Multi-Agent Debate for Machine Translation: A Case Study on English–Japanese Translation

Zhan Shen[1] **Jason Naradowsky**[1] **Xiaotian Wang**[1] **Yusuke Miyao**[1,2]

[1]Department of Computer Science, The University of Tokyo

[2]Research and Development Center for Large Language Models, National Institute of Informatics

shenzhan@g.ecc.u-tokyo.ac.jp, jason.narad@gmail.com

xtwang@is.s.u-tokyo.ac.jp, yusuke@is.s.u-tokyo.ac.jp

Abstract

As machine translation increasingly requires deeper contextual, linguistic, and cultural understanding, multi-agent collaboration has emerged as a promising approach. Multi-agent debate (MAD) frameworks, in which multiple agents deliberate to produce a final output, have shown strong performance on objective tasks, but remain underexplored in translation, where multiple valid renderings often exist. We adapt three MAD frameworks for English-Japanese translation and evaluate them against strong generative baselines, reasoning-capable LLMs, and a prompt-based self-reflection method. Across general-domain and culturally grounded datasets, the Society of Mind (SoM) variant yields the strongest results in the English-to-Japanese direction, showing that zero-shot translations leave substantial room for improvement through structured deliberation. Yet the gains of debate are front-loaded: later rounds do not reliably improve quality and often reintroduce translation errors. Diagnostic and error-span analyses show that hand-designed debate protocols tend to over-revise already strong translations, leading to semantic drift and process-induced degradation. These findings highlight both the promise and the limitations of debate-based agentic translation, and suggest that effective iterative improvement requires mechanisms for preserving high-quality intermediate translations while limiting unnecessary revisions.

1 Introduction

With advances in large language models (LLMs) (Ouyang et al., 2022; Touvron et al., 2023; Yang et al., 2025), machine translation has moved beyond sentence-level adequacy toward better handling discourse context, implicit meaning, and culturally conditioned nuances. Such phenomena often require contextual and pragmatic reasoning. To strengthen multi-step inference during generation, prior work has explored prompt-based reasoning strategies such as chain-of-thought (CoT) prompting (Wei et al., 2022), self-reflection (Shinn et al., 2023; Madaan et al., 2023), and self-consistency (Wang et

https://github.com/ZhanShenYo/madmt

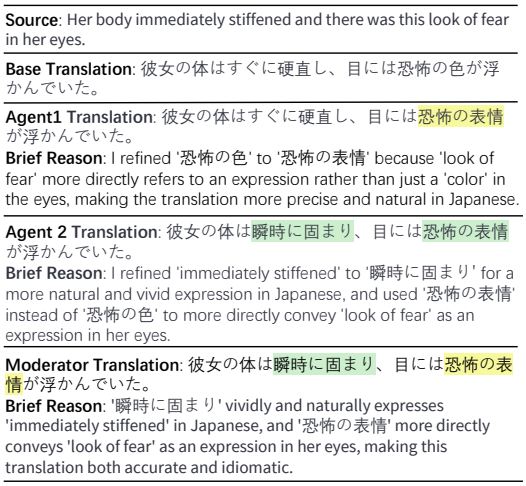

Figure 1: An intermediate round of the Society of Mind (SoM) framework illustrating how agents refine a translation collaboratively.

al., 2022). In parallel, recent reasoning-capable LLMs, such as OpenAI o4-mini (OpenAI, 2024) and DeepSeek-R1 (Guo et al., 2025), embed stronger reasoning ability through model architecture or training.

Building on these developments, multi-agent debate (MAD) has emerged as an orthogonal strategy to improve LLM reasoning through explicit interaction among multiple agents. In MAD, agents exchange intermediate views during inference and iteratively revise their outputs without any gradient-based updates. Recent studies have shown that MAD is effective on objective tasks such as question answering and factual verification (Xiong et al., 2023; Khan et al., 2024; Estornell et al., 2024). However, its role in machine translation remains underexplored, especially in settings where multiple valid translations may exist and consensus is less straightforward.

English-Japanese (En-Ja) translation provides a particularly challenging testbed for this question. The two languages differ substantially in word order, ellipsis patterns, honorific usage, and stylistic register (Mori, 2020). As a result, translation quality often depends not only on lexical adequacy, but also on speaker intent, discourse context, and cultural convention. These characteristics make En-Ja translation a compelling setting for evaluating whether structured multi-agent deliberation can improve context-sensitive and culturally informed translation.

In this work, we adapt three MAD frameworks, originally proposed for reasoning (Liang et al., 2024; Du et al., 2024) and translation evaluation (Feng et al., 2025), to En-Ja translation. We compare them against strong generative baselines, reasoning-capable LLMs, and a prompt-based self-reflection baseline. Translation quality is evaluated using BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2019), and xCOMET (Guerreiro et al., 2024), together with human evaluation and diagnostic analyses that capture debate dynamics and error patterns.

Figure 1 shows an intermediate round of the debate framework. Given a source sentence, multiple agents maintain and refine translation hypotheses by incorporating useful insights from one another. This design supports context-sensitive stylistic adaptation, collaborative refinement of translation errors or weak segments, and a more transparent decision process than single-pass generation.

Our contributions are as follows:

- To our knowledge, we present the first systematic study of MAD for En-Ja machine translation across both general-domain and culturally grounded settings.
- We show that the SoM variant is the most effective MAD framework overall, particularly in the En→Ja direction, while the benefits of debate are concentrated in the early rounds.
- We provide diagnostic analyses of SoM that characterize later-round degradation

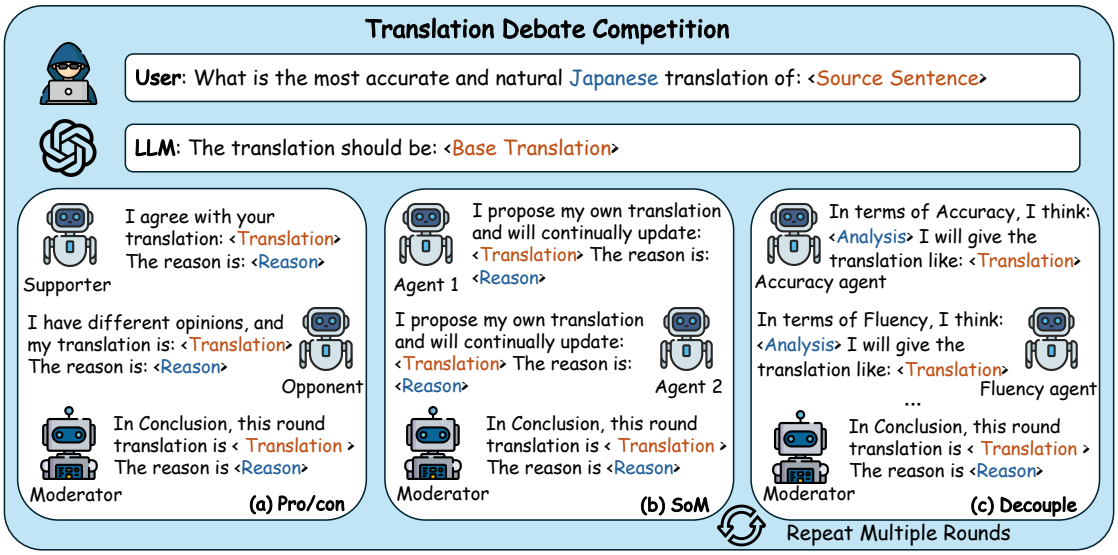

Figure 2: Illustration of three multi-agent debate (MAD) frameworks adapted for machine translation: (a) Pro/Con Debate, (b) Society of Mind (SoM), and (c) Decoupled Debate.

and identify process-level limitations of iterative debate for translation.

