Can ChatGPT Rival Neural Machine Translation? A Comparative Study

Zhaokun Jiang[†], **Ziyin Zhang**[‡*]
[†] School of Foreign Languages, Shanghai Jiao Tong University
[‡] Department of Computer Science and Engineering, Shanghai Jiao Tong University

Abstract

Inspired by the increasing interest in leveraging large language models for translation, this paper evaluates the capabilities of large language models (LLMs) represented by ChatGPT in comparison to the mainstream neural machine translation (NMT) engines in translating Chinese diplomatic texts into English. Specifically, we examine the translation quality of ChatGPT and NMT engines as measured by four automated metrics and human evaluation based on an error-typology and six analytic rubrics. Our findings show that automated metrics yield similar results for ChatGPT under different prompts and NMT systems, while human annotators tend to assign noticeably higher scores to ChatGPT when it is provided an example or contextual information about the translation task. Pairwise correlation between automated metrics and dimensions of human evaluation produces weak and non-significant results, suggesting the divergence between the two methods of translation quality assessment. These findings provide valuable insights into the potential of ChatGPT as a capable machine translator, and the influence of prompt engineering on its performance.

1 Introduction

Against the backdrop of translation automation and the advances in translation technology, neural machine translation (hereafter NMT) has been widely investigated under both academic and professional scenarios (Gaspari et al., 2015). Scholars have discussed its applications across different domains, and found it to fare well in both non-literary and literary texts, demonstrating satisfying quality and even creativity to some degree (Hu and Li, 2023).

In recent years, the emergence of large language models (LLMs) has further revolutionized the relationship between translation technologies and translators, and has opened up new possibilities for interactive translation using customized prompts, making proficiency in using AI tools an integral part of digital literacy (Godwin-Jones, 2022). Among them, ChatGPT, an AI-powered LLM designed and trained by OpenAI1, has demonstrated remarkable capabilities in a wide range of tasks, including translation among others (Hendy et al., 2023; Lai et al., 2023; Hendrycks et al., 2021; Srivastava et al., 2023; Zhao et al., 2023; Zhang et al., 2023). Recent studies have shown that ChatGPT is in par with or surpasses mainstream NMT engines in translating news, user comments, conversation, social media posts, and scientific texts (Jiao et al., 2023; Hendy et al., 2023; Wang et al., 2023; Peng et al., 2023; Lu et al., 2023; Chen et al., 2023; He et al., 2023; Kocmi and Federmann, 2023; Jiang et al., 2023).

The earlier explorations focusing on translation quality assessment (TQA) of machine translation have demonstrated the promising role of ChatGPT in generating quality translations. However, most of these studies are conducted on huge translation corpora, and the scholars mostly rely on automated metrics and scarcely resort to human evaluation, which only serves as a complement to automated metrics.

As NMT systems and LLMs keep making improvements, these measures may fail to fully capture the quality of machine translation. In particular, we have limited understanding of what aspects or dimensions other than content overlap can really define high-quality machine translations and distinguish them from those low-quality ones. There is a lack of metrics that are capable of measuring translation quality as humans do, who pay attention to dimensions such as intended audience, cultural sensitivity, adherence to translation norms, textual coherence, and practicality. Also, previous research is often conducted on extensive, publicly available datasets that are crawled from the internet

[*] daenerystargaryen@sjtu.edu.cn

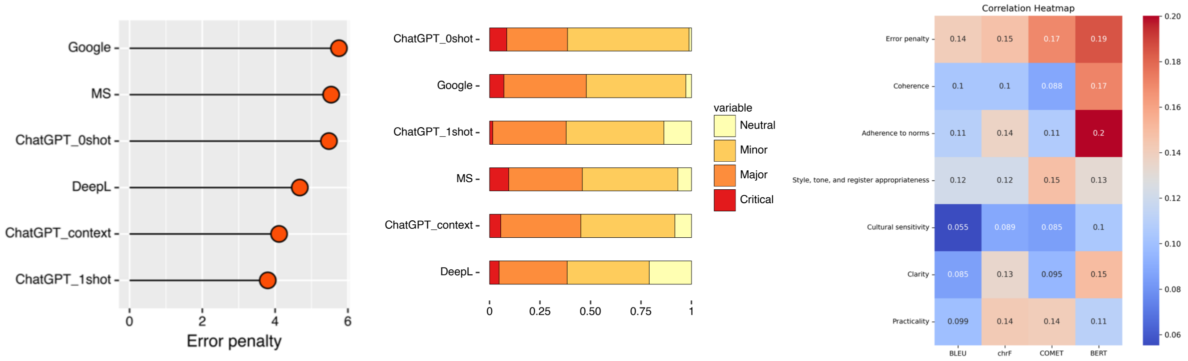

Figure 1: Total error penalty (left) and proportion of error severity (middle) assigned by human annotators to different translations. Right: correlation coefficients between human evaluation and automated metrics.

and encompass diverse domains[1], genres, and text types. There seems to be a scarcity of high-quality, domain-specific parallel corpora that would enable more focused and specialized quality assessment in machine translation.

Thus, we focus on a specific domain, and construct our major research questions as follows:

1. How well do NMT systems and ChatGPT perform under automated assessment?

2. How well do NMT systems and ChatGPT perform under human evaluation?

3. To what degree do automated metrics align with human evaluation?

Essentially, RQ1 and RQ2 aim to provide a well-rounded assessment of translations from NMT systems and ChatGPT using both automated metrics and human evaluation. Based on RQ1 and RQ2, RQ3 taps into the inter-correlation between automated metrics and human scores. For RQ1 and RQ2, we find that ChatGPT excels under human evaluation and semantic-aware automatic evaluation (see Figure 1 and Table 2). For RQ3, we find that the two methods may exhibit divergences when approaching machine-translated texts, contrary to Lu and Han (2023)'s findings that automated metrics can show moderate to strong correlations with human-assigned scores in assessing interpreting outputs, possibly due to the inherent differences between interpreting and translation.

2 Related Work

2.1 Improving the Translations of LLM via Prompt Engineering

Large Language Models (LLMs) represented by OpenAI's ChatGPT have advanced remarkably in recent years, showcasing emergent abilities such as in-context learning and Chain-of-Thought reasoning (Brown et al., 2020; Chowdhery et al., 2022; Wei et al., 2022a,b; Wang et al., 2022), and several studies have explored the influence of prompting strategies on the translation performance of LLMs (Jiao et al., 2023; Hendy et al., 2023; Wang et al., 2023; Peng et al., 2023; Lu et al., 2023; Chen et al., 2023; He et al., 2023). It has been found that with carefully written prompts, LLMs' translation quality can gain noticeable improvement, though such better performance is not seen in every scenario (Kocmi and Federmann, 2023; Lu et al., 2023). For instance, to inject cultural awareness into LLMs such as ChatGPT, Yao et al. (2023) propose "cultural knowledge prompting". Likewise, Mu et al. (2023) use translation memories as part of the prompts to enhance the translation performance of LLMs.

