Scholar Publishing Group
International Journal of Educational Curriculum Management and Research
https://doi.org/10.38007/IJECMR.2024.050121
ISSN 2790-0967 Vol. 5, Issue 1: 153-163

A Study on the Translation Quality of ChatGPT

Min Zhang[1,a*]

[1]*Graduate School of Translation and Interpretation, Tianjin Foreign Studies University, Tianjin, 300204, China*
[a]*arine.zhang302@gmail.com*
[*]*corresponding author*

Keywords: ChatGPT, Translation Quality Evaluation, Literary Translation, English-to-Chinese, English-to-Japanese

Abstract: The rapid development of natural language processing technology has led to a surge in interest in the application of large language models in the field of translation. This study employs the large language model ChatGPT as a translation tool, utilizing the American short story "A Service of Love" as a translated text. The objective is to compare and evaluate the translation quality and performance of ChatGPT in the language combinations of English-to-Chinese and English-to-Japanese with that of Google Translate and DEEPL Translator. Additionally, the study aims to explore the accuracy, fluency, fidelity, and adaptability of ChatGPT in the literary translation, with the intention of providing reference and inspiration for the translation strategy of AI, particularly in the context of literary translation in the language pairs of English and Asian languages.

1 Introduction

Large Language Models (LLMs) have recently demonstrated remarkable capabilities in natural language processing tasks and beyond (Naveed et al., 2023). Among the numerous LLMs, ChatGPT is widely regarded as a model of significant importance, having demonstrated exceptional performance in language translation, text generation, and semantic analysis tasks. Additionally, it has demonstrated considerable potential for application in machine translation. Moreover, the emergence of AI, like ChatGPT, paved the way for proficient MT tools enabling translations between various languages (Sanz-Valdivieso & López-Arroyo, 2023). The existing research on machine translation has focused more on practical applications and representations in technical documents and conventional texts in the more common languages. Nevertheless, it is possible that the evaluation of quality in literary translation, particularly in language pairs involving English and Asian languages such as Chinese and Japanese, remains a relatively understudied area. Should the translation model based on ChatGPT be able to achieve a high standard in the translation of literary works, the wide application of this automated translation tool will play a significant role in motivating international cultural communication. Concurrently, the methodology will furnish

translators with an innovative instrument that may facilitate the enhancement of translation efficiency and the assurance of translation quality.

This study proposes a novel evaluation criterion for literary translation. It considers not only the accuracy of the machine translation but also on the transformation of cultural context, the spread of emotions, and the readability and aesthetics of the translation. This multifaceted evaluation strategy offers a novel approach to the study of machine translation quality, providing insights that can inform future research. This study employs the American short story "A Service of Love" as its research subject, combining automatic and human evaluation methods to analyze the quality of the translation produced by ChatGPT in the English-to-Chinese and English-to-Japanese language combinations. The study compares this translation with those produced by Google Translate and DEEPL Translator. The objective of this study is to provide empirical evidence in support of large language models in the field of machine translation and to further investigate their potential advantages and avenues for improvement.

2 Large Language Model ChatGPT

Generative Pre-trained Transformer (GPT), a renowned advanced language model created by OpenAI, has acquired considerable scrutiny for its capacity to comprehend and produce coherent and logical text (Hendy et al., 2023; Sahari et al., 2024). Since its initial release, ChatGPT has undergone numerous version updates and optimizations, and has been gradually and universally adopted by the international community, becoming one of the most influential language models. However, the translation generated needs thorough evaluation because of the need for more understanding of domain terminologies and the cultural context of the model (Khoshafah, 2023). Considering this, Chowdhery et al. (2022) emphasize the importance of thoroughly evaluating the translation produced by MT rather than relying solely on automated metrics.

Consequently, ChatGPT can be employed as a tool to facilitate translation tasks. However, its translation quality must be evaluated.

3 Translation Quality Evaluation Methods

3.1 Overview of Translation Quality Evaluation

Translation quality evaluation has emerged as a significant area of research, with the objective of assessing the quality of translation works from a more scientific and objective perspective. The various means of translation quality evaluation demonstrate the depth and complexity of translation as a cross-cultural communication activity. A rigorous and organized evaluation process is regarded as the core of ensuring high translation quality.