2 Related Work

2.1 General Trends in Machine Translation

Machine translation has long been a fundamental research task in natural language processing (Tsujii, 1986). Unlike traditional neural machine translation systems that rely heavily on parallel corpora (Kocmi et al., 2022; Vaswani et al., 2017; Castilho et al., 2017), LLMs (Ouyang et al., 2022; Wei et al., 2021; Touvron et al., 2023; Yang et al., 2025) have advanced translation through few-shot translation and broader domain generalization (Brown et al., 2020; Chowdhery et al., 2023). Despite these advances, machine translation still struggles with pragmatic and cultural phenomena. In this work, we investigate whether structured multi-agent deliberation can surface diverse cultural hypotheses and collaboratively resolve sociolinguistic ambiguities.

2.2 Reasoning-Augmented and Multi-Agent Approaches

Recent work has improved LLM reasoning through both prompt-level scaffolding and advances in model training. Prompt-based methods, such as chain-of-thought prompting (Wei et al., 2022), tree of thoughts (Yao et al., 2023), self-reflection (Shinn et al., 2023; Madaan et al., 2023), MAPS (He et al., 2024), and self-consistency (Wang et al., 2022), encourage models to generate more effective intermediate reasoning steps at inference time. Meanwhile, recent reasoning-oriented LLMs, including DeepSeek-R1 (Guo et al., 2025) and Kimi K1.5 (Team et al., 2025), strengthen reasoning through post-training strategies.

Multi-agent methods instead provide explicit, structured deliberation by distributing reasoning across agents. Beyond general applications to generation (Chan et al., 2023), negotiation (Fu et al., 2023), and reasoning (Du et al., 2024; Liang et al., 2024), recent agentic machine translation work has explored long-form literary translation (Wu et

WMT2023	**En:** Even if it's true that such facts exist in science it's still possible to argue that scientific facts are theory-laden. **Ja:** 科学にそのような事実が本当に存在すると仮定しても、科学的事実が理論に裏打ちされたものだと議論する余地はある。
KFTT	**En:** Soujun IKKYUU was a Zen monk in the Daitokuji branch of the Rinzai sect, during the Muromachi period. **Ja:** 一休宗純（いっきゅうそうじゅん）は、室町時代の臨済宗臨済宗大徳寺派の禅僧である。
IWSLT2017	**Ctx.:** Thank you so much, Chris. And it's truly a great honor to have the opportunity to come to this stage twice; I'm extremely grateful. I have been blown away by this conference, and I want to thank all of you for the many nice comments about what I had to say the other night. And I say that sincerely, put yourselves in my position. **En:** I flew on Air Force Two for eight years. **Ja:** 8年間私はエアフォースツーで飛んでいました。

Figure 3: Representative examples from the three evaluation datasets, illustrating their stylistic differences and the context provided in our evaluation setup. *Ctx.* denotes the provided context information.

al., 2025), general agentic machine translation workflows (Briva-Iglesias, 2025), and legal-domain systems (Sin et al., 2025). We extend this line by evaluating debate-style interaction for En-Ja translation and diagnosing when deliberation improves or degrades quality.

3 MAD for Machine Translation

Applying MAD to machine translation begins with an initial translation, after which candidate translations are iteratively refined or replaced through multi-round agent interaction. Unlike objective tasks, where majority voting over discrete answers can yield a reliable consensus, translation often admits multiple valid renderings. To account for this open-endedness, all MAD variants in our setup use a moderator-based summarization strategy rather than voting-based aggregation.

Each debate runs for N rounds with M agents assigned distinct roles or functions. In each round, one agent serves as the moderator, aggregating the current outputs into a single candidate translation, either for the next round or as the final output. Agents retain access to prior interaction history, allowing them to incorporate earlier context into subsequent revisions. Figure 2 summarizes the three MAD frameworks considered in this work. Full prompts are provided in the Appendix (Tables 8, 9, 10, 11, and 12).

3.1 Pro/Con Debate

Inspired by formal argumentation, the Pro/Con framework structures deliberation through explicit opposition (Figure 2(a)). Following the framework of Liang et al. (2024), we adapt this format to translation by first generating an initial translation, having one agent defend it and another critique it by proposing a preferred alternative, and then letting a moderator select, revise, or synthesize the next-round candidate. This variant allows us to examine whether explicitly opposing viewpoints improve translation quality.

3.2 Society of Mind Debate

The Society of Mind (SoM) framework (Du et al., 2024), inspired by Minsky's *The Society of Mind* (Minsky, 1986), replaces direct opposition with parallel hypothesis refinement (Figure 2(b)). Unlike Pro/Con, SoM does not start from a single shared translation. Instead, each agent independently maintains a separate translation hypothesis and iteratively refines it by selectively incorporating insights from other agents, rather than through direct critique or replacement. The moderator aggregates the resulting hypotheses into the next-round candidate. This design encourages diversity without direct confrontation and keeps multiple translation alternatives active throughout deliberation.

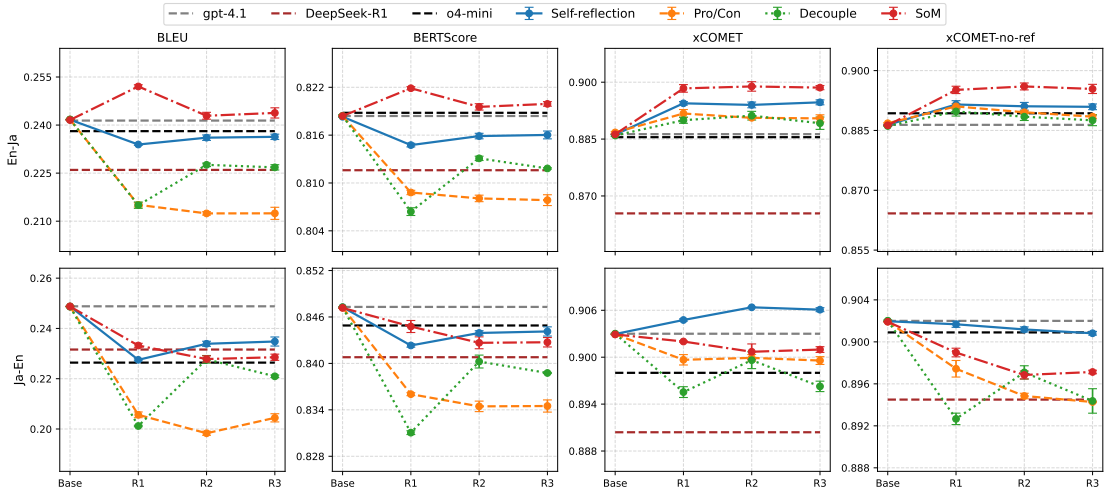

Figure 4: Overall performance on WMT 2023. All MAD variants use GPT-4.1 as the backbone model and are compared against zero-shot baselines.

3.3 Decoupled Debate

The Decoupled variant is adapted from M-MAD (Feng et al., 2025), which builds on the Multidimensional Quality Metrics (MQM) framework (Freitag et al., 2021) for fine-grained translation evaluation. Because MQM decomposes translation quality into interpretable dimensions such as accuracy, fluency, terminology, and style, it provides a natural scaffold for structured translation refinement. In our adaptation (Figure 2(c)), a base agent first produces an initial translation, after which four specialized agents independently provide dimension-specific feedback. A moderator then aggregates their suggestions into a revised translation for the next round. This design closely aligns the debate process with translation evaluation while preserving the iterative coordination of MAD.

4 Experiments

4.1 Datasets

We use the **WMT 2023** (Freitag et al., 2023) test set as our primary benchmark for general-domain machine translation, as it covers diverse domains and genres and includes human judgments of overall translation quality. The Kyoto Free Translation Task **KFTT** (Neubig, 2011) is a Japanese-English parallel corpus derived from Wikipedia articles about Kyoto, many of which concern its history, culture, and landmarks. Its professionally produced and verified translations make it well suited to evaluating culturally grounded translation. **IWSLT 2017** (Cettolo et al., 2017) consists of parallel sentences from TED talks. We use it to evaluate robustness under varying context lengths. For computational efficiency, we evaluate on the first 1,000 examples from each dataset. Figure 3 presents representative examples from the three datasets, illustrating their stylistic differences and the context provided in our evaluation setup.