2.2 Translation Quality Assessment

Much attention has been devoted to the field of Translation Quality Assessment (TQA) in the past few decades, with both empirical and theoretical discussions increasing markedly. Particularly, many studies have emphasized the importance of MT evaluation since MT has already become an integral instrument for translation practitioners, in both translator training and translation education. Coupled with this process is the rapid development of NMT engines and latelt the emergence of chat-

[1] In this work, the terms "domain" and "'register" are used interchangeably. The former is preferred by the AI community. The later is preferred by linguistics literature.

bots represented by ChatGPT, which poses great challenges to the applicability and validity of various assessment metrics and methods.

Translation quality, however, has in fact no commonly agreed definition, as "theorists and professionals overwhelmingly agree there is no single objective way to measure quality" (Drugan, 2013). For those who share the belief that a translation should resemble the source text as closely as possible, quality is necessarily entangled with the concept of "equivalence". However, as translation studies experienced the cultural turn, this approach faces much criticism from the descriptive approach, which holds that emphasis should be laid on observing and describing the way translations are actually carried out, and how they are influenced by socio-cultural factors. The fit-for-purpose approach advocated by functionalists stresses that the notion of quality should be decided by how translations are received by end-users in real-life contexts.

When MT becomes the subject of TQA, it can also draw on the above conceptions of quality. However, since MT is application-oriented in nature, calculability and measurability should be at the core of its quality assessment. One widely adopted view of MT quality is proposed by Koby et al. (2014), which regards accuracy and fluency as the two determinants of translation quality. Alignment with human translation is also stressed, and Hassan et al. (2018) argue that only when MT shows no significant difference from human translations as measured by some scores can it be counted as high-quality translation.

Despite the diversity of frameworks to address the evaluation of MT quality, we follow Chatzikoumi (2020) to roughly divide approaches to TQA for MT into two lines: automated metrics and human evaluation. Automated metrics are of high efficiency, speed, consistency, and cost-effectiveness, and thus are widely adopted in both the academia and the industry. Within this line, metrics that compare a given translation to one or more reference translations are **reference-based**. They measure the similarity between the candidate MT and the reference(s) based on various linguistic and statistical features. Examples of reference-based metrics include BLEU (Bilingual Evaluation Understudy, Papineni et al., 2002), TER (Translation Edit Rate, Snover et al., 2006), ChrF (Character n-gram F-score, Popovic, 2015), and METEOR (Metric for Evaluation of Translation with Explicit Ordering, Banerjee and Lavie, 2005).

Reference-free metrics, unlike reference-based metrics that rely on a set of high-quality reference translations, assess the quality of translations based solely on the characteristics of the translated text itself. These metrics aim to measure translation quality in a more independent and self-contained manner. They often focus on linguistic properties, statistical patterns, or other features of the translation, and are particularly useful in scenarios where high-quality reference translations are not readily available or when evaluating translations with specific requirements that do not align with existing references. Typical reference-free metrics include YiSi-2 (Lo et al., 2018), COMET-QE (Rei et al., 2021), and QuestEval (Scialom et al., 2021).

Human evaluation can also be categorized into different types based on the methodology used. In rubric scoring (Bentivogli et al., 2018), human evaluators directly assign a score for translations within a fixed rating scale according to predefined criteria such as fluency, adequacy, accuracy, grammar, style, and overall coherence. Their judgments are typically combined to provide a final quality score. Another way of human evaluation is based on ranking (Bojar et al., 2016), which involves comparing and ranking multiple translations to decide which is superior or more suitable for a given purpose or context. As for error-based evaluation, errors or issues present in translations are identified and categorized according to pre-determined error typology. Human evaluators analyze the translations and annotate specific errors, taking into consideration dimensions that include accuracy, fluency, terminology, style, and so on. The most representative error typologies are DQF (Dynamic Quality Framework) proposed by TAUS in 2011, MQM (Multidimensional Quality Metrics) developed by QTLaunchPad, and the harmonized DQF-MQM taxonomy (Popović, 2018). Finally, post-editing can also be used for evaluation purposes, since the posting-editing efforts by a human translator to render a MT "good enough" and "deliverable" implicitly reflects the quality of the raw translation (Massardo et al., 2016).

2.3 Comparive Studies of LLM Translations and NMT

As a promising contender to dedicated NMT systems, LLMs have attracted much attention from TQA-related studies that compare their performance against NMT engines using both automated metrics and human assessment. For example, Jiao et al. (2023) compare the translation performance of ChatGPT and GPT4 against three NMT systems: DeepL, Google Translate, and Tencent TranSmart using four automated metrics. They find that ChatGPT and GPT4 perform comparably well to NMT engines in specific European languages, but not in others. They also find that ChatGPT and GPT4 excel in translating spoken texts, but not in some specialized fields, such as abstracts of academic papers. Similarly, Hendy et al. (2023) show that for high-resource language pairs such as English and French, ChatGPT could exhibit state-of-the-art translation capabilities matching or even surpassing the mainstream NMT engines, while Jiang et al. (2023) extract a wide range of linguistic features of translations in political domains, and find that the translations of ChatGPT are closer to NMT than to human translations. Karpinska and Iyyer (2023) show that when translating paragraph-level texts, ChatGPT produces fewer mistranslations, grammatical errors, and stylistic inconsistencies compared to Google Translate.

The existing researches have confirmed the strong capacity of LLMs in translating high-resource languages such as English and German. However, we are not fully aware of their competence in understanding and translating middle and low-resource languages, such as Chinese. Also, most of the assessments are conducted on publicly available corpora such as WMT datasets, leaving more specialized translations - such as diplomatic discourse - as under-investigated domains awaiting more investigation (Jiang et al., 2023).

3 Methodology

3.1 Corpus

The corpus in this study consists of 6,878 pieces of Spokesperson's Remarks on regular press conferences, which contains questions proposed by foreign reporters and answers from the Chinese spokespersons centering several foreign affairs at a range of press conferences. The corpus includes 17,837 parallel sentences in total, which amount to 642K Chinese tokens and 450K English tokens. We focus on Chinese-to-English discourse out of three considerations: (1) data availability, since all the textual materials are easily found online; (2) high quality, as the human translation is performed by professional institutional translators; (3) relative complexity, because the sources texts contain many idiomatic expressions and political terms only seen in Chinese political contexts, which may pose challenge to NMT engines and ChatGPT.