The criteria for evaluating translation quality have undergone a transformation from a subjective to an objective approach. In the early stages of the field, most evaluation criteria were based on translators' personal experience and intuition, and no unified evaluation system was established based on these criteria and real data. As research in the field of translation progressed, the academic community came to recognize the necessity of establishing a set of systematic and scientific evaluation criteria. In 1965, Catford proposed the "Translation Shifts Theory," which posits that the essential meaning of the source language should be translated into the target language. This translation should not only be linguistically accurate but also achieve the equivalence of meanings between the two languages. In his theory of "functional equivalence," Nida (1986) posits that the

effect of the translated text on readers of the target language should aim to maintain the same effect as that of readers of the source language. This set of theories not only expands the theoretical framework of translation research but also provides clear evaluation criteria for evaluating translation quality.

Translation quality evaluation methods are typically classified into two categories: human evaluation and automated evaluation. Human process typically relies on the expertise and experience of the evaluator. One of the advantages of this method is that it can comprehensively consider a variety of factors, such as context and culture, so that the output results of the evaluation are both accurate and authoritative. However, even human translations are potentially biased and subjective, considering the possibility of several translations for an original text that could be deemed accurate (Rivera-Trigueros, 2022). To address the limitations of human evaluation techniques, several automated evaluation tools and techniques have been developed, including the BLEU metric and the METEOR metric, among others. The BLEU metric is employed to assess the quality of a translation by estimating the degree of overlap between the translated text and a set of references. The METEOR metric integrates the knowledge of word pairing, word order, sentence structure and its meaning, which enhances the transparency of the evaluation output. It evaluates a translation by computing a score based on explicit word-to-word matches between the translation and a reference translation (Banerjee & Lavie, 2005).

Both human and automated evaluations have inherent limitations. Although human evaluation can consider elements of cultural traditions and language usage, the results tend to exhibit a certain degree of subjectivity. The automatic evaluation method is based on pre-set algorithms and rules, which makes it challenging to effectively address complex and diverse translation scenarios and cultural background differences. Consequently, a significant current trend in the research direction of translation quality evaluation is the integration of the respective advantages of human and automatic evaluation, with the objective of proposing a multi-dimensional and comprehensive evaluation methodology.

The advent of big data and artificial intelligence technology in recent years has created novel circumstances for the assessment of translation quality. The application of machine learning to a vast corpus of translation materials and evaluation data enables the construction of a more precise and expedient evaluation model. For instance, Google employs a neural network-based NMT (Neural Machine Translation) model in its translation system, resulting in a notable enhancement in translation quality. Researchers are developing more sophisticated and user-friendly evaluation tools, such as PLATO (Probabilistic and Logic-based Annotation Tool), which can accurately assess the quality of translated content and provide targeted suggestions for improvement based on the results.

The evaluation of translation quality is a multifaceted and complex process that considers several factors, including language, cultural background, context, and others. Considering rapid advancement of technological tools, the evaluation of translation quality is evolving towards a more scientific and systematic approach. As artificial intelligence and big data technology continue to develop, it is anticipated that the methods and tools for evaluating the quality of translation will become more sophisticated and comprehensive, thereby providing more reliable support for cross-cultural interaction and collaboration.

3.2 Automatic evaluation methods

Automatic evaluation techniques occupy a central place in the evaluation of translation quality.

Most evaluations are implemented based on computer technology and statistical modeling. The most important advantages of automatic evaluation systems are reflected in their operational efficiency, objective accuracy and repeatability, which render them particularly valuable in large-scale translation projects.

Among the most utilized automatic evaluation techniques are BLEU, METEOR, TER, and ROUGE metrics. Each of these metrics possesses distinctive properties and is employed in specific contexts. BLEU is the earliest automatic quality evaluation technique to be widely accepted by the general readership. It determines the translation quality by calculating the consistency between the translated text and the reference translation at different positions of the n-grams. Studies have indicated that BLEU scores in the range of 0.3 to 0.6 are regarded as a relatively good level of translation in comparisons across languages. However, BLEU does not capture fluency, semantic similarity, or word order variations and can penalize correct translations with different phrasing (Segonne & Mickus, 2023; Haque et al., 2022).

