Article

CADERNOS de TRADUÇÃO

The potential of ChatGPT in translation evaluation: A case study of the Chinese-Portuguese machine translation

Lili Jiang
Macao Polytechnic University
Macao, China
p2209219@mpu.edu.mo
https://orcid.org/0009-0001-3297-4892

Yunxiao Jiang
Macao Polytechnic University
Macao, China
p1707643@mpu.edu.mo
https://orcid.org/0000-0003-4938-8854

Lili Han
Macao Polytechnic University
Macao, China
hanlili@mpu.edu.mo
https://orcid.org/0000-0002-8995-2301

Abstract: The integration of artificial intelligence (AI) in translation assessment represents a significant evolution in the field, transcending traditional human-only scoring approaches. This study specifically examines the role of ChatGPT, a multilingual, transformer-based large language model developed by OpenAI, in the automated evaluation of machine translations between Portuguese and Mandarin. Despite ChatGPT's burgeoning reputation for its advanced Natural Language Processing (NLP) capabilities, research on its application in translation evaluation, particularly for this language pair, remains unexplored. To fill this gap, our research employed three prevalent machine translation tools to translate a set of twenty sentences from Chinese into Portuguese. Translated target text versions provided by professional Chinese-Portuguese translators were also included to estimate if the machine-translated target texts have achieved a certain degree of human parity. We then assessed these translations using both GPT models (ChatGPT 3.5 and ChatGPT 4.0) and five human raters to offer a comprehensive scoring analysis. The study's findings reveal that, particularly ChatGPT 4.0, exhibits substantial promise in evaluating translations across varied text types. However, this potential is tempered by notable inconsistencies and limitations in its performance. Through both quantitative analysis and qualitative insights, this research highlights the synergy between ChatGPT's automated scoring and traditional human assessment. It uncovers some key benefits of this automated approach: (1) exploring viability of automated translation evaluation, particularly in Chinese-Portuguese language pair; (2) fostering critical supplement to human evaluation, and (3) uncovering the astonishing capability of cutting-edge machine translation tools in

Cadernos de Tradução, 44(1), 2024, e98613
Graduate Program in Translation Studies
Federal University of Santa Catarina, Brazil. ISSN 2175-7968
DOI https://doi.org/10.5007/2175-7968.2024.e98613

Chinese-Portuguese language pair. Our findings contribute to a more detailed comprehension of ChatGPT's role in translation assessment and underscore the need for a balanced approach that leverages both human expertise and AI capabilities.
Keywords: ChatGPT; machine translation (MT); automatic scoring; human assessment; evaluation metric.

1. Introduction

With the development of globalization, the studies on Chinese and Portuguese translation and interpreting have been catching growing attention among the academic community (Han, L., 2022b; Lu *et al*., 2022; Sun & Ye, 2023; Guo & Han, 2024). In this context, the elaboration of a reliable assessment framework in this language pair has also been the focus of many scholars (Han, L., 2022a). Regarding the area of translation and interpreting quality assessment the research on this topic has been tremendously influenced by the evolution of artificial intelligence. No longer limited to human evaluations, an integrated combination of human and machine assessments has contributed to a comprehensive final score. This evolution holds significant relevance in the field of machine translation evaluation, as there is an increasing need for efficient and precise assessment of translation quality. In this context, the utilization of AI technologies for automated translation evaluation has garnered attention from stakeholders in the translation industry as well as researchers in translation studies.

With the launch of ChatGPT by OpenAI, researchers of the translation community have explored its application on natural language processing tasks, such as automated translation (Hendy *et al*., 2023; Jiao *et al*., 2023), machine translation evaluation for languages with high resources (Lu & Han, 2023) and low resources (Kadaoui *et al*., 2023), language proficiency evaluation (Ghafar, 2023) and rating accuracy in interpreting quality (Han, C., 2020, 2021, 2022a, 2022b). Nevertheless, a noticeable research gap exists in the realm of automated machine translation evaluation, specifically in terms of applying ChatGPT to the Portuguese-Mandarin language pair.

Drawing upon the existing knowledge, this study aims to explore the feasibility of utilizing ChatGPT (versions 3.5 and 4.0) for the purpose of automatic translation evaluation between Chinese and Portuguese. The experiment aims to leverage the advanced capabilities of ChatGPT in assessing the quality of machine translations, so as to bridge the research gap in this area and potentially offer new insights into the efficacy of AI-driven translation evaluation.

2. Literature review

Translation quality assessment (TQA) is an important aspect of evaluating translations. There are two main approaches to TQA: qualitative and quantitative assessment. Qualitative assessment focuses on in-depth analysis and models, while quantitative assessment aims to provide specific scores for translations (Han, 2020).

In terms of qualitative assessment, various models have been proposed by scholars. Reiss (2000), for example, categorized texts into different types and emphasized the need to differentiate assessment criteria based on text type. House (1997, 2001) constructs the first systematic and

comprehensive TQA model in the area of Translation Criticism internationally by using the register analysis model of Systemic Functional Linguistics and drawing on the results of Comparative Pragmatics and Intercultural Studies. Similarly, Williams (2001) proposed a TQA model grounded in argumentation theory, assuming that different types of discourse exhibit distinct argumentative structures, which should serve as the primary basis for TQA, whereas the quality of translation mainly depends on whether it can accurately reflect the argumentative structure of the original text. Yang (2019), on the other hand, conducted a systematic summary and generalisation of both Chinese and Western models of TQA.

However, qualitative assessment can be time-consuming and complex, making it less suitable for routine teaching or large-scale translation tests. This has led to the emergence of quantitative assessment methods. Scholars have proposed different methods, such as error analysis, analytical scoring, and mixed-methods scoring (Colina, 2009; Williams, 2009; Mu, 2006; Xiao, 2012; Yang, 2019). These methods aim to provide objective scores for translations and improve the efficiency of assessment.

In addition, different from traditional TQA methods, comparative judgement is a relatively new method in TQA (Han *et al*., 2019; Han, C., 2020, 2022a; Han *et al*., 2022). Instead of assigning specific scores, this method involves comparing different versions of translations and selecting the one with better quality. While the reliability, validity, and practicality of this method have been confirmed (Han *et al*., 2019; Han, C., 2021, 2022a; Han *et al*., 2022), a significant drawback arises from the heightened cognitive burden placed on the assessors during the evaluation process.

Moreover, despite the exploration of diverse indicators and methods, such as rater consensus indices and rater consistency indices, to calculate the reliability of TQA (Han, 2018, 2020), challenges persist in the recruitment of competent raters and the assurance of scoring results' objectivity and fairness (Han, C., 2022b).