All of the questions are asked in English, and answers delivered by the spokespersons are in Chinese and then translated into English by institutional translators. Texts in our dataset are in fact transcripts of these press conferences, with some adjustments of wordings and contents, as well as corrections of speaking errors conducted by professional editors to be in line with the requirements of government websites. Since what the Chinese spokespersons express to the outside world concerns China's national interest and stance, their English translations are carefully curated and of premiere quality. All the materials can be directly downloaded from the official website of the Ministry of Foreign Affairs of the People's Republic of China[2].

3.2 Translation Tools

For NMT systems, we consider Microsoft Translate, Google Translate, and DeepL. After the completion of machine translation, we conducted a manual examination of the outputs to verify that they were system errors.

For LLM, we use GPT-3.5-Turbo, more commonly known as ChatGPT[3]. We handcraft three prompts: 0-shot, 1-shot, and a third prompt with additional information about the domain and context, as shown in Table 1.

3.3 Evaluation Methods

We adopt both automated metrics and human evaluation to balance efficiency and quality. Four widely-adopted automated metrics were selected. For human evaluation, we choose the integrated MQM-DFQ error typology and a six-dimensional analytic rubric scoring method.

3.3.1 Automated Metrics

BLEU involves a simple calculation of the number of n-gram matches between the MT output and one or multiple reference translations, with a penalty for unreasonably short translations. **ChrF** computes the F-score of character n-gram overlap instead. **BERTScore** leverages representations from pre-trained language models to compute the cosine similarity between translations and references, while **COMET** is similarly based on pre-trained

[2] https://www.mfa.gov.cn/eng/

[3] We use gpt-3.5-turbo-0613 API, accessed on November 2nd, 2023.

Method	Prompt
0-shot	Please translate the following Chinese sentence {source} into English. You should only output the translation.

1-shot	Please translate the following Chinese sentence {source} into English. You should only output the translation.
	Here is an example for you:
	{source}
	{reference}

Context	Please translate the following Chinese sentence {source} into English. It comes from a press conference in which a Chinese official interact with foreign reporters. The Chinese sentence is in the domain of politics. You should only output the translation.

Table 1: Prompts for querying ChatGPT.

language models, but also takes the source sentence into consideration during the scoring process.

We calculate BLEU, chrF, and BERTScore using torchmetrics[4]. COMET is computed using Unbabel-COMET[5].

3.3.2 Human Evaluation Based on Error Typology

Multidimensional Quality Metrics (MQM) is a hierarchical and specifications-based framework for TQA (Lommel et al., 2013) that considers different dimensions of translation quality simultaneously in a comprehensive and systematic manner. Similarly, the DQF error typology developed by TAUS further contributes to this endeavor by providing a standardized classification system for identifying and categorizing errors in translations.

The similarity between MQM and DQF makes their marriage possible. The integrated DQF-MQM typology aims to provide a unified approach to TQA. It comprises seven primary error types, several subtypes under each primary type, and 4 severity levels (critical, major, minor, neutral) to address translation issues related to accuracy, locale convention, terminology, style, design, fluency, verity, and other issues.

We choose accuracy, fluency, style, and terminology to form a subset of error typology to conduct human evaluation. To ensure the evaluation quality, annotators are provided with a detailed guideline, which can be found in the Appendix A. It covers specifications of the textual materials, error typologies used in this study, severity levels, and penalty points. Our guideline is drafted in accordance with the official file provided by TAUS[6].

3.3.3 Human Evaluation Based on Analytic Rubric Scoring

Analytic rubric scoring is another method widely adopted in TQA research. It is founded on the assumption that the overall concept of quality can be broken down into individual components, and typically comprises several sub-scales addressing separate dimensions of translation (Lu and Han, 2023). To complement the error typology-based evaluation, we propose six analytic rubrics to capture translation quality from different perspectives, encompassing dimensions of (1) coherence, (2) adherence to norms, (3) style, tone, and register appropriateness, (4) cultural sensitivity, (5) clarity, and (6) practicality. We use a 7-point Likert Scale and asked annotators to assign scores to the translation outputs for each rubric.

Considering time and costs, for each origin of translation (translations by ChatGPT under three prompts and by three NMT systems), we randomly sample 50 pieces of text to conduct human evaluation. To avoid pre-conceived judgements or biases on the quality of these translation tools, we do not inform the annotators of the origin of these translation samples beforehand. We recruited 6 annotators in total, all post-graduate students majoring in translation and interpreting (MTI). Each annotator is assigned 100 pieces of texts from all the six origins, and thus each text is annotated by two annotators. In terms of the error typology-based evaluation, the two annotations yield an average Cohen's Kappa of 0.73. Cases of disagreement (including the error type and the severity level) are settled through discussion and come to unanimous results. For the analytic rubric scoring, we average the scores assigned by the two annotators in each dimension.

[4] https://github.com/Lightning-AI/torchmetrics

[5] https://github.com/Unbabel/COMET

[6] https://info.taus.net/dqf-mqf-error-typology-template-download

Model	BLEU	Chrf	BERTScore	COMET
ChatGPT 0-shot	23.82	55.83	96.03	84.19
ChatGPT 1-shot	25.08	55.83	96.14	84.64
ChatGPT with context	23.98	56.19	96.38	85.30
Google Translate	25.28	55.86	96.12	83.02
MS Translate	29.16	59.17	96.15	84.41
DeepL	26.39	57.39	96.18	83.92

Table 2: Automated metrics of ChatGPT and NMT systems.

3.4 Analysis of Scores from Automated Metrics and Human Annotators

To answer RQ1, we compute the average scores of the four automated metrics. Their descriptive statistics can be found in Appendix B. To answer RQ2, we first provide annotation results guided by the MQM-DQF error typology, including the average error penalties, the frequencies of error types and subtypes, and the distribution of error severities. Then we show the results of analytic rubric scoring on six dimensions. The detailed descriptive statistics of human evaluation are listed in Appendix C. To answer RQ3, we calculate the Pearson's correlation coefficients at the significance level of 0.05 to examine the inter-relationship between automated metrics and human scores.

4 Results and Analysis

This section provides results of automated metrics and human evaluation. Section 4.1 shows the performance of ChatGPT and NMT engines measured by four automated metrics. The first part of Section 4.2 centers on the results of human evaluation, displaying error penalties, the distribution of error severities and error types in different translation outputs. The second part displays scores assigned by human annotators in six analytic rubrics. Section 4.3 shows the inter-correlation between automated metrics and human evaluation.