In recent years, the advent of deep learning techniques and large language models has led to the emergence of novel evaluation tools such as BLEURT and COMET. BLEURT integrates the features of pre-trained language models, which enables the effective capture of complex semantic structures in translated texts, thereby significantly improving performance. COMET employs large-scale cross-language pre-determined training models to more efficiently understand and evaluate the translation standards of translation across languages, especially when dealing with under-resourced language pairs.

Automated evaluation techniques play a pivotal role in the assessment of translation quality, having demonstrated their efficacy in a multitude of ways.

In this study, the quality of translations was evaluated using the BLEU and TER scores calculated by the Shiyibao Translation Evaluation Tool.

3.3 Human Evaluation Methods

In the field of translation quality research, human evaluation methods occupy a central place. Although modern automated evaluation tools such as BLEU and METEOR can provide quantitative evaluations to some extent, the role of human evaluation is unique and indispensable in dealing with the challenges of the detail, cultural context, and contextual dimensions of translation.

3.3.1 Criteria for Evaluation

It is of paramount importance to select appropriate evaluation criteria when conducting human evaluation. The metrics employed in the present study for the purposes of human evaluation are as follows, as detailed in Table 1.

Table 1. Human Evaluation Metrics

Criteria	Description
Accuracy	To access whether the translated content accurately and effectively conveys the original meaning of the source text.
Fluency	To assess whether the translated work follows the grammatical rules of the target language and reads smoothly. Research has confirmed that fluency has a significant impact on readers' comprehension and receptivity.
Fidelity	To assess the fidelity of the translated text to the content of the original text, thereby ensuring that the translated text is devoid of deletions, additions, or alterations. The accuracy of a translation necessitates a meticulous examination of the degree to which the content, depth of information, and tone align. For instance, when translating quotations, technical terms, or common phrases, it is essential to guarantee that the original meaning and artistic style of the original text are preserved.
Adaptability	To assess the capacity of a translation to be effectively adapted to a specific linguistic and cultural environment. Adaptability encompasses the accurate representation and alignment with cultural, historical, and social contexts. When translating a literary work, it is crucial to accurately convey the time and cultural context in which the story is set.

3.3.2 Evaluation Strategies

To conduct a more efficient human evaluation, the following different strategies will be employed during the study, as detailed in Table 2.

Table 2. Human Evaluation Strategies

Evaluation Strategies	Description
Evaluator Questionnaire Method	This strategy permits translators or industry professionals with specialized skills to complete the survey and then evaluate the translations. This strategy facilitates the compilation of comprehensive and detailed quality feedback, which can then be used to quantitatively measure the quality of translations.
Comparative Reading Method	This method entails the analysis and comparison of the translated text with the original text, with the objective of identifying and quantifying the differences in terms of information transfer, style of use, and language expression. This technique is particularly well-suited to the detection of information omissions or translation errors in translated texts.
Error Analysis Method	This method identifies and analyzes errors in the translation in a systematic manner, subsequently evaluating the overall quality of the translation.
Triangulated Review Method	This evaluation method comprises three or more reviewers independently examining the same translation and comparing the two evaluations to arrive at a comprehensive evaluation of the conclusion. This technique is designed to significantly reduce the negative impact of personal preconceptions on the evaluation results.

3.4 Machine Translation Tool Selection

This study aims to compare and analyze the translation quality of ChatGPT-3.5 with that of Google Translate and DEEPL. Google Translate is a well-known online translation tool that has been developed over many years. It has amassed a substantial corpus and a wealth of user feedback data, which has enabled it to become a dependable and precise translation tool. Additionally, DEEPL is renowned for its sophisticated deep learning technology, which enables the generation of more natural and fluid translations.

4 Results Evaluation and Analysis

4.1 Overview of the Source Text

"A Service of Love" is a short literary work by the celebrated American author O. Henry, renowned for his compelling narrative structures and intricate emotional portrayals. His works often captivate readers through their exceptional narration, meticulous structural designs, and unexpected endings. This piece depicts the immense power of love, while simultaneously illustrating the struggles and limitations faced by the common folk in 19th-century America. The novel is not only one of O. Henry's most celebrated works but also a widely acclaimed classic in the American short story circle, with a profound and lasting impact. From a literary perspective, it is a novel with a concise and clear language, DEEPLy colorful and emotional, but also contains thoughtful philosophical meanings. O. Henry's use of subtle dialogue techniques and fine psychic descriptions brings the characters to life, allowing the reader to experience the changes in their hearts and minds in depth.

