Prompt-Based Interventions and Evaluation Challenges in Culturally Aware Dialogue Translation

Hei Man, Haley Fong[1] Katsuhiko Hayashi[1]
[1]The University of Tokyo
{fonghmhaley, katsuhiko-hayashi}@g.ecc.u-tokyo.ac.jp

Abstract

While culture-aware machine translation research has largely focused on culture-specific items, cross-cultural communicative norms in dialogue remain underexplored. This study examines whether prompt-based interventions can improve the cultural and interactional appropriateness of dialogue translation between Japanese and English, as well as the extent to which existing evaluation frameworks capture such quality. Results show that prompting can yield pragmatically improved outputs, but these effects are inconsistent and limited. More importantly, both automatic and manual evaluation frameworks fail to adequately assess culturally grounded dialogue quality, underscoring the need for culturally sensitive evaluation taxonomies for dialogue translation between culturally distant languages.

1 Introduction

As machine translation reached a new paradigm with large language models (LLMs), traditional challenges such as long contexts and parallel data requirements have been largely alleviated [1]. With outstanding generative and in-context learning abilities, LLMs carry potential to produce translations beyond string-level equivalence [2], renewing interest in document-level coherence, interactive translation, and culturally adaptive output.

Translation extends beyond linguistic forms, operating as culturally grounded mediation rather than code conversion [3]. Culture manifests in language through conceptual meaning-making, linguistic encoding, and contextualized use [4], with translation functioning as *recontextualization*— removing culturally situated meanings from their original setting and re-embedding them in a new cultural frame [5].

Recent culture-aware MT research has focused on culture-specific items (CSIs), with [6] finding that LLMs handle CSIs better than neural MT, particularly when prompted with CSI definitions or cultural reasoning.

Prompt engineering has been tested for MT and cultural alignment, with research showing that task and domain details can improve translation quality [7], while cultural and demographic specifications in prompts aim to mitigate bias [8, 9].

Existing prompt engineering research for MT has mainly varied task description details, with output evaluation occurring solely at the linguistic level through both automatic and manual means [10, 7, 11]. Despite the scarcity of studies concerning cultural and contextual language use, [12] provides the closest example, employing specific prompting techniques to adapt American English dialogues to Indian norms. Beyond CSI-handling and naturalness, they evaluated contextual aspects including offensiveness and stereotypical behaviour, demonstrating that prompts can encourage cultural localization while maintaining naturalness.

Japanese and English provide a particularly suitable test case for investigating culturally grounded dialogue translation given their substantially different communicative norms. For instance, Japanese communication prioritizes indirectness and relational harmony through honorific systems, hedging, and strategic omission, whereas North American English favours directness and explicit reference [13]. Translation systems often generate culturally inappropriate results when they fail to recognize these different communication logics.

This study investigates whether prompt-based interventions can improve the cultural and interactional appropriateness of dialogue translation between Japanese and English, and whether current evaluation frameworks can reliably assess such quality. Experiments with multiple

prompting conditions reveal that while prompting occasionally produces pragmatically appropriate translations, these improvements remain inconsistent, and existing metrics prove insufficient.

2 Methodology

2.1 Datasets

The dialogues used in the experiments were borrowed from BPersona-Chat [14][1] (hereby the BPC corpus) and the Business Scene Dialogue Corpus (BSD) [15][2]. The former is an evaluation dataset designed to assess machine translation systems' handling of colloquial, multi-turn conversations. Only dialogues with human translations where all instances were marked as acceptable were included in this project. Since the model appeared to have been exposed to the BSD corpus, back translation was performed for this dataset.

Dataset	# of Dia.	Turns/Dia.	SL Turn Length
BPC (j → e)	99	10.925	45.665 chars
BPC (e → j)	110	14.605	12.598 words
BSD (e ← j)	140	16.513	34.478 chars
BSD (j ← e)	106	19.413	15.578 words

Table 1 Basic statistics on the data used in the experiments.