4.1 Automated Metrics

The results of ChatGPT and NMT systems' translations are presented in Table 2. Compared with the 0-shot scenario, providing ChatGPT with an example or contextual information only brings slight increase in BLEU score. There is no noticeable change on other evaluative metrics, indicating that ChatGPT under the 0-shot condition has already demonstrated strong capability in handling this specific translation task. The high BERTScore and COMET scores prove that in terms of semantic accuracy, ChatGPT has no difficulty in understanding the characteristic expressions in the domain of diplomatic discourse, and can deliver relatively faithful English translations even under the 0-shot condition. Among the three NMT systems, MS Translate gains the highest scores on 3 out of 4 metrics, only slightly lagging behind DeepL on BERTScore. However, the differences among these systems are not marked.

4.2 Human Assessment

The total error penalty assigned by human annotators to each translation system is plotted in the left subfigure of Figure 1. We observe that ChatGPT under the 1-shot condition is assigned the lowest penalties, indicating its highest translation capacity among the translation outputs listed, while ChatGPT under the 0-shot condition performs much worse. This suggests that providing even a small amount of training data, in this case only one example, significantly improves the translation capability of ChatGPT. Error penalty in translations by ChatGPT when it is given contextual information is the second lowest, showing that incorporating context, possibly through the use of context-aware prompts, can improve the translation quality provided by ChatGPT. On the other hand, Google Translate, MS Translate, and DeepL have comparably high error penalties, indicating similar translation performance among these widely used machine translation systems.

The middle subfigure of Figure 1 displays the proportion of errors in each level of severity. Among NMT engines, DeepL contains the highest percentage of neutral errors and least percentage of major and critical errors, indicating that it demonstrates better capability compared with the other two NMT engines. In contrast, Google Translate sees the highest proportion of major errors in its translations. As for ChatGPT, most of its errors

		ChatGPT			Google	MS	DeepL
		0-shot	1-shot	w. context			
Accuracy	Addition	1	0	1	0	0	0
	Omission	1	0	2	3	0	0
	Mistranslation	4	7	4	6	6	10
	Over-translation	1	1	1	2	2	2
	Under-translation	5	1	2	0	3	4
	Total	11	8	9	12	11	15
Fluency	Punctuation	0	0	1	2	0	0
	Grammar	6	5	8	5	8	8
	Total	6	5	8	7	8	8
Terminology	Wrong terms	9	6	6	7	4	7
	Inconsistent usage	0	0	0	0	0	0
	Total	9	7	7	7	5	8
Style	Inconsistent style	0	1	0	1	1	1
	Awkward	16	2	3	2	3	2
	Unidiomatic	3	13	7	9	4	9
	Total	19	17	12	13	9	13
Other Errors		0	1	0	1	0	0

Table 3: Major types and subtypes of errors in different translations.

under the 0-shot condition are minor and major errors. Giving it an example significantly decreases the proportion of its critical errors and increases its neutral errors. Providing it with contextual information yields a similar effect.

Table 3 shows the occurrence of errors under each major and sub-category. Overall, stylistic errors occur most frequently in most translations, followed by errors related to accuracy. While most errors of ChatGPT are attributed to style, accuracy poses the most challenge to NMT systems. Particularly, translations by ChatGPT are identified significantly more awkward and unidiomatic expressions compared with NMT systems.

In terms of analytic rubrics, Figure 2 shows that ChatGPT under the 1-shot setting consistently receives the highest scores across all the rubrics, indicating its overall superiority under human evaluation. In contrast, ChatGPT in the 0-shot scenario ranks last in five out of six dimensions (adherence to norms, practicality, clarity, cultural sensitivity, as well as style, tone, and register appropriateness), which suggests that providing it with an example containing high-quality human translation contributes to boosting its translation performance markedly. Exposing ChatGPT to contextual information also serves to improve its translation quality in most cases (four out of six dimensions). The three NMT systems receive similar scores, showing that human annotators do not observe notable differences among their translation outputs.

4.3 Correlation between Automated Metrics and Human Assessment

Based on the results in Section 4.1 and ref 4.2, we calculate the correlation coefficients between human-assigned scores and scores from automated metrics. The results are shown in the right subfigure of Figure 1. We observe that in general, BERTScores align with human-assigned scores most closely, showing an average correlation coefficient of 0.16. BLEU scores deviate most from human evaluation, with a mere 0.10 correlation coefficient on average. However, the majority of these correlations are not statistically significant (the p values can be found in Appendix D), suggesting that human evaluation and automated metrics tend to focus on different aspects of translation quality. In terms of aspects of human evaluation, error penalty shows the strongest correlation with automated scores in general, while cultural sensitivity demonstrates the weakest correlation.