4.2 Models

We experiment with **general-purpose LLMs**, including DeepSeek-V3 (Liu et al., 2024), GPT-4.1 (OpenAI, 2024), and Gemini 1.5 Pro (Team et al., 2024). These models are used either as standalone baselines or as backbone models for individual agents. We also eval-

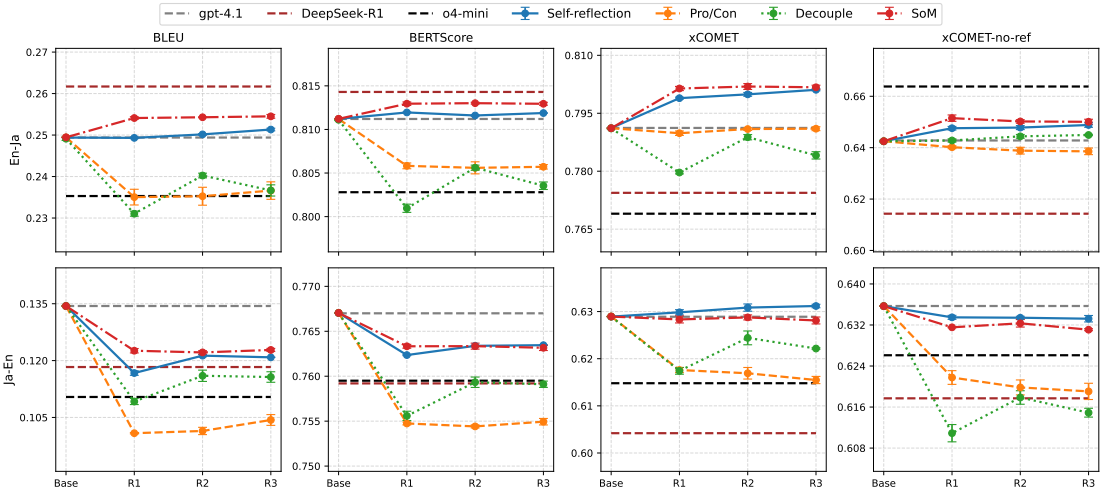

Figure 5: Overall performance on KFTT. All MAD variants use GPT-4.1 as the backbone model and are compared against zero-shot baselines.

uate **reasoning-capable models**, including DeepSeek-R1 (Guo et al., 2025) and o4-mini (OpenAI, 2024), as strong zero-shot baselines.

4.3 Metrics

General Metrics We evaluate translation quality using three automatic metrics. BLEU (Papineni et al., 2002) measures n-gram overlap with reference translations, while BERTScore (Zhang et al., 2019) captures semantic similarity. We further use xCOMET (Guerreiro et al., 2024), a state-of-the-art neural metric for overall translation quality that supports both reference-based and reference-free evaluation.

Additional Analyses To analyze debate dynamics, we use the **Net Gain Rate (NGR)** and perform LLM-based **Diagnostic Analysis** and xCOMET **Error-Span Analysis**.

4.4 Baselines

We compare MAD-based translation against the following baselines. **Zero-shot translation**: a model directly translates the source input without additional deliberation. **Self-Reflection translation**: following recent work on introspective generation (Shinn et al., 2023; Madaan et al., 2023), a model first generates an initial translation, evaluates its own output, and then iteratively refines it through self-feedback (see Table 9 in Appendix). **Reasoning-capable LLMs**: high-capacity reasoning-oriented models, including DeepSeek-R1 and o4-mini, are used in standard zero-shot settings as strong reasoning baselines.

4.5 Human Evaluation

To assess how translation quality evolves across debate rounds, we conduct a small-scale human evaluation. Seven evaluators affiliated with Japanese universities participate in the study, including four native speakers of Japanese and three highly proficient near-native speakers, all with graduate-level or higher education. We randomly sample 100 En→Ja examples, ensuring that each example is rated by at least one native speaker and that approximately 75% receive double ratings. Evaluators rank the Base and Round 1-3 translations, presented in random order,

Systems	R1	R2	R3
En→Ja			
Self-reflection	0.071	0.000	0.024
Pro/Con	-0.004	-0.014	0.015
SoM	**0.170**	-0.016	-0.020
Decouple	0.019	0.001	-0.006

Table 1: Net Gain Rate (NGR) under xCOMET for the SoM framework on WMT 2023 (↑ higher is better).

along four dimensions: adequacy/faithfulness, fluency/naturalness, terminology/consistency, and style/register.

4.6 Implementation Details

All models were accessed via official APIs. For decoding, we use a temperature of 0, a top-p of 1.0, and a maximum output length of 6000 tokens, except for GPT-4 o4-mini, for which the temperature is fixed at 1.0 due to API constraints. System-level comparisons were repeated three times with fixed hyperparameters to improve reproducibility and assess run-to-run stability. Full experimental results, including significance test results, are reported in the Appendix. A small number of API failures (timeouts or non-responses) occurred, but the overall failure rate remained below 1%; metrics were computed over completed outputs.

5 Results and Analysis

5.1 Overall Translation Performance

Figure 4 summarizes overall performance on WMT 2023 (Full results with paired t-tests are reported in Appendix Tables 4 and 5). Performance differs substantially across translation directions.

English→Japanese In this direction, the SoM framework consistently outperforms all baselines and other MAD variants. Performance peaks in the first round and then gradually declines in later rounds. The decline is most pronounced in BLEU and BERTScore, reflecting increasing divergence from the reference translations. By contrast, xCOMET and xCOMET-no-ref remain relatively stable, suggesting that reduced reference overlap does not necessarily correspond to lower overall quality. These results indicate that zero-shot En→Ja translations leave substantial room for improvement, but additional debate rounds tend to introduce semantic drift that can ultimately offset the gains.

	Base	R1	R2	R3
Avg. Rank ↓	2.052	1.925	**1.615**	1.655

Table 2: Human evaluation results for WMT 2023, reported as average rank across rounds (↓ lower is better).

Japanese→English This direction appears unstable. SoM is only comparable to Self-Reflection in the first round and subsequent rounds lead to consistent declines across all metrics, suggesting that the effectiveness of MAD is sensitive to translation direction.

Pro/Con and Decoupled consistently underperform in both translation directions, suggesting that debate design materially affects translation outcomes. Moreover, implicit reasoning models (e.g., DeepSeek-R1 and o4-mini) do not show a consistent advantage, indicating that gains on reasoning tasks do not directly transfer to bilingual generation. The culturally nuanced KFTT dataset shows a similar pattern (Figure 5).

In addition to average metric scores, we compute the **Net Gain Rate** (NGR) to quantify the net effect of debate across instances. NGR is defined as the difference between the number of improved and degraded translations, normalized by the total number of evaluated instances. It captures whether a system produces more improved than degraded out-

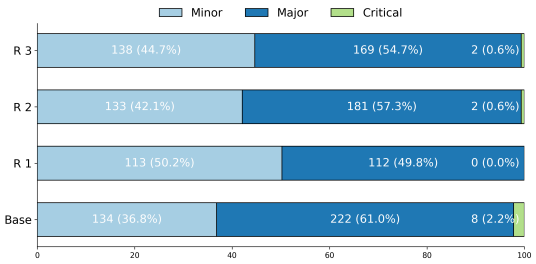

Figure 6: Dynamics of xCOMET error spans across debate rounds by severity level.

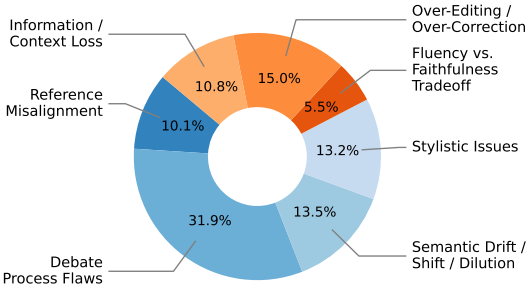

Figure 7: Error type distribution from LLM-based Diagnostic Analysis.

puts, regardless of the magnitude of individual score changes.

$$\text{NGR} = \frac{N_{\text{improved}} - N_{\text{degraded}}}{N_{\text{total}}} \quad (1)$$

Table 1 reports Net Gain Rate (NGR) under xCOMET. For En→Ja, SoM achieves a strongly positive NGR in the first round, consistent with the improvements observed in the main results. The lower NGR in the second round suggests that these gains become less uniform across instances, even when average scores remain stable or continue to improve. Full results are shown in Appendix (Table 6).