To overcome these challenges, researchers have started to explore the application of computer science technology, particularly the Generative Pre-trained *large language models (LLMs)*. In recent years, LLMs have been applied in various natural language processing (NLP) tasks. They have demonstrated capabilities in understanding, reasoning, and generating human-like text. Researchers have explored their applications in text simplification, cultural heritage, healthcare, low-resource languages, and other domains LLMs have shown potential in improving content similarity assessment, enhancing visitor engagement in cultural spaces, facilitating clinical text translation, standardizing radiology reports, and more (Beauchemin *et al*., 2023; Guerreiro *et al*., 2023; Hasani *et al*., 2024; Trichopoulos *et al*., 2023; Yang *et al*., 2023).

The evolution of *automatic translation evaluation (ATE)* has also been influenced by LLMs. While traditional evaluation metrics like BLEU and METEOR offer objective measures, they may not capture language nuances and contextual appropriateness. LLMs, with their vast datasets and sophisticated algorithms, promise a more nuanced understanding of language semantics and context. They have been employed to assess machine translation quality and offer new methodologies and insights in text evaluation tasks.

The capacity of LLMs in text evaluation tasks has been studied increasingly. Leiter *et al*. (2023) introduced a competition focusing on prompting LLMs for MT and text summarization evaluation. Their study revealed that even with restrictions like disallowing fine-tuning, LLMs achieved results

comparable to recent reference-free metrics, showcasing their utility in MT quality assessment. Fernandes *et al*. (2023) proposed AutoMQM, a novel prompting technique using LLMs for detailed MT evaluation. This study focused on identifying and categorizing errors in translations, leveraging the reasoning capabilities of models like PaLM and PaLM-2.

The utility of LLMs has also been examined under cross-cultural contexts. Cao *et al*. (2023) introduced a novel approach to MT, focusing on the cultural adaptation of recipes between Chinese and English. Utilizing LLMs, traditional machine translation, and information retrieval techniques, the study underscored the importance of nuanced understanding in cross-cultural contexts. GPT-4 even demonstrated impressive abilities in adapting Chinese recipes into English, though it was less effective for English to Chinese translations, expanding the scope of MT beyond linguistic accuracy to cultural relevance. This is particularly important in an increasingly globalized world where cultural nuances play a critical role in effective communication.

To sum up, preliminary studies employing generative pre-trained models such as GPT in translation evaluation have shown promising results, suggesting that such AI models could offer evaluations that closely mirror human judgment and expand the scope of translation evaluation beyond linguistic accuracy to cultural relevance. However, research focusing on specific language pairs, especially less commonly paired languages like Chinese and Portuguese, remains understudied. This gap highlights the need for further exploration into the application of ChatGPT in translation evaluation for diverse language pairs.

3. Research questions

The research questions designed for the present study are as follows:

i. Does ChatGPT have the potential to implement automatic assessment of translations from Chinese to Portuguese? The verification of this question is carried out in two phases. On the one hand, the results of ChatGPT's scoring are analysed and compared to the human scoring results in terms of correlation; on the other hand, the reliability and validity of all scoring results are examined. In order to augment the assessment's validity, two distinct scoring methods, namely comparative judgement and holistic scoring, are employed in this study to ascertain the consistency of the scoring outcomes with each other.

ii. Is ChatGPT capable of performing the most basic comparison of the translation quality between Machine and human? Three distinct tasks are designed in this study to validate this hypothesis. In addition to the scoring task and the ranking task, a third identification task is added, which aims to detect the sole version translated by human among a collection of mixed versions comprising both human and machine translations.

iii. Are there significant improvements in ChatGPT 4.0 compared to ChatGPT 3.5 in terms of conducting automated assessments of Chinese-Portuguese translations? Both versions 3.5 and 4.0 of ChatGPT are employed in this study to score the identical set of translation texts for the three tasks mentioned above. The results are analysed and compared to the human scoring results, considering factors such as stability, similarity, and other relevant aspects in order to discern the discrepancies between versions 3.5 and 4.0 of ChatGPT.

4. Methodology and method

To achieve the objectives and address the research questions of the present study, a comprehensive approach combining quantitative and qualitative methods are adopted. The experiment is meticulously designed according to the following methods:

4.1 Translation samples

Source text

In this study, a total of 20 sentences of Chinese text are selected as the source text, among which 10 sentences are political text extracted from *Classical Words Quoted by President Xi Jinping*, officially translated and published by the Ministry of Foreign Affairs of the People's Republic of China[1]. The remaining 10 sentences are classical Chinese poems, selected from the Chinese textbooks used in primary and secondary schools for the nine-year compulsory education system, which are required to be memorized by all students.

Target text

In order to ensure the quality of the human translations of the selected texts, we chose a rigorous approach. For the political text, we utilize the official translation published by the Chinese Ministry of Foreign Affairs, available on the website of *Yizhiyoudao* as the target text. For the classical poems, we source the translations from the published book "*Poemas Clássicos Chineses*" by Capparelli and Sun (2012). Both translators are experienced bilingual scholars and university professors proficient in Chinese and Portuguese. It is worth noting that the translation provided by the two experts are in Brazilian Portuguese variation, which aligns with the predominant variation used by the current machine translation engine and ChatGPT.

Regarding the machine translation versions, we select the most commonly used machine translation engines, namely Google Translate, DeepL and ChatGPT, for the target texts in this paper. Therefore, for each sentence of the source text, we have one human translation and three machine translation versions, resulting in a total of four corresponding translation versions.

4.2 Raters

The raters involved in this study are categorized into 2 parts, i.e., human raters and GPT(s) as automatic raters, and 3 groups, i.e. human, GPT-3 and GPT-4. The human raters consist of five individuals who are bilingual Chinese-Portuguese speakers. Among them, three are professional translators and interpreters with at least 5 years of experience and master's degree majored in Chinese-Portuguese translation or intercultural studies, while the remaining two are Chinese PhD students specializing in Portuguese with language level no less than C1 within the *Common European framework of reference for languages* (2001) standard. All of the human raters are fluent in Mandarin

[1] https://yizhiyoudao.kuaizhan.com/v2/categories/post-list?post_category_id=4695925051, consulted in 11th of May 2023.

Chinese as their first language and Portuguese as their second language. In addition, the GPT evaluator scores each of the 20 sets of target text pairs (comprising a total of 80 translated sentences) five times using both the models of ChatGPT 3.5 and 4.0.

