Woody International Publish Limited
An Multidisciplinary Academic Journal Publisher

Journal of Theory and Practice of Humanities Science, Volume 1, Issue 3, 2024
https://www.woodyinternational.com/

Exploring the Potential of ChatGPT-4o in Translation Quality Assessment

Jingjing Wang[1,2]

[1]Kunshan Huaqiao Senior High School, Suzhou, Jiangsu, China
[2]School of Foreign Languages, Nanjing Normal University, Nanjing, Jiangsu, China
**Author to whom correspondence should be addressed.*

Abstract: ***The advancement of large language models (LLMs) has demonstrated significant potential in the domains of foreign language teaching and research. By evaluating two translated works of MTI students, this study discusses the application effect of ChatGPT-4o in the evaluation of human translation based on Multidimensional Quality Metrics (MQM). The research involves literary texts and non-literary texts and conducts human translations before human and ChatGPT-4o modifications. Subsequently, the versions will be evaluated and compared in accordance with MQM standards. Through the score comparison and the qualitative analysis, the results show that ChatGPT-4o demonstrates high consistency with human evaluators in evaluating translations based on the MQM method. The scores and suggested modifications significantly enhance the translation quality, particularly in terms of maintaining terminology consistency and ensuring grammatical accuracy.***

Keywords: Multidimensional Quality Metrics; ChatGPT-4o; Translation quality assessment; Large language models; Language teaching and research.

Cited as: Wang, J. (2024). Exploring the Potential of ChatGPT-4o in Translation Quality Assessment. *Journal of Theory and Practice in Humanities and Social Sciences*, *1*(3), 19–30. Retrieved from https://woodyinternational.com/index.php/jtphss/article/view/27

1. Introduction

With the rise of generative artificial intelligence, large language models (LLMs) have demonstrated powerful capabilities across various fields (Chang et al. 2024). Nowadays, educational informatization has become an inevitable choice in the digital age (Wang and Liu 2023). Numerous studies have shown that LLMs like ChatGPT can empower education and research (e.g. Kasneci et al. 2023; Dai et al. 2023; de Winter 2023; Hellas et al. 2023; Wang and Demszky 2023; Lin and Chen 2024). Due to their strong natural language understanding and generation abilities (Liang et al. 2022; Bubeck et al. 2023; Wu et al. 2023; Bang et al. 2023; Wang et al. 2023), their application domains are continuously expanding. Therefore, foreign language teachers should organically integrate intelligent technologies into translation teaching. In addition to empowering teaching implementation, teaching resources, teaching evaluation, and teaching management (Wang and Xie 2024; Wang and Liu 2023), LLMs like ChatGPT can also perform translation practice tasks such as machine translation and terminology management (Ye 2024; Zhang et al. 2023), significantly improving translation efficiency. This includes comparative studies of ChatGPT translation results with neural network systems (Hidayati and Nihayah 2024; Roza and Zulhirawati 2023; Mohsen 2024; Pang and Wang 2023; Brewster et al. 2024), other LLMs (Rachid 2024), iteration of ChatGPT (Hendy et al. 2023; Siu 2023) or human translation results (Khoshafah 2023; Cao and Zhong 2023) to analyze their strengths and weaknesses. However, only a few studies focus on its use as an evaluator to assess human translation results and provide improvement suggestions (ÜNLÜ 2023).

Assessing translation quality has always been a challenging issue in translation research and practice. Generally, evaluation methods are divided into human evaluation and automatic evaluation. Automatic evaluation typically uses standard metrics and evaluation tools to evaluate model performance (Bang et al. 2023; Lin and Chen 2023; Qin et al. 2023; Wang et al. 2023), and sometimes more comprehensive and accurate feedback is needed through human evaluation (Bubeck et al. 2023; Liang et al. 2022). Both evaluation methods require certain metrics for assessing translation quality, such as translation principles, the LISA QA model, and Multidimensional Quality Metrics (MQM). Among them, MQM is a systematic tool designed to evaluate translation quality comprehensively and objectively based on multidimensional criteria. The MQM, developed jointly by the German Language

Research Center (DFKI) and the Translation Automation User Society (TAUS) (Lommel 2018), has become one of the important standards in the field of Translation Quality Assessment (TQA). MQM consists of a flexible catalogue of 182 error types, therefore, not all of the types are to be covered in the translation assessment, but to ensure a translation "meets specifications" (Lommel 2018). This study aims to explore the similarities and differences between LLMs represented by ChatGPT-4o and human evaluation based on MQM, and to determine whether ChatGPT-4o's scores and modification suggestions can effectively improve translation quality. The results of this study hope to provide a new perspective on TQA, validate the potential of ChatGPT-4o in translation evaluation, and offer valuable references for translation teaching and practice.

2. Literature Review

Due to the wide application of LLMs, such as OpenAI's GPT-4o, in translation teaching and studies, the performance of these models in translation tasks not only includes the generation of high-quality translated texts, but also provides assistance in translation quality assessment (TQA). This article will review the application of LLMs in TQA, and explore its performance in different assessment dimensions such as accuracy, fluency, and fidelity.

2.1 Large Language Models (LLMs)

The application of large language models (LLMs) in natural language processing (NLP) has sparked considerable debate. LLMs have broad applicability and can handle various language tasks, especially in the areas of language understanding and generation (Liang et al. 2022; Bubeck et al. 2023; Wu et al. 2023). However, LLMs, including ChatGPT, yield a lower performance for lower resource languages (e.g. Chowdhery et al. 2022; Muennighoff et al. 2022; Bang et al. 2023; Hendy et al. 2023), some elementary mathematical (Gilson et al. 2022; Frieder et al. 2023; Davis 2023b) and commonsense reasoning tasks (Guo et al. 2023; Davis 2023b). This study views LLMs as tools with broad applicability that require further optimization for translation teaching and assessment.

