counts (#) and the percentage (%) are given.

version, however, mitigates these issues through improved prompt engineering.

6 Conclusion

Our experiments on the multi-agent pipeline indicate that higher initial translation quality leads to better final outcomes. Furthermore, enforcing the Postedit Agent to identify more errors is essential for ensuring meaningful revisions within the workflow. In this way, IR_Multi-agentMT demonstrates the potential to achieve translation quality comparable to that of larger models, while operating at roughly half the cost. A more detailed discussion of these findings will be provided in the main system paper.

Acknowledgment

This research was supported by G-LAMP Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. RS-2025-25441317)."

References

Vincent Briva-Iglesias. 2025. Are ai agents the new machine translation frontier? challenges and opportunities of single-and multi-agent systems for multilingual digital communication. *arXiv preprint arXiv:2504.12891*.

DeepL. Deepl translator. https://www.deepl.com/translator. Accessed: 2025-07-05.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Ahrii Kim. 2025. RUBRIC-MQM : Span-level LLM-as-judge in machine translation for high-end models. In *ACL 2025 Industry Track*. Associations for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the wmt24 general machine translation shared task: The llm era is here but mt is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*, pages 1–46, Online / Virtual. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Gemba-mqm: Detecting translation quality error spans with gpt-4. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the Conference on Machine Translation (WMT), Shared Task Papers, Volume 2*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Ricardo Rei, Alon Lavie Farinha, Luisa Coheur, 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)*. ACL.

Stuart Russell and Peter Norvig. 2010. *Artificial Intelligence: A Modern Approach*, 3rd edition. Prentice Hall, Upper Saddle River, NJ.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.