4.3 Scoring methods

Considering the absence of standardized scoring criteria and rubric scales for Chinese-Portuguese translation in the academic world (Han, L., 2022a, 2022b), this study adopts Information Completeness and Correctness as the fundamental scoring criteria, which are commonly used for assessing the quality of interpreting and translating (Mu, 2006; Zou, 2005). In the present study, we implemented the criteria adopted by the Test for English Majors Grade Eight (TEM-8) exam for their translation test section, which referred Information Completeness and Correctness as two core scale for the evaluation of translation (TEM 8 Syllabus Revision Group, 1998; Zou, 2005; Mu, 2006; Xiao, 2012). TEM-8 is a large-scale national examination administered by the National Advisory Committee for Foreign Language Teaching under the Ministry of Education of China, designed to assess the English proficiency of undergraduate English majors at the end of their four-year program (TEM 8 Syllabus Revision Group, 1998; Zou, 2005; Zou & Xu, 2017). For the information Completeness, or fidelity, it refers to how well the translation retains all the information from the source text, while the Correctness, or adequacy, includes lexical, syntactical, semantic, and stylistic accuracy.

As for scoring methods, the present study adopts two approaches: holistic scoring using general impression scoring technique (Mu, 2006; Xiao, 2012; Yang, 2019) and comparative judgement (Han *et al*., 2019; Han, C. 2020, 2022a; Han *et al*., 2022). For holistic scoring, we take the 10-point scale of the TEM-8 exam (Zou, 2005; Mu, 2006; Xiao, 2012; Zou & Xu, 2017); while for the comparative judgement, we expand the translation versions within each group from two to four for the evaluators' ranking task.

4.4 Data analysis

4.4.1 Data analysis on the correlation between GPT models and human raters

For the analysis conducted in this study, it is pivotal to determine the normality of data. In cases when the data follows a normal distribution, *Pearson's coefficient* is employed for analysis. Conversely, when the data deviates from the normal distribution, *Spearman's coefficient* is utilized.

Based on the specific characteristics of the data, we select different correlation coefficients to estimate the statistical relationship between the human evaluations and the GPT model evaluation results. The results of these correlations are presented in the following Table 1.

Table 1: Correlation coefficient used to compare the human and machine evaluations' results

Task Type	Texture type	GPT-3.5	GPT-4
Ranking	P	r_p	r_s
	L	r_s	r_s
	P+L	r_s	r_s

Scoring	P	r_p	r_p
	L	r_s	r_s
	P+L	r_s	r_s
Identification	P	r_p	r_p
	L	r_s	r_s
	P+L	r_p	r_s

Note: P = Political text, L= Literary text.

Source: Authors (2024)

[Description] As mentioned in the text [End of description].

4.4.2 Reliability analysis

In pursuit of understanding evaluator reliability across various assessment tasks, we employed a range of statistical methodologies that are tailored to each task's inherent characteristics. In this subsection, we explain how the rationale behind the selection of these indicators, taking into consideration the data characteristics.

4.4.2.1 Kendall's Tau (τ) in the context of ranking evaluations

For the evaluative task that requires ranking candidates on a scale from 1 to 5, Kendall's Tau is our statistical measure of choice. Kendall's Tau is a non-parametric tool that offers insights into the strength and direction of ordinal associations between two sets of data. It juxtaposes the number of pairings that retain their order with those that reverse it. By calculating the average of the coefficients from all possible pairs of evaluation results, we can estimate the inherent consistency. Given the ranking nature of the task, Kendall's Tau is aptly suited for assessing the reliability of the evaluations.

4.4.2.2 Application of Cronbach's Alpha (α) for scoring evaluations

When evaluators are tasked with scoring candidates on a continuum from 1 to 10, Cronbach's Alpha is employed as the statistical measure. This metric gauges the internal consistency of a test and ensures that all test items measure the same construct. Our objective is to gauge the homogeneity of ratings across different human evaluators and GPT evaluators, making Cronbach's Alpha an appropriate choice.

However, for GPT-4, Cronbach's Alpha is deemed inapplicable due to a substantial overlap in ratings. This overlap results in a lack of consistent variance between raters, which is a prerequisite for utilizing Cronbach's Alpha as a reliability measure.

4.4.2.3 Intraclass Correlation Coefficient (ICC) for evaluating GPT-4's inherent reliability

To gauge GPT-4's internal reliability, we invoke the Intraclass Correlation Coefficient (ICC). The ICC is a statistical measure used to evaluate the consistency and uniformity of measurements

made by different observers on the identical quantity. In the context of our study, where diverse instances of the GPT-4 model are activated at distinct times, these instances can act as "multiple observers". This makes the ICC an ideal metric for assessing inter-observer reliability in this context.

4.4.2.4 Fleiss' Kappa (κ) for categorical distinguishment evaluations

For the task that mandates evaluators to categorically distinguish between human and machine translation candidates, we leverage the Fleiss' Kappa statistic. Fleiss' Kappa is an advancement on Cohen's Kappa and is specifically designed to assess the reliability of agreement among a fixed set of raters when attributing categorical ratings. The categorical nature of this task makes Fleiss' Kappa the most appropriate statistical measure for reliability evaluation.

Table 2 below demonstrates the selected statistical indicators used to evaluate the reliability of the evaluations conducted by human or large language model evaluators.

Table 2: Statistical indicators selected for the evaluation of the rater's reliability

	Ranking Task	Scoring Task	Identification Task
Ranking	τ	α	κ
Scoring	τ	α	κ
Identification	τ	ICC	κ

Source: Authors (2024)

[Description] As mentioned in the text [End of description].

5. Results

5.1 Average duration

During the experiment, we record the length/duration of time required for both human and machine scoring separately. The comparison of these durations, as illustrated in Table 3, reveals that the machine evaluators outperformed human raters significantly in terms of both total time and average time per question. Even after excluding human rater no.5, which took the longest time for scoring, the average total time spent by human raters for scoring 20 translated sentence pairs is still 48.6 minutes. In contrast, ChatGPT 4.0 only took 19.9 minutes, and ChatGPT 3.5 completed the task in a mere 2.6 minutes. The average total time expended by human raters was 2.4 times that of ChatGPT 4.0 and 18.6 times that of ChatGPT 3.5, while the time difference between the two generations of GPT models was around 7.6 times. When observing the average length of duration, the GPT model took to answer each question, ChatGPT 4.0 require an average time of 46 to 88 seconds, while ChatGPT 3.5 take only 5 to 12 seconds on average to complete all three scoring tasks for a set of sentence pairs.

Table 3: Record of the duration for assessment tasks[2]

	Total (sec.)	Total (min.)	Average /question (sec.)
Human rater no. 1	2347.00	39.12	117.35
Human rater no. 2	2505.00	41.75	125.25
Human rater no. 3	2668.00	44.47	133.40
Human rater no. 4	4150.00	69.17	207.50
Human rater no. 5	25359.00	422.65	1267.95
Average of all human raters (no. 1-5)	7405.80	123.43	370.29
Average of human raters no. 1-4	2917.50	48.63	145.88
ChatGPT 3.5	157.00	2.62	7.85
ChatGPT 4.0	1191.00	19.85	59.55

Source: Authors (2024)

[Description] As mentioned in the text [End of description].