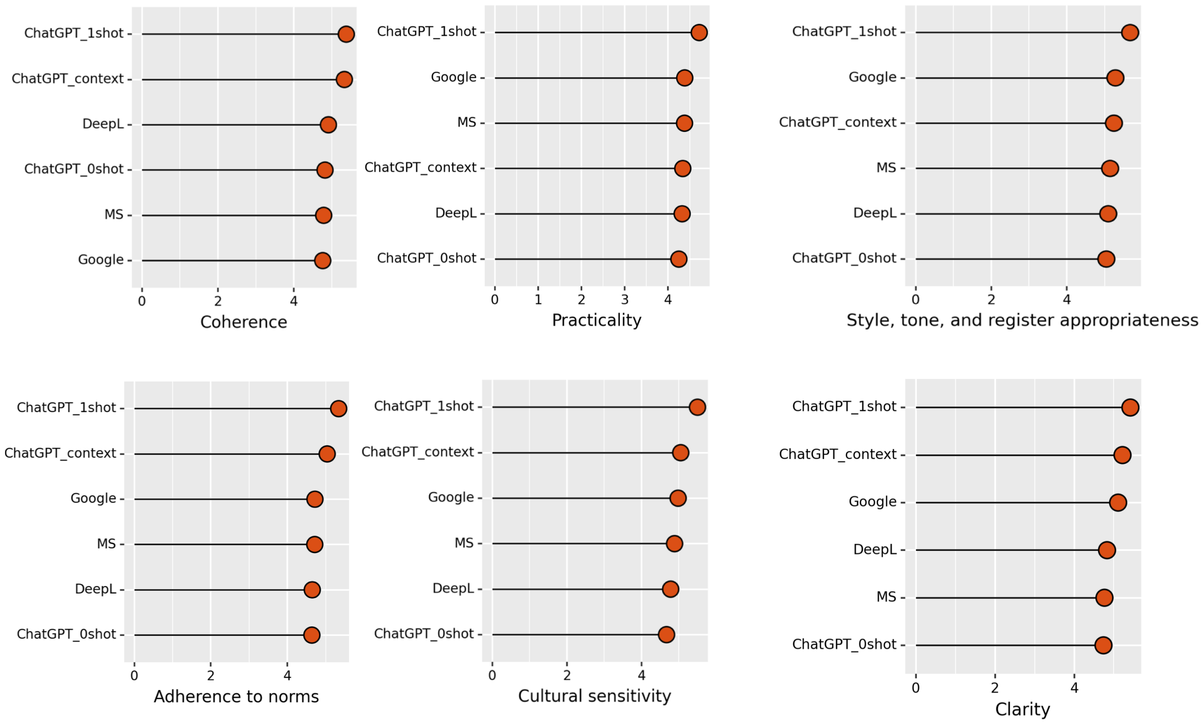

Figure 2: Human-assigned scores for each analytic rubric.

5 Discussion

From the results of automatic assessment, we can see that when measured by BLEU and Chrf, ChatGPT is overshadowed by NMT systems. This suggests that translations generated by ChatGPT tend to deviate more from the references in terms of the n-gram matches than NMT system, and ChatGPT may struggle to produce translations that precisely replicate the wording and phrasing of the references. However, its performance shines when evaluated using BERTScore and COMET score. These metrics take into account the semantic similarity and fluency of translations, rather than focusing solely on n-gram overlap. ChatGPT's ability to challenge or even surpass NMT systems when measured by these two metrics suggests that despite the deviations from the reference translations, its translations exhibit a strong semantic closeness to the references.

This observation aligns with previous researches on generative LLMs, which have shown that these models excel at capturing semantic relationships and generating coherent text. The underlying architecture of ChatGPT grants it exceptional context awareness and ability of language understanding, which enables its to produce translations that are more semantically aligned with the references, even if they may differ in specific word choices or phrase structure. It is worth noting that although BLEU and Chrf scores are widely adopted for evaluating translation quality, they have limitations, particularly when applied to language models like ChatGPT, which prioritize semantic coherence over exact phrasing. On the other hand, BERTScore and COMET provide a more nuanced evaluation by considering semantic aspects and fluency. These metrics are better at reflecting the strengths of ChatGPT in translating context-dependent texts.

Another finding is that different prompting strategies do not exert significant boost to ChatGPT's translation performance, which can be attributed to two possible reasons. First, ChatGPT in the 0-shot scenario can already understand the semantic meaning of the original text perfectly, so that providing it with either extra information will not significantly affect its translation capability. Second, these automated metrics fail to capture the improvement brought by customized prompts. Drawing from results of human evaluation, the second explanation holds more feasibility, which will be further discussed later.

In terms of human evaluation, we observe from the overall error penalties (see Figure 1) that ChatGPT tends to demonstrate strong plasticity in its translation performance: providing it with only one example or relevant contextual information can greatly reduce error penalties in its translation outputs and lead to a significant improvement in translation quality.

This is further corroborated by evidence from scores of six analytic rubrics (see Figure 2), which demonstrates the variability of the performance of ChatGPT under different conditions. When exposed to a single example, ChatGPT consistently receives the highest scores across all the analytic rubrics, indicating its superior performance as evaluated by human annotators. In contrast, ChatGPT without any example (0-shot condition) ranks last in five out of six dimensions: adherence to norms, practicality, clarity, cultural sensitivity, and style, tone, and register appropriateness. This indicates that ChatGPT without the guidance of an example struggles in these areas and relies on the presence of a reference translation to enhance its performance. Similarly, exposing ChatGPT to contextual information improves its translation quality across all the six rubrics, showing that the inclusion of context aids ChatGPT in generating more contextually appropriate translations. This informs us the importance of crafting appropriate and relevant prompts to fully uncover the potential of ChatGPT in generating high-quality translations. On the other hand, the three NMT systems receive similar scores across the assessed dimensions, implying that human annotators do not observe significant differences in translation quality among these NMT systems.

In addition, NMT engines are more prone to mistakes compared with ChatGPT. This corresponds to previous findings reported by Manakhimova et al. (2023), who investigate the ability of GPT4 and NMT systems in handling challenging translation issues. Their study shows that GPT4 outperforms NMT engines in most cases, displaying fewer mistakes and higher accuracy. This is further supported by the distribution of errors (see Table 3), which shows that accuracy poses the greatest difficulty for NMT systems in general.

In contrast, ChatGPT exhibits style errors as its primary challenge. In particular, translations by ChatGPT are noted for containing more awkward and idiomatic expressions compared to NMT systems, but providing relevant context notably reduced its style-related errors. ChatGPT's style errors may stem from its language generation capabilities, as it tends to prioritize semantic coherence over precise phrasing, leading to stylistic variations. The awkward and idiomatic expressions may arise from its tendency to generate creative and diverse outputs, which can sometimes result in less conventional or less fluent translations.

Finally, the Pearson's correlations between human-assigned scores and automated metrics are weak and non-significant, though some specific dimensions show slightly higher correlations than others. This is somehow contradictory to findings in Lu and Han (2023), possibly because assessing interpreting is different from assessing translation, with the latter having higher requirements on aspects other than accuracy. Our results highlight the challenges in relying solely on automated scores to assess translation quality, as they may not fully capture the intricacies of dimensions such as adherence to norms, cultural sensitivity, clarity, and practicality. Human evaluation as a more contextual and culture-specific way of assessment is necessary to obtain a comprehensive understanding of translation quality.

What implications do these results have for future research in TQA? First, we should acknowledge the limitations of traditional metrics. Our findings highlight that metrics like BLEU and chrF, which primarily focus on n-gram overlap and exact phrasing, cannot effectively capture the superior performance of language models such as ChatGPT in a range of critical dimensions for high-quality human-like translations. This calls for the development and adoption of more nuanced evaluation metrics that consider cultural aspects and contextual appropriateness. Next, carefully crafted and customized prompts are needed to fully unleash the great potential of ChatGPT as a capable machine translator. Our results show that providing a single example or some relevant contextual information can greatly reduce its errors and drive up its scores in all the analytic rubrics. This highlights the practical significance of tailoring prompts to guide the generation process and enhance the translation quality of ChatGPT.

