MQM-Chat: Multidimensional Quality Metrics for Chat Translation

Anonymous ACL submission

Abstract

The complexities of chats pose significant challenges for machine translation models. Recognizing the need for a precise evaluation metric to address the issues of chat translation, this study introduces Multidimensional Quality Metrics for Chat Translation (MQM-Chat). Through the experiments of five models using MQM-Chat, we observed that all models generated certain fundamental errors, while each of them has different shortcomings, such as omission, overly correcting ambiguous source content, and buzzword issues, resulting in the loss of stylized information. Our findings underscore the effectiveness of MQM-Chat in evaluating chat translation, emphasizing the importance of stylized content and dialogue consistency for future studies.

1 Introduction

Neural machine translation (NMT) has experienced significant development in recent years (Bahdanau et al., 2014), leading to notable improvements in the performance of machine translation systems, especially in translating formatted documents such as news, academic papers, and else (Maruf and Haffari, 2018; Barrault et al., 2019, 2020; Nakazawa et al., 2019; Ma et al., 2020). However, despite the success of translating documents, current methods still face substantial challenges when translating chats (Tiedemann and Scherrer, 2017; Maruf et al., 2018; Farajian et al., 2020) due to their higher degrees of ambiguity and speaker-specific stylized contents, including sentiments, personalities, and cultural nuances (Baldwin et al., 2013; Eisenstein, 2013; Uthus and Aha, 2013; Xu et al., 2014; Läubli et al., 2018; Toral et al., 2018; Farajian et al., 2020).

To enhance chat translation, it is important to thoroughly understand the qualities and limitations of existing translation models in handling chats. Traditional automatic evaluation metrics such as BLEU (Papineni et al., 2002) and COMET (Rei et al., 2022a,b) focus primarily on accuracy but fail to capture the nuances of chats. Thus, a refined error categorization framework assessing semantic accuracy while preserving the speaker's stylized nuances is better suited for identifying the specific problems of chat translation (Gehman et al., 2020).

To address this need, we propose the Multidimensional Quality Metrics for Chat Translation (MQM-Chat) in this research. Based on the existing Multidimensional Quality Metrics (MQM) framework[1] (Burchardt, 2013; Mariana, 2014), MQM-Chat encloses seven error types: mistranslation, omission or addition, terminology or proper noun issue, unnatural style, ambiguity and disambiguation, buzzword or loanword issue, and dialogue inconsistency, where the latter three are designed specifically for chats. We applied MQM-Chat to evaluate the chat translation abilities of five models in the experiments: the large language models (LLMs) GPT-4 (Achiam et al., 2023) and LLaMA3 (Touvron et al., 2024), the commercialized model DeepL[2], the bilingual model produced by team SKIM in WMT23 (Kudo et al., 2023), and the multilingual model produced by Facebook in WMT21 (Tran et al., 2021). The experiments, held in Chinese⇒English and Japanese⇒English, included short but noisy chats to ensure the buzzwords and ambiguous content, and longer but cleaner chats to provide a comparison. Proficient bilingual annotators were invited to label the translations using the error types and severity levels from MQM-Chat.

The Overall Quality Score (OQS) calculations indicate that GPT-4 outperformed the other models. On the other hand, the severity percentage of each error type shows that there are usually more severe mistranslations, buzzword or loanword issues, and dialogue inconsistency errors in chat translations.

[1] https://themqm.org/

[2] https://www.deepl.com/translator

Without the severity penalties, the number of errors shows that all five models exhibited common tendencies toward mistranslation errors. Buzzword issues and dialogue inconsistency are considerably important, especially in short chats. MQM-Chat helped to qualify the strengths and weaknesses of the five models, emphasizing the significance of preserving the stylized contents in chats.

In summary, this research contributes to chat translation with a novel evaluation metric designed to assess the quality of chat translations, MQM-Chat. Five state-of-the-art translation models were evaluated with MQM-Chat in handling chat content. The experiments also helped to build annotated Chinese⇒English and Japanese⇒English chat translation data. These contributions enhance the understanding of chat translation, providing valuable resources for further advancements.

2 Related Work

Chat Translation Tasks While formal documents follow standardized structures, chats often include slang, idiomatic expressions, and personalized styles, adding complexity to translation (Baldwin et al., 2013; Eisenstein, 2013; Xu et al., 2014). High accuracy in translating chats is important, but preserving speaker-specific content, like buzzwords and speaking style, is sometimes even more crucial (Hovy, 2015; Salganik, 2020).

The first workshop specifically focused on chat translation was WMT2020 (Barrault et al., 2020; Farajian et al., 2020), which laid the groundwork for further exploration in this domain. It was followed by WMT2022 (Kocmi et al., 2022; Farinha et al., 2022) and continues by WMT2024. WMT shared tasks have primarily concentrated on customer service chats, which are relatively structured and standardized. The emphasis has been on evaluating the overall performance of chat translation models with a strong focus on syntax accuracy. WMT2022 shared task started to address chat-specific issues, while Liang's team, as a continuation of WMT2020, improved models for chat translation, highlighting the importance of coherence, fluency, and speaker personalities (Liang et al., 2021a,b, 2022).