GPT-4 may respond slower than GPT-3.5 primarily due to its increased complexity and advanced capabilities. As a more sophisticated model with more parameters, GPT-4 requires more computational power to process these parameters, which can lead to longer response times. Additionally, improvements in accuracy and contextual understanding in GPT-4, while beneficial for performance, further contribute to this slower response. External factors like server load and network infrastructure can also impact response times.

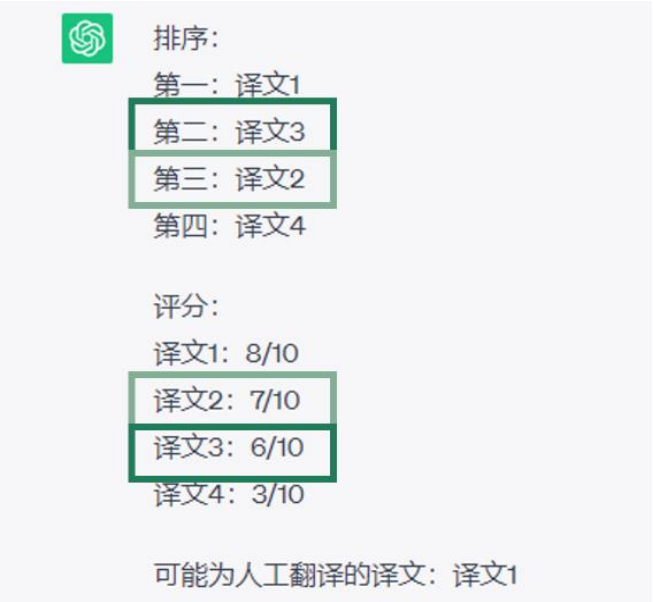
5.2 Stability and self-consistency

During the experiment, we noticed that both GPT-3.5 and GPT-4 exhibit a lack of stability in their responses, and the answers provided by the models are generally inconsistent or varied across each different enquiry. However, relatively speaking, ChatGPT 4.0 performed more stability than ChatGPT 3.5, showing less volatility when comparing the five different assessments for each set of texts.

In some instances, ChatGPT provides contradictory evaluation results across the three different tasks of the same set of texts within the same round. These scores are not self-consistent with the ranking results, as depicted in Figure 1. In contrast, we observe no such self-contradiction among the human raters during their evaluation.

[2] Consulted during 17th to 22nd of May 2023.

Figure 1: Screenshot of assessment by ChatGPT 3.5 with contradictions



Source: https://chat.openai.com/chat, consulted in 20th of May 2023

[Description] The figure shows the evaluation results given by ChatGPT 3.5 in Chinese. It can be translated in English is as follows:
Ranking Result:
1st: Target text 1
2nd: Target text 3
3rd: Target text 2
4th: Target text 4
Scoring Result:
Target text 1: 8/10
Target text 2: 7/10
Target text 3: 6/10
Target text 4: 3/10
As can be seen, according to the ranking result, the score of target text 3 should be higher than target text 2, however, the scores given by ChatGPT 3.5 shows apparently the opposite results. [End of description].

From a statistical standpoint, GPT-4 demonstrates a relatively high level of consistency in its evaluations across all three genres of assessment tasks comparing to GPT-3.5 and to human raters. For the ranking task, GPT-4 shows a Kendall's tau of 0.82 (versus 0.4 of GPT-3.5 and 0.57 of human raters), indicating a strong positive correlation between its rankings and the human evaluations. In the scoring task, GPT-4 demonstrates an ICC1k coefficient of 0.98 (versus 0.51 of GPT-3.5 and 0.70 of human raters), indicating a high degree of agreement among different instances of GPT-4 in their scoring evaluation. Lastly, for the identification of human versions of translation, GPT-4 achieves a Fleiss' kappa coefficient of 0.90 (versus 0.41 of GPT-3.5 and 0.67 of human raters), suggesting substantial agreement among multiple GPT-4 evaluators in categorizing translations as either human or machine-generated. These statistical indicators, as summarized in Table 4, demonstrate the robust consistency exhibited by GPT-4 in its evaluation across various assessment tasks.

Table 4: Reliability of human raters and GPT raters[3]

	Ranking (Kendall's Tau)	Scoring (Cronbach's Alpha)	Identification (Fleiss' kappa)
Human	$\tau = 0.57^{**}$	$\alpha = 0.70$	$\kappa = 0.67$
GPT-3.5	$\tau = 0.40^{**}$	$\alpha = -0.51$	$\kappa = 0.41$
GPT-4	$\tau = 0.82^{**}$	ICC1k (Average raters absolute) = 0.98	$\kappa = 0.90$

Note: ** $p<0.01$.

Source: Authors (2024)

[Description] As mentioned in the text [End of description].

5.3 Correlations between GPT scores and human-assigned scores

As evident from Table 5 and Figures 2 and 3 below, the overall correlation between GPT-3.5 and human raters' scoring in the three different scoring tasks is relatively lower, whereas the overall correlation between GPT-4 and human scoring is notably high. These findings suggest that GPT-4 exhibits a stronger alignment with human raters in terms of scoring evaluations, while GPT-3.5 shows a comparatively weaker correlation.

Table 5: Correlation of GPT evaluations with human evaluations[4]

GPT Model	Task Type	Texture Type	Coefficient Type	Coefficient Value	P-Value
3.5	R	P	r_p	0.6059	3.43E-05
4	R	P	r_s	0.7191	1.72E-07
3.5	R	L	r_s	0.0988	5.44E-01
4	R	L	r_s	0.8353	2.05E-11
3.5	R	P+L	r_s	0.3317	2.64E-03
4	R	P+L	r_s	0.7841	7.90E-18
3.5	S	P+L	r_s	0.2879	0.0096
4	S	P+L	r_s	0.8387	2.76E-22

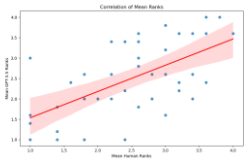
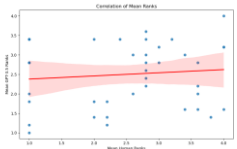
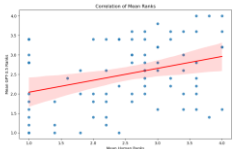
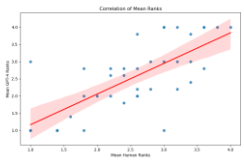
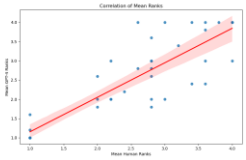
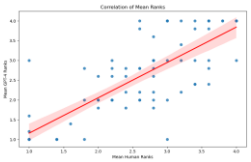
Note. R = Ranking, S = Scoring, P = Political text, L= Literary text.