6 Conclusion

This study compares the translation quality of ChatGPT and three NMT engines by using both automated metrics and human evaluation. For the former, we compute four widely-adopted metrics - BLEU, chrF, COMET, and BERTScore - and find that they fail to distinguish high-quality and lower-quality translations. For the latter, we conduct annotation based on both the integrated MQM-DQF error typology and six analytic rubrics. Results show that exposing ChatGPT to one example or context-relevant information greatly boosts its performance under all dimensions of human evaluation. To examine the correlation between automated metrics and human evaluation, we calculate the pairwise Pearson's correlation coefficients. The weak and non-significant results overall demonstrate that human understanding of translation quality is significantly different from what is captured by automated metrics.

In light of the findings discussed above, we suggest two directions to advance researches in TQA: Firstly, further exploration is needed to develop more effective evaluation metrics that can better capture the nuances of translation quality, particularly in terms of coherence, clarity, practicality adherence to norms, cultural sensitivity, as well as appropriateness of style, tone, and register. Secondly, it is crucial to continue investigating the role of proper prompts and contextual information in improving the performance of language models like ChatGPT, which exhibit strong context awareness and language understanding capabilities.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Luisa Bentivogli, Mauro Cettolo, Marcello Federico, and Christian Federmann. 2018. Machine translation human evaluation: an investigation of evaluation based on post-editing and its relation with direct assessment. In *Proceedings of the 15th International Conference on Spoken Language Translation, IWSLT 2018, Bruges, Belgium, October 29-30, 2018*, pages 62–69. International Conference on Spoken Language Translation.

Ondrej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia. 2016. Ten years of wmt evaluation campaigns: Lessons learnt.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Eirini Chatzikoumi. 2020. How to evaluate machine translation: A review of automated and human metrics. *Nat. Lang. Eng.*, 26(2):137–161.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023. Iterative translation refinement with large language models. *CoRR*, abs/2306.03856.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Joanna Drugan. 2013. *Quality In Professional Translation: Assessment and Improvement.* Bloomsbury Academic.

Federico Gaspari, Hala Almaghout, and Stephen Doherty. 2015. A survey of machine translation competences: Insights for translation technology educators and practitioners. *Perspectives Studies in Translatology*, 23:1–26.

Robert Godwin-Jones. 2022. Partnering with ai: Intelligent writing assistance and instructed language learning. *Language Learning & Technology*, 26(2):5–24.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring human-like translation strategy with large language models. *CoRR*, abs/2305.04118.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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. *CoRR*, abs/2302.09210.

Kaibao Hu and Xiaoqian Li. 2023. The creativity and limitations of ai neural machine translation: A corpus-based study of deepl's english-to-chinese translation of shakespeare's plays. *Babel*, 69(4):546–563.

Zhaokun Jiang, Qianxi Lv, and Ziyin Zhang. 2023. Distinguishing translations by human, nmt, and chatgpt: A linguistic and statistical approach. *CoRR*, abs/2312.10750.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *CoRR*, abs/2301.08745.

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, WMT 2023, Singapore, December 6-7, 2023*, pages 419–451. Association for Computational Linguistics.

Geoffrey S. Koby, Daryl R. Hague, Arle Lommel, and Alan K. Melby. 2014. Defining translation quality. In *Tradumatica*, 12, pages 413–420.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023, Tampere, Finland, 12-15 June 2023*, pages 193–203. European Association for Machine Translation.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 13171–13189. Association for Computational Linguistics.

Chi-kiu Lo, Michel Simard, Darlene A. Stewart, Samuel Larkin, Cyril Goutte, and Patrick Littell. 2018. Accurate semantic textual similarity for cleaning noisy parallel corpora using semantic machine translation evaluation metric: The NRC supervised submissions to the parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 908–916. Association for Computational Linguistics.

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

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. *CoRR*, abs/2303.13809.

Xiaolei Lu and Chao Han. 2023. Automatic assessment of spoken-language interpreting based on machine-translation evaluation metrics: A multi-scenario exploratory study. *Interpreting*, 25(1):109–143.

Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2023. Linguistically motivated evaluation of the 2023 state-of-the-art machine translation: Can chatgpt outperform nmt? In *Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023*, pages 224–245. Association for Computational Linguistics.

Isabella Massardo, Jaap van der Meer, Sharon O'Brien, Fred Hollowood, Nora Aranberri, and Katrin Drescher. 2016. MT post-editing guidelines.

Yongyu Mu, Abudurexiti Reheman, Zhiquan Cao, Yuchun Fan, Bei Li, Yinqiao Li, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2023. Augmenting large language model translators via translation memories. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10287–10299. 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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

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, Singapore, December 6-10, 2023*, pages 5622–5633. Association for Computational Linguistics.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation, WMT@EMNLP 2015, 17-18 September 2015, Lisbon, Portugal*, pages 392–395. The Association for Computer Linguistics.

Maja Popović. 2018. *Error Classification and Analysis for Machine Translation Quality Assessment*, pages 129–158. Springer International Publishing, Cham.

Ricardo Rei, Ana C. Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro G. Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-ist 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation, WMT@EMNLP 2021, Online Event, November 10-11, 2021*, pages 1030–1040. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. Questeval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6594–6604. Association for Computational Linguistics.

Matthew G. Snover, Bonnie J. Dorr, Richard M. Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, AMTA 2006, Cambridge, Massachusetts, USA, August 8-12, 2006*, pages 223–231. Association for Machine Translation in the Americas.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 16646–16661. Association for Computational Linguistics.

Thomas Wang, Adam Roberts, Daniel Hesslow, Teven Le Scao, Hyung Won Chung, Iz Beltagy, Julien Launay, and Colin Raffel. 2022. What language model architecture and pretraining objective works best for zero-shot generalization? In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 22964–22984. PMLR.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

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

Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. 2023. Empowering llm-based machine translation with cultural awareness. *CoRR*, abs/2305.14328.

Ziyin Zhang, Chaoyu Chen, Bingchang Liu, Cong Liao, Zi Gong, Hang Yu, Jianguo Li, and Rui Wang. 2023. Unifying the perspectives of nlp and software engineering: A survey on language models for code. *CoRR*, abs/2311.07989.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

A Guidelines for Human Assessment under the MQM Error Typology

A.1 General Instructions

You will be assessing translations at the sentence level. Each translated sentence is aligned with its corresponding source text. You have the flexibility to revise previous annotations as needed.

There are two tasks for you to finish.

The first one is error-analysis-based. Your task is to identify errors within each translated sentence, with a maximum limit of five errors. If there are more than five errors, focus on marking the five most severe ones. In cases where the translation is severely distorted or unrelated to the source, mark a single Non-translation error that covers the entire segment.