Due to the powerful capabilities of LLMs, LLMs can also serve as auxiliary tools for evaluating students' translations, providing detailed feedback to help students improve their translation skills. Current research generally agrees that a prompt-based fine-tuning approach can offer optimal performance in the classification tasks (Jung et al. 2023). And the effect of prompting strategies matters on the performance of GPT models for machine translation (Hendy et al. 2023). Despite the great technical potential of LLMs, their application also raises ethical and fairness concerns, moral biases, issues of factuality and hallucination (Bang et al. 2023), which researchers should focus on.

2.2 LLMs in Translation Quality Assessment

Translation quality assessment is a crucial aspect of translation studies, aiming to ensure that the translated text maintains semantic, stylistic, and functional consistency with the source text. Traditional methods of translation evaluation include both automatic and human assessments. As previously mentioned, large language models hold significant potential for application in translation teaching and research, particularly in the realm of machine translation. This study explores how to evaluate and score large language models when used as evaluators.

For automatic evaluation, a human-labeled test dataset was used, followed by a statistical analysis to compare the accuracy, precision, recall, and F1-score of GPT-3.5, GPT-4, and PandaLM against human annotations. Notably, the performance of PandaLM-70B even surpasses that of GPT-4 (Wang et al. 2024). To investigate the potential of large language models, specifically GPT-3.5 Turbo and GPT-4, in assessing interpreting performance (ÜNLÜ 2023), the study employed specific translation evaluation standards. These standards included accuracy, fidelity, and completeness; textual integrity; terminology; and disfluency markers. The analysis also involved the automatic evaluation of interpreters' practice records. Additionally, significant improvements were observed when ChatGPT post-edited the output of DeepL Translator, a paid version of machine translation, in contexts where lexical variety was essential (Farrell, 2023).

Compared with automatic evaluation, human evaluation can provide more comprehensive and accurate feedback. A series of human-crafted tests using GPT-4 were conducted, and researchers found that GPT-4 performs comparably to, or even exceeds, human performance on multiple tasks (Bubeck et al. 2023). Beyond ChatGPT, the MAPS method, which enables large language models (LLMs) to mimic human translation strategies to achieve high-quality translations, was introduced (He et al. 2024). In this experiment, human preference studies and

Multidimensional Quality Metrics (MQM) evaluations were conducted as part of the human evaluation process. The results showed that MAPS is generally more preferred by humans and provides more favorable translations by reducing mistranslations, awkward style, untranslated text, and omission errors. Additionally, a comprehensive analysis and human evaluation, combining source-based sentence-level contrastive Direct Assessment and Scalar Quality Metric, were conducted to further understand the characteristics of GPT translation (Hendy et al. 2023).

Since evaluation criteria are fundamental components of the human assessment process (Chang et al. 2024), the main metrics are summarized as follows based on the literature that employed both automatic and human evaluation:

(1) Accuracy: This metric assesses whether the translation accurately conveys the meaning of the original text, including correctness in vocabulary, grammar, and syntactic structure. Most evaluations focus on semantic consistency and the correct use of specialized terminology.

(2) Fluency: A fluent text is not only grammatically correct but also ensures readability and a seamless user experience (Chang et al. 2024). The translation should read smoothly and naturally (Lee et al. 2019), adhering to the linguistic norms of the target language and ensuring cohesion and coherence.

(3) Fidelity: This criterion evaluates whether the translation remains faithful to the style, register, and intent of the original text. It assesses whether the translation retains the emotional tone and cultural nuances of the original.

2.3 The Present Research

Previous studies have shown that large language models (LLMs) have demonstrated promising results in machine translation (MT). However, there is limited research on applying LLMs to Translation Quality Assessment (TQA) and Automatic Post-Editing (APE). Most existing studies focus on ChatGPT's performance as a translation tool, lacking systematic exploration of its potential and effectiveness in evaluating translation quality and automatically correcting translation errors. In particular, studies using the Multidimensional Quality Metrics (MQM) method to assess ChatGPT's evaluation capabilities are rare.

This research aims to fill this gap by exploring the effectiveness of ChatGPT-4o in assessing human translations based on the MQM method and comparing its evaluations with those of human evaluators. Specifically, this study will evaluate ChatGPT-4o's performance in terms of accuracy, fluency (especially cohesion and coherence), and fidelity, analyzing how its scores and modification suggestions impact translation quality. Additionally, the study will examine the advantages and limitations of ChatGPT-4o in translation quality assessment, providing new insights and methods for translation teaching and practice. Through this research, we aim to systematically delve into ChatGPT's potential in translation quality assessment. The following two questions are addressed in this research:

RQ1: To what extent are the ChatGPT-4o evaluation results consistent with human evaluation results?
RQ2: Can the ratings and recommendations of ChatGPT-4o effectively improve the quality of translations?

3. Evaluation Criteria and Prompt Generation

This section will explain the mechanism by which the GPT-4o model generates feedback and the assessment criteria used to evaluate student translations.

3.1 Evaluation Criteria

As shown in Figure 1, to simplify the work of translation evaluation, MQM provides a "core" of 19 questions applicable to many generic evaluations of plain text translation (Burchardt et al. 2013). Besides, the following basic formula is used for calculating MQM quality scores in an error-count environment:

$$TQ = 100 - AP - (FP_T - FP_S) - (VP_T - VP_S)$$

Where:

TQ = quality score. The overall rating of quality
AP = penalties for Accuracy. Sum of all weighted penalty points assigned in the Accuracy branch

FP_T = Fluency penalties for the target. Sum of all weighted penalty points in the target text assigned to the Fluency branch. (Note: for computational purposes, Design and Internationalization are treated with Fluency.)
FP_S = Fluency penalties for the source. Sum of all weighted penalty points in the source text assigned to the Fluency branch. If the source is not assessed $FP_S = 0$.
VP_T = Verity penalties for the target. Sum of all weighted penalty points in the target text assigned to the Verity branch.
VP_S = Verity penalties for the source. Sum of all weighted penalty points in the source text assigned to the Verity branch. If the source is not assessed $VP_S = 0$.