WMT and derivative studies have gradually recognized the importance of source content issues and preserving the speaker's style in chat translations. MQM was adapted in the WMT2022 shared task, but it remained too broad with 31 error types, most of which were about accuracies, and relatively superficial analyses. Liang's studies focused on personality and sentiment but did not consider source issues. Previous research has typically utilized binary classification for chat translation evaluation, focusing on coherence (Li et al., 2022, 2023), which did not capture the complexity of chat translations either.

With the foundations, we have refined the evaluation by differentiating the source issues within chat translations into ambiguity issues and cultural nuances issues such as buzzwords, and emphasizing the importance of dialogue consistency. Additionally, we de-emphasized grammatical accuracy, as it is not always the highest priority in everyday conversations. To make MQM-Chat broadly applicable to general chats, we chose data covering a wide range of topics, including news, sports, hobbies, daily life, social media, and others. Additionally, we included Japanese data, a language not extensively studied in chat translation tasks. The comparison between our research and previous studies is shown in Table 1.

Translating with LLMs Several studies have demonstrated that GPT performs well in translation tasks (Hendy et al., 2023; Zhang et al., 2023), particularly in scenarios involving literary translation (Thai et al., 2022; Karpinska and Iyyer, 2023). These studies suggest that LLM translations might be favored over traditional neural machine translation (NMT) models when the input domain is likely to contain noisy, ill-formed sentences. Despite these promising findings, no dedicated research specifically addresses chat translation using LLMs. This gap highlights the need for focused studies on applying LLMs to the unique challenges of chat translation.

3 Multidimensional Quality Metrics for Chat Translation (MQM-Chat)

In this research, we define high-quality chat translation as maintaining accuracy while capturing and conveying the speaker's personality, styles, and cultural nuances. We refined the Multidimensional Quality Metrics (MQM) framework and introduced customized categories for chat translations. MQM-Chat focuses on seven error types: mistranslation, omission or addition, terminology or proper noun issues, unnatural style, ambiguity and disambiguation, buzzwords or loanwords issues, and dialogue inconsistency. The latter three (*) are customized

	Chat Domain	**Human Evaluation Method**	**Evaluation Focus**	**Fine-grained Analysis**	**Language Pairs**
WMT 2020 Chat Translation	Custom Service	Segment Rating + Document Context	Pronoun (*it*).	△	en⇔de
WMT 2022 Chat Translation	Custom Service	Adapted MQM*	Accuracy, Linguistic Conventions, Terminology, ... MT Hallucination, Source Issue.	△	en⇔de, en⇔fr, en⇔pt_br
CPCC	Custom Service, TV series	Customized	Preference, Coherence, Consistency, Fluency.	○	en⇔de, en⇔zh
CSA-NCT	Custom Service, TV series	Customized	Coherence, Speaker, Fluency.	○	en⇔de, en⇔zh
SML	Custom Service, TV series	Question-based	Coherence, Fluency.	○	en⇔de, en⇔zh
MQM-Chat Annotation	**Various** (news, sports, hobbies, daily life, social media, etc.)	MQM-Chat	Source Issue→**Disambiguation**, Consistency→**Dialogue Consistency**, Speaker→**Stylized Contents**, **Cultural Contents**, **Buzzwords and Loanwords**.	○	zh⇒en **ja**⇒en

Table 1: Comparison of our research with previous studies across several dimensions: data domain, human evaluation method, evaluation focus, granularity of results, and language pairs studied. WMT2020 and WMT2022 analyses are considered less detailed due to the Segment Rating + Document Context method and lack of fine-grained explanations on terminal nodes.

typologies tailored for chat translation. These error types are evaluated with three severity levels for a detailed and accurate assessment.

3.1 Error Types

Mistranslation Mistranslation denotes fundamental inaccuracies in the translation process, including untranslated source segments, incorrect lexical choice or grammar that distorts the meaning, under-translation, and over-translation. These errors are critical as they directly impact the comprehensibility and accuracy of the translation.

Omission or Addition Missing source contents (omission) or additional content not present in the source (addition) are Omission or Addition errors. Such errors can significantly mistake the intended message and disrupt the coherence of the translated text, leading to potential misunderstandings.

Terminology and Proper Noun Issues Terminology and Proper Noun Issues pertain to inaccuracies in translating specialized vocabulary, inherent terms, and proper nouns from the source text. Misinterpretations in this category can undermine the reliability of the translation, especially in professional and academic contexts.

Unnatural Style Unnatural Style refers to grammatically correct translations that are not natural in the target language. These errors affect the readability and acceptability of the translation, making it appear awkward or stilted to native speakers.

Ambiguaty and Disambiguation* Ambiguity and Disambiguation errors occur when the ambiguities or errors in the source text, such as typographical errors, omissions, unclear abbreviations, and erroneous punctuation, are not faithfully reflected in the translation. Deviations from this principle are considered errors, highlighting the need to accurately translate the speaker-specific stylized content into corresponding errors in the target language. This category emphasizes the importance of maintaining the authenticity of the source text, including its imperfections. Examples are shown in Table 2.

Buzzword or Loanword Issues* Buzzword or Loanword Issues arise when such terms are not translated accurately according to their usage in the source and target languages. This includes the incorrect translation of popular sayings, newly created words, internet slang, and memes. If there is no corresponding term in the target language, the original pronunciation should be retained and written in the target language. Failure to do so results in error translations that obscure the source text's intended meaning and cultural nuance. Examples are shown in Table 2.