To assess whether the observed metric trends are consistent with human preferences, we conduct a small-scale human evaluation of the SoM variant. As shown in Table 2, the results are consistent with the automatic metrics: the SoM framework substantially improves translation quality, whereas additional rounds do not necessarily lead to further gains.

5.2 Analyze Performance Decline of SoM

Because SoM performs best among the MAD variants, we focus our analysis on its debate dynamics. While SoM initially improves En→Ja translation quality in Round 1, it often deteriorates in later rounds. To better understand this behavior, we analyze cases that improve in Round 1 (R1) but degrade in Round 2 (R2), using xCOMET error-span analysis and LLM-based diagnostics.

Error-Span Analysis draws on xCOMET's error-span outputs, trained on human-annotated data, to classify errors by severity (i.e., **Minor**, **Major**, and **Critical**). Figure 6 shows that errors at all three severity levels decline substantially after Round 1 (R1). **Critical** errors are eliminated entirely at this stage, but errors at **Critical** severity level are partially reintroduced in later debate rounds.

For the **Diagnostic Analysis**, sampled degraded cases are submitted to GPT-4.1 for open-ended diagnosis without predefined error categories, the full prompt can be found in Appendix (Table 7), and the resulting outputs are manually clustered into coarse-grained failure types. As shown in Figure 7, a substantial portion of later-round degradation stems from **debate process flaws**, including Debate Stagnation, Translation Over-Convergence, Lack of Critical Reassessment, and Debate Process-Induced Semantic Drift. These failure types often co-occur: early gains are followed by premature convergence, while continued "polishing" introduces unnecessary edits that cause meaning shifts, information loss, or tonal dilution.

Figure 8 presents a representative case of semantic over-adjustment and idiomatic drift in an intermediate debate round. The agents produce an idiomatic, near-optimal translation in R1, but introduce unnecessary revisions in

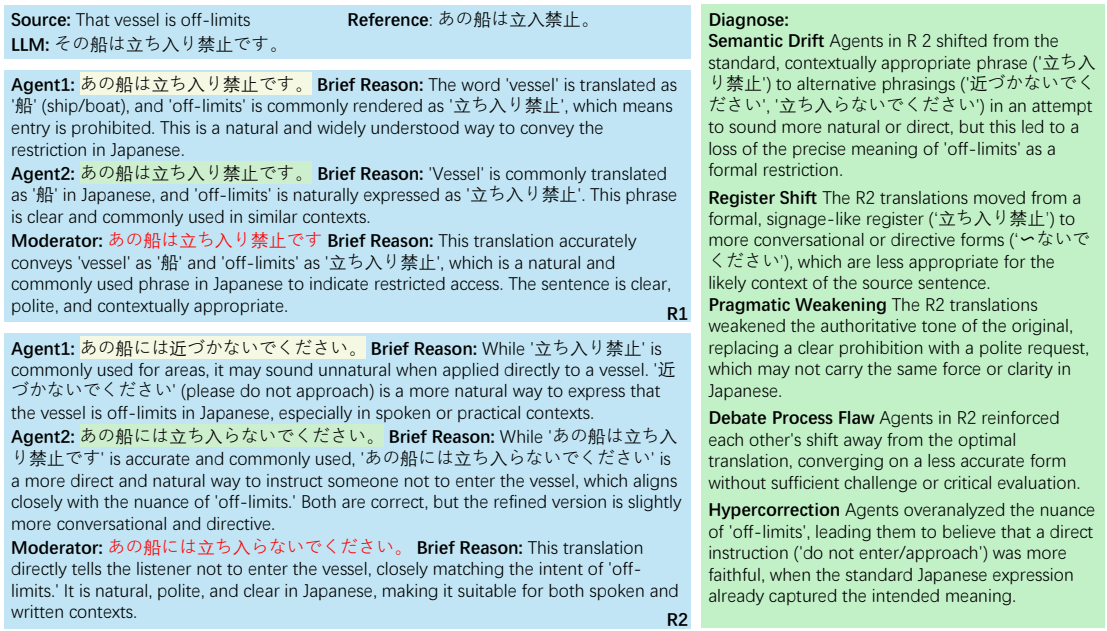

Figure 8: Open-ended diagnostic analysis of a single translation case in the SoM framework.

R2 under the assumption that further changes are inherently beneficial. Although the revised translation remains grammatically valid, it departs from natural usage and signage conventions, resulting in semantic drift, register shift, and reduced idiomaticity. The agent rationales further reveal speculative justifications, such as questioning standard expressions, suggesting a bias toward change over stability. This case illustrates a more fundamental limitation of SoM: without a mechanism for recognizing translation sufficiency, agents may continue revising outputs that are already near-optimal, thereby degrading quality.

Taken together, these findings suggest that the effectiveness of multi-round debate in translation depends not only on the ability to revise, but also on preserving strong intermediate outputs. Post-hoc refinement can be beneficial, as zero-shot translations may underspecify contextual or cultural meaning. However, repeatedly revising a single translation, as in Pro/Con and Decoupled debate, is often associated with semantic drift and reduced translation quality. By contrast, the stronger performance of SoM is consistent with its design intuition: maintaining multiple translation hypotheses may preserve a broader search space than repeatedly refining a single trajectory. At the same time, the benefits of debate appear to be concentrated in the early rounds, as later iterations do not reliably produce further gains and often introduce degradation. These results indicate that effective agentic translation may require not only collaborative refinement, but also mechanisms for preserving high-quality intermediate outputs and limiting unnecessary revision.

5.3 Ablation Studies of SoM Framework

To probe the factors underlying SoM's performance, we conduct ablation studies on model diversity, agent plurality, reasoning language, and contextual visibility under the full experimental setup. Full ablation results are reported in the Appendix (Tables 10, 11, 12, and 13).

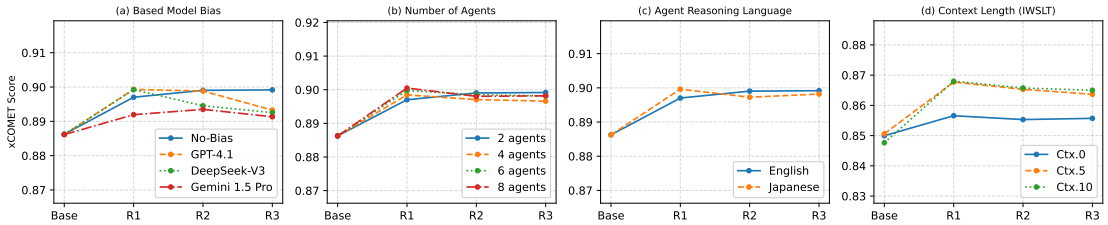

Figure 9: Ablation studies of the SoM framework. (a) Different models alternate between the debater and moderator roles. (b) The effect of agent plurality over debate rounds. (c) The effect of using English or Japanese as the debate language. (d) The effect of different context lengths on translation quality.

Model Diversity To assess the role of model diversity, we incorporate other strong LLMs (e.g., DeepSeek-V3, GPT-4.1, and Gemini 1.5 Pro) as debaters and moderators in a role-swapping setup. As shown in Figure 9(a), hybrid configurations are competitive in the first round, suggesting that model-specific biases can be beneficial at the most informative stage of debate. However, they deteriorate more sharply in later rounds and are overall less stable than the single-model setup.

Agent Plurality We vary the number of debating agents from 2 to 4, 6, and 8. Consistent with the main experiments, the effect of plurality is clearest in Round 1, the most informative stage of debate: xCOMET increases steadily as the number of agents grows. In later rounds, this trend becomes less consistent (Figure 9(b)).