Given the textual characteristics of the student translations selected for this experiment, a crucial aspect of fluency assessment involves evaluating the cohesion, coherence, and naturalness of the text. In addition, the term "verity" in the evaluation formula has been replaced with "fidelity," denoted as (fP_T), to emphasize whether the translation retains the emotional tone and cultural nuances of the original text. Since the fluency and verity of the source text are not assessed in this study, FP_S and VP_S are set to zero. Thus, the formula for Translation Quality (TQ) is transformed into:

$$TQ = 100 - AP - FP_T - fP_T.$$

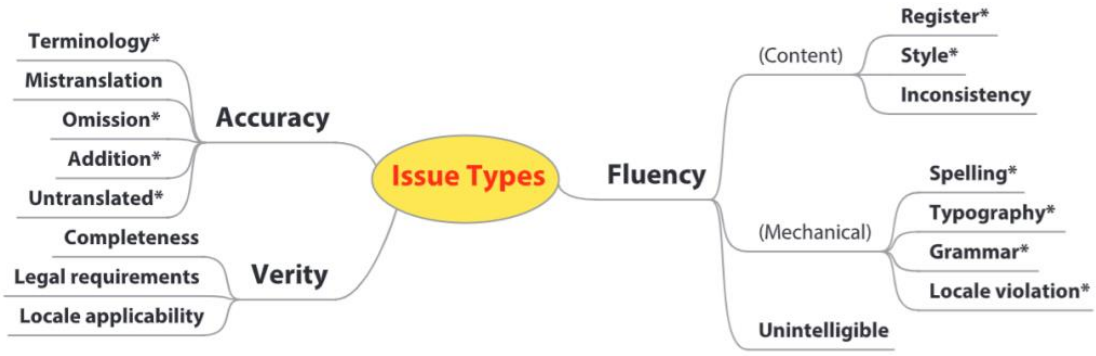

Figure 1: MQM Core

3.2 Prompting

When giving a ChatGPT prompt, users need to define the goal of the conversation, be careful to use words accurately and concisely, provide context, and specify exact roles and keep the conversation on track (Atlas 2023). In our case, we adopted the BROKE frame (Wang and Xie 2024), namely:

(1) State the background: I am a translation teacher at the university, please make a detailed and objective assessment of students' translations.

(2) Define the Role: You will be an expert translator.

(3) Define Objectives: Please evaluate the translations by answering specific questions according to the evaluation method of MQM indicators, including accuracy, fluency and fidelity. The number of errors in the three categories should be pointed out respectively, and the final score was subtracted by 100 points from the number of errors in the three categories.

(4) Define the Key Result:

Accuracy

-a) Terminology: Is the terminology in the translation consistent with that in the source text?
-b) Mistranslation: Are there any translation errors, such as mistranslations or misinterpretations of the source text?
-c) Omission: Are there any omissions of information from the source text in the translation?
-d) Addition: Does the translation include any extraneous information not present in the source text?
-e) Untranslated: Are there any untranslated parts in the translation?

Fluency

-a) cohesion & coherence: Is the text logically smooth and coherent?
-b) naturalness: Is the text naturally readable and smooth?

Fidelity

-a) emotional tone: Does the translation retain the emotional tone of the original text?
-b) cultural nuances: Does the translation maintain the cultural nuances of the original while complying with the requirements of the target language country or region?

(5) Test and Evolve: Please revise the evaluation results according to my feedback.

4. Case Study

4.1 Methodology

The case study primarily evaluates two translations by an MTI student: the Chinese-to-English translation of the literary work The Peach Colony and the English-to-Chinese translation of Global Wage Report 2020–21: Wages and minimum wages in the time of COVID-19 (excerpt of 838 words). These translations were then evaluated by three human evaluators and ChatGPT-4.0. The source and target texts from the student are presented in Appendix I and Appendix II, respectively.

Two translation teachers from a university and a translator from a translation company were provided with the source text and reference translations, which were evaluated and scored according to the Multidimensional Quality Metrics (MQM) framework. The teachers input the source text and the student's translation into ChatGPT-4o, which then evaluated and scored the translations based on the proposed prompts.

Given the limited number of assessors and the small sample size, this case study employs score comparison and qualitative analysis to compare the advantages and disadvantages of human assessment and ChatGPT-4o in terms of accuracy, fluency, and fidelity.

4.2 Results and Discussion

The experimental results show that ChatGPT4o can provide detailed and accurate feedback for the translation, and in terms of scoring, ChatGPT4o is consistent with the results of human evaluation (see Appendix III&IV). Table 1 and 2 present the number of errors and the overall rating of quality (TQ) assigned by three human evaluators and ChatGPT-4o.

Table 1: Error Counts and TQ of Literary text

Evaluator	Accuracy	Fluency	Fidelity	TQ
Evaluator 1	6	2	4	88
Evaluator 2	7	2	5	86
Evaluator 3	8	3	4	85
ChatGPT4o	7	3	4	86

Table 2: Error Counts and TQ of Non-literary text

Evaluator	Accuracy	Fluency	Fidelity	TQ
Evaluator 1	5	1	1	93
Evaluator 2	4	2	1	93
Evaluator 3	3	2	1	94
ChatGPT4o	4	1	0	95

a) Accuracy

For both texts, accuracy scores show minor discrepancies among evaluators, but overall they align closely. For the literary text, omission and minor mistranslations are the primary issues in terms of accuracy, resulting in errors between 6 and 8. For the non-literary text, accuracy is slightly higher, ranging from 3 to 5, indicating better consistency in terminology and fewer mistranslations, less errors in omission, addition, and untranslated. ChatGPT-4o's errors count (7 for the literary text and 4 for the report) demonstrate its ability to identify and

evaluate different types of texts effectively.