2.2 Model and Prompts

`Llama-3.1-8B-Instruct` [16] was employed to carry out the experiments with zero-shot prompting to ensure that any changes in the output is due to using different prompts. Below is the **B**aseline prompt, referenced from [10], was expanded with instructions to leverage inter-sentential context for **D**ocument-level translation[12], consider **F**ormality [7], and adhere to designated target **C**ultures [8].

```
Dialogue:
{srctxt}
####
Translate each line in the dialogue into {tl},
ensuring coherence across turns in the dialogue.
This is a {colloquial | business} conversation.
Make sure your translation aligns with
{Japanese | North American} culture and audience.
Translated Dialogue:
```

2.3 Evaluation and Error Analysis

Automatic evaluation using COMET[17][3] identified dialogues for manual analysis.Following [10], COMET was applied to each turn, with 0.796 serving as the threshold to distinguish well-translated from poorly translated lines. Dialogues were ranked by their proportion of low-scoring utterances within each prompting condition. After excluding problematic outputs (missing, untranslated, or romanized Japanese), dialogues with the highest and lowest average COMET scores were selected for manual inspection, referred to as high-COMET and low-COMET dialogues.

For manual evaluation, the MQM-Chat framework[18], an extension of the Multidimensional Quality Metrics (MQM) [4] specifically designed for dialogue translation, was employed. Error types include Mistranslation, Omission or Addition, Terminology, Unnatural Style, and dialogue-specific categories (Ambiguity and Disambiguation, Buzzword or Loanword Issues, Dialogue Inconsistency). Although MQM-Chat does not explicitly encode cultural dimensions, annotation incorporated culturally salient features including honorific choice, backchanneling, and register consistency [12]. While discourse-level properties such as indirectness and interpersonal stance are not independent categories, they are realized through these observable features and reflected in MQM-Chat error labels.

3 Results

3.1 Automatic Evaluation

Table 2 reports COMET scores after the aforementioned filtering process. Echoing the findings of [10], the results indicate that the model is generally capable of producing semantically adequate translations. Across both corpora, ja → en translation repeatedly achieved higher COMET scores than en → ja, indicating a systematic directional asymmetry in dialogue translation performance.

With respect to prompting, document-level instructions improved performance for casual dialogues, particularly with ja → en translation. Supplementing hints on formality yielded similar gains for casual dialogues, this time with better results in the opposite direction. In contrast, busi-

1) https://github.com/cl-tohoku/BPersona-chat
2) https://github.com/tsuruoka-lab/BSD
3) https://github.com/Unbabel/COMET
4) https://themqm.org

Setup	B	D	D+F	D+C	D+F+C
BPC (j → e)	0.797 (0.99)	**0.806 (1.00)**	0.798 (0.97)	0.796 (0.98)	0.796 (0.98)
BPC (e → j)	0.727 (0.70)	0.791 (0.85)	0.803 (0.88)	**0.805 (0.87)**	0.776 (0.81)
BSD (e ← j)	**0.831 (1.00)**	0.830 (0.96)	0.822 (0.96)	0.820 (0.94)	0.817 (0.94)
BSD (j ← e)	**0.773 (0.92)**	0.762 (0.88)	0.756 (0.72)	0.736 (0.72)	0.736 (0.72)

Table 2 Percentage of successful output with corresponding COMET scores upon preliminary filtering (COMET shown first, pass rate in parentheses).

Error Type	en-ja	ja-en
Mistranslation	104 (0.27)	**43 (0.29)**
Omission/Addition	40 (0.10)	37 (0.25)
Terminology & Proper Noun	52 (0.13)	19 (0.13)
Unnatural Style	**179 (0.46)**	39 (0.26)
Ambiguity & Disambiguation	4 (0.01)	4 (0.03)
Buzzwords & Loanwords	4 (0.01)	6 (0.04)
Dialogue Inconsistency	7 (0.02)	0 (0.00)
Total	**390**	**148**

Table 3 Frequency distribution of MQM-Chat error types across different translation directions

ness dialogues did not benefit from additional prompting, with baseline configurations reaching the highest scores. Culture-aware prompting produced measurable improvements only for casual en → ja dialogues. However, combining all prompting components did not yield further gains across datasets, suggesting that prompt over-specification may introduce noise instead.