Agent Reasoning Language We compare SoM under English- and Japanese-language reasoning in En→Ja translation. As shown in Figure 9(c), Japanese reasoning yields slightly better first-round results, but overall translation quality remains largely unchanged.

Contextual Visibility To assess contextual visibility, we evaluate SoM on IWSLT 2017 with 5 and 10 preceding sentences as context (Figure 9(d)); baselines receive the same context for fair comparison. Although longer context introduces noise for zero-shot translation, it appears more beneficial for deliberative reasoning. Providing contextual sentences substantially improves translation quality, with the 10-sentence setting achieving the highest xCOMET score. These results suggest that SoM can effectively leverage extended context during deliberation.

6 Conclusion

This paper presents a systematic study of multi-agent debate (MAD) for English-Japanese machine translation. We adapt three MAD frameworks to translation and evaluate them against strong baselines on both general-domain and culturally grounded datasets. Our results show that the Society of Mind (SoM) variant performs best overall, especially in the En→Ja direction, whereas the other debate variants offer little to no improvement, highlighting the importance of debate structure in translation. We further find that the gains of SoM are concentrated in the early rounds, but later rounds frequently introduce degradation. Our analyses suggest that this later-round decline reflects process-level limitations of iterative deliberation, especially once strong intermediate translations have already been reached. Overall, our findings suggest that effective agentic translation depends not only on collaborative exploration, but also on mechanisms for preserving high-quality intermediate outputs and limiting unnecessary revision.

Carbon Impact Statement This work relies on existing large language models accessed through official APIs and does not involve pre-training or fine-tuning new models. Its environmental impact therefore arises primarily from inference-time computation, particularly in multi-agent settings that require multiple model calls per example. Computational efficiency is therefore an important consideration for future work on agentic translation, especially for debate-based systems with iterative inference.

Acknowledgements This work was supported by JST SPRING, Grant Number JPMJSP2108 and by the "R&D Hub Aimed at Ensuring Transparency and Reliability of Generative AI Models" project of the Ministry of Education, Culture, Sports, Science and Technology. We gratefully thank all annotators who participated in the human evaluation study.

References

Briva-Iglesias, Vicent. 2025. Are ai agents the new machine translation frontier? challenges and opportunities of single-and multi-agent systems for multilingual digital communication. *Proceedings of Machine Translation Summit XX: Volume 1*, pages 365–377.

Brown, Tom, 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.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*.

Cettolo, Mauro, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the iwslt 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14.

Chan, Chi-Min, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of machine learning research*, 24(240):1–113.

Du, Yilun, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning*.

Estornell, Andrew, Jean-Francois Ton, Yuanshun Yao, and Yang Liu. 2024. Acc-debate: An actor-critic approach to multi-agent debate. *arXiv preprint arXiv:2411.00053*.

Feng, Zhaopeng, Jiayuan Su, Jiamei Zheng, Jiahan Ren, Yan Zhang, Jian Wu, Hongwei Wang, and Zuozhu Liu. 2025. M-mad: Multidimensional multi-agent debate for advanced machine translation evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7084–7107.

Freitag, Markus, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Freitag, Markus, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, et al. 2023. Results of wmt23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628.

Fu, Yao, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*.

Guerreiro, Nuno M, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Guo, Daya, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

He, Zhiwei, 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.

Khan, Akbir, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*.

Kocmi, Tom, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, et al. 2022. Findings of the 2022 conference on machine translation (wmt22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45.

Liang, Tian, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 17889–17904.

Liu, Aixin, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Madaan, Aman, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594.

Minsky, Marvin. 1986. *Society of mind*. Simon and Schuster.

Mori, Junko. 2020. The cambridge handbook of japanese linguistics ed. by yoko hasegawa. *The Journal of Japanese Studies*, 46(2):536–540.

Neubig, Graham. 2011. The kyoto free translation task. http://www.phontron.com/kftt.

OpenAI. 2024. Openai research. `https://openai.com/research`. Accessed: 2025-07-25.

Ouyang, Long, 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.

Papineni, Kishore, 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.

Shinn, Noah, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652.

Sin, King-kui, Xi Xuan, Chunyu Kit, Clara Ho-yan Chan, and Honic Ho-kin Ip. 2025. Solving the unsolvable: Translating case law in hong kong. *arXiv preprint arXiv:2501.09444*.

Team, Gemini, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Team, Kimi, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tsujii, Jun'ichi. 1986. Future directions of machine translation. In *Coling 1986 Volume 1: The 11th International Conference on Computational Linguistics*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Xuezhi, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Wei, Jason, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Wu, Minghao, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wan, Weihua Luo, and Kaifu Zhang. 2025. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *Transactions of the Association for Computational Linguistics*, 13:901–922.

Xiong, Kai, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7572–7590.

Yang, An, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yao, Shunyu, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Appendix

Human evaluation agreement and significance testing On the 74 doubly-rated items, inter-annotator agreement for the 4-way ranking is modest but non-trivial: mean per-pair Kendall $\tau_b = 0.221$ (95% bootstrap CI [0.033, 0.395]) and Krippendorff α (ordinal, 7 raters) $= 0.291$ (95% CI [0.143, 0.453]). For the direction of the key Base-vs-R2 contrast, 66.1% of double-rater pairs agree. Diagnosis agreement is lower: Cohen's $\kappa = 0.102$ on the binary "R2 is worse than Base" label and mean Jaccard $= 0.266$ on cause tags, reflecting the well-known subjectivity of fine-grained error attribution in MT evaluation. Despite this noise, the ranking differences are statistically detectable. For significance testing, we aggregate annotations at the item level and use all 100 evaluated items. A Friedman test across Base/R1/R2/R3 is significant ($\chi^2(3) = 21.22$, $p = 9.5 \times 10^{-5}$, Kendall's $W = 0.071$, $n = 100$). Post-hoc Wilcoxon signed-rank tests with Holm–Bonferroni correction confirm that R2 and R3 are both significantly better than Base ($p_{\text{Holm}} = 2.1 \times 10^{-3}$ and 4.3×10^{-3}, respectively) and that R2 is significantly better than R1 ($p_{\text{Holm}} = 4.3 \times 10^{-3}$), while Base-vs-R1 and R2-vs-R3 differences are not significant. Taken together, the debate procedure yields statistically significant, albeit modest, ranking improvements by Round 2 in this small-scale test, with the main contrasts remaining significant under conservative multiple-comparison correction.

LLM Call Analysis Table 3 reports per-segment LLM invocations: every translation, evaluation, or critique/rewrite call counts as one inference unit per sentence, independent of physical batching. Let n be the number of sentences and R the number of rounds. Each method (except zero-shot) first generates a base translation, then iterates R rounds whose per-round cost is determined by the agent topology: $2R$ for REFLECTION (eval + rewrite), $3R$ for PRO/CON (affirmative + negative + moderator), $(m{+}1)R$ for SOM (m society agents + moderator; $m{=}2$ in our setup), and $5R$ for DECOUPLE (four specialist critics + moderator). BLEU, COMET, and XCOMET are computed offline and do not contribute LLM calls.

Method	Per segment	Total
Zero-shot	1	n
Reflection	$1 + 2R$	$n(1 + 2R)$
Pro/Con	$1 + 3R$	$n(1 + 3R)$
SoM	$1 + (m{+}1)R$	$n(1{+}(m{+}1)R)$
Decouple	$1 + 5R$	$n(1 + 5R)$

Table 3: Per-segment LLM inference units. n: #sentences; R: #rounds; m: SoM society size ($m{=}2$ in our experiments).

Supplementary Experimental Details This section provides additional experimental details and prompts used in our study.