b) Fluency

Fluency scores indicate minor variation, where errors range from 2 to 3 in terms of literary text. The evaluation result of evaluator 3 is similar to that of ChatGPT-4o. In contrast, the non-literary report shows higher fluency scores where all human evaluators counted similar errors, reflecting the consistency between assessing on this kind of text. ChatGPT-4o's fluency assessments (3 for the literary text and 1 for the report) indicate the student has a strong ability to maintain readability and cohesion, particularly in the more structured non-literary text.

c) Fidelity

Compared with the non-literary report, errors in fidelity are consistently high in The Peach Colony, indicating the difficulty in accurately conveying the source content and cultural nuances in literary texts. For instance, ChatGPT-4o suggested that the translation captures the serene and idyllic tone of the original text. However, some emotional subtleties are slightly lost, such as the awe and wonder of the fisherman. For the wage report, ChatGPT-4o did not find any errors, suggesting a high level of recognition of the translation than human evaluators.

Overall, the MQM evaluation reveals that while both human evaluators and ChatGPT-4o provide valuable assessments, ChatGPT-4o demonstrates a consistent ability with humans to identify and address translation issues across different text types, which is in line with the results of relevant researches (Bubeck et al. 2023; He et al. 2024). The results highlight the challenges in producing high-quality translations that are accurate, fluent, and faithful to the original literary texts. Despite the good alignment between human judgement, the usability of such a tool for scoring is still questionable (ÜNLÜ 2023). Further optimization in prompting and fine-tuning of the model can generate useful output (Jung et al. 2023). Therefore, the combined results can help students to modify the translation according to the assessment, accurately identify the types of errors, and improve their translation ability.

5. Conclusion and Future Developments

This study explored the application of ChatGPT-4o in evaluating MTI students' translations based on the Multidimensional Quality Metrics (MQM) method. Score comparisons and qualitative analyses revealed that ChatGPT-4o's ratings were highly correlated with those of human evaluators and were often more objective in certain dimensions. The modification suggestions provided by ChatGPT-4o significantly improved translation quality. These findings address the research questions of this study. Although ChatGPT-4o generally provided effective assessments and suggestions in specific fields and cultural contexts, its evaluation results need to be complemented by human evaluators to ensure comprehensive translation quality (Dai et al. 2023).

The significance of this study lies in validating the application value of ChatGPT-4o in translation quality assessment and providing new insights and methods for translation teaching and practice. Specifically, ChatGPT-4o can serve as an auxiliary tool in translation teaching, helping teachers evaluate students' translations. Additionally, students can conduct self-assessments and rectify translation errors, thereby improving their translation skills. Translation companies and professional translators can utilize ChatGPT-4o to assist human reviewers with preliminary assessments and modifications, increasing efficiency and translation quality.

However, this study has limitations that should be considered in future research. First, the sample size was small, evaluating only two translation works by MTI students, which may affect the generalizability of the results. More empirical research with fine-tuned LLM models using larger datasets is recommended for future studies. Second, the study focused solely on two types of texts and Chinese-English language pairs. The potential of LLMs in assessing translations should be explored in other high-resource and low-resource languages (Bang et al. 2023; Hendy et al. 2023). Third, this study focused on the feasibility of LLMs in evaluating written translation, while similar studies could be adapted for evaluating multimodality in LLMs, sentiment analysis, and language identification tasks (Bang et al. 2023).

Due to the randomness, state dependencies, fine-tuning of the model, and external factors within the large language model, ChatGPT-4o has certain limitations and disadvantages, such as scoring the same text with different results. Additionally, it is essential to ensure that task-specific cues keep the model free of hallucinations, ethical biases, and dangerous content. By further optimizing and refining ChatGPT-4o's evaluation algorithms, we expect it to

drive comprehensive improvements in translation quality.

References

[1] Atlas, S. 2023. "ChatGPT for higher education and professional development: A guide to conversational AI". https://digitalcommons. uri.edu/cba_facpubs/548.

[2] Burchardt, A. 2013. "Multidimensional quality metrics: a flexible system for assessing translation quality." In Proceedings of Translating and the Computer 35: 1-7.

[3] Bang, Y. et al. 2023. "A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity." arxiv preprint arxiv: 2302.04023. https://doi.org/10.48550/arXiv.2302.04023.

[4] Brewster, R. C. et al. 2024. "Performance of ChatGPT and Google Translate for Pediatric Discharge Instruction Translation." Pediatrics. 154(1): e2023065573. https://doi.org/10.1542/peds.2023-065573.

[5] Bubeck, S. et al. 2023. "Sparks of artificial general intelligence: Early experiments with gpt-4." ArXiv preprint, abs/ 2303.12712. https://doi.org/10.48550 /arXiv.2303.12712

[6] Chang, Y. et al. 2024. "A survey on evaluation of large language models." ACM Transactions on Intelligent Systems and Technology. 15 (3): 1-45. https://doi.org/10.1145/3641289.

[7] Chowdhery, A. et al. 2023. "Palm: Scaling language modeling with pathways." Journal of Machine Learning Research. 24(240): 1-113.

[8] Cao, S. and Zhong, L. 2023. "Exploring the effectiveness of ChatGPT-based feedback compared with teacher feedback and self-feedback: Evidence from Chinese to English translation." arXiv preprint arXiv:2309.01645. https://doi.org/10.48550/arXiv.2309.01645

[9] Dai, W. et al. 2023. "Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT." 2023 IEEE International Conference on Advanced Learning Technologies (ICALT). pp.323-325. doi: 10.1109/ICALT58122.2023.00100.

[10] Davis, E. 2024. "Mathematics, word problems, common sense, and artificial intelligence." Bulletin of the American Mathematical Society. 61(2): 287-303.

[11] de Winter, J.C. 2023. "Can ChatGPT Pass High School Exams on English Language Comprehension?" International Journal of Artificial Intelligence in Education. 1-16. https://doi.org/10.1007/s40593-023-00372-z.