"A Service of Love" is an in-depth study of the intrinsic connection between love and sacrifice. It posits that sincere love comes from selfless giving and sacrifice, not from material wealth. This central theme has not only gained widespread recognition in American literature but has also inspired far-reaching resonance on an international scale. As an example of the American short story, it has an academic value that cannot be ignored in terms of its translation and dissemination. The novel has been translated into numerous languages and has been universally recognized in

various cultural contexts. In this study, the Chinese translation published by Foreign Language Teaching and Research Press and translated by Nankai University Translation Group and the Japanese translation by Toyama Translation Farm are selected as the reference translations, as they are the most widely disseminated and the most acclaimed, and thus have reference value.

4.2 Evaluation of the Translation from English to Chinese

Table 3 illustrates that the ChatGPT English-to-Chinese translation outperforms Google Translate and DEEPL.

Table 3. BLEU and TER scores for the translations from English to Chinese

Metrics	ChatGPT	Google Translate	DEEPL
BLEU	0.3307	0.1668	0.2690
TER	0.5169	0.7724	0.6297

As illustrated in Table 3, the BLEU score of ChatGPT is higher than that of Google Translate and DEEPL, indicating that the quality of ChatGPT's translation is superior to that of Google Translate and DEEPL. Regarding the TER score, given that the source text is a literary text with a more colloquial content and no definitive translation standard, there is minimal difference between the three in terms of paragraph correspondence and grammatical accuracy. In conclusion, it can be stated that ChatGPT has already demonstrated a certain degree of competitiveness in the field of literary translation in comparison with other translation software. The following is an in-depth analysis of the translation standard of ChatGPT in the test text based on the three aspects of translation accuracy, fluency and readability, integrating the automated evaluation method and the human evaluation method.

First, the translation's accuracy is evaluated to determine whether the translated text is highly consistent with the original text. As demonstrated in Table 3, the BLEU score of ChatGPT is 0.3307, indicating that the translated text exhibits a high degree of consistency with the original text. Nevertheless, it would be advantageous to conduct a more thorough examination of the application of specific words in sentence translation. As demonstrated in Table 4, the translation produced by Google Translate of opening sentence of the novel, is the most accurate. However, it is evident that ChatGPT and DEEPL exhibit varying degrees of mistranslation. Both incorrectly translate the word "service" as "fúwù".

Table 4. Example No.1 of translation comparison from English to Chinese

Source Text	When one loves one's Art, no **service** seems too hard.
Reference Translation	dāng yí gè rén ài zhuó zì jǐ de yì shù shí, zuò chū shén me **xī shēng** dū bù nán.
ChatGPT	dāng yí gè rén rè ài zì jǐ de yì shù shí, sì hū méi yǒu shén me **fú wù** shì tài kùn nan de.
Google Translate	dāng yí gè rén rè ài zì jǐ de yì shù shí, méi yǒu shén me **xī shēng** shì nán yǐ chéng shòu de.
DEEPL	dāng yí gè rén rè ài zì jǐ de yì shù shí, rèn hé **fú wù** dū bú huì xiǎn de tài nán.

From the perspective of fluency, the Gunning Fog Index tool was employed to assess the readability of the text. The results indicated that the original text was scored at 7.8, while the Chinese translation was scored at 8.2. This suggests that the Chinese translation is relatively

suitable for the Chinese context. However, there are still instances where certain sentences do not fully align with the intended Chinese expression. To illustrate, as demonstrated in Table 5, Whether it is produced by ChatGPT, Google Translate or DEEPL, the translation does not fully capture the nuances of the original sentence, particularly regarding the specific details provided.