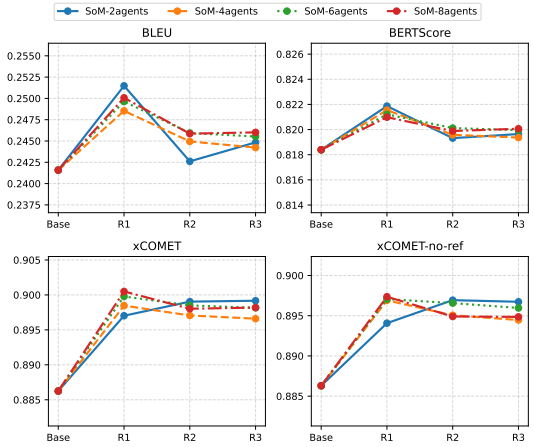

Figure 10: the Impact of Agent Plurality Across All Metrics

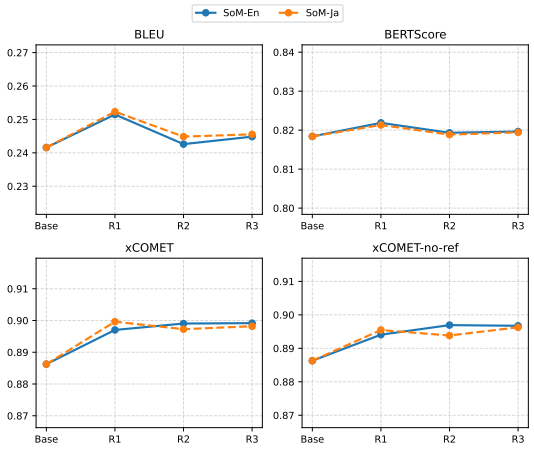

Figure 11: Impact of agent reasoning language

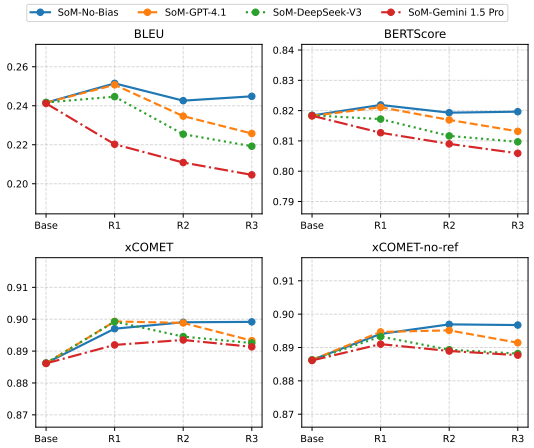

Figure 12: Introduce bias from other models. SoM-GPT-4.1: GPT-4.1 as moderator; diverse LLMs as agents

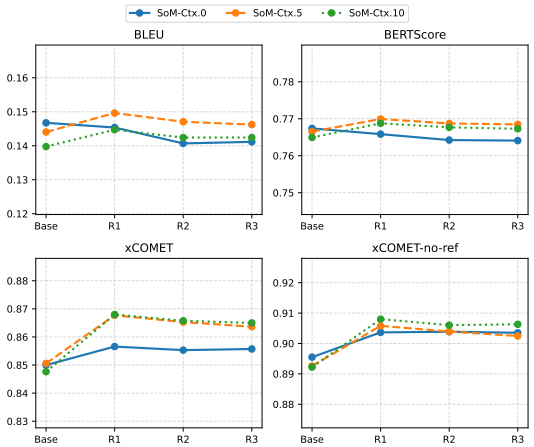

Figure 13: Comparison of distinct context lengths (0, 5 and 10 preceding sentences)

BLEU

System	Base	Round 1	Round 2	Round 3
Self-reflection	24.14	23.38 *(-0.76, p<.001)*	23.58 *(-0.56, p<.001)*	23.61 *(-0.53, p=.001)*
Pro/Con	24.15	21.51 *(-2.64, p<.001)*	21.25 *(-2.90, p<.001)*	21.23 *(-2.91, p<.001)*
Decouple	24.15	21.48 *(-2.67, p<.001)*	22.74 *(-1.42, p<.001)*	22.66 *(-1.49, p<.001)*
SoM	24.13	25.18 *(+1.05, p<.001)*	24.29 *(+0.16, p=.435)*	24.38 *(+0.25, p=.211)*

BERTScore

System	Base	Round 1	Round 2	Round 3
Self-reflection	81.83	81.48 *(-0.36, p<.001)*	81.59 *(-0.24, p<.001)*	81.60 *(-0.23, p<.001)*
Pro/Con	81.84	80.88 *(-0.96, p<.001)*	80.81 *(-1.03, p<.001)*	80.78 *(-1.06, p<.001)*
Decouple	81.84	80.64 *(-1.20, p<.001)*	81.31 *(-0.53, p<.001)*	81.18 *(-0.66, p<.001)*
SoM	81.84	82.19 *(+0.35, p<.001)*	81.95 *(+0.11, p=.075)*	81.99 *(+0.15, p=.019)*

xCOMET

System	Base	Round 1	Round 2	Round 3
Self-reflection	88.62	89.44 *(+0.83, p<.001)*	89.40 *(+0.78, p<.001)*	89.47 *(+0.85, p<.001)*
Pro/Con	88.68	89.17 *(+0.50, p<.001)*	89.06 *(+0.39, p=.008)*	89.05 *(+0.37, p=.011)*
Decouple	88.59	89.00 *(+0.40, p=.009)*	89.12 *(+0.52, p<.001)*	88.92 *(+0.32, p=.016)*
SoM	88.63	89.84 *(+1.21, p<.001)*	89.89 *(+1.26, p<.001)*	89.86 *(+1.23, p<.001)*

xCOMET-no-ref

System	Base	Round 1	Round 2	Round 3
Self-reflection	88.64	89.15 *(+0.51, p<.001)*	89.11 *(+0.47, p<.001)*	89.09 *(+0.45, p<.001)*
Pro/Con	88.68	89.10 *(+0.42, p=.003)*	88.96 *(+0.27, p=.066)*	88.85 *(+0.16, p=.271)*
Decouple	88.61	88.98 *(+0.37, p=.022)*	88.84 *(+0.23, p=.077)*	88.75 *(+0.14, p=.296)*
SoM	88.63	89.52 *(+0.89, p<.001)*	89.60 *(+0.97, p<.001)*	89.54 *(+0.91, p<.001)*

Table 4: Overall performance on the WMT En→Ja setting. Values in parentheses indicate the change from the Base system and the p-value from a paired t-test.

BLEU

System	Base	Round 1	Round 2	Round 3
Self-reflection	24.80	22.68 *(-2.12, p<.001)*	23.32 *(-1.48, p<.001)*	23.41 *(-1.39, p<.001)*
Pro/Con	24.80	20.50 *(-4.30, p<.001)*	19.77 *(-5.03, p<.001)*	20.39 *(-4.41, p<.001)*
Decouple	24.79	20.06 *(-4.73, p<.001)*	22.70 *(-2.09, p<.001)*	22.03 *(-2.76, p<.001)*
SoM	24.80	23.25 *(-1.55, p<.001)*	22.73 *(-2.07, p<.001)*	22.80 *(-2.00, p<.001)*

BERTScore

System	Base	Round 1	Round 2	Round 3
Self-reflection	84.73	84.23 *(-0.49, p<.001)*	84.39 *(-0.33, p<.001)*	84.41 *(-0.31, p<.001)*
Pro/Con	84.72	83.60 *(-1.12, p<.001)*	83.44 *(-1.28, p<.001)*	83.45 *(-1.27, p<.001)*
Decouple	84.73	83.11 *(-1.62, p<.001)*	84.02 *(-0.70, p<.001)*	83.88 *(-0.85, p<.001)*
SoM	84.72	84.48 *(-0.24, p=.006)*	84.27 *(-0.45, p<.001)*	84.27 *(-0.44, p<.001)*

xCOMET

System	Base	Round 1	Round 2	Round 3
Self-reflection	90.30	90.48 *(+0.18, p=.003)*	90.64 *(+0.34, p<.001)*	90.61 *(+0.31, p<.001)*
Pro/Con	90.29	89.97 *(-0.33, p<.001)*	89.99 *(-0.30, p=.002)*	89.96 *(-0.34, p<.001)*
Decouple	90.30	89.55 *(-0.75, p<.001)*	89.96 *(-0.34, p<.001)*	89.63 *(-0.67, p<.001)*
SoM	90.29	90.20 *(-0.09, p=.301)*	90.07 *(-0.22, p=.015)*	90.10 *(-0.19, p=.032)*

xCOMET-no-ref

System	Base	Round 1	Round 2	Round 3
Self-reflection	90.20	90.17 *(-0.03, p=.686)*	90.12 *(-0.08, p=.258)*	90.08 *(-0.11, p=.104)*
Pro/Con	90.19	89.74 *(-0.45, p<.001)*	89.48 *(-0.71, p<.001)*	89.43 *(-0.77, p<.001)*
Decouple	90.20	89.27 *(-0.94, p<.001)*	89.71 *(-0.49, p<.001)*	89.44 *(-0.76, p<.001)*
SoM	90.19	89.90 *(-0.29, p<.001)*	89.68 *(-0.51, p<.001)*	89.71 *(-0.48, p<.001)*

Table 5: Overall performance on the WMT Ja→En setting. Each cell reports the score, with values in parentheses indicating the difference from the Base system and the p-value from a paired t-test.