[12] Frieder, S. et al. 2024. "Mathematical capabilities of chatgpt." Advances in neural information processing systems. 36.

[13] Gilson, A. et al. 2022. "How well does ChatGPT do when taking the medical licensing exams? The implications of large language models for medical education and knowledge assessment." MedRxiv. 23.

[14] Guo, B. et al. 2023. "How close is chatgpt to human experts? comparison corpus, evaluation, and detection." arxiv preprint arxiv:2301.07597. https://doi.org/10.48550/arXiv.2301.07597.

[15] Hendy, A. et al. 2023. "How good are gpt models at machine translation? a comprehensive evaluation." arXiv preprint arXiv: 2302.09210. https://doi.org/10.48550 /arXiv.2302.09210.

[16] Hellas, A. et al. 2023. "Exploring the responses of large language models to beginner programmers' help requests." In Proceedings of the 2023 ACM Conference on International Computing Education Research. (1): 93-105. https://doi.org/10.1145/3568813.3600139.

[17] Hidayati, N. N. and Nihayah, D. H. 2024. "Google Translate, ChatGPT or Google Bard AI: A Study toward Non-English Department College Students' Preference and Translation Comparison." Inspiring: English Education Journal, 7(1), 14-33. https://doi.org/10.35905/inspiring.v7i1.8821

[18] Jung, D. et al. 2023. "Enhancing Machine Translation Quality Estimation via Fine-Grained Error Analysis and Large Language Model." Mathematics. 11(19). 4169. https://doi.org/10.3390/math11194169.

[19] Kasneci, E. et al. 2023. "ChatGPT for good? On opportunities and challenges of large language models for education." Learning and individual differences. 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274.

[20] Khoshafah, F. 2023. "Chatgpt for arabic-english translation: Evaluating the accuracy." 1-20. https://doi.org/10.21203/rs.3.rs-2814154/v2.

[21] Lommel, A. 2018. Metrics for Translation Quality Assessment: A case for standardising error typologies. In Moorkens, J., et al. (eds.) Translation Quality Assessment, Machine Translation: Technologies and Applications vol. 1, pp. 109–128. Springer, Switzerland.

[22] Liang, P. et al. 2022. "Holistic evaluation of language models." arxiv preprint arxiv:2211.09110. https://doi.org/10.48550/arXiv.2211.09110.

[23] Lin, Z., & Chen, H. 2024. "Investigating the capability of ChatGPT for generating multiple-choice reading comprehension items." System. 123: 103344. https://doi.org/10.1016/j.system.2024.103344.

[24] Farrell. M. 2023. “Preliminary evaluation of ChatGPT as a machine translation engine and as an automatic post-editor of raw machine translation output from other machine translation engines.” Proceedings of the International Conference HiT-IT 2023, pages 108–113. https://doi.org/10.26615/issn.2683-0078.2023_007
[25] Mohsen, M. 2024. “Artificial Intelligence in Academic Translation: A Comparative Study of Large Language Models and Google Translate.” PSYCHOLINGUISTICS, 35(2), 134-156. https://doi.org/10.31470/2309-1797-2024-35-2-134-156.
[26] Muennighoff, N. et al. 2022. “Crosslingual generalization through multitask finetuning.” arxiv preprint arxiv: 2211.01786. https://doi.org/10.48550/arXiv.2211.01786.
[27] Pang, Y. and Wang, X. 2023. “A Study on the Translation Quality of ChatGPT in the Context of Large Language Model——A Case Study of Shaanxi Local LiteratureLife(Excerpt).” Modern English. (22): 67-70. doi:CNKI:SUN:XDYM.0.2023-22-019.
[28] Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., and Yang, D. 2023. “Is ChatGPT a general-purpose natural language processing task solver?” arxiv preprint arxiv:2302.06476. https://doi.org/10.48550/arXiv.2302.06476
[29] Rachid. E. D. 2024. “Comparative Analysis of Copilot 4 and Chatgpt 4 for Literary Translation: A Comprehensive Evaluation.” Available at SSRN 4782157. http://dx.doi.org/10.2139/ssrn.4782157.
[30] Roza, V. and Zulhirawati, Z. 2023. “Higher Students’ Perception of Using Chat GPT in Translating English Texts.” BiCED Proceeding, 1: 64–73. Retrieved from https://proceedings.uinbukittinggi.ac.id /biced/article/view/278
[31] Siu, S. C. 2023. “ChatGPT and GPT-4 for Professional Translators: Exploring the Potential of Large Language Models in Translation.” 10.2139/ssrn.4448091.
[32] ÜNLÜ, C. 2023. “Interpretutor: Using large language models for interpreter assessment.” Proceedings of the International Conference HiT-IT 2023, pages 78–96. https://doi.org/10.26615/issn.2683-0078.2023_007
[33] Van Der Lee, C., Gatt, A., Van Miltenburg, E., Wubben, S., & Krahmer, E. 2019. “Best practices for the human evaluation of automatically generated text.” In Proceedings of the 12th International Conference on Natural Language Generation (pp. 355-368). DOI: 10.18653/v1/W19-8643
[34] Widiatmika, P. W., et al. 2023. “Examining the result of machine translation for linguistic textbook from English to Indonesian.” In proceeding the second english national seminar “exploring emerging technologies in english education”. 54-65. LPPM Press STKIP PGRI PACITAN.
[35] Wang, H. S and Xie, F. 2024. “A Study on the Innovation of Translation Education Practice Models Driven by Large Language Model Technology.” Chinese Translators Journal, (02), 70-78.
[36] Wang, R. E. and Demszky, D. 2023. “Is chatgpt a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction.” arxiv preprint arxiv: 2306.03090. https://doi.org/10.48550/arXiv.2306.03090.
[37] Wang, Z. et al. 2023. “Is ChatGPT a good sentiment analyzer? A preliminary study.” arxiv preprint arxiv:2304.04339. https://doi.org/10.48550/arXiv.2304.04339.
[38] Wu, H. et al. 2023. “Chatgpt or grammarly? evaluating chatgpt on grammatical error correction benchmark.” arxiv preprint arxiv:2303.13648. https://doi.org/10.48550/arXiv.2303.13648.
[39] Wang, Y. D. et al. 2023. “PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization.” arXiv preprint arXiv:2306.05087.
[40] https://doi.org/10.48550/arXiv.2306.05087
[41] Ye, L. 2024. “The Feasibility Study of Artificial Intelligence ChatGPT in Translation Field.” Frontiers in Computing and Intelligent Systems, 8(1), 52-57.
[42] Zhang, B., Haddow, B., and Birch, A. 2023. “Prompting large language model for machine translation: A case study.” In International Conference on Machine Learning (pp. 41092-41110). PMLR.