Dialogue Inconsistency* Dialogue Inconsistency occurs when translations fail to maintain consistency based on context, particularly when the

Ambiguity and Disambiguation

Source (zh, ja)	Possible Good Translation (en)	Bad Translation (en)
队啊！你应该试试！	Yaas! You should try!	Team ah! You should try!
知ってｒ？昨日、ヘレンとあったよ！	u know waht, I saw Helen yesterday!	You know what, I saw Helen yesterday!

Buzzword or Loanword Issues

Source (zh, ja)	Possible Good Translation (en)	Bad Translation (en)
鼠的，真的累死了	Yaap, I’m really tired	Damn it, I’m really exhausted
草wwwwww	lol	grass

Table 2: Examples of ambiguity and disambiguation errors, buzzword or loanword issues. Translations in blue are possibly expected, and translations in red are bad.

speakers change within the chat. This includes inappropriate handling of demonstrative pronouns, personal references, or definite articles. Maintaining consistency in dialogue is crucial to ensure coherence and avoid confusing the reader.

3.2 Error Severity Levels

We provided three levels of severity for each error to evaluate the translations comprehensively: **major** for errors that significantly impact the understandability of the content; **minor** for errors that do not impact the overall understandability but detract from the quality; **neutral** for errors requiring additional revision but do not pose significant risks. Severity penalty multipliers are 5 for major, 1 for minor, and 0 for neutral.

4 Experiments

We conducted experiments to evaluate the effectiveness of MQM-Chat by translating chats from Japanese (ja) and Chinese (zh) into English (en) and having proficient bilingual annotators evaluate the translations using MQM-Chat.

4.1 Datasets

We selected 200 chat data from the Open 2ch Dialogue Corpus (Inaba, 2019) to be the short but noisy data for the ja⇒en translations, which features ambiguous content and popular sayings from Japan’s well-known online community 2channel. Similarly, we chose 200 data from the LCCC-base dataset (Wang et al., 2020) for the zh⇒en translations. To provide a comparison and a broader range of contents, we included 100 longer and cleaner chat data from BPersona-chat (Sugiyama et al., 2021; Li et al., 2022) for ja⇒en, and 100 from the NaturalConv (Wang et al., 2021) for zh⇒en. Statistics of selected chats are listed in Table 6 of Appendix C.

4.2 Translation Models

We employed four models for each language pair: GPT-4, LLaMA3 (70B-Instruct), DeepL and Facebook@WMT21 for zh⇒en; GPT-4, LLaMA3, DeepL and SKIM@WMT23 for ja⇒en. The models represented diverse approaches, including sentence-to-sentence transformers-based models (Vaswani et al., 2017), large language models (LLMs), and commercialized systems. GPT-4 and LLaMA3 were used in zero-shot learning configurations (Romera-Paredes and Torr, 2015; Wang et al., 2019) with prompts designed on methodologies proposed by Hendy et al. (2023) and recent studies (Farinhas et al., 2023; Peng et al., 2023)[3].

4.3 Crowdsourcing and Annotating Tasks

We recruited six professional annotators proficient in Japanese and English and six in Chinese and English through crowdsourcing. Annotators identified translation errors and assigned severity levels based on MQM-Chat specifications. We chose Label Studio[4] (Tkachenko et al., 2020-2022) as the online annotation tool due to its user-friendly interface and robust functionality. Annotators were provided with detailed guidelines to ensure error labeling and severity assessment consistency. Details of the annotation tasks could be found in Appendix B.

4.4 Overall Quality Scores

$$OQS = \left(1 - \frac{APT}{EWC}\right) \times 100 \qquad (1)$$

As shown in Equation 1, we calculated the Overall Quality Scores (OQS) by the Evaluation Word Count (EWC) and the Absolute Penalty Total (APT) to provide a quantifiable measure of the translation quality and a comprehensive evaluation

[3] Prompts and parameters of the models are in Appendix A.

[4] https://labelstud.io/

Chinese→English

		Short	**Long**
GPT-4	OQS	88.99	96.66
	Error Counts	246	434
LLaMA3	OQS	79.70	96.54
	Error Counts	416	345
DeepL	OQS	77.08	91.18
	Error Counts	460	756
Facebook	OQS	55.93	89.50
	Error Counts	658	851

Japanese→English

		Short	**Long**
GPT-4	OQS	86.36	94.17
	Error Counts	495	807
LLaMA3	OQS	58.71	85.44
	Error Counts	994	940
DeepL	OQS	76.83	89.24
	Error Counts	761	1030
SKIM	OQS	49.80	73.75
	Error Counts	1097	1365

Table 3: The overall quality score (OQS) and number of errors (error counts) of translation models for different datasets and language pairs.

of different models, highlighting their strengths and weaknesses in chat translations.

5 Results and Analysis

5.1 Overall Performance

We calculated the average OQS (eq. 1) and counted the total number of errors, as shown in Table 3. OQS and error counts suggest that models perform better when translating longer chats than shorter ones since selected long chats have fewer buzzwords and ambiguities, making the translation task closer to traditional document translation. The results demonstrate that zh⇒en translations have higher overall quality and considerably fewer errors than ja⇒en. GPT-4 surpassed all other models, while the NMT models performed the worst in their respective languages. LLaMA3 is slightly better than DeepL when translating Chinese but has significantly lower scores than DeepL when translating Japanese, especially in short chats. Possible reasons could be the lack of Japanese training data and language transfer capabilities.