Table 5. Example No.2 of translation comparison from English to Chinese

Source Text	Joe Larrabee came out of the post-oak flats of the Middle West **pulsing with a genius for pictorial art**.
Reference Translation	qiáo · lā là bǐ lái zì měi guó zhōng xī bù shèng chǎn xīng máo lì de dà píng yuán, **tā hún shēn sàn fà zhuó yì gǔ huì huà yì shù de tiān cái qì xī**.
ChatGPT	qiáo · lā là bǐ lái zì zhōng xī bù de bǎi shù píng yuán, **dài zhuó duì huì huà yì shù de tiān fù**.
Google Translate	qiáo · lā lā bǐ lái zì zhōng xī bù xiàng shù chéng yīn de píng yuán, **yǒu zhe huì huà yì shù de tiān fù**.
DEEPL	qiáo · lài ruì bì cóng zhōng xī bù de xiàng shù hòu píng dì zǒu chū lái shí, **chōng mǎn liǎo huì huà yì shù de tiān fù**.

In the human evaluation section, five native Chinese translators were invited to evaluate the accuracy, fluency, fidelity, and adaptability of the text in depth. The individual evaluations of each participant were then aggregated and presented in Table 6.

Table 6. Human translation evaluation from English to Chinese

Metrics	ChatGPT	Google Translate	DeepL
Accuracy	6.9	7.7	6.8
Fluency	6.8	8.7	7.5
Fidelity	6.9	8.0	7.5
Adaptability	8.8	9.3	8.9

As illustrated in Table 6, the mean score of Google Translate is higher than that of ChatGPT. The mean score for translation accuracy for ChatGPT is 6.9, while that for Google Translate is 7.7. In terms of language fluency, ChatGPT achieved a score of 6.8, while Google Translate attained a score of 8.7. Regarding fidelity, the scores for ChatGPT and Google Translate are 6.9 and 8.0, respectively. In terms of adaptability, ChatGPT achieved an average score of 8.8, while Google Translate scored 9.3. Although there are minor omissions in certain details of the translation, the translation performance of ChatGPT is excellent overall. Nevertheless, the translation of certain cultural metaphors, such as "to keep the chafing dish bubbling," is somewhat underdeveloped in the translation method employed (as demonstrated in Table 7). The translation of ChatGPT " yǐ wéi chí dùn guō de fèi téng " is not without flaws. A more efficacious translation would be " yǐ què bǎo jiā lǐ néng jiē dékāi guō," which more effectively conveys the metaphor's underlying meaning.

Table 7. Example No.3 of translation comparison from English to Chinese

Source Text	to keep **the chafing dish bubbling**.
Reference Translation	yǐ què bǎo jiā lǐ néng **jiē dé kāi guō**.
ChatGPT	yǐ wéi chí **dùn guō de fèi téng**.
Google Translate	cái néng ràng **huǒ guō lǐ rè qì téng téng**.
DEEPL	ràng **zhè pán " dà zá huì " jì xù mào pào**.

In summary, while ChatGPT's BLEU score is higher than those of Google Translate and DEEPL, its human evaluation results are less satisfactory. ChatGPT exhibits a high degree of accuracy and fluency in its English-to-Chinese translations. However, there is room for improvement in its ability to convey emotions and to handle cultural metaphors. Future research should aim to integrate disciplinary knowledge and contextual insights to enhance the model's translation accuracy.

4.3 Evaluation of the Translation from English to Japanese

Table 8 illustrates that the ChatGPT English-to-Japanese translation underperforms Google Translate and DEEPL.

Table 8. BLEU and TER scores for the translations from English to Japanese

Metrics	ChatGPT	Google Translate	DEEPL
BLEU	0.1469	0.1871	0.2434
TER	0.7953	0.7431	0.6734

The automatic evaluation results indicate that ChatGPT performs less effectively in Japanese-to-English translation than in Chinese-to-English translation. DEEPL has the highest BLEU value of 0.2434, but this is still below the 0.3 threshold, indicating that the quality of the translated text is suboptimal.

In the human evaluation section, five experts in Japanese translation assess the translated text in terms of semantic communication accuracy, linguistic fluency, fidelity, and cultural adaptability. Their evaluation results are presented in Table 9.