Appendix I The Peach Colony

Source text	Target text
《桃花源记》 晋太元中，武陵人捕鱼为业. 缘溪行，忘路之远近. 忽逢桃花林，夹岸数百步，中无杂	Fairyland of Peach Blossom/A Story of Peach Blossom Source During the period of Taiyuan (A.D.376-396) in Eastern Jin Dynasty, a man in Wuling County made his living by fishing. When he was rowing down a stream, he forgot to notice how far he had rowed. Suddenly a peach blossom

树, 芳草鲜美, 落英缤纷, 渔人甚异之, 复前行, 欲穷其林.	forest that grew along the banks of the stream came into view. Hundreds of paces long, there were no other trees mingled in it where boasted fragrant flowers and fresh grass. The falling peach blossom was scattering on the ground where covered with fallen petals. Surprised and charmed by the scene, the fisherman went on rowing to explore the end of the forest.
林尽水源, 便得一山, 山有小口, 仿佛若有光. 便舍船, 从口入. 初极狭, 才通人. 复行数十步, 豁然开朗. 土地平旷, 屋舍俨然, 有良田美池桑竹之属. 阡陌交通, 鸡犬相闻. 其中往来种作, 男女衣着, 悉如外人. 黄发垂髫, 并怡然自乐.	The peach grove ended at the source of the stream where he saw a hill with a small cave in it. And there seemed to be a light looming up in the cave. So he left his boat and went into the cave. At first the cave was too narrow for only one person to pass through. After a few more steps, it suddenly became wide and bright. What came into sight were a flat expanse of land with neat rows of houses, fertile fields, beautiful ponds, mulberries and bamboo groves etc. The paths in fields intersected each other and crowing and barking among hamlets can be heard everywhere. Besides, farmers were busy with their farm work; what men and women dressed were similar with people outside the Peach Blossom; the old and the young were all at ease.
见渔人, 乃大惊, 问所从来. 具答之. 便要还家, 设酒杀鸡作食. 村中闻有此人, 咸来问讯. 自云先世避秦时乱, 率妻子邑人来此绝境, 不复出焉, 遂与外人间隔. 问今是何世, 乃不知有汉, 无论魏晋. 此人一一为具言所闻, 皆叹惋. 余人各复延至其家, 皆出酒食. 停数日, 辞去. 此中人语云: “不足为外人道也.”	When the villagers saw the fisherman, they were quite amazed and asked him where he had come from. The fisherman answered in detail and was invited to their home with wine, chicken and staple food. When other villagers heard of the arrival of such a man, they came to ask for information. They said that this was how their ancestors led their wives, children and neighbors to this place isolated from the outside world without going out to escape the wars and chaos during the Qin period (221BC-208BC). Thus, they terminated their contact with people outside. They asked the fisherman what dynasty it was now, but they didn’t even know there once had been Han Dynasty (206BC-220AD), not to mention the Wei and Jin dynasties (220AD-420AD). Therefore, the fisherman told them all the things he knew about the outside world in detail, which they sighed with amazement and lament after hearing. Then the fisherman was invited by the rest of the villagers and treated with food and wine. After a few days, he took leave of the villagers who begged him not to tell people outside this place and what he had experienced.
既出, 得其船, 便扶向路, 处处志之. 及郡下, 诣太守, 说如此. 太守即遣人随其往, 寻向所志, 遂迷, 不复得路.	Having came out of the peach blossom, the fisherman found his boat and rowed according to how he had come with careful mark. Soon he arrived at Wuling County and went to the prefecture. He reported his experience to the prefect who immediately sent his men to follow him. They looked for the marks he had made before to reach the grove. Finally, they got lost and failed to find the way to the Peach Blossom.
南阳刘子骥, 高尚士也, 闻之, 欣然规往. 未果, 寻病终, 后遂无问津者.	Liu Ziji, a virtuous and ambitious hermit from Nanyang, heard this story and planned to find the place with pleasure and enthusiasm. But he didn’t achieve his goal and he died of illness soon after. Since then, no one dare to explore the way to the Peach Blossom.

Appendix II Global Wage Report 2020–21: Wages and minimum wages in the time of COVID-19 (excerpt)

Source Text	Target Text
Part I. Recent trends in wages	**第一部分. 近期工资趋势**
In the four years preceding the COVID-19 pandemic (2016–19), global wage growth fluctuated between 1.6 and 2.2 per cent; when China is excluded from the sample, real wage growth in those four years fluctuated at a lower level, between 0.9 and 1.6 per cent. In advanced G20 economies, real wage growth fluctuated between 0.4 and 0.9 per cent, while rising more rapidly – between 3.5 and 4.5 percent annually – in emerging G20 countries. Between 2008 and 2019, real wages more than doubled in China. Among advanced G20 economies, wage growth accelerated the most (by 22 per cent) in the Republic of Korea, followed by	**在新冠肺炎疫情爆发的前四年 (2016—19), 全球工资增长在 1.6%至 2.2%之间; 如果把中国排除在外, 在这四年间, 实际工资增长则处于较低水平, 0.9%至 1.6%.** 在二十国集团发达经济体中, 实际工资增长在 0.4%至 0.9 之间波动, 然而二十国集团新兴经济体的增速更快, 年增长率在 3.5%至 4.5%之间. 2008 年至 2019 年间, 中国的实际工资增长了一倍. 在二十国集团发达经济体中, 韩国的工资增速最快