3.2 Manual Evaluation

To begin with, a substantial proportion of dialogue turns contained errors regardless of their COMET scores. Only around a third of the turns were marked as error-free or acceptable across all settings. Although high-COMET dialogues contained markedly more acceptable turns than low-COMET ones, the proportion of error-free lines in the high-COMET group averaged only about 50% per dialogue, compared to approximately 23% in the low-COMET group. This indicates that while COMET successfully distinguishes severely degraded outputs from more usable translations, even semantically strong outputs exhibit widespread interactional and pragmatic problems.

Across all experimental conditions, Unnatural Style was the most prevalent error category, accounting for approximately 41% of all identified errors, followed by Mistranslation (28%), Omission or Addition (14%), and Terminology or Proper Noun errors (13%). Dialogue-specific categories altogether accounted for less than 5% of the total. Translation direction reveals a clear asymmetry: Unnatural Style errors constituted nearly half of all errors in en → ja settings, whereas ja → en translation exhibited a more balanced distribution across Mistranslation, Omission or Addition, and Unnatural Style. This pattern suggests that generating contextually appropriate translations, especially in Japanese, remains a challenge for the model.

Comparing different prompts, the relative error profile remained highly stable. Although high-COMET outputs contained fewer total errors than low-COMET ones, the proportions of different error types barely changed, indicating that prompting primarily affected error quantity rather than error composition. This stability held across both domains and translation directions. Casual and business dialogues both maintained stable error profiles under different prompts, and the same broad patterns observed in the aggregate analysis persisted within each subset. In other words, while prompting sometimes improved or degraded individual translations, it did not produce a consistent rebalancing of error types toward fewer interactional or cultural problems.

While the error type distributions show prompting primarily affects error quantity rather than composition, examining specific dialogues reveals how prompt variations manifest in practice. Referring to **??**, Example 1 demonstrated how prompt design affected how the model handled deferential expressions. While more affirmative expressions would be preferred in American contexts, the Japanese source encodes deference that should not be entirely discarded. With the full set of contextual information, the model arrived at a translation that struck a good balance of humbleness and directness. For Example 2, it was noticed that the output from the full-stacked prompt is in fact better than the reference, which is rather literal. Notably, this more contextually appropriate output received a lower COMET score than the baseline, suggesting that COMET may not adequately capture improvements in discourse coherence and pragmatic appropriateness.

On the other hand, Example 3 from displayed how providing detailed prompts remedies problematic honorific use and word choice, where the model successfully returned conventional collocations and a register appropriate

for workplace interactions. Yet, such effects of prompting seem to be inconsistent, as terminology and stylistic issues are not solved in other turns of the same dialogue raised previously, as seen in Examples 4 and 5.

4 Discussion

Experimental results indicate that prompt-based interventions produce inconsistent improvements in cultural and interactional quality. While prompting occasionally improved conventional collocations and honorific use, these gains did not generalize across turns or dialogue types. Moreover, both automatic and manual evaluation frameworks demonstrate significant limitations in capturing culturally grounded dialogue quality.

A recurring phenomenon was turn restructuring—splitting long source turns into multiple shorter utterances. This often made dialogues sound more natural and conversational, suggesting turn insertion functions as pragmatic adaptation rather than error. However, it was frequently accompanied by content redistribution, where information from later turns appeared prematurely, creating referential and temporal inconsistencies. This produced a tension between communicative quality and evaluation reliability. Splitting turns often increased perceived naturalness and interactional plausibility, yet both turn insertion and content redistribution caused substantial mismatches during automatic scoring because metrics such as COMET assume one-to-one sentence alignment. Consequently, dialogues that were pragmatically improved by turn restructuring could receive low automatic scores, revealing a fundamental mismatch between conversational naturalness and sentence-based evaluation frameworks.