[3] Consulted in 25th of May 2023.
[4] Consulted in 25th of May 2023.

Source: Authors (2024)
[Description] As mentioned in the text [End of description].

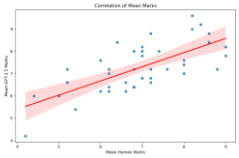
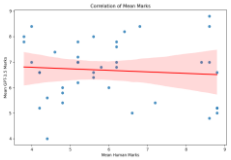
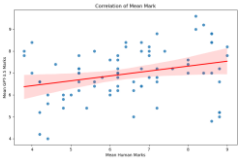
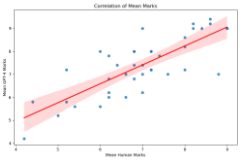
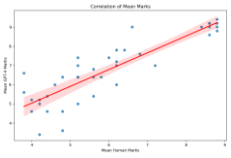
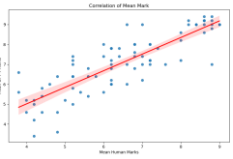
It is noteworthy that the correlations between human raters' scoring and both GPT-3.5 and GPT-4 are both high in the evaluation of the political text. However, a significant discrepancy is observed in the evaluation of the literary text. Specifically, when completing Task 1 (Ranking) and Task 2 (Scoring), GPT-3.5 shows a substantial discrepancy from human rated scores in the literary text. This disparity results in a relatively low overall correlation between GPT-3.5 and human rated scores.

Figure 2: Visualization of the correlation of GPT evaluations with human evaluations in ranking tasks[5]

	Political text	Literary text	Overall correlation
GPT-3.5	 (1)	 (2)	 (3)
GPT-4	 (4)	 (5)	 (6)

Source: Authors (2024)
[Description] As mentioned in the text [End of description].

Figure 3: Visualization of the correlation of GPT evaluations with human evaluations in scoring tasks[6]

	Political text	Literary text	Overall correlation
GPT-3.5	 (1)	 (2)	 (3)
GPT-4	 (4)	 (5)	 (6)

[5] Consulted in 26th of May 2023.
[6] Consulted in 26th of May 2023.

Source: Authors (2024)
[Description] As mentioned in the text [End of description].

The figures above offer a visual representation of the correlation between automatic and manual scoring, with the fitted line to illustrate the relationship. We can visualize that the scores of the GPT-4 model are highly correlated with the scores of the human raters across all evaluation tasks and text genres. For both ranking (Task 1) and scoring (Task 2), GPT-4 outperforms GPT-3.5 in terms of its similarity with human raters. Specifically, the overall marks given by GPT-4 has a correlation coefficient of 0.84 with human scoring, demonstrating its great potential in scoring tasks as an alternative for human scoring.

5.4 Identification of human translation from Machine Translation

The third evaluation task in this study is to distinguish and identify the human-translated version from the machine-translated versions. The results of this task reveal that GPT-4 achieves a Spearman coefficient of 0.9002 (p=6.44E-08), indicating a notably superior performance compared to GPT-3.5, which obtains a Pearson coefficient of 0.3369 (p=0.1464). As shown in Table 6 and in Appendix A, when identifying political texts, the GPT models have high accuracies, with GPT-4 outperforming GPT-3.5. However, in the cases of literary texts, GPT-3.5 exhibits a limited cognitive range, perceiving sometimes the machine-translated version as preferable to the human translation. This leads to lower identification accuracies for GPT-3.5. Conversely, GPT-4 consistently achieves significant higher accuracies in the identification of literary texts, showcasing superior performance in this task.

6. Discussion

The discussion in the text revolves around the research questions and primarily centres on comparing the performance of ChatGPT, specifically GPT-3.5 and GPT-4, with that of human raters in the context of ranking, scoring, and identification tasks, addressing also the cost-effectiveness of using ChatGPT models for assessment tasks compared to employing human raters.

6.1 ChatGPT vs human raters in assessment of ranking and scoring tasks

Upon careful analysis of our experimental findings, we have identified four key advantages that ChatGPT offers. These advantages stem from the exceptional performance exhibited by the models and have significant implications for their potential applications in the next few years:

6.1.1 Reliability

In general, we observe that GPT-4 exhibits greater stability in terms of response consistency compared to GPT-3.5 during the experiment. Although both GPT-3.5 and GPT-4 provide different results for each query throughout Task 1 (ranking) and Task 2 (scoring), GPT-4, especially in Task 1, demonstrate a more cohesive set of answers for each group.

It is important to note that the current GPT models may not exhibit complete stability in their responses, but our findings indicate the presence of a certain degree of internal consistency. This suggests that ChatGPT, with its rapid iteration with technological advancement, holds the potential for automatic assessment tasks.

6.1.2 Rapidness

In 5.1 we have found out that the machine raters are much more faster than the human raters. Even though our study involves a relatively small number of text samples, human raters take averagely more than twice as long as the machine to complete the tasks. In the case of qualification exams for translators, which typically involve a considerable larger number of translation samples, the duration of human assessment would be substantially extended. Our post-interviews with the five human raters further confirm this observation. They report experiencing fatigue and a decrease of concentration during the assessment process. In some cases, they have to take proper rest breaks before they are able to continue with the rest of the assessment tasks.

6.1.3 Fatigue-free

As mentioned in the preceding paragraph, machine raters, such as ChatGPT, do not get exhausted. In contrast, human raters, are susceptible to show rater's effects and scoring errors under fatigue, leading to potential biases, such as excessive severity/leniency in their assessment. This becomes particular relevant when dealing with a large number of text samples, such as those encountered in translation qualification exams, as human raters are more likely to be negatively affected by overload. However, machines, like ChatGPT, do not get tired or require rest, and remain unaffected by rater's effects due to personal preferences or fatigue-related inconsistencies.

6.1.4 Low cost

The current GPT-3.5 model is completely free, with no limit on the number of requests per day. On the other hand, GPT-4 uses a profit model with a monthly fee of $20 with a maximum of 25 requests for each three hours, i.e., it can answer a maximum of 200 requests or conduct 200 sets of assessments per day. In comparison, recruiting human raters is far more costly in terms of both money, time and effort than machine evaluators. Furthermore, the recruitment of qualified raters can be quite challenging due to the 6 characteristics of an ideal rater concluded by Han Chao (2022b), not to mention the further process of selection, and training of the raters. To be considered ideal, a rater must possess 6 key attributes: (1) a high level of proficiency in the relevant language pairs, (2) academic background in interpreting, ideally at a postgraduate level, (3) professional experience in the field of interpreting, (4) a track record of teaching interpreting, (5) extensive participation in interpreting evaluations, and (6) a thorough understanding of the fundamental principles of reliable assessment practices (Han, C., 2022b). All of these factors support the fact that the automatic assessment is less costly.