Germany (15 per cent). By contrast, real wages declined in Italy, Japan and the United Kingdom.	(22%), 其次是德国(15%). 相比下，意大利、日本和英国的实际工资急剧下降.
In the first half of 2020, as a result of the COVID-19 crisis, a downward pressure on the level or growth rate of average wages was observed in two thirds of the countries for which recent data are available; in other countries average wages increased, largely artificially as a reflection of the substantial job losses among lower-paid workers. In times of crisis, average wages can be significantly skewed by sharp changes in the composition of employment – what is known as the "composition effect". In Brazil, Canada, France, Italy and the United States, average wages have been rising markedly because of job losses mainly affecting those at the lower end of the wage scale. In contrast, a downward pressure on average wages has been observed in Japan, the Republic of Korea and the United Kingdom. In countries where strong job retention measures have been introduced or extended to preserve employment, surges in unemployment have been moderated, such that the effects of the crisis may have been more apparent through a downward pressure on wages than through massive job losses.	**2020 年上半年，由于疫情，在有最新数据的国家中，三分之二的国家平均工资水平或增长率面临下行压力；在其他国家，平均工资的增长反映出低薪工人的大量失业.** 在危机时期，平均工资可能会因就业结构的急剧变化而大幅扭曲，这就是所谓的"结构效应". 在巴西、加拿大、法国、意大利和美国，平均工资一直在显著上升，因为失业主要影响那些工资水平较低的人. 相比之下，日本、韩国和英国注意到平均工资的下行压力. 为维持就业，一些国家已经采取或提供强有力的工资补助措施，因而缓解了失业率的激增. 因此，危机的影响可能更多地体现在对工资的下行压力上，而不是大规模失业上.
The impacts of the crisis on total wages have fallen differently on men and women, the latter being disproportionately affected. Looking at a selection of European countries, the report estimates that without the payment of wage subsidies, workers would have lost 6.5 per cent of their total wage bill between the first and second quarters of 2020. For women, the total wage bill would have declined by 8.1 percent, compared to a decline of 5.4 percent for men. Such a discrepancy was mainly caused by reduced working hours, more than by the difference in the number of lay-offs. The wage bill lost as a result of the drop in working hours was 6.9 per cent for women compared to 4.7 per cent for men.	**危机对男性和女性工资总额的影响差异较大，女性的工资总额受到严重的影响.** 纵观选定的欧洲国家，该报告估计，如果不支付工资补贴，工人在 2020 年第一和第二季度的工资总额将减少 6.5%. 对女性而言，工资总额将下降 8.1%，男性的工资总额将下降 5.4%. 主要原因是工作时数的减少，而不是裁员人数的差异. 由于工作时数减少，女性的工资额下降 6.9%，而男性的工资额下降 4.7%.
The crisis disproportionately affected lower-paid workers, thereby increasing wage inequalities. Studies have shown that in many countries, reductions in hours worked have impacted lower-skilled occupations – in particular those in elementary work – more than higher-paying managerial and professional jobs. For selected European countries, the report estimates that without wage subsidies the lowest-paid 50 per cent of workers would have lost an estimated 17.3 per cent of their wages, which is much more than the estimated 6.5 per cent decline for all workers. Consequently, the share of the total wage bill received by those in the bottom 50 per cent of the wage distribution – a measure of inequality – would have fallen by about 3 percentage points, from 27 to 24 per cent on average of the total wage bill, while the share of the upper half of the distribution would have risen from 73 to 76 per cent.	**疫情异常严重地影响了低薪工人，从而加剧了工资不平等.** 研究表明，许多国家，工作时数减少影响了低技能职业工作者，特别是那些基层工作者，高于高薪管理者和专业人士. 在选定的欧洲国家中，如果没有工资补贴，收入最低的 50%工人的工资将减少 17.3%，这比所有工人 6.5%的工资降幅要大得多. 因此，处在工资分配 (一种衡量不平等程度的指标) 后 50%的群体工资总额所占的份额将下降约 3 个百分点，即平均工资总额从 27%降至 24%，而处在前 50%的群体工资总额将从 73%升至 76%.
However, temporary wage subsidies have enabled many countries to compensate part of the wage bill that would have been lost, and to lessen the effect of the crisis on wage inequality. Many countries across the world have either introduced or expanded existing wage subsidies in order to safeguard jobs during the crisis. In a selection of ten European countries for which data are available, the report estimates that wage subsidies have permitted to compensate 40 per cent of the	**然而，许多国家已通过临时工资补贴补偿可能损失的部分工资额，并因此减轻危机对工资不平等的影响.** 为了在危机期间保障就业，全球许多国家已推行或扩大现有的工资补贴. 报告估计，在选定的 10 个欧洲国家中，工资补贴可弥补工资总额损失的 40%，其中包括因工作时数减少而导致

total wage bill loss, including 51 per cent of the wage bill loss caused by the reduction in working hours. Wage subsidies have also permitted to moderate the effects of the crisis on earnings inequalities because the main beneficiaries were those who have been more severely hit by the crisis, namely workers in lower-paying jobs.	51%的工资损失. 工资补贴也有助于缓解疫情对收入不平等现象的影响, 因为该政策的主要受益者是受疫情影响较严重的低收入工作者.
With a view to supporting low-paid workers, many countries with regular minimum wage adjustments went ahead with planned increases in the first half of 2020. Analysis reveals that in the 60 countries that adjust minimum wages on a regular basis, all the adjustments scheduled for the first quarter of 2020 took place as expected, whereas 6 out of 9 countries that usually adjust in the second quarter kept to the scheduled adjustment date, in the midst of the crisis. Among the 87 countries that adjust minimum wages irregularly, 12 increased their minimum wages in the first half of 2020 – a lower number than in the previous year. This suggests that the COVID-19 crisis may have induced some governments to postpone potential adjustments this year.	**为了支持低薪工人, 许多定期调整最低工资的国家在 2020 年上半年按计划提高最低工资.** 分析显示, 在 60 个定期调整最低工资标准的国家中, 所有的调整计划于 2020 年第一季度如期进行, 而通常于第二季度调整的 9 个国家中, 有 6 个在疫情期间仍按计划日期调整. 在 87 个不定期调整最低工资标准的国家中, 有 12 个国家在 2020 年上半年提高了最低工资, 低于前一年. 这表明, 疫情可能已经导致一些政府推迟了调整.