The error analysis reveals significant limitations in MQM-Chat when applied to Japanese–English dialogue translation. Although MQM-Chat introduces dialogue-specific error categories, only a small proportion of observed errors fell under chat-oriented types such as Dialogue Inconsistency or Ambiguity and Disambiguation. More critically, the vast majority of culturally and interactionally salient problems were collapsed under the broad label of Unnatural Style, despite representing fundamentally different types of issues. For instance, overly literal rendering (for example *Oh cool, do you speak Japanese then?* rendered as おおすごい日本語が話すことができるね rather than the more natural へえ、素敵、じゃあ、日本語が話せるの？) would traditionally fall under locale convention in standard MQM, as it concerns whether linguistic choices align with target-culture norms. Meanwhile, errors in backchanneling— such as collapsing diverse Japanese acknowledgements (はい, ええ, そうですね, なるほど) into repetitive *yea* or *yeah* responses, or rigid mapping of forms like すばらしい or いいですね for *sounds great* or *that's nice*—concern dialogue-specific interactional features like the variety and appropriateness of response tokens that maintain conversational engagement. MQM-Chat could benefit from adding finer-grained subcategories that distinguish locale convention violations from dialogue style deficiencies so that it could better capture the distinct types of pragmatic infelicities observed in culturally distant language pairs, drawing on traditional MQM dimensions to build naturally on the existing dialogue-aware framework.

Future work should incorporate multiple annotators to assess inter-annotator agreement and to better capture subjective judgments of cultural and interactional appropriateness. Exploring stronger models with greater capacity may reveal different patterns in how cultural and pragmatic information is handled in dialogue translation. Additionally, testing alternative prompting strategies—such as few-shot examples with culturally marked utterances or explicit metalinguistic guidance on register and politeness—could provide further insights into whether prompt design can more reliably activate culturally appropriate translation choices across diverse conversational contexts.

5 Conclusion

This study demonstrates that while prompt-based interventions can occasionally improve the cultural and interactional appropriateness of Japanese–English dialogue translation, these improvements remain localized and unpredictable. More critically, existing evaluation frameworks—both COMET and MQM-Chat—prove inadequate for assessing culturally grounded dialogue quality, with COMET penalizing pragmatically appropriate outputs and MQM-Chat collapsing distinct cultural errors under a single category. These findings underscore the need for culturally sensitive evaluation taxonomies and stronger models to advance MT systems capable of producing genuinely appropriate cross-cultural communication.

References

[1] Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. Salute the Classic: Revisiting Challenges of Machine Translation in the Age of Large Language Models. **Transactions of the Association for Computational Linguistics**, Vol. 13, pp. 73–95, 2025.

[2] Duygu Ataman, Alexandra Birch, Nizar Habash, Marcello Federico, Philipp Koehn, and Kyunghyun Cho. Machine Translation in the Era of Large Language Models:A Survey of Historical and Emerging Problems. **Information**, Vol. 16, No. 9, p. 723, 2025.

[3] Anthony Pym. Translation and text transfer. 2010.

[4] Claire Kramsch. Language and culture. **AILA review**, Vol. 27, No. 1, pp. 30–55, 2014.

[5] Juliane House. Translation studies and pragmatics. In **Pragmatics and its interfaces**, pp. 143–162. John Benjamins Publishing Company, 2018.

[6] Yong Cao, Min Chen, and Daniel Hershcovich. Bridging cultural nuances in dialogue agents through cultural value surveys. In Yvette Graham and Matthew Purver, editors, **Findings of the Association for Computational Linguistics: EACL 2024**, pp. 929–945, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[7] Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. Towards making the most of ChatGPT for machine translation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 5622–5633, Singapore, December 2023. Association for Computational Linguistics.

[8] Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. Cultural bias and cultural alignment of large language models. **PNAS nexus**, Vol. 3, No. 9, p. pgae346, 2024.

[9] Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. Investigating cultural alignment of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 12404–12422, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[10] Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. **arXiv preprint arXiv:2302.09210**, 2023.