Table 9. Human translation evaluation from English to Japanese

Metrics	ChatGPT	Google Translate	DeepL
Accuracy	3.7	6.5	5.2
Fluency	5.3	8.1	6.9
Fidelity	5.4	7.2	6.0
Adaptability	9.6	9.6	9.1

The evaluation results indicated that the average score for semantic transfer was 3.7 out of 10. This suggests that there is still room for improvement in conveying information accurately in the translated text. The translated text was found to be significantly less effective than the English-to-Chinese translation. The score of language fluency was 5.3. Experts posited that nearly half of the sentences were translated in a word-for-word manner, and numerous English words were directly translated into Japanese in the form of loanwords, which resulted in unnatural Japanese expressions or even mistranslations in the translated texts. Conversely, it is encouraging to observe that the cultural adaptability score is relatively high, with an average of 9.6. This indicates that in the translation process, ChatGPT has more fully considered the appropriate conversion of certain specific expression patterns in American culture into Japanese expression.

For example, as demonstrated in Table 10, all three translation engines translate it as "Rarabii husai ha huratto de kaji wo haji me ta", which simply means to start doing housework and does not

accurately convey the message of the original text. Moreover, the word "flat" in Japanese should be translated as " apaato ", rather than the English loanword "huratto".

Table 10. Example No.1 of translation comparison from English to Japanese

Source Text	Mr. and Mrs. Larrabee began **housekeeping** in a ***flat***.
Reference Translation	Rarabii husai ha aru ***apaato*** de **ie wo ta i ta**.
ChatGPT	Rarabii husai ha ***huratto*** de **kaji** wo haji me ta.
Google Translate	Rarabii husai ha ***huratto*** de **kaji** wo haji me ta.
DEEPL	Rarabii husai ha ***huratto*** de **kaji** wo haji me ta.

Another example is demonstrated in Table 11. It is evident that the translation of ChatGPT was a direct, word-for-word replication of the source text, which fails to accurately convey the original message. Also, "Art" in Japanese should be translated as " geizyutsu ", instead of the English loanword "aato".

Table 11. Example No.2 of translation comparison from English to Japanese

Source Text	for they had their ***Art***, and they had each other.
Reference Translation	hutari ni ha sorezore no ***geizyutsu*** ga ari, soshite o taga i ga i ta kara de aru.
ChatGPT	kare ra ha jibun tachi no ***aato*** to o taga i wo motte i mashi ta.
Google Translate	kare ra ni ha ***geizyustu*** ga ari, o taga i ga i ta kara da
DEEPL	hutari ni ha ***geizyustu*** ga ari, o taga i ga i ta kara da

In the process of cross-cultural translation, it is important to note that the adaptability and flexibility of cultural background knowledge must be adapted, a process that is still not perfect in ChatGPT-3.5. To gain a more nuanced understanding of readers' receptivity, a non-massive reader questionnaire was administered. The target group for this research consisted of 20 native Japanese speakers who were required to read and evaluate the Japanese translation of ChatGPT in this study. The results of the data analysis indicated that 54% of the readers were able to comprehend and accept the translation results of ChatGPT. Nevertheless, a number of readers expressed the opinion that certain translated content was described in a somewhat formal and unnatural manner that did not fully align with the typical Japanese linguistic style.

Through the analysis of the detailed data and the review by experts, we can make it clear that although ChatGPT has relatively good English-to-Japanese translation functions with the support of large language modeling, there are still some urgent questions to be answered in real application scenarios, such as semantic details and cultural adaptability. Therefore, in-depth research and optimization of large language models, especially the translation methods in line with expressions and background of target languages, will become the core focus of future research.

The analysis of the detailed data and the review by experts have revealed that although ChatGPT has relatively good English-to-Japanese translation functions with the support of large language modeling, there are still some urgent questions to be answered in real application scenarios. These include the need to address semantic details and cultural adaptability. Consequently, future research will concentrate on in-depth analysis and optimization of large language models, with a particular focus on translation methods that align with the expressions and cultural context of target languages.

It is also noteworthy that the translations produced by ChatGPT demonstrate the potential for active learning and continuous improvement. For instance, incorporating a substantial quantity of Japanese cultural background materials and translation cross-reference materials into the future

translation model will facilitate the effective enhancement of translations in cross-cultural contexts. Concurrently, the fine-tuning and special optimization of the model based on the expert opinions and feedback from readers will become a crucial pathway to achieving high-quality translations.