Appendix III ChatGPT4o Evaluation (literary text)

(1) Accuracy
a) Terminology: The term “Peach Blossom Source” could be more accurately translated as “Peach Blossom Spring,” which is a more commonly accepted translation. “Wuling County” is accurately translated, maintaining the original place name.
Error Count: 1
b) Mistranslation: “During the period of Taiyuan” accurately translates the time period but could be more precise in historical context. “A man in Wuling County made his living by fishing” translates well but “rowing down a stream” should be “following the stream.” The phrase “Suddenly a peach blossom forest that grew along the banks of the stream came into view” could be more accurately rendered to emphasize the suddenness and exclusivity of peach blossoms.
Error Count: 3
c) Omission: Some descriptive elements and nuances are slightly downplayed, such as the vivid description of the peach blossom forest being free of other trees.
Error Count: 2
d) Addition: The addition of explanatory phrases like “a virtuous and ambitious hermit” for Liu Ziji adds context not explicitly in the source text.
Error Count: 1
e) Untranslated: The text is fully translated with no parts left untranslated.
Error Count: 0

(2) Fluency
a) Cohesion & Coherence: The translation maintains logical flow and coherence. The story progresses smoothly with clear connections between sentences and paragraphs.
Some sentences could be better connected to enhance flow.
Error Count: 1
b) Naturalness: The text reads naturally for the most part but has occasional awkward phrasings that disrupt the flow, such as “Surprised and charmed by the scene” which could be simplified to “Amazed by the scene.”
Error Count: 2

(3) Fidelity
a) Emotional Tone: The translation captures the serene and idyllic tone of the original text. However, some emotional subtleties are slightly lost, such as the awe and wonder of the fisherman.
Error Count: 2

b) Cultural Nuances: The translation does a good job of retaining cultural elements, though some finer details about historical context and lifestyle might need more precise wording, such as the significance of "阡陌交通" (crisscross paths in fields).
Error Count: 2

Summary
Accuracy:
Terminology: 1 error
Mistranslation: 3 errors
Omission: 2 errors
Addition: 1 error
Untranslated: 0 errors
Fluency:
Cohesion & Coherence: 1 error
Naturalness: 2 errors
Fidelity:
Emotional Tone: 2 errors
Cultural Nuances: 2 errors
Final Score Calculation:
100 - (14 total errors) = 86

Appendix IV ChatGPT4o Evaluation (non- literary text)

(1) Accuracy
a) Terminology:
- "Global wage growth" is accurately translated as "全球工资增长".
- "Real wage growth" is correctly translated as "实际工资增长".
- "G20 economies" and "emerging G20 countries" are accurately translated as "二十国集团发达经济体" and "二十国集团新兴经济体".
- The term "composition effect" is appropriately translated as "结构效应".
- Error Count: 0

b) Mistranslation:
- "全球工资增长在 1.6%至 2.2%之间" do not accurately translates "global wage growth fluctuated between 1.6 and 2.2 percent".
- "如果把中国排除在外" accurately translates "when China is excluded from the sample".
- "在二十国集团发达经济体中, 实际工资增长在 0.4%至 0.9 之间波动" accurately translates "In advanced G20 economies, real wage growth fluctuated between 0.4 and 0.9 percent".
- "在二十国集团发达经济体中, 韩国的工资增速最快(22%)" accurately translates "Among advanced G20 economies, wage growth accelerated the most (by 22 percent) in the Republic of Korea".
- Error Count: 1

c) Omission:
- The translation includes all key information from the original text but omitting some significant details.
- Error Count: 1

d) Addition:
- The translation does not include extraneous information that is not present in the original text.
- Error Count: 0

e) Untranslated:
- The entire text is fully translated generally, but some details are not fully translated.
- Error Count: 2

(2) Fluency
a) Cohesion & Coherence:
- The translation maintains logical flow and coherence, ensuring that the text is easy to follow.
- Sentences and paragraphs are well connected, preserving the structure and meaning of the original text.
- Error Count: 0

b) Naturalness:

- The text reads naturally and smoothly, with minor awkward phrasings that could be improved for better readability.
- Example:"结构效应" (composition effect) might be better contextualized, but it is understandable.
- Error Count:1

(3) Fidelity

a) Emotional Tone:

- The translation captures the analytical and factual tone of the original text accurately.
- The technical and formal style of the original is well maintained in the translation.
- Error Count: 0

b) Cultural Nuances:

- The translation appropriately retains cultural and contextual nuances of the original, particularly in references to economic conditions and government policies.
- The specific details about wage trends and impacts in various countries are accurately conveyed.
- Error Count: 0

Summary

1) Accuracy:
- Terminology: 0 errors
- Mistranslation: 1 errors
- Omission: 1 errors
- Addition: 0 errors
- Untranslated: 2 errors

2) Fluency:
- Cohesion & Coherence: 0 errors
- Naturalness: 1 error

3) Fidelity:
- Emotional Tone: 0 errors
- Cultural Nuances: 0 errors

Final Score Calculation: 100 - (5 total error) = 95