6.2 ChatGPT vs human raters in assessment of identification task

6.2.1 GPT-3.5 vs GPT-4

In the performance of all three assessment tasks we designed, GPT-4 has a significant advantage over GPT-3.5 in terms of stability of responses, internal consistency, correlation with human scoring, and accuracy in identifying human translations. Although GPT-3.5 shows certain assessment ability in ranking and scoring tasks, its performance in the identification task is relatively mediocre, as highlighted in Table 6. The overall accuracy of GPT-3.5 is only 46% in identifying human-translated versions of source texts, which means the limited capability in accurately distinguishing human translations. In contrast, GPT-4 showcases a superior performance in the identification tasks. It achieves an accuracy rate of over 90% in both political and literary texts, with an exceptional overall accuracy rate of 96%. The only drawback of GPT-4 in comparison to GPT-3.5 is that it takes a relatively long time to complete the assessment process.

Table 6: Performance of human raters and GPT raters[7]

	Political Text	**Literary Text**	**Overall**
ChatGPT 3.5	56%	36%	46%
ChatGPT 4.0	94%	98%	96%
ChatGPT 3.5&4.0	75%	67%	71%
Human	70%	100%	85%

Source: Authors (2024)

[Description] As mentioned in the text [End of description].

6.2.2 Human Errors

The findings presented in Table 6 are remarkable. GPT-4 has an impressive accuracy rate in identifying human-translated versions, surpassing even the accuracy of human raters. While human raters achieve a perfect 100% accuracy in literary texts, their accuracy drops to 70% when identifying political texts, which is significantly lower than the machine's 94% accuracy. Consequently, the human raters' accuracy ends up being approximately 10% below GPT-4's accuracy on the overall average accuracy. Even when calculating the mean accuracies of GPT-3.5 and GPT-4, machines still achieve higher identification accuracies than humans in the case of political texts.

Furthermore, a comparison of the 5 sets of assessment results given by human raters and GPT (see Appendix) shows that the highest accuracy rate achieved by human raters is 90% (with two raters both achieving this score). In contrast, GPT-4 consistently achieves 100% accuracy for all 20 sets of texts across 3 sets of assessment. This suggests that as technology advances, it is highly likely that machines will be able to accurately identify humans' translations, while humans may err and struggle to identify machine-generated translations.

Alexander Pope's line (1711), 'To err is human,' aptly reflects the challenges in translation evaluation. Human raters, despite their expertise, are prone to errors. In contrast, GPT

[7] Consulted in 28th of May 2023.

demonstrates consistent accuracy, identifying issues that may be missed by human evaluators. While GPT cannot replace human judgment, it serves as a valuable complement, enhancing the overall reliability of TQA.

6.2.3 Improvement of Machine Translation Quality

The experiment reveals significant controversy among human raters regarding the identification and assessment of certain groups of sentences, especially in the case of political texts. As shown in the Appendix, for the second set of texts, 4 out of 5 human raters identify the Google-translated version as the human translations, while 1 choose the DeepL-translated version as the human translation. None of the raters correctly identify the actual human-translated version. In the present experiment, only 2 out of 10 sets of political texts are completely uncontroversial in the identification task. This highlights the challenges faced by human raters and shows that the quality of machine translation is gradually improving.

Interestingly, the experiment also indicates, out of the 3 different machine translations evaluated, GPT most frequently mistake the DeepL translation for a human translation, followed by the GPT translation, and finally the Google version. Human raters, on the other hand, most often mistake the DeepL translation as human, followed by the Google and finally the GPT translation. The comparison of results reveals that the machine translation of DeepL is recognized as the best machine translation among the evaluated options. In addition, we can also infer that GPT's scoring logic tends to favour its own translations over other machine translation engines, indicating certain degree of subjectivity in the scoring. However, it is worthy to be noted that there still exists a big gap between machine and human in terms of ancient poetry translation in terms of content precision. As can be seen in Table 7, under an ancient poem context, all of the machine translators explained the word "单车" as "bicycle" in Portuguese (*bicicleta*), while only the human translator used "solitary carriage" in Portuguese (*carro solitário*) to express the precise meaning.

Table 7: Performance of human raters and GPT raters[8]

Source text (in Chinese)	.	Translator	Target text (in Portuguese)
《使至塞上》 <shǐ zhì sài shàng> 单车欲问边， 属国过居延。 dān chē yù wèn biān shǔ guó guò jū yán	1	**ChatGPT 3.5**	**Levar para a fronteira** Quer saber a fronteira da **bicicleta,** pertence ao país de Juyan.
	2	**DeepL**	**Fazer uma viagem a Seaside** Quando quis pedir a fronteira de **bicicleta**, atravessei Juyan pelo campo.
	3	**Google Translation**	**Fazer para ligar** Se você quiser perguntar sobre a lateral da **bicicleta**, ela pertence ao país para passar por Juyan.

[8] Consulted in 20th of May 2023.

	4	**Human translators**	**Missão na Fronteira** **Carro solitário** passa pela estradas da fronteira. Juyan ficou para trás: eis o país ocupado.

Source: Authors (2024)
[Description] As mentioned in the text. [End of description].

The main point is that the term "单车" (dān chē) in ancient Chinese poems referred to a carriage, not a bicycle as it does in modern Chinese. This demonstrates the importance of intralingual translation (translation between different time periods of the same language) when translating ancient Chinese poems (Han, 2023). Simply doing an interlingual translation (between languages) would not properly convey the original meaning and content. This example supports the view put forward by Pöchhacker (2022) that future trends in translation will involve both intralingual and interlingual translation, with intralingual translation playing a more important role. The experiment described here shows that machine translation still faces significant limitations in performing intralingual translation accurately. Since intralingual translation requires understanding subtle differences in meaning over time within a language, this remains a challenging task for MT.

7. Conclusions

The objective of the present research is to investigate the viability of applying ChatGPT to the automatic assessment of Chinese-Portuguese translations. Three different evaluation tasks are designed, followed by a comparison of the results by machine and human evaluations to assess the effectiveness of ChatGPT in automatic assessment of translation.

While this study has still certain limitations, such as a relatively small sample size and limited experimental scope, the results obtained provide valuable insights: GPT models, particularly ChatGPT 4.0, demonstrate a high potential in generating evaluations for translations of different textual types and shows promising results in accurately assessing the quality of translated texts. This highlights the powerful capabilities of GPT models in the field of automatic translation evaluation.