5.2 Severity Analysis

To investigate the severity distribution for each error type, we analyzed the number of errors at each severity level across different models, data types, and languages. The results are presented as heatmaps in Figure 1.

Accuracy The results show that there are usually more severe mistranslations, omissions, and additions. Mistranslations in the translations of ja⇒en long chats tend to be minor, while zh⇒en translations have fewer omission or addition errors than ja⇒en, likely due to the omission of subjects, objects, and sub-sentences in the Japanese language. Translations of terminologies and proper nouns also show major to minor issues, indicating the need for better translation of proper nouns and specialized terminology for general chats crossing various topics.

Stylized Nuances Errors categorized as unnatural style are primarily neutral to minor, related to the definitions where errors are grammatically correct but not natural. Ambiguous content issues are usually average in zh⇒en translations but tend to be more neutral in ja⇒en, suggesting that disambiguation's significance may differ according to different languages. Notably, translations of buzzwords or loanwords consistently contain a high proportion of major errors in all cases, highlighting the critical challenge in chat translation. For Japanese short data, buzzwords are either major issues or neutral issues.

Dialogue inconsistency errors are usually major or minor errors, indicating that sentence reference within dialogues remains a significant issue. However, this problem is less generated in the translations of Chinese long chats, possibly due to the well-generated data in the NaturalConv dataset.

The heatmaps also point out that the overall performance of ja⇒en is worse than zh⇒en, especially for short and noisy chats. The distribution shows that not all error types mainly contain major errors, with many neutral errors present.

5.3 Error Counts Analysis

Since neutral errors are not calculated for the OQS, a detailed analysis is required based on the number of errors. In this section, we counted the number of errors for each type without severity levels to comprehensively understand each model's strengths and weaknesses, shown in Figure 2.

Similarly to the analysis in previous sections, all models generated the most mistranslations, with ja→en translations performing worse than zh→en. GPT-4 consistently generated the fewest omissions and additions and was good at resolving dialogue consistency issues, terminology, proper nouns, buzzwords, and loanwords, relating to its extensive

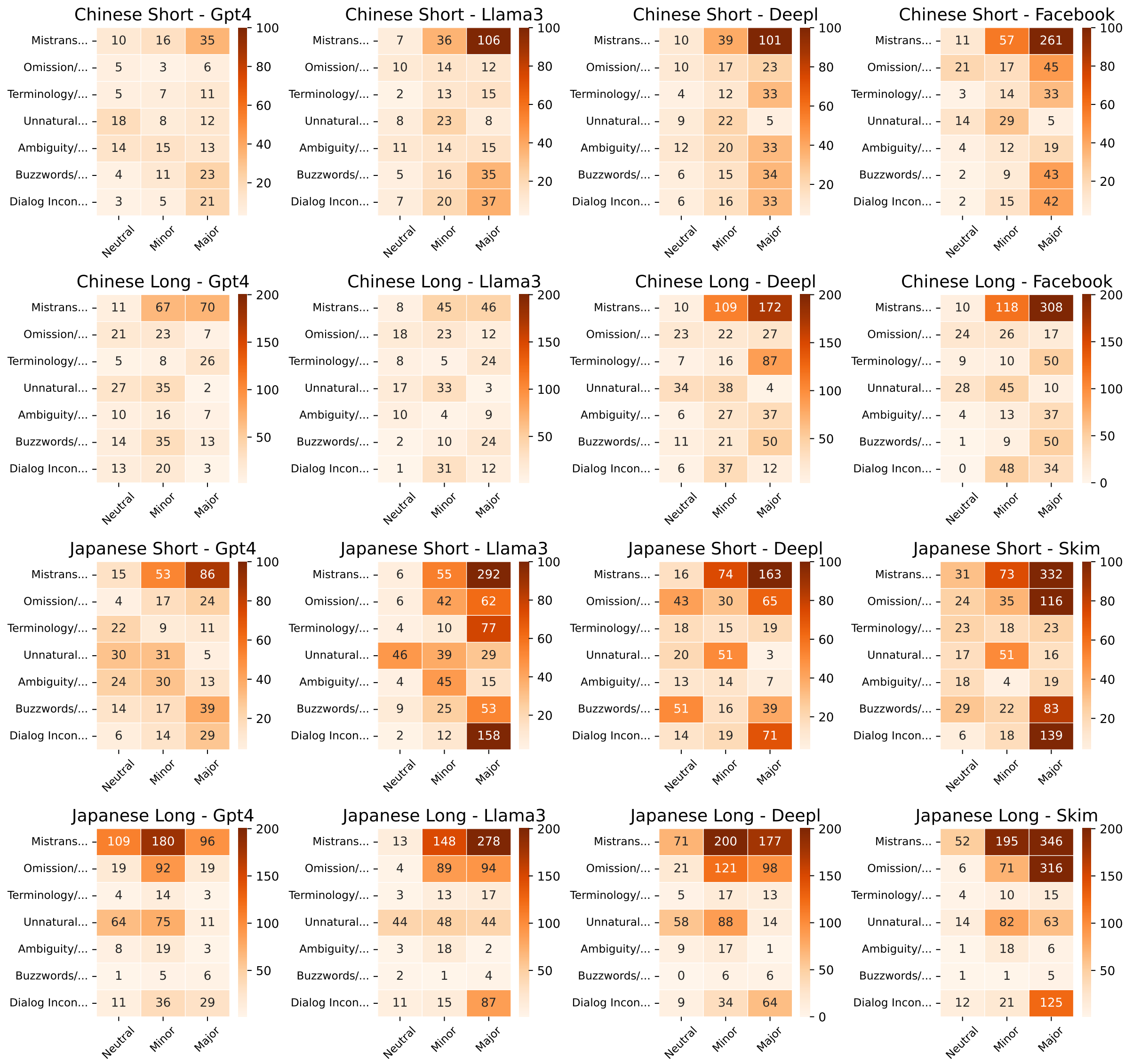

Figure 1: The heatmap showing distributions of major, minor, and neutral errors for different error types, across various language pairs and data types.

