Investigating the Effects of Translation Quality on LLM Performance in Machine-Translated Theory of Mind Benchmarks

Haruhisa Iseno[1,2], Atsumoto Ohashi[1] , Tetsuji Ogawa[3], Shinnosuke Takamichi[4], Ryuichiro Higashinaka[1,2]

[1]Graduate School of Informatics, Nagoya University [2]NII LLMC [3]Department of Communications and Computer Engineering, Waseda University [4]Department of Information and Computer Science, Keio University
{iseno.haruhisa.h4@s.mail, ohashi.atsumoto.c0@s.mail}.nagoya-u.ac.jp
ogawa.tetsuji@waseda.jp, shinnosuke_takamichi@keio.jp, higashinaka@i.nagoya-u.ac.jp

Abstract

In recent years, a variety of benchmarks have been proposed to evaluate the Theory of Mind (ToM) of large language models (LLMs). However, most of these benchmarks are constructed in English, and there is still a shortage of ToM benchmarks for other languages. A straightforward approach to creating non-English ToM benchmarks is to machine-translate existing English benchmarks, but it remains unclear how translation errors affect ToM evaluation results. In this study, we machine-translated two English ToM benchmarks, ToMBench and FANToM, into Japanese and examined how translation quality influences the ToM evaluation of LLMs. Our experiments show that the impact of translation quality differs across benchmarks: for ToMBench, machine translation consistently decreased the evaluation scores of all models, while for FANToM the impact was limited. Furthermore, our analysis indicates that translation errors related to accuracy are a major factor in the degradation of evaluation scores.

Introduction

Most existing Theory of Mind (ToM) benchmarks (Chen et al. 2025) are constructed only in English (Le, Boureau, and Nickel 2019; Gandhi et al. 2023; Kim et al. 2023; Xu et al. 2024; Shinoda et al. 2025), and evaluation benchmarks in languages other than English remain insufficient. As a method to address this issue, benchmark creation through machine translation of existing datasets (Chen et al. 2024) can be considered; however, the extent to which machine translation errors affect ToM evaluation remains unclear.

In this study, we investigate the impact of translation quality on ToM evaluation of Large Language Models (LLMs) in machine-translated ToM benchmarks. First, we machine-translate existing ToM benchmarks (ToMBench and FANToM) into Japanese and evaluate the extent to which LLM performance changes compared to the English versions. Additionally, we manually post-edit a subset of the translated questions to investigate whether correcting translation errors improves accuracy rates. We then analyze what types of translation errors affect ToM evaluation. Through these analyses, we clarify the impact of translation quality on ToM evaluation results and identify the underlying factors.

Approach

We first select multiple ToM benchmarks with different characteristics as evaluation targets for generalizability. We then machine-translate the selected benchmarks into Japanese and investigate changes in LLM evaluation results through comparison with the original English versions. Multiple LLMs are selected from recent state-of-the-art models for evaluation. Furthermore, we conduct manual post-editing on a subset of the translated benchmarks to correct translation errors. By comparing LLM accuracy rates before and after correction, we investigate the impact of translation quality on benchmark performance. We then analyze the relationship between the amount of translation errors in questions (context, question text, and answer choices) and their correctness using correlation analysis.

Datasets

In this study, we selected two benchmarks for translation: ToMBench (Chen et al. 2024) and FANToM (Kim et al. 2023). ToMBench uses narrative texts as context, while FANToM uses dialogues; this difference results in the two benchmarks having distinct characteristics.

Translated datasets

We machine-translated all data from ToMBench and FANToM into Japanese with Llama4-Scout, which showed superior performance over DeepL for sampled questions. The translation prompt was simply "Please translate the following English text into Japanese," and translation was performed in a zero-shot manner. As a result of translation, about 1.6 million characters were translated for ToMBench and about 10 million characters for FANToM.

Post-edited datasets

We sampled questions from the translated benchmarks and conducted manual translation quality evaluation and post-editing.

For ToMBench, we sampled a total of 60 questions (556 sentences) from 8 major question types out of the 20 types included in the benchmark. For FANToM, we randomly sampled 20 dialogues (1,563 sentences) containing a total of 250 question (including 72 questions from the three main

Table 1: Accuracy rates on English and Japanese benchmarks. Bold indicates the higher accuracy rate.