System	Round 1	Round 2	Round 3
En-Ja			
SoM	**0.088**	-0.069	0.012
Self-reflection	-0.068	0.027	0.001
Pro/Con	-0.198	-0.033	0.012
Decouple	-0.200	0.129	0.004
Ja-En			
SoM	-0.129	-0.067	0.020
Self-reflection	-0.165	0.020	0.008
Pro/Con	-0.288	-0.032	0.038
Decouple	-0.320	0.211	-0.023

(a) BLEU

System	Round 1	Round 2	Round 3
En-Ja			
SoM	**0.127**	-0.031	-0.000
Self-reflection	-0.088	0.031	-0.001
Pro/Con	-0.237	-0.025	-0.003
Decouple	-0.260	0.156	-0.018
Ja-En			
SoM	-0.080	-0.078	-0.002
Self-reflection	-0.163	0.044	0.001
Pro/Con	-0.266	-0.029	0.006
Decouple	-0.316	0.216	-0.034

(b) BERTScore

System	Round 1	Round 2	Round 3
En-Ja			
SoM	**0.170**	-0.016	-0.020
Self-reflection	0.071	0.000	0.024
Pro/Con	-0.004	-0.014	0.015
Decouple	0.019	0.001	-0.006
Ja-En			
SoM	0.006	-0.055	0.020
Self-reflection	0.059	0.044	-0.002
Pro/Con	-0.050	-0.018	-0.011
Decouple	-0.122	0.068	-0.035

(c) xCOMET

System	Round 1	Round 2	Round 3
En-Ja			
SoM	**0.076**	0.030	-0.016
Self-reflection	0.069	-0.025	-0.005
Pro/Con	0.036	-0.027	-0.011
Decouple	0.040	-0.040	0.024
Ja-En			
SoM	-0.057	-0.066	0.012
Self-reflection	0.013	-0.008	-0.022
Pro/Con	-0.100	-0.040	-0.025
Decouple	-0.142	0.055	-0.063

(d) xCOMET-no-ref

Table 6: Net Gain Rate (NGR) after each debate round across systems and metrics. Each subtable shows one evaluation metric for both En-Ja and Ja-En directions.

Prompt

You are an expert in machine translation and multi-agent debate analysis. Your task is to analyze why translation quality declined in a multi-agent debate scenario, **without being constrained by predefined error categories**.
Context: - In a translation debate, multiple AI agents debate to improve translation quality
- First round improved translation quality over the base translation (xCOMET score increased)
- Subsequent round caused quality to decline (xCOMET score decreased)
- Your goal is to freely identify and analyze what went wrong
Input Data: - Source: {source}
- Reference: {reference}
- Base Translation: {base_translation}
- xCOMET Scores: {xcomet_scores}
- Score Pattern: {score_pattern}
- Score Trend Change: Improvement={improvement:.4f}, Decline={decline:.4f}
Round-by-Round Analysis: {rounds_analysis}
Your Task: Please perform a comprehensive, open-ended analysis:
1. **Identify all issues you observe** – Don't limit yourself to predefined categories
2. **Create your own error classifications** based on what you actually see
3. **Analyze the debate process itself** – How did agent interactions contribute to the problem?
4. **Examine the progression** – What specific changes occurred between rounds?
5. **Consider multiple perspectives** – Translation quality, debate dynamics, agent reasoning
Instructions: - Be completely free in your analysis – identify whatever problems you see
- Create your own error type names that best describe the issues
- Provide specific evidence from the translations and agent reasoning
- Analyze both the translation quality **and** the debate process
- Don't feel constrained by any predefined framework
Expected Output (in JSON format): { "identified_problems": [{ "problem_name": "your own descriptive name for this issue", "problem_type": "your own category/classification", "description": "detailed explanation of what went wrong", "evidence": "specific quotes or examples from the data", "severity": "high/medium/low", "round_affected": "which round(s) this appeared in" }], "debate_process_analysis": { "overall_effectiveness": "assessment of how well the debate worked", "agent_behavior_issues": "problems with how agents interacted", "reasoning_quality": "quality of agent reasoning and justifications", "consensus_building": "how well agents reached agreement" }, "translation_progression": { "round_0_assessment": "analysis of initial translation", "round_1_changes": "what improved and why", "round_2_problems": "specific issues that caused decline", "overall_pattern": "summary of the quality trajectory" }, "root_cause_analysis": { "primary_cause": "main reason for quality decline", "contributing_factors": ["other factors that played a role"], "systemic_vs_specific": "whether this seems like a systemic issue or case-specific" }, "recommendations": { "immediate_fixes": ["specific suggestions for this type of case"], "process_improvements": ["suggestions for improving the debate process"], "prevention_strategies": ["how to prevent similar issues in future"] } }
Only return the JSON object, no additional text.

Table 7: Prompt template for open-ended error diagnosis in multi-agent translation debates

Prompt	Prompt Content
base_prompt	Translate the following text from <src_lng> to <tgt_lng>: <source>. <context> Please return only the <tgt_lng> translation. Do not include JSON, metadata, explanations, or any additional content.

Table 8: Prompt templates for the base translation

Prompt	Prompt Content
eval_prompt	Label the translation quality as "Good", "Medium", or "Bad". Now please output your answer in JSON format, strictly following this structure: "label": "Good" Only return the JSON object. Do not include any other content.
reflection_prompt	You are refining your previous <tgt_lng> translation to make it more accurate, fluent, and natural. Original sentence: <source> Previous translation (labeled as <label>): <pre_tran> <context> Return your improved version. If no better version exists, return the same one. Only return the final <tgt_lng>. translation. No explanations, no metadata.

Table 9: Prompt templates for the self-reflection framework

Prompt	Prompt Content
player_meta_prompt	You are a debater participating in a translation debate competition. Welcome to this constructive and thoughtful debate. It is entirely appropriate to offer different perspectives, as our shared objective is to discover the most accurate and natural translation. The debate topic is: What is the appropriate <tgt_lng> translation of the following <src_lng> text: "<source>" <context>
moderator_meta_prompt	You are acting as both a moderator and a linguistics expert. In this translation debate, two debaters will present their proposed translations and explain their reasoning regarding the correct <tgt_lng> version of the given <src_lng> text: "<source>". <context> At the conclusion of each round, your role is to evaluate the translations based on: 1. Accuracy; 2. Fluency; 3. Grammar & Syntax; 4. Lexical Choice; 5. Sentence Structure & Voice.
affirmative_prompt	You believe the correct translation is: <base_translation>. Please restate your translation and explain your reasoning.
negative_prompt	<aff_ans> You have a different interpretation of the translation. Please provide your version and explain your reasoning.
moderator_prompt	The <round> round has concluded. Affirmative side presented: <aff_ans>. Negative side presented: <neg_ans>. As a linguist and the moderator, evaluate both and select only one final Japanese translation. If both are equally good, choose one or slightly revise one of them to produce the final result, but do not merge. Strictly output only one Japanese translation. Do not include Romaji, alternatives, or explanations. Return this JSON format: {"Whether there is a preference": "Yes or No", "Supported Side": "Affirmative or Negative", "Reason": "", "debate_translation": ""}.
debate_prompt	<oppo_ans> Do you agree with the reasoning above? Please explain your position and present your own refined translation. You may refer to the moderator translation from the previous round: <pre_mod>.