training data and contextual learning capabilities.

LLaMA3 produced fewer errors in Chinese long chats than GPT-4, but its OQS indicates that its translations contained fewer but more severe errors. LLaMA3 and DeepL performed similarly, with DeepL exhibiting more disambiguation and terminology errors. In zh→en translations, DeepL showed more buzzword and ambiguity issues, whereas, in ja→en translations, LLaMA3 struggled with terminology, proper nouns, dialogue consistency, and natural style.

The Facebook model had significantly more mistranslations, omissions, and additions than others. Similarly, the SKIM model exhibited the fewest ambiguity and buzzword errors but had the highest mistranslation and omission errors, suggesting these contents may be mistranslated or omitted, not having a chance to be considered as other errors.

In conclusion, LLMs performed better but struggled with ambiguous source contents, especially in short and noisy chats. The analysis underscores GPT-4's strengths in handling various error types across contexts and emphasizes the need to improve traditional NMT models.

5.4 Tone Words

Based on feedback from annotators, we identified models' behavior where declarative or exclama-

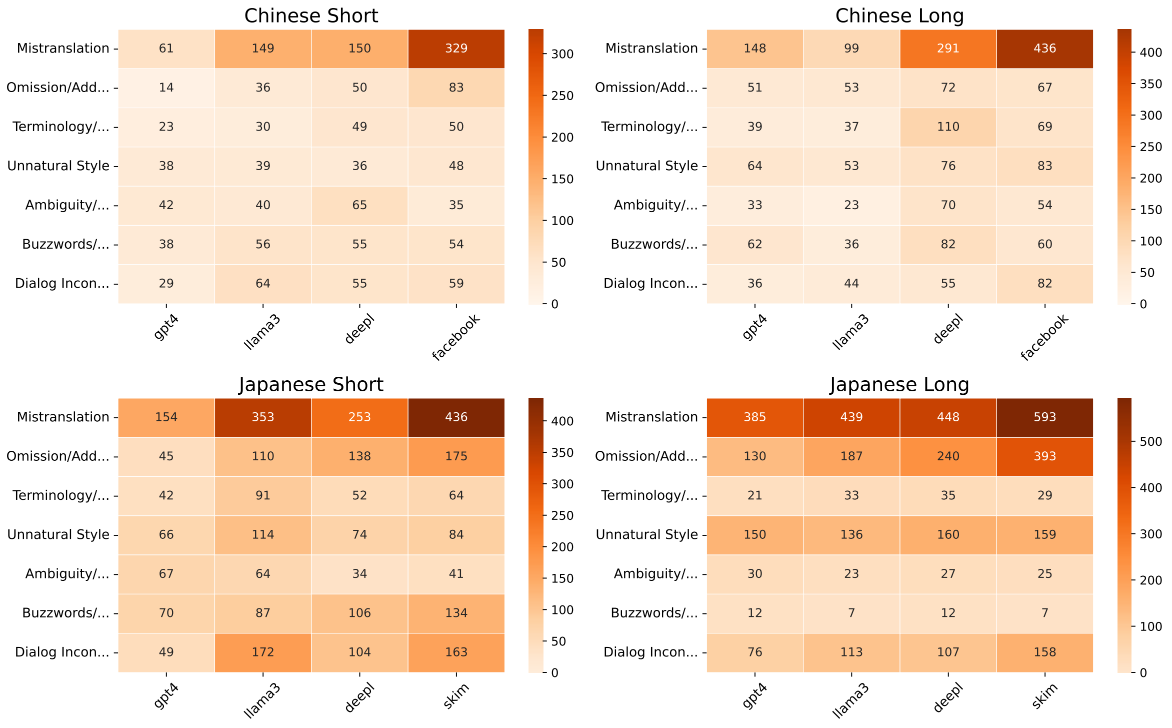

Figure 2: The heatmap of the number of errors for different error types, across various language pairs and types of data. The darker the color, the higher the number of errors.

Tone Word Count				
	GPT-4	**LLaMA3**	**DeepL**	**NMT**
zh→en	6	7	9	3
ja→en	13	61	13	6
Explanation Count				
	GPT-4	**LLaMA3**	**DeepL**	**NMT**
zh→en	2	2	0	0
ja→en	5	5	1	1

Table 4: The number of tone words and additional explanations of models for different language pairs.

tory sentences were translated as interrogative sentences. This typically involved adding interrogative questions like “right?” or “isn’t?” at the end of translations. The results of the counted occurrence of this error are presented in Table 4, suggesting that LLaMA3 exhibits a significantly higher frequency of this behavior when translating short Japanese chats than any other model. We consider this largely due to the lack of sufficient Japanese data, which likely impairs LLaMA3’s ability to accurately comprehend and represent Japanese expressions and sentence structures, leading to this misinterpretation. Examples are shown in Appendix D.