	ToMBench		FANToM	
	English	Japanese	English	Japanese
GPT-4o	**78.8**	71.8	**64.4**	59.6
Claude4	**79.4**	77.6	59.7	**61.4**
Gemini2.5	**79.7**	73.3	73.3	**73.7**
Llama3.3	**76.5**	70.5	**50.6**	45.4
Qwen3	**73.2**	62.0	45.4	**48.5**

Table 2: Accuracy rates before and after correction on the sampled questions for post-editing. Bold indicates the higher accuracy rate.

	ToMBench		FANToM	
	Before	After	Before	After
GPT-4o	71.6	**75.0**	56.3	**57.7**
Claude4	**80.0**	75.0	**61.9**	60.5
Gemini2.5	68.3	**75.0**	**61.9**	**61.9**
Llama3.3	61.6	**66.6**	**39.4**	**39.4**
Qwen3	65.0	**71.6**	60.5	**61.9**

question categories: BeliefQ, InfoAccessibilityQ, and AnswerabilityQ).

Workers evaluated and corrected the sampled translated sentences based on Multidimensional Quality Metrics (MQM) (Lommel et al. 2014). Specifically, following the JTF Translation Quality Assessment Guidelines[1] established by the Japan Translation Federation (JTF) based on MQM, workers checked for translation errors on a sentence-by-sentence basis. Sentences containing errors were corrected to appropriate Japanese sentences. As a result, 354 translation errors were corrected in ToMBench and 768 in FANToM. The average number of errors per 100 characters was 3.1 for ToMBench and 3.4 for FANToM.

Experiment

Procedure

In the evaluation, we presented LLMs with context (narrative or dialogue text) and had them answer multiple-choice questions. The prompts for solving the questions were designed for this study, with the same content provided in English and Japanese versions respectively (details in Appendix). We used five state-of-the-art LLMs as evaluation targets (GPT-4o, Claude-4-Sonnet, Gemini-2.5-Flash, Llama-3.3-70B, Qwen-3-32B).

We evaluated LLM performance on ToM using both the original English benchmarks and the Japanese-translated versions. For ToMBench, we used all 2,860 questions. For FANToM, we used 3,280 questions, covering BeliefQ, InfoAccessibilityQ, and AnswerabilityQ.

To examine the extent to which translation affects benchmark performance, we evaluated LLM performance on ToM before and after correction on the corrected question sets. For ToMBench, we used all 60 corrected questions. For FANToM, we used 72 questions from the three main question categories; the remaining questions in the dataset were not used because they were sub-questions derived from the main questions.

[1]https://www.jtf.jp/pdf/jtf_translation_quality_guidelines_v1.pdf

Table 3: Point-biserial correlation coefficients between the number of errors in a question and the correctness for that question. Bold indicates the higher correlation. * indicates statistical significance at $p < 0.05$.

	Accuracy errors	Fluency errors
GPT-4o	$\mathbf{+0.230}^*$	$+0.048$
Claude4	$\mathbf{+0.184}^*$	-0.072
Gemini2.5	$\mathbf{+0.105}$	$+0.031$
Llama3.3	$\mathbf{+0.235}^*$	$+0.189^*$
Qwen3	$\mathbf{+0.045}$	$+0.040$

We used point-biserial correlation coefficients to analyze the relationship between the number of errors in a question and the correctness for that question. The analysis was conducted for all the samples of ToMBench and FANToM. In this analysis, a positive correlation indicates that questions with more translation errors are more likely to be answered incorrectly by the models. This analysis was performed separately for accuracy errors and fluency errors based on MQM classification (the distribution of translation errors in each benchmark is provided in Appendix).

Result

Table 1 shows a comparison of LLM performance on ToM between the original English benchmarks and the Japanese-translated versions. For ToMBench, performance decreased in the Japanese version across all models. In contrast, for FANToM, results were mixed, with some models showing decreased performance and others showing improved performance. This indicates that the impact of translation on benchmark performance varies depending on the type of benchmark.