Table 10: Prompt templates for the pro/con framework.

Prompt	Prompt Content
player_meta_prompt	You are a debater participating in a translation debate competition. Welcome to this constructive and thoughtful debate. It is entirely appropriate to offer different perspectives, as our shared objective is to discover the most accurate and natural translation. The debate topic is: What is the appropriate <tgt_lng> translation of the following <src_lng> text: "<source>" <context>
moderator_meta_prompt	You are acting as a moderator of a translation debate competition. In this translation debate, debaters will present their proposed translations and explain their reasoning regarding the correct <tgt_lng> version of the given <src_lng> text: "<source>".<context> At the conclusion of each round, your role is to synthesize and optimize the translations
initial_prompt	Please translate the following sentence to <tgt_lng>: "<source>" Afterwards, please briefly explain your reasoning. Strictly return JSON: { "Translation": "", "Brief Reason": ""} Only return the JSON object. Do not include anything else.
moderator_prompt	The <round> round of the translation debate has concluded. If it is helpful, here are some suggested translations from other agents: <candidate_solutions> you can select the best translation from the candidates, or you can refine one to improve it if you think it can be improved. You may also choose to reject all candidates and provide your own translation. But you must return only one final translation to represent this round. State the answer FIRST, then your reasoning. strictly return JSON: { "Translation": "", "Brief Reason": ""} Only return the JSON object. Do not include anything else.
debate_prompt	Please translate the following sentence to <tgt_lng>: "<source>" If it is helpful, here are some suggested translations from other agents: <candidate_solutions> <pre_mod> If an agent's reasoning is incorrect, point it out. Using the suggestions from other agents as additional information, please refine your translation if necessary. State the answer FIRST, then your reasoning. Strictly return JSON: {"Translation": "", "Brief Reason": ""} Only return the JSON object. Do not include anything else.

Table 11: Prompt templates for the SoM framework.

Prompt	Prompt Content
player_meta_prompt	You are an annotator for the quality of machine translation. You will be assessing translations at the segment level, where a segment may contain one or more sentences. Each segment is aligned with a corresponding source segment. Please identify all errors within the specified category within each translated segment. To identify an error, highlight the relevant span of text, and select a category/sub-category and severity level from the available options. (The span of text may be in the source segment if the error is a source error or an omission.) When identifying errors, please be as fine-grained as possible. For example, if a sentence contains two words that are each mistranslated, two separate mistranslation errors should be recorded. If a single stretch of text contains multiple errors, you only need to indicate the one that is most severe. For severity, Major represents errors that may confuse or mislead the reader due to a significant change in meaning or because they appear in a visible or important part of the content; Minor represents errors that don't lead to loss of meaning and wouldn't confuse or mislead the reader but would be noticed, would decrease stylistic quality, fluency or clarity, or would make the content less appealing. In case when it is not possible to reliably identify distinct errors because the translation is too badly garbled or is unrelated to the source, then mark a special category, called non-translation, that spans the entire segment. There can be at most one non-translation error per segment, and it should span the entire segment. No other errors should be identified if non-translation is selected. <context> Only return the JSON object. Do not include anything else.
moderator_meta_prompt	You are a moderator of a translation debate competition. You will be provided with the <src_lng> source text, candidate <tgt_lng> translations and the annotations from the accuracy, fluency, terminology, and style experts. <context> Your task is to comprehensively consider the annotations and provide a final translation.

Prompt (cont.)	Prompt Content (cont.)
accuracy_agent	You are an accuracy errors detection expert for translations. Please check the translation for the following subcategories of accuracy errors: 1. Accuracy Addition: The translation includes information not present in the source. 2. Omission translation: The translation is missing content from the source. 3. Mistranslation: The translation does not accurately represent the source. 4. Untranslated Text: Source text has been left untranslated. Please analyze the following segment pair and annotate errors. <src_lng> source: <source> <tgt_lng> translation:<target_segment> Provide your annotations in JSON format as follows: annotations":["error_span":(if non-translation error is selected, provide 'all'; otherwise, the error_span must be chosen from within the translated segment), "category":"(accuracy/subcategory or non-translation)", "severity":"(minor or major)", "is_source_error":"(yes or no)",...],"suggested_translation": "(Please provide a revised translation that corrects the identified accuracy issues while preserving fluency and meaning.)"
fluency_agent	You are a fluency errors detection expert for translations. Please check the translation for the following subcategories of fluency errors: 1. **Punctuation**: Incorrect punctuation (for locale or style). 2. **Spelling**: Incorrect spelling or capitalization. 3. **Grammar**: Problems with grammar, other than orthography. 4. **Register**: Wrong grammatical register (e.g., inappropriately informal pronouns). 5. **Inconsistency**: Internal inconsistency (not related to terminology). 6. **Character Encoding**: Characters are garbled due to incorrect encoding. Please analyze the following segment pair and annotate errors. <src_lng> source: <source> <tgt_lng> translation: <target_segment> Provide your annotations in JSON format as follows: "annotations":["error_span":(if non-translation error is selected, provide 'all'; otherwise, the error_span must be chosen from within the translated segment), "category":"(fluency/subcategory or non-translation)", "severity":"(minor or major)", "is_source_error":"(yes or no)",...], "suggested_translation": "(Please provide a revised translation that corrects the identified accuracy issues while preserving fluency and meaning.)"

Prompt (cont.)	Prompt Content (cont.)
term_agent	You are a terminology errors detection expert for translations. Please check the translation for the following subcategories of terminology errors: 1. **Inappropriate for context**: Terminology is non-standard or does not fit context. 2. **Inconsistent use**: Terminology is used inconsistently. Please analyze the following segment pair and annotate errors. <src_lng> source: <source> <tgt_lng> translation: <target_segment> Provide your annotations in JSON format as follows: "annotations":["error_span":(if non-translation error is selected, provide 'all'; otherwise, the error_span must be chosen from within the translated segment), "category":"(terminology/subcategory or non-translation)", "severity":"(minor or major)", "is_source_error":"(yes or no)",...], "suggested_translation": "(Please provide a revised translation that corrects the identified accuracy issues while preserving fluency and meaning.)"
style_agent	You are a style errors detection expert for translations. Please check the translation for the style error: **Awkward**: translation has stylistic problems. Please analyze the following segment pair and annotate errors. <src_lng> source: <source> <tgt_lng> translation: <target_segment> Provide your annotations in JSON format as follows: "annotations":["error_span":(if non-translation error is selected, provide 'all'; otherwise, the error_span must be chosen from within the translated segment), "category":"(style/subcategory or non-translation)", "severity":"(minor or major)", "is_source_error":"(yes or no)",...], "suggested_translation": "(Please provide a revised translation that corrects the identified accuracy issues while preserving fluency and meaning.)"

Prompt (cont.)	Prompt Content (cont.)
moderator_prompt	The <round> round of the annotation has concluded. From previous annotations, we have the accuracy errors detection expert annotation: <accuracy_annotation>; the fluency errors detection expert annotation: <fluency_annotation>; the terminology errors detection expert annotation: <term_annotation>; and the style errors detection expert annotation: <style_annotation..Based on the above information, output your comprehensive translation, you can select the best translation from the candidates, or you can refine one to improve it if you think it can be improved. You may also choose to reject all candidates and provide your own translation. But you must return only one final translation as the best translation to represent this round.State the answer FIRST, then your reasoning. Strictly return JSON: "Translation": "", "Brief Reason": "" Only return the JSON object. Do not include anything else.

Table 12: Prompt templates for the Decouple framework.