5 Conclusion

This study employs an in-depth and comprehensive analysis and evaluation of ChatGPT's translation quality in a large language modeling environment. The American short story "A Service of Love" serves as a case study, with a primary focus on its translation quality performance in English to Chinese and English to Japanese. Following a quantitative analysis of the BLEU scores for the three language models, it was found that ChatGPT's BLEU scores for English to Chinese and English to Japanese were superior to those of the traditional machine translation models. However, in terms of human evaluation, the mistranslation rate of ChatGPT is higher than that of Google Translate and DEEPL, particularly for English to Japanese. This indicates that there is still room for improvement in terms of sentence structure and syntactic accuracy. The aforementioned factors, namely syntactic accuracy, lexical filtering, and semantic truthfulness, serve to further substantiate the assertion that ChatGPT is currently incapable of attaining the same level of translation proficiency as traditional translation engines and humans, particularly in the context of complex texts and multilingual translation.

A comprehensive examination of the translated texts reveals that ChatGPT exhibits limited proficiency in acquiring cultural and contextual knowledge. In the context of sentences with complex cultural backgrounds and emotional expressions, it is evident that ChatGPT is still unable to accurately convey the depth of emotion and context present in the source text.

A comparison of the performance of ChatGPT, Google Translate, and DEEPL reveals that ChatGPT does not demonstrate a clear advantage in semantic understanding or language generation. In terms of quality and accuracy, ChatGPT is significantly inferior to professional translation engines, particularly in the context of non-Western languages such as Chinese and Japanese.

In summary, while progress has been made in the field of ChatGPT translation quality research, there are still some shortcomings and problems as previously mentioned. Future research in this field should focus on refining evaluation methods for translation quality, expanding the scope of samples, conducting in-depth discussions on the types and causes of errors, and comprehensively evaluating them in conjunction with the latest computational technologies and practical application contexts. These optimization strategies will facilitate the enhancement of the translation capabilities of large language models, while simultaneously providing a robust theoretical foundation for the continued advancement of machine translation capabilities within the context of large language models.

References

[1] Naveed, H., Khan, K U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2024). A Comprehensive Overview of Large Language Models. https://arxiv.org/abs/2307.06435

[2] Sanz-Valdivieso, L., & López-Arroyo, B. (2023). Google Translate vs. ChatGPT: Can non-language professionals trust them for specialized translation? Proceedings of the International Conference onHuman-Informed Translation and Interpreting Technology 2023. https://doi.org/10.26615/issn.2683-0078.2023_008

[3] Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita,H., Kim, Y.J., Afify, M., & Awadalla H.H. (2023). How good are GPT models at machine translation? A comprehensive Evaluation. https://doi.org/10.48550/arXiv.2302.09210

[4] Sahari, Y., Qasem, F., Asiri, E., Alasmri, I., Assiri A., & Mahdi, H. (2023). Evaluating the Translation of Figurative Language: A Comparative Study of ChatGPT and Human Translators. https://doi.org/10.21203/rs.3.rs-3921149/v1

[5] Khoshafah, F. (2023). ChatGPT for Arabic-English Translation: Evaluating the Accuracy (2023). https://doi.org/10.21203/rs.3.rs-2814154/v1

[6] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., ..., Fiedel, N. (2023). Palm: Scaling language modeling with pathways. https://doi.org/10.48550/arXiv.2204.02311

[7] Rivera-Trigueros, I. (2022). Machine translation systems and quality assessment: a systematic review. Language Resources & Evaluation, 56, 593–619. https://doi.org/10.1007/s10579-021-09537-5

[8] Banerjee, S., & Lavie, L. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization (pp. 65–72). https://www.cs.cmu.edu/~alavie/METEOR/pdf/Banerjee-Lavie-2005-METEOR.pdf

[9] Segonne, V., & Mickus, T. (2023). " Definition Modeling: To model definitions. " Generating Definitions With Little to No Semantics. https://doi.org/10.48550/arXiv.2306.08433

[10] Haque, S., Eberhart, Z., Bansal, A., & McMillan, C. (2022). Semantic similarity metrics for evaluating source code summarization. Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension. https://doi.org/10.1145/3524610.3527909