Table 2 shows a comparison of LLM performance on ToM before and after post-editing. For all models except Claude-4, performance improved or was maintained after correction in both benchmarks. This suggests that the inclusion of translation errors tends to decrease performance on ToM evaluation. We observed that the improvement after correction was larger for ToMBench compared to FANToM.

Table 3 shows the point-biserial correlation coefficients. Examining the overall trend, we found that accuracy errors showed higher correlation than fluency errors across all models. In particular, for GPT-4o, Claude-4, and Llama-3.3, we confirmed a significantly high correlation between the number of accuracy errors and correctness ($p < 0.05$).

Conclusion

In this study, we translated ToMBench and FANToM into Japanese and analyzed performance comparisons with the English versions as well as the effects of post-editing. We also investigated how translation errors affect benchmark performance.

A limitation of this study is that the evaluation was restricted to two benchmarks. Future research should investigate more diverse benchmarks to analyze this causal relationship in detail.

Acknowledgments

This work was supported 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.

References

Chen, R.; Jiang, W.; Qin, C.; and Tan, C. 2025. Theory of Mind in Large Language Models: Assessment and Enhancement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 31539–31558.

Chen, Z.; Wu, J.; Zhou, J.; Wen, B.; Bi, G.; Jiang, G.; Cao, Y.; Hu, M.; Lai, Y.; Xiong, Z.; and Huang, M. 2024. ToMBench: Benchmarking Theory of Mind in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15959–15983.

Gandhi, K.; Fränken, J.-P.; Gerstenberg, T.; and Goodman, N. 2023. Understanding social reasoning in language models with language models. *In Proceedings of the 37th International Conference on Neural Information Processing Systems*, 36: 13518–13529.

Kim, H.; Sclar, M.; Zhou, X.; Bras, R.; Kim, G.; Choi, Y.; and Sap, M. 2023. FANToM: A Benchmark for Stress-testing Machine Theory of Mind in Interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 14397–14413.

Le, M.; Boureau, Y.-L.; and Nickel, M. 2019. Revisiting the Evaluation of Theory of Mind through Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 5872–5877.

Lommel, A.; Burchardt, A.; Popović, M.; Harris, K.; Avramidis, E.; and Uszkoreit, H. 2014. Using a new analytic measure for the annotation and analysis of MT errors on real data. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, 165–172.

Shinoda, K.; Hojo, N.; Nishida, K.; Mizuno, S.; Suzuki, K.; Masumura, R.; Sugiyama, H.; and Saito, K. 2025. ToMATO: Verbalizing the mental states of role-playing LLMs for benchmarking theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1520–1528.

Xu, H.; Zhao, R.; Zhu, L.; Du, J.; and He, Y. 2024. OpenToM: A Comprehensive Benchmark for Evaluating Theory-of-Mind Reasoning Capabilities of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8593–8623.

Appendix

Prompts for Theory of Mind Tasks

This section presents the English versions of the prompts given to LLMs to solve ToMBench and FANToM. When solving questions in Japanese, these prompts were manually translated into Japanese and provided to the LLMs.

The prompt for solving ToMBench is as follows. {context} contains the context that serves as the basis for inference, {question} contains questions about characters' mental states, and {a}, {b}, {c}, and {d} are the answer choices.

```
Please read the passage and the question I will ask.
Choose the correct answer from options A, B, C, and D
.
{context}
{question}
A: {a}
B: {b}
C: {c}
D: {d}
Please answer with the letter of the option that you
think is correct and do not output anything other
than a single letter.
```

The following are the prompts used to solve FANToM, which is used to solve BeliefQ, InfoAccessibilityQ, and AnswerabilityQ. {context} contains the dialogue text that serves as the basis for inference, and {BeliefQ}, {InfoQ}, and {AnsQ} contain question texts defined for each task by the dataset. Additionally, {factQ} and {factA} contain the facts asked in BeliefQ, and {candidates} lists the names of the characters.

```
{context}
Question: {BeliefQ}
{ans_a}
{ans_b}
Please choose either a or b as the correct answer.
Output only a or b.
```

```
{context}
Information: {factQ} {factA}
Question: {InfoQ}
Characters: {candidates}
Choose the characters who correctly answer the
question from the list above.
Separate names with commas.
Answer:
```

```
{context}
Target: {factQ}
Question: {AnsQ}
Characters: {candidates}
Choose the characters who correctly answer the
question from the list above.
Separate names with commas.
Answer:
```

Distribution of Translation Errors

This section presents the distribution of translation errors contained in the questions from ToMBench and FANToM that were subject to post-editing. Translation errors are classified into three severity levels: critical, major, and minor, based on their impact on comprehension.