However, it is important to acknowledge the limitations and possible instabilities associated with the GPT models. These limitations should be considered when interpreting the results and when further developing and refining automatic evaluation systems utilizing GPT models.

Overall, this study contributes to our understanding of the potential of GPT models, particularly ChatGPT 4.0, in the automatic evaluation of translations from Chinese to Portuguese across various textual genres. It also emphasizes the need for continued research and refinement to address the observed limitations and instabilities, ultimately improving the effectiveness and reliability of automatic translation evaluation systems.

Building upon the findings of this study, there are several directions for future research to consider. One direction involves adjusting various variables, such as exploring different language pairs, different translation directionality, and utilizing different machine translation metrics.

Additionally, examining the use of reference texts or not in the assessment process could also be explored.

By exploring these research directions, we can further enhance our understanding of the capabilities and limitations of automatic translation assessment tools like ChatGPT. This will contribute to the advancement of Translation Studies and the development of more effective and comprehensive assessment methodologies in the field.

References

Beauchemin, D., Saggion, H., & Khoury, R. (2023). MeaningBERT: Assessing Meaning Preservation between Sentences. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1223924

Cao, Y., Kementchedjhieva, Y., Cui, R., Karamolegkou, A., Zhou, L., Dare, M., Donatelli, L., & Hershcovich, D. (2023). Cultural Adaptation of Recipes. *arXiv* preprint. https://doi.org/10.48550/arXiv.2310.17353

Capparelli, S., & Sun, Y. (2012). *Poemas clássicos chineses*. L&PM.

Colina, S. (2009). Further Evidence for a Functionalist Approach to Translation Quality Evaluation. *Target*, *21*(2), 235–264. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge University Press.

Fernandes, P., Deutsch, D., Finkelstein, M., Riley, P., Martins, A. F., Neubig, G., Garg, A., Clark, J. H., Freitag, M., & Firat, O. (2023). The Devil is in the Errors: Leveraging Large Language Models for Fine-grained Machine Translation Evaluation. *arXiv preprint*. https://doi.org/10.18653/v1/2023.wmt-1.100

Ghafar, Z. N. (2023). ChatGPT: A New Tool to Improve Teaching and Evaluation of Second and Foreign Languages A Review of ChatGPT: the Future of Education. *International Journal of Applied Research and Sustainable Sciences, 1*(2), 73–86. https://doi.org/10.59890/ijarss.v1i2.392

Guerreiro, N. M., Alves, D. M., Waldendorf, J., Haddow, B., Birch, A., Colombo, P., & Martins, A. F. (2023). Hallucinations in Large Multilingual Translation Models. *Transactions of the Association for Computational Linguistics*, *11*, 1500–1517. https://doi.org/10.1162/tacl_a_00615

Guo, M., & Han, L. (2024). From Manual to Machine: Evaluating Automated Ear–voice Span Measurement in Simultaneous Interpreting. *Interpreting*, *26*(1), 24–54. https://doi.org/10.1075/intp.00100.guo

Han, C. (2018). Quantitative Research Methods in Translation and Interpreting Studies. *The Interpreter and Translator Trainer*, *12*(2), 244–247. https://doi.org/10.1080/1750399X.2018.1466262

Han, C. (2020). Translation Quality Assessment: A Critical Methodological Review. *The Translator*, *26*(3), 257–273. https://doi.org/10.7202/037044ar

Han, C. (2021). Analytic Rubric Scoring versus Comparative Judgment: A Comparison of Two Approaches to Assessing Spoken-Language Interpreting. *Meta*, *66*(2), 337–361. https://doi.org/10.7202/1083182ar

Han, C. (2022a). Assessing Spoken-Language Interpreting: The Method of Comparative Judgement. *Interpreting*, *24*(1), 59–83. https://doi.org/10.1075/intp.00068

Han, C. (2022b). Interpreting Testing and Assessment: A State-of-the-art Review. *Language Testing*, *39*(1), 30–55. https://doi.org/10.1177/02655322211036100

Han, C., Chen, S., & Fan, Q. (2019). *Rater-mediated Assessment of Translation and Interpretation*: Comparative Judgement versus Analytic Rubric Scoring. *5th International Conference on Language Testing and Assessment*, Guangzhou, China.

Han, C., Hu, B., Fan, Q., Duan, J., & Li, X. (2022). Using Computerised Comparative Judgement to Assess Translation. *Across Languages and Cultures*, *23*(1), 56–74. https://doi.org/10.1556/084.2022.00001

Han, L. (2022a). 中葡交替传译教程 *Interpretação Consecutiva Chinês-Português*. Universidade Politécnica de Macau.

Han, L. (2022b). 中葡口譯教學歷史、現狀與展望——兼及澳門的貢獻 [Portuguese Interpreting Teaching in China: Past, Present, and Future - Macao's Contribution]. 澳門理工學報 [*Revista da Universidade Politécnica de Macau*], *25*(2), 52–61.

Han, L. (2023). Tradução de poemas de Adriana Lisboa para o chinês: uma breve reflexão. *Cadernos de Tradução*, *43*(esp. 3), 388–397. https://doi.org/10.5007/2175-7968.2023.e97182

Hasani, A. M., Singh, S., Zahergivar, A., Ryan, B., Nethala, D., Bravomontenegro, G., ... & Malayeri, A. (2024). Evaluating the performance of Generative Pre-trained Transformer-4 (GPT-4) in standardizing radiology reports. *European Radiology, 34*(6), 3566-3574. https://doi.org/10.1007/s00330-023-10384-x

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. *arXiv preprint*. https://doi.org/10.48550/arXiv.2302.09210

House, J. (1997). *Translation Quality Assessment: A Model Revisited*. Gunter Narr Verlag.