5.5 Additional Explanations

We observed that when the model translates culturally specific terms from the source language, it occasionally adds corresponding explanations in parentheses to aid understanding. For example, when translating “Yu E Bao” , the translation included “savings” in parentheses to clarify the term, as it is a unique saving method currently prevalent in China. We have also quantified the occurrence of these additional explanations. Results are illustrated in Table 4, with examples in Appendix D.

According to the results, although this additional explanatory behavior is present to some extent, it is not overly prevalent overall. DeepL and SKIM added the same explanation for “Shogi” as “Japanese chess” in long Japanese chats, which may be related to the training data used by DeepL and SKIM. On the other hand, the additional explanations provided by LLMs varied significantly. It is important to note that we did not explicitly instruct the LLMs to include such explanations when prompting them. We believe that the additional explanatory behavior of LLMs stems from their contextual learning abilities and the extensive training data they have been exposed to.

	GPT-4	LlaMA3	DeepL	NMT
ja→en short	2	6	39	17

Table 5: The number of translations of buzzwords or loanwords that are omitted during the Japanese to English translation process for short chat data.

5.6 Lost Buzzwords

The Japanese Open2ch data used in this study contains considerably more buzzwords than other datasets, many of which are not Japanese characters but emoticons, emojis, or the "w" character used in Japanese internet culture to denote laughter ("lol"). For the Japanese short chats, we specifically quantified the number of buzzwords lost in translation, as illustrated in Table 5, with examples in Appendix D.

The figure shows that DeepL lost the most buzzwords during translation compared to other models. We hypothesize that this phenomenon may be due to DeepL's translation process, which potentially omits non-source language characters after identifying the source text language. Meanwhile, the traditional NMT model produced by SKIM also ignored buzzwords in its translating process. Noticing that this model was training for news translations, we consider that it prefers to read formatted contexts that do not contain non-Japanese words. The Chinese short chats from the LCCC-base do not contain many buzzwords in non-Chinese characters; further experiments are needed to confirm whether DeepL and traditional NMT models systematically filter out non-Chinese characters.

5.7 Discussions

In conclusion, models perform worse on Japanese data compared to Chinese data. GPT-4 demonstrates the best performance among all the models. Conversely, the traditional NMT models, SKIM and Facebook, exhibit the worst performance, which is expected because the NMT models are not specifically trained for chat translation. DeepL's performance falls in the mid-range. Meanwhile, LLaMA3 shows varied performance across different languages. For zh⇒en translations, LLaMA3 performs slightly better than DeepL but worse than GPT-4; however, for ja⇒en, LLaMA3 performs worse than GPT-4 and DeepL. All models generate the most mistranslations, but their strengths vary depending on different experimental settings. The refined error types from MQM-Chat—ambiguity issues, buzzword problems, and dialogue inconsistency—provided deeper insights into the shortcomings of chat translation.

The differing models' performances on these aspects suggest potential solutions for further chat translation tasks. For instance, GPT-4's success in handling terminologies, proper nouns, buzzwords, and loanwords indicates that training with more diverse and real-life conversational data and translating with the support of common knowledge may improve the performance of chat translation. Its good performance of resolving dialogue consistency and including buzzwords indicates the importance of understanding the source contents. Using prompts may help LLMs improve the stylized content issues in the future; from this point of view, LLMs may be better suited for chat translation at this moment than existing NMT models. These insights guide future improvements in chat translation, aiming to develop models that better capture the intricacies of everyday conversations.

6 Conclusion

This research evaluated chat translation models using the Multidimensional Quality Metrics for Chat Translation (MQM-Chat). The zh⇒en and ja⇒en experiments on GPT-4, LLaMA3, DeepL, SKIM from WMT23, and Facebook from WMT21 showed that GPT-4 consistently outperforms other models, particularly in handling dialogue inconsistencies and managing buzzwords or loanwords. Traditional NMT models SKIM and Facebook performed the worst, while DeepL performed intermediately. LLaMA3 performed well for zh⇒en but struggled with ja⇒en translations.The severity of errors varies in languages and data types. LLMs sometimes added explanations for culturally specific terms, reflecting their contextual learning abilities, while DeepL and NMT models ignore buzzwords when translating Japanese short chats.

Our findings highlight the need for tailored training for chat translation models, especially in handling culturally specific content and maintaining dialogue consistency with the usage of MQM-Chat. Overall, this study provides valuable insights into the capabilities and limitations of current chat translation models, laying a foundation for future research and development in this field.

Limitations

With data limited to translations from Chinese and Japanese to English, the result of our experiments is relatively narrow. Future research may extend the MQM-Chat evaluation to more language pairs and bidirectional translations to better understand chat translation across different languages. The high frequency of mistranslation errors in our results indicates that this error type needs further refinement. We plan to conduct more detailed reviews of the annotations to identify if additional nodes of mistranslation are necessary.

In summary, MQM-Chat has laid a solid foundation for this type of research, opening up many possible directions for improving and expanding chat translation evaluation.