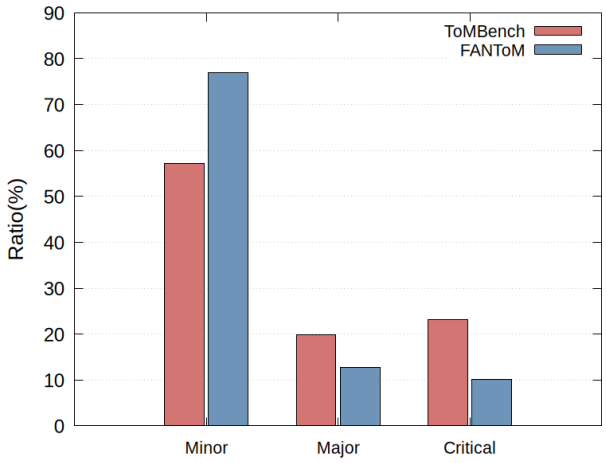

Figure 1: Distribution of error severity in the sampled questions.

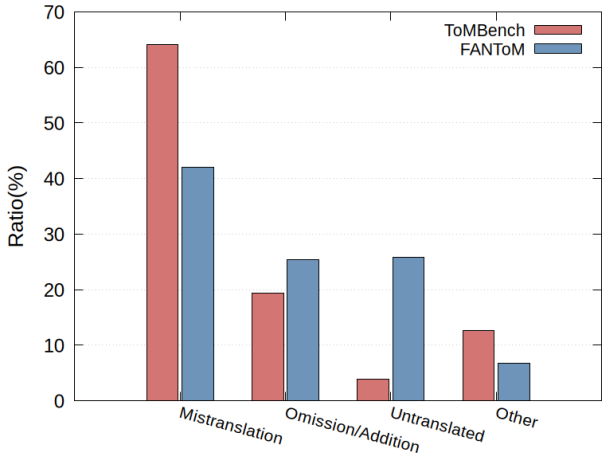

Figure 2: Distribution of accuracy errors in the sampled questions.

Figure 1 shows the distribution of translation error severity in both benchmarks. FANToM has a higher proportion of minor errors compared to ToMBench.

Figure 2 shows the distribution of accuracy errors in both benchmarks. Mistranslations (errors where the wrong word choice or expression changes the intended meaning) were most frequently observed in both benchmarks. Additionally, while untranslated portions were rarely observed in ToMBench, FANToM showed a relatively large proportion of untranslated portions. This is likely because the context in FANToM is longer than in ToMBench, making it more prone to leaving portions untranslated in long text translation.

Figure 3 shows the distribution of fluency errors in both benchmarks. Inconsistency was the main error factor in both benchmarks. This is mainly attributed to character names not being translated consistently.

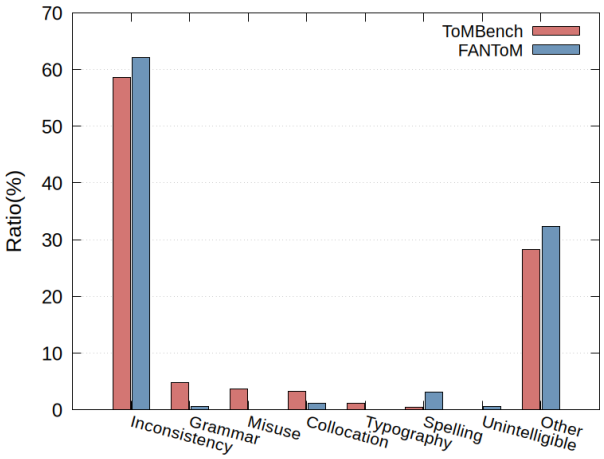

Figure 3: Distribution of fluency errors in the sampled questions.