House, J. (2001). Translation Quality Assessment: Linguistic Description versus Social Evaluation. *Meta*, *46*(2), 243–257. https://doi.org/10.7202/003141ar

Jiao, W., Wang, W., Huang, J.-t., Wang, X., & Tu, Z. (2023). Is ChatGPT a Good Translator? A Preliminary Study. *arXiv preprint*. https://doi.org/10.48550/arXiv.2301.08745

Kadaoui, K., Magdy, S. M., Waheed, A., Khondaker, M. T. I., El-Shangiti, A. O., Nagoudi, E. M. B., & Abdul-Mageed, M. (2023). Tarjamat: Evaluation of Bard and ChatGPT on Machine Translation of Ten Arabic Varieties. *arXiv preprint*. https://doi.org/10.48550/arXiv.2308.03051

Leiter, C., Opitz, J., Deutsch, D., Gao, Y., Dror, R., & Eger, S. (2023). The Eval4NLP 2023 Shared Task on Prompting Large Language Models as Explainable Metrics. *arXiv preprint*. https://doi.org/10.48550/arXiv.2310.19792

Lu, J., Han, L., & André, C. A. (2022). Tradução portuguesa de referências culturais extralinguísticas no Manual de Chinês Língua Não Materna: aplicação de estratégias de tradução propostas por Andrew Chesterman. *Cadernos de Tradução*, *42*(1), 1–39. https://doi.org/10.5007/2175-7968.2022.e82416

Lu, X., & Han, C. (2023). Automatic Assessment of Spoken-language Interpreting based on Machine-translation Evaluation Metrics: A Multi-scenario Exploratory Study. *Interpreting*, *25*(1), 109–143. https://doi.org/https://doi.org/10.1075/intp.00076.lu

Mu, L. (2006). 翻译测试及其评分问题 [Translation Testing and Grading]. *Foreign Language Teaching and Research*, *38*(6), 466–471. https://doi.org/10.3969/j.issn.1000-0429.2006.06.010

Pöchhacker, F. (2022). Interpreters and Interpreting: Shifting the Balance? *The Translator*, *28*(2), 148–161. https://doi.org/10.1080/13556509.2022.2133393

Pope, A. (1711). *Sound and Sense*. Sound and Sense.

Reiss, K. (2000). *Translation Criticism: The Potentials and Limitations - Categories and Criteria for Translation Quality Assessment*. St. Jerome Publishing.

Sun, Y., & Ye, Z. (2023). Tradução de metáforas verbo-pictóricas para páginas *web* do *smartphone* Huawei P40 Pro à luz da teoria de necessidades. *Cadernos de Tradução*, *43*(esp. 3), 272–302. https://doi.org/10.5007/2175-7968.2023.e97183

TEM 8 Syllabus Revision Group. (1998). 高校英语专业八级考试大纲 [*Syllabus for TEM 8*]. Shanghai Foreign Language Education Express.

Trichopoulos, G., Konstantakis, M., Alexandridis, G., & Caridakis, G. (2023). Large language models as recommendation Systems in Museums. *Electronics, 12*(18), 3829. https://doi.org/10.3390/electronics12183829

Williams, M. (2001). The Application of Argumentation Theory to Translation Quality Assessment. *Meta, 46*(2), 326–344. https://doi.org/10.7202/004605ar

Williams, M. (2009). Translation Quality Assessment. *Mutatis Mutandis*: *Revista Latinoamericana de Traducción*, *2*(1), 3–23.

Xiao, W. (2012). *Research on the Test of Undergraduate Translation Majors*. People's Publishing House.

Yang, X., Yun, J., Zheng, B., Liu, L., & Ban, Q. (2023). Oversea Cross-lingual Summarization Service in Multilanguage Pre-trained Model through Knowledge Distillation. *Electronics*, *12*(24), 5001. https://doi.org/10.3390/electronics12245001

Yang, Z. (2019). 翻译测试与评估研究 [*Studies on Translation Testing and Assessment*]. Foreign Languages Teaching and Research Press.

Zou, S. (2005). 语言测试 [*Studies on Translation Testing and Assessment*]. Shanghai Foreign Language Education Express.

Zou, S., & Xu, Q. (2017). A Washback Study of the Test for English Majors for Grade Eight (TEM8) in China—From the Perspective of University Program Administrators. *Language Assessment Quarterly*, *14*(2), 140–159.

Appendix A Answer records and average accuracy of identification task

Rater	No.	Political Text										Literary Text										Average Accuracy		
		1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	Political Text	Literary Text	Overall
ChatGPT 3.5	**1**	H	D	D	G	D	H	H	C	H	D	G	C	H	H	D	C	H	C	D	H	40%	40%	40%
	2	H	H	D	H	D	H	H	C	H	D	H	H	H	H	D	C	H	H	H	H	60%	80%	70%
	3	H	H	D	H	C	H	H	C	H	H	C	H	G	D	C	C	D	C	G	H	70%	20%	45%
	4	H	D	D	H	D	H	H	C	H	D	C	G	G	D	C	D	H	C	G	H	50%	20%	35%
	5	H	G	D	H	D	H	H	C	H	H	C	G	G	D	C	D	H	C	G	H	60%	20%	40%
ChatGPT 4.0	**1**	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	100%	100%	100%
	2	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	100%	100%	100%
	3	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	100%	100%	100%
	4	H	H	H	H	H	H	H	C	H	G	H	H	H	H	C	H	H	H	H	H	80%	90%	85%
	5	H	H	H	H	H	H	H	H	H	G	H	H	H	H	H	H	H	H	H	H	90%	100%	95%
Human	**1**	H	G	H	H	D	H	H	D	H	H	H	H	H	H	H	H	H	H	H	H	70%	100%	85%
	2	H	G	H	H	H	H	H	H	G	D	H	H	H	H	H	H	H	H	H	H	70%	100%	85%
	3	H	G	D	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	80%	100%	90%
	4	C	D	H	H	D	H	C	H	H	D	H	H	H	H	H	H	H	H	H	H	50%	100%	75%
	5	H	G	H	H	H	H	D	H	H	H	H	H	H	H	H	H	H	H	H	H	80%	100%	90%

Note: H = Human translation, D = DeepL translation, G = Google Translate, C = ChatGPT 3.5 translation.

Notes

Authorship contribution

Conception and preparation of the manuscript: L. Jiang, Y. Jiang, L. Han
Data collection: L. Jiang, Y. Jiang, L. Han
Data analysis: L. Jiang, Y. Jiang, L. Han
Discussion of results: L. Jiang, Y. Jiang, L. Han
Review and approval: L. Jiang, Y. Jiang, L. Han

Research dataset
Not applicable.

Funding
Not applicable.

Image copyright
Not applicable.

Approval by ethics committee
This research is part of the project number RP/FLT-11/2022 of Macao Polytechnic University and fully adheres to ethical standards in research.

Conflict of interests
The authors declare no conflicts of interest.

Data availability statement
The data from this research, which are not included in this work, may be made available by the author upon request.

Publisher
Cadernos de Tradução is a publication of the Graduate Program in Translation Studies at the Federal University of Santa Catarina. The journal *Cadernos de Tradução* is hosted by the Portal de Periódicos UFSC. The ideas expressed in this paper are the responsibility of its authors and do not necessarily represent the views of the editors or the university.

Section editors
Andréia Guerini – Willian Moura

Technical editing
Alice S. Rezende – Ingrid Bignardi – João G. P. Silveira – Kamila Oliveira

Article history
Received: 22-02-2024
Approved: 12-08-2024
Revised: 22-09-2024
Published: 09-2024