[11] Sui He. Prompting ChatGPT for translation: A comparative analysis of translation brief and persona prompts. In Carolina Scarton, Charlotte Prescott, Chris Bayliss, Chris Oakley, Joanna Wright, Stuart Wrigley, Xingyi Song, Edward Gow-Smith, Rachel Bawden, Víctor M Sánchez-Cartagena, Patrick Cadwell, Ekaterina Lapshinova-Koltunski, Vera Cabarrão, Konstantinos Chatzitheodorou, Mary Nurminen, Diptesh Kanojia, and Helena Moniz, editors, **Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)**, pp. 316–326, Sheffield, UK, June 2024. European Association for Machine Translation (EAMT).

[12] Pushpdeep Singh, Mayur Patidar, and Lovekesh Vig. Translating across cultures: LLMs for intralingual cultural adaptation. In Libby Barak and Malihe Alikhani, editors, **Proceedings of the 28th Conference on Computational Natural Language Learning**, pp. 400–418, Miami, FL, USA, November 2024. Association for Computational Linguistics.

[13] June Ock Yum. The impact of confucianism on interpersonal relationships and communication patterns in east asia. **Communications Monographs**, Vol. 55, No. 4, pp. 374–388, 1988.

[14] Yunmeng Li, Jun Suzuki, Makoto Morishita, Kaori Abe, Ryoko Tokuhisa, Ana Brassard, and Kentaro Inui. Chat translation error detection for assisting cross-lingual communications. In **Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems**, pp. 88–95, Online, November 2022. Association for Computational Linguistics.

[15] Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. Designing the business conversation corpus. In **Proceedings of the 6th Workshop on Asian Translation**, pp. 54–61, Hong Kong, China, November 2019. Association for Computational Linguistics.

[16] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.

[17] 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. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, **Proceedings of the Seventh Conference on Machine Translation (WMT)**, pp. 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[18] Yunmeng Li, Jun Suzuki, Makoto Morishita, Kaori Abe, and Kentaro Inui. MQM-chat: Multidimensional quality metrics for chat translation. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, **Proceedings of the 31st International Conference on Computational Linguistics**, pp. 3283–3299, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

A Appendix

The examples below are drawn from the BSD corpus and dialogues from which the examples were extracted all belong to the high-COMET subset.

Example 1		COMET
src	はい、そのように思っております。	
ref	Yes I believe so.	
B	Yes, I do.	0.830
D+F	Yes, I do.	0.830
D+C	Yes, I do.	0.830
D+F+C	Yes, I believe I do.	0.921
Example 2 (previous turn: 日本で産まれましたが、父の仕事のため海外に住んでいました。)		
src	へぇ、どちらへ行かれたんですか？	
ref	Oh, where did you go?	
B	Wow, where did you go?	0.948
D+F	Wow, where did you go?	0.948
D+C	Wow, where did you go?	0.948
D+F+C	Wow, where did you move to?	0.847
Example 3		
src	Yes. Also, I'm planning to go buy paper plates and paper cups.	
ref	はい。あとは紙皿とか、紙コップとかを買いに行く予定です。	
D	そうだ。紙の皿や紙のカップも買いに行く予定だ。	0.892
D+F	そうです。さらに、紙の皿や紙のカップを買いに行きます。	0.877
D+F+C	そうですね。紙皿や紙カップを買いに行きます。	0.877
Example 4		
src	では、手短に自己紹介をして頂けますか？	
ref	So can you do a quick introduction of yourself?	
B	Can you make a brief self-introduction, please?	0.919
D+(F|C)	Could you please make a brief introduction?	0.919
D+F+C	Let's keep this brief. Can you give me a quick introduction?	0.846
Example 5		
src	ええ、御社は国際貿易に関わっておられますので、…	
ref	Well, your company is involved in international trade, …	
B	Yes, I thought that since your company is involved in international trade, …	0.910
D	Well, our company is involved in international trade, …	0.909
D+C	Well, our company is involved in international trade, …	0.915
D+F+C	Well, our company is involved in international trade, …	0.914