Ethical Considerations

The crowdsourcing experiments employed in this study adhere to stringent ethical guidelines to ensure participant privacy and data protection. The experiments deliberately avoid collecting any personally identifiable information from the participants. No restrictions or enforcement of work hours were imposed upon participants, thereby eliminating undue influence or coercion. Given the absence of personal data collection and voluntary participation, the data is not subject to ethics review at the organization. Consequently, the data collection procedures adhere to the ethical standards and regulations governing research practices.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Timothy Baldwin, Marie-Catherine De Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2013. Noisy user-generated text: Impact on natural language processing. In *Proceedings of the 25th International Conference on Computational Linguistics: Tutorials*, pages 5–11.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Aljoscha Burchardt. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Jacob Eisenstein. 2013. The birth of sociolinguistic expectations: Optimality-theoretic approaches to production and perception. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 572–582.

M. Amin Farajian, António V. Lopes, André F. T. Martins, Sameen Maruf, and Gholamreza Haffari. 2020. Findings of the WMT 2020 shared task on chat translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 65–75, Online. Association for Computational Linguistics.

Ana C Farinha, M. Amin Farajian, Marianna Buchicchio, Patrick Fernandes, José G. C. de Souza, Helena Moniz, and André F. T. Martins. 2022. Findings of the WMT 2022 shared task on chat translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 724–743, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

António Farinhas, José de Souza, and Andre Martins. 2023. An empirical study of translation hypothesis ensembling with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11956–11970, Singapore. Association for Computational Linguistics.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762.

Michimasa Inaba. 2019. A example based dialogue system using the open 2channel dialogue corpus. In *Proceedings of SIG-SLUD-B902-33*, pages 129–132.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Keito Kudo, Takumi Ito, Makoto Morishita, and Jun Suzuki. 2023. SKIM at WMT 2023 general translation task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 128–136, Singapore. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Yunmeng Li, Jun Suzuki, Makoto Morishita, Kaori Abe, and Kentaro Inui. 2023. An investigation of warning erroneous chat translations in cross-lingual communication. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 10–16, Nusa Dua, Bali. Association for Computational Linguistics.

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

Yunlong Liang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021a. Modeling bilingual conversational characteristics for neural chat translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5711–5724, Online. Association for Computational Linguistics.

Yunlong Liang, Fandong Meng, Jinan Xu, Yufeng Chen, and Jie Zhou. 2022. Scheduled multi-task learning for neural chat translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4375–4388, Dublin, Ireland. Association for Computational Linguistics.

Yunlong Liang, Chulun Zhou, Fandong Meng, Jinan Xu, Yufeng Chen, Jinsong Su, and Jie Zhou. 2021b. Towards making the most of dialogue characteristics for neural chat translation. In *Proceedings of* *the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 67–79, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3505–3511, Online. Association for Computational Linguistics.

Valerie R Mariana. 2014. *The Multidimensional Quality Metric (MQM) framework: A new framework for translation quality assessment*. Brigham Young University.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284, Melbourne, Australia. Association for Computational Linguistics.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2018. Contextual neural model for translating bilingual multi-speaker conversations. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 101–112, Brussels, Belgium. Association for Computational Linguistics.

Toshiaki Nakazawa, Nobushige Doi, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Yusuke Oda, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2019. Overview of the 6th workshop on Asian translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 1–35, Hong Kong, China. Association for Computational Linguistics.

Kishore Papineni, 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of ChatGPT for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633, Singapore. Association for Computational Linguistics.

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. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Bernardino Romera-Paredes and Philip Torr. 2015. An embarrassingly simple approach to zero-shot learning. In *International conference on machine learning*, pages 2152–2161. PMLR.

Matthew J Salganik. 2020. *Bit by Bit: Social Research in the Digital Age*. Princeton University Press.

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2021. Empirical analysis of training strategies of transformer-based japanese chit-chat systems.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2024. Llama: Open and efficient foundation language models. `https://ai.facebook.com/blog/large-language-models-llama-3`.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook AI's WMT21 news translation task submission. In *Proceedings of the Sixth Conference on Machine*

Translation, pages 205–215, Online. Association for Computational Linguistics.

David C Uthus and David W Aha. 2013. Multiparticipant chat analysis: A survey. *Artificial Intelligence*, 199:106–121.

Ashish Vaswani, 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.

Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. 2019. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37.

Xiaoyang Wang, Chen Li, Jianqiao Zhao, and Dong Yu. 2021. Naturalconv: A chinese dialogue dataset towards multi-turn topic-driven conversation.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *NLPCC*.

Weidi Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2014. An empirical study on generalization and robustness of deep convolutional neural networks. In *arXiv preprint arXiv:1411.1924*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study.

A Machine Translation Parameters

A.1 GPT-4 and LLaMA3

The prompts used in GPT-4 and LLaMA3's requests were structured as follows:

> You are a professional Chinese to English translator. This is a Chinese to English chat translation task. Please translate each line of the chat from Chinese into English. Each line of the chat is considered a message sent by a different speaker.

Notice that the source language would change to Japanese in the Japanese to English translation requests. Models were set to `gpt-4` and `Meta-Llama-3-70B-Instruct`, respectively. Other parameters were set to `temperature=1`, `top_p=1.0`, `max_token=500`, and defaults.