To identify an error, highlight the relevant portion of the translation and choose a category/subcategory and severity level from the available options. When identifying errors, please be as fine-grained as possible. For instance, if a sentence contains two mistranslated words, record them as separate mistranslation errors. If a single section of text has multiple errors, indicate the most severe one.

Please pay particular attention to the context when annotating. If a translation may be questionable on its own but fits within the context, it should not be considered erroneous. Conversely, if a translation might be acceptable in some contexts, but not for the current sentence, mark it as incorrect.

There is a special error category called Non-translation, which can only be used once per sentence and should encompass the entire sentence. If Non-translation is selected, no other errors should be identified.

The second task is impression-based. You need to report your level of agreement with 7 statements on a 7-point Likert scale, in which 1 means “completely disagree”, and 7 means “completely agree”. The statements are the following:

1. **Coherence**: The translation flow is mostly smooth and coherent. There is no logical disconnection or meaning inconsistency.
2. **Adherence to norms**: The translation fulfills the common standards, requirements, and norms of political translation.
3. **Style, tone, and register appropriateness**: The translation is consistent in style, tone, and register with the source text. For example, if the source text has a formal tone and sophisticated style, the translation also reflects that formality and sophistication.
4. **Cultural sensitivity**: The translation demonstrates cultural sensitivity. It suitably conveys culture specific items (CSI), humor, and other cultural nuances in a way that is understandable and relatable to the target audience.
5. **Clarity**: The translation is clear and easily understandable to the target audience. It does not contain ambiguities, excessive jargon, or overly complex language that may hinder comprehension.
6. **Practicality**: The translation can be directly put for actual use. In this case, it can be put on the government website for people across the world to read.

A.2 Specifications

We select a portion of parameters from the ASTM F2575-14 specifications[7] to describe what is expected of the translation.

[7] https://www.astm.org/f2575-14.html

Source-content information

Source language: Chinese
Text type: political and diplomatic texts remarks from the Chinese spokesmen.
Audience: Chinese readers who intend to know China's stance on a range of important foreign affairs (e.g., reporters, politicians, and diplomats).
Purpose: to deliver China's stance and attitudes on a range of important foreign affairs.
Volume: you will be given 100 sentences.
Complexity: usually written in a relatively complex and formal style, commonly found in official statements or diplomatic contexts.
Origin: official website of the Ministry of Foreign Affairs of The People's Republic of China (https://www.fmprc.gov.cn/fyrbt_673021/dhdw_673027/index_1.shtml)

Target content requirements

Target language: English
Audience: international readers who intend to know China's stance on a range of important foreign affairs (e.g., reporters, politicians, and diplomats).
Purpose: to deliver China's stance and attitudes on a range of important foreign affairs.
Format: written texts displayed on the government website, which are transcribed and carefully edited from spoken remarks.
Style: in a formal, official, and often assertive tone commonly found in official statements or diplomatic discourse.

A.3 Error Typology, Severity Levels and Penalty Levels

Error Typology

Error Type	Subtype	Definition
Accuracy		The target text does not accurately reflect the source text, allowing for any differences authorized by specifications.
	Addition	The target text includes text not present in the source.
	Omission	Content is missing from the translation that is present in the source.
	Mistranslation	The target content does not accurately represent the source content.
	Over-translation	The target text is more specific than the source text.
	Under-translation	The target text is less specific than the source text.
Fluency		Issues related to the form or content of a text, irrespective as to whether it is a translation or not.
	Punctuation	is used incorrectly (for the locale or style).
	Spelling	Issues related to spelling of words.
	Grammar	Issues related to the grammar or syntax of the text, other than spelling and orthography.
	Inconsistency	The text shows internal inconsistency.
	Link/cross-reference	Links are inconsistent in the text.
Terminology		A term (domain-specific word) is translated with a term other than the one expected for the domain or otherwise specified.
	Wrong terms	Use of term that it is not the term a domain expert would use or because it gives rise to a conceptual mismatch.
	Inconsistent use of terminology	Terminology is used in an inconsistent manner within the text.
Style		The text has stylistic problems.
	Inconsistent style	Style is inconsistent within a text.
	Awkward	A text is written with an awkward style.
	Unidiomatic	The content is grammatical, but not idiomatic.
Other		Any other issues.

Severity Levels

Non-translation	The translation is severely distorted or unrelated to the source.
Critical	Errors that may carry health, safety, legal or financial implications, violate geopolitical usage guidelines, damage the entities' reputation, or which could be seen as offensive.
Major	Errors that may confuse or mislead the readers due to significant change in meaning or because errors appear in a visible or important part of the content.
Minor	Errors that don't lead to loss of meaning and wouldn't confuse or mislead the readers but would be noticed, would decrease stylistic quality, fluency or clarity, or would make the content less appealing.
Neutral	Used to log additional information, problems or changes to be made that don't count as errors, e.g., they reflect a reviewer's choice or preferred style, they are repeated errors or instruction/glossary changes not yet implemented, a change to be made that the translator is not aware of.

Penalty Levels

Non-translation	100	Deduct the penalty points for non-translation errors
Critical errors	25	Deduct the penalty points for critical errors
Major errors	10	Deduct the penalty points for major errors
Minor errors	1	Deduct the penalty points for minor errors
Neutral errors	0	Deduct the penalty points for neutral errors

B Descriptive Statistics of Scores Using Automated Metrics

The descriptive statistics of automated metrics (BLEU, chrF, COMET, and BERTScore) are given in Table 4, 5, 6, and 7 respectively.

	Mean	Std.	Min	Max	Kurtosis	Skewness
ChatGPT 0-shot	0.228	0.196	0.000	1.000	0.211	0.605
ChatGPT 1-shot	0.251	0.213	0.000	1.000	0.317	0.662
ChatGPT w. context	0.235	0.200	0.000	1.000	0.172	0.607
Google Translate	0.243	0.209	0.000	1.000	0.306	0.654
MS Translate	0.264	0.217	0.000	1.000	-0.089	0.528
DeepL	0.292	0.236	0.000	1.000	-0.039	0.562

Table 4: BLEU statistics.

	Mean	Std.	Min	Max	Kurtosis	Skewness
ChatGPT 0-shot	0.558	0.152	0.076	1.000	0.214	-0.333
ChatGPT 1-shot	0.586	0.092	0.173	1.000	0.357	-0.178
ChatGPT w. context	0.563	0.154	0.067	1.000	0.215	-0.314
Google Translate	0.559	0.167	0.000	1.000	0.754	-0.421
MS Translate	0.592	0.169	0.065	1.000	0.039	-0.147
DeepL	0.574	0.163	0.065	1.000	0.077	-0.282

Table 5: ChrF statistics.