A.2 Facebook@WMT21

The neutral machine translation model used for Chinese to English translations was the multilingual model submitted to WMT 2023 by Facebook (Tran et al., 2021; Akhbardeh et al., 2021). The model can directly translate text from 7 languages: Hausa (ha), Icelandic (is), Japanese (ja), Czech (cs), Russian (ru), Chinese (zh), German (de) to English. For Chinese to English, it was trained on 166M bitext data from the WMT shared task, and 123M monolingual data from Commoncrawl[5] which are news-domain. The model consists of a 24-layer encoder/decoder with an embedding size of 2,048 and a feedforward size of 16,384 and 32 attention heads, resulting in 4.7B total parameters. Trainings were taken on 32 Volta 32GB GPUs. Fore more details, please refer to the original paper (Tran et al., 2021).

A.3 SKIM@WMT23

We used a neural machine translation system submitted to WMT 2023 by team SKIM (Kudo et al., 2023), who achieved the best accuracy among the participants in WMT23 (Kocmi et al., 2023). The model was trained on publicly available Japanese-English parallel data of around 31M sentences and a synthetic parallel corpus of 681M sentences. The model consists of a 9-layer encoder/decoder with an embedding size of 1,024 and a feedforward size of 8,192, and 16 attention heads, resulting in 547M total parameters. Training took around four days with eight NVIDIA RTX A6000 GPUs. For more details of training settings, please refer to the original paper (Kudo et al., 2023).

B Crowdsourcing Annotation Tasks

Crowdsourcing Annotators Considering that chat translation requires not only proficiency in two languages but also an understanding of the source text, we called for native Chinese or Japanese speakers who are fluent in English to be the annotators through crowdsourcing platforms. We prepared qualifications for the candidates to determine their suitability for the task, which consisted of five short chats and three long chats. Participants who showed a better understanding of both the source and target languages were considered to meet our expectations better and were selected as annotators. All annotators are aware that their annotations will be used for academic research, not commercial.

Annotating Instructions Annotators were provided with detailed instructions in English, Chinese, and Japanese. The instructions include the labeling descriptions with Label Studio and the definitions of error types and severities. Each error type and severity level was provided with 1-5 detailed examples to help annotators understand. Annotators are instructed and required to report offensive data when the source contains extremely offensive content as well. The reported data are removed to avoid having toxic contents in the annotated dataset.

Annotating Payments We paid each annotator an extra 30-35 USD to familiarize them with the instructions and operations. Being familiar with the instruction and operation of Label Studio, the annotator took about 3-5 minutes to complete one short chat and about 5-8 minutes for a long chat. Depending on the length of data, each annotator was paid about 0.5-1.5 USD per short chat and 0.7-2 USD per long chat, with the final payment fluctuating according to the exchange rate. In conclusion, every annotator was paid around 18-22 USD per hour.

C Datasets

Translation Data The statistical information of the selected monolingual data and their translations are shown in Table 6. The NLTK package (Bird

[5] http://data.statmt.org/cc-100/

	LCCC-base	NaturalConv	Open2ch Dialogue	BPersona-chat
Source Language	Chinese	Chinese	Japanese	Japanese
# of chats selected	200	100	200	100
Avg.turns	5	21	5	12
Avg. # of char (src)	52	423	101	490
Avg. # of words (GPT-4)	38	248	52	214
Avg. # of words (LLaMA3)	37	235	50	203
Avg. # of words (DeepL)	36	247	49	218
Avg. # of words (NMT)	37	272	45	182

Table 6: The number of average turns, average source words, average worse in translations of the selected data.

and Loper, 2004) was used to calculate the word counts.

Licenses All four datasets used in this research come with a license allowing non-commercial and academic usage. To be specific, the licenses are: MIT License for LCCC-base (Wang et al., 2020); Tecent AI Lab NaturalConv Dataset Terms and Conditions for NaturalConv (Wang et al., 2021); CC BY-SA 4.0 for Open2ch Dialogue Corpus (Inaba, 2019); and CC BY-NC 4.0 for BPersona-chat (Sugiyama et al., 2021; Li et al., 2022). The annotated data of this research will also be published in CC BY-NC 4.0 for non-commercial usage in the future.

D Error Examples

Table 7 shows the examples of annotated errors.

Error Type	**Source (zh, ja)**	**Translation (en)**
Tone Word Issue	怎么可能那么快啊反正这个周五之前啦	How is it possible that it's so fast? Anyway, it's before this Friday, *right?*
Tone Word Issue	それとこれとは別だろう	That's different from this, *isn't it?*
Additional Explanation	不要以为我不知道你余额宝里的巨款！	Don't think I don't know about the large sum of money in your 余额宝 *(savings)*!
Additional Explanation	排了两个小时队弄到的。。。。这个七夕礼物喜欢不	I lined up for two hours to get it..... I like this Qixi gift!* *(Note: "七夕" (Qixi) refers to the Qixi Festival, which is a traditional Chinese festival celebrating the mythological story of the Weaving Maiden and the Cowherd.)*
Additional Explanation	よくアニメを見てますよ。まとまった休みがとれたら、聖地巡礼ってやつをするのが楽しくてね。あなたは？	I often watch anime. If I get a big chunk of time off, I enjoy going on what they call "pilgrimages to holy sites" *(visiting real-life locations of anime scenes)*. How about you?
Lost Buzzword Issue	やったことあるw	I've done it before.

Table 7: Examples of annotated errors. * indicates another type of error.