	Mean	Std.	Min	Max	Kurtosis	Skewness
ChatGPT 0-shot	0.842	0.059	0.441	0.986	5.068	-1.512
ChatGPT 1-shot	0.846	0.061	0.441	0.986	3.902	-1.296
ChatGPT w. context	0.841	0.062	0.368	0.986	6.591	-1.714
Google Translate	0.830	0.084	0.269	0.986	13.968	-2.824
MS Translate	0.844	0.068	0.341	0.986	4.771	-1.386
DeepL	0.839	0.064	0.441	0.986	3.549	-1.166

Table 6: COMET statistics.

	Mean	Std.	Min	Max	Kurtosis	Skewness
ChatGPT 0-shot	0.960	0.007	0.938	0.992	0.394	0.302
ChatGPT 1-shot	0.961	0.008	0.940	0.999	0.992	0.579
ChatGPT w. context	0.961	0.008	0.938	0.995	0.470	0.471
Google Translate	0.961	0.008	0.939	0.995	0.313	0.328
MS Translate	0.962	0.008	0.938	0.996	0.832	0.540
Tengxun Translate	0.962	0.008	0.939	0.995	0.283	0.408

Table 7: BERTScore statistics.

C Descriptive Statistics of Scores Assigned by Human Annotators

The descriptive statistics of annotators' assigned scores to the six systems are given in Table 8-13.

	Error penalty	Coherence	Adherence to norms	Style, tone, and register appropriateness	Cultural sensitivity	Clarity	Practicality
Mean	-4.110	4.830	4.890	5.250	4.850	4.970	4.340
Std.	-6.302	1.640	1.406	1.351	1.591	1.611	1.713
Max	10.000	2.000	2.000	2.000	1.000	1.000	1.000
Min	-35.000	7.000	7.000	7.000	7.000	7.000	7.000
Kurtosis	-4.373	-0.973	-1.028	-0.603	-0.917	-0.728	-1.066
Skewness	-1.225	-0.444	-0.023	-0.560	-0.357	-0.563	-0.101

Table 8: Human annotation statistics of ChatGPT 0-shot.

	Error penalty	Coherence	Adherence to norms	Style, tone, and register appropriateness	Cultural sensitivity	Clarity	Practicality
Mean	-3.808	5.380	5.340	5.680	5.470	5.420	4.717
Std.	-7.675	1.285	1.121	1.024	1.259	1.365	1.385
Max	10.000	1.000	2.000	2.000	2.000	2.000	1.000
Min	-25.000	7.000	7.000	7.000	7.000	7.000	7.000
Kurtosis	-1.268	0.708	-0.138	0.297	-0.233	-0.413	-0.303
Skewness	-1.218	-0.937	-0.438	-0.578	-0.604	-0.598	-0.386

Table 9: Human annotation statistics of ChatGPT 1-shot.

	Error penalty	Coherence	Adherence to norms	Style, tone, and register appropriateness	Cultural sensitivity	Clarity	Practicality
Mean	-4.110	4.830	4.890	5.250	4.850	4.970	4.340
Sd.	-6.302	1.640	1.406	1.351	1.591	1.611	1.713
Max	10.000	2.000	2.000	2.000	1.000	1.000	1.000
Min	-35.000	7.000	7.000	7.000	7.000	7.000	7.000
Kurtosis	-4.373	-0.973	-1.028	-0.603	-0.917	-0.728	-1.066
Skewness	-1.225	-0.444	-0.023	-0.560	-0.357	-0.563	-0.101

Table 10: Human annotation statistics of ChatGPT with context.

	Error penalty	Coherence	Adherence to norms	Style, tone, and register appropriateness	Cultural sensitivity	Clarity	Practicality
Mean	-5.660	5.260	5.220	5.290	5.170	5.110	4.384
Std.	-7.907	1.474	1.345	1.328	1.464	1.645	1.658
Max	5.000	1.000	2.000	3.000	2.000	1.000	1.000
Min	-25.000	7.000	7.000	7.000	7.000	7.000	7.000
Kurtosis	-0.210	0.215	-0.980	-1.116	-1.221	-0.754	-0.890
Skewness	-1.066	-0.911	-0.231	-0.179	-0.277	-0.588	-0.058

Table 11: Human annotation statistics of Google Translate.

	Error penalty	Coherence	Adherence to norms	Style, tone, and register appropriateness	Cultural sensitivity	Clarity	Practicality
Mean	-5.540	4.788	4.717	5.152	4.889	4.768	4.384
Std.	-9.440	1.728	1.597	1.466	1.684	1.778	1.748
Max	10.000	1.000	1.000	1.000	1.000	1.000	1.000
Min	-35.000	7.000	7.000	7.000	7.000	7.000	7.000
Kurtosis	-1.262	-0.959	-0.502	0.210	-0.554	-0.868	-1.055
Skewness	-1.407	-0.397	-0.452	-0.752	-0.559	-0.489	-0.358

Table 12: Human annotation statistics of Microsoft Translate.

	Error penalty	Coherence	Adherence to norms	Style, tone, and register appropriateness	Cultural sensitivity	Clarity	Practicality
Mean	-4.900	4.950	4.677	5.121	4.790	4.830	4.300
Std.	-6.870	1.438	1.268	1.118	1.266	1.443	1.501
Max	0.000	2.000	2.000	2.000	2.000	1.000	1.000
Min	-25.000	7.000	7.000	7.000	7.000	7.000	7.000
Kurtosis	-1.411	-0.276	-0.311	-0.191	-0.283	-0.216	-0.554
Skewness	-1.477	-0.547	-0.160	-0.196	-0.081	-0.410	-0.395

Table 13: Human annotation statistics of DeepL.

D Average Correlation Coefficients between Human Evaluation and Automated Metrics

See Table 14.

	BLEU	chrF	COMET	BERTScore
Error penalty	0.141	0.147*	0.170	0.185*
Coherence	0.103	0.103	0.088	0.171
Adherence to norms	0.109	0.137	0.106	0.200*
Style, tone, and register appropriateness	0.121	0.179	0.148	0.192
Cultural sensitivity	0.055	0.089	0.085	0.104
Clarity	0.085	0.134	0.095	0.150
Practicality	0.099	0.143	0.141	0.110
Average	0.102	0.133	0.119	0.159

Table 14: Correlation coefficients between human evaluation and automated metrics. *: $p < 0.005$, significant in all six systems.