

RESEARCH ARTICLE

A COMPARATIVE ANALYSIS OF THE RELIABILITY OF CHATGPT AND HUMAN TRANSLATION WITH REFERENCE TO LITERARY TEXTS

Rania Shokri Qasem Shaher^{1,*}, and Adel Abdulkhaleq Abdulkareem Dulul²¹ Dept. of Translation, Faculty of Languages and Translation, University of Aden, Yemen² Dept. of English, Faculty of Languages and Translation, University of Aden, Yemen; E-mail: adel.dulul@gmail.com

* Corresponding author: Rania Shokri Qasem Shaher; E-mail: shahrramy29@gmail.com

Received: 10 August 2026 / Accepted 06 September 2026 / Published online: 30 September 2026

Abstract

The study entitled, *A comparative Analysis of the Reliability of ChatGPT and Human Translation with Reference to Literary Texts*, aimed to assess the reliability of ChatGPT and human translation with respect to literary texts. It explores the effectiveness of both translation procedures in terms of accuracy, fluency, coherence, cultural relevance and general acceptability. The study is descriptive in nature and was conducted with mixed-method approach, which is a combination of qualitative and quantitative methods. This study has also taken into account a corpus of three literary texts. The source texts were translated by ChatGPT and compared with human translations. The identification and categorization of translation errors were based on a modified Multidimensional Quality Metrics (MQM) framework for the quantitative analysis, as well as the evaluators' assessment of the translation acceptance provided by a ranking questionnaire. Descriptive statistical evaluations were used to compare the acceptance of the two translation techniques. The qualitative analysis was based on feedback provided by six evaluators as well. The results indicate that both the human and ChatGPT translations were of acceptance in most cases. However, the human translations were always deemed more acceptable, and showed higher inter-rater agreement. It has been found that Human translation achieved a mean score of 8.42, compared with 7.89 for CHT, showing that human translation was significantly better in the translation of literary texts in terms of better preserving style, rhetorical devices, emotional nuances and cultural embedded meanings. While the translation capability of the ChatGPT was effective, it had limitations when it came to stylistic and cultural nuances. The study concludes that ChatGPT has considerable potential to support professional translation tasks but cannot replace human translators in tasks that require creativity, cultural understanding and literary subtlety. The results are a contribution to the emerging field of AI-assisted translation, and they provide tangible benefits for translators, educators and scholars.

Keywords: AI tools; Artificial intelligence; ChatGPT; Human translation; Reliability; Translation quality; MQM.

Introduction

New machine translation technology has a significant impact on intercultural communication. Although machines can better translate languages, there is still scope for improvement as technology and globalization drive the need for global information exchange in various fields. The world's swift growth brings both challenges and benefits. Translation is fundamental for the transmission of knowledge and for the promotion of intercultural dialog. Despite the recent developments in machine learning and artificial intelligence translation, especially in ChatGPT, doubts still linger in terms of accuracy, context and technical terms. According to translation scholars like Catford (1965), Nida (1964), Newmark (1988) and Ghazala (1995), a good translation should convey the intended meaning of the source text and preserve proper linguistic and communicative equivalence. This is especially true for literary and specialized writing, where it is imperative

to retain meaning, style, cultural nuances, and terminology. ChatGPT may struggle with ambiguous, historical, legal, theological, and literary content, especially when specialized terminology and contextual meanings are involved, even though it can process large amounts of text and produce fluid translations. Therefore, human expertise, supervision and editing are required to ensure the preservation of the accuracy and the original meaning and stylistic features of the source material.

Statement of the Problem

The rapid expansion of artificial intelligence (AI) in translation, particularly through large language models like ChatGPT, presents a significant challenge to traditional human translation practices. While AI-powered systems offer unprecedented speed, cost-effectiveness, and accessibility, their reliability and effectiveness in translating complex and nuanced text genres, such as literary, legal, and political documents remain subjects of critical inquiry. The central issue lies in the tension between the automated efficiency of AI and the sophisticated, context-dependent interpretation that has historically been the domain of professional human translators. This study is motivated by the need to understand the practical implications of this tension, particularly as AI becomes more integrated into professional workflows.

Newmark (1988) and Ghazala (1995) argued that the problem is further compounded by the unique demands of specialized text genres. Literary translation requires a deep appreciation of aesthetic, emotive, and cultural nuances. Foundational translation theories emphasize that these elements are central to the original work's artistic integrity and communicative power. Recent studies (Al Rousan, Jaradat, & Malkawi, 2025; Wang, 2025) indicate that while ChatGPT shows promise in creative tasks, it often fails to capture the subtle cultural and stylistic layers required for high-quality literary translation, leading to a loss of the original's "soul". Despite the increasing use of AI in professional translation, a significant research gap persists in the direct comparative analysis of AI and human translation specifically for literary texts. Much of the prior research has focused on either broad assessment of machine translation or on technical and functional documents, leaving a void in our understanding of how AI handles genres that demand high levels of contextual sensitivity, cultural awareness, and specialized knowledge. While some recent studies have begun to explore these areas individually (Al Rousan, Jaradat, & Malkawi, 2025; Blend, 2025; Motoki, 2025), there is a lack of comprehensive, comparative research that evaluates ChatGPT's reliability across all three genres simultaneously. This study seeks to address this gap by systematically evaluating the performance of ChatGPT against that of professional human translators across literary texts. The findings aim to provide translation specialists with practical insights into when AI can be a reliable tool and when human expertise remains indispensable. Furthermore, the results will provide AI developers with empirical evidence of current system limitations. By providing a rigorous comparative analysis, this study will contribute to a more nuanced understanding of the evolving relationship between AI and human expertise in the field of translation. This study provides tangible benefits for translators, educators and scholars in literary field.

Research Objectives

This study tries to achieve the following objectives:

1. To compare the accuracy of ChatGPT-generated translations and professional human translations across selected literary texts in English and Arabic, using predefined criteria of accuracy, fluency, consistency, and cultural appropriateness.
2. To analyse the extent to which ChatGPT translations preserve cultural references, idiomatic expressions, and contextual meanings, and to compare these findings with human translators' performance in the same genre.
3. To categorize the most frequent types of translation errors (e.g., omissions, additions, distortions, mistranslations) in ChatGPT outputs, and to contrast them with error patterns found in human translations.

Research Questions

1. How do ChatGPT-generated translations compare to professional human translations in terms of accuracy, fluency, consistency, and cultural appropriateness across literary texts in English and Arabic?
2. To what degree do ChatGPT translations preserve cultural references, idiomatic expressions, and contextual meanings in literary texts, when evaluated against human translations using content analysis and expert ranking questionnaires?
3. What types of translation errors (e.g., omissions, additions, distortions, and mistranslations) occur most frequently in ChatGPT outputs, and how do these error patterns differ from those identified in human translations of the same texts?

Significance of the Study

The purpose of this study is to conduct a comparative analysis of the reliability of ChatGPT and human translations of literary texts; it stems from its potential to answer key questions about the reliability of AI-based translation tools like ChatGPT in comparison to human translation, especially when dealing with complex text genres like literary documents. These areas of translation necessitate not only linguistic correctness, but also a thorough grasp of context, culture, and intent, making the research both topical and significant. The increasing dependence on AI in translation raises ethical and social concerns concerning the possible displacement of human translators, as well as the risks of relying too heavily on faulty AI systems. This study has addressed these problems by highlighting the situations in which human assistance is critical. Offering a balanced view on AI's role in translation, emphasizing its use as a tool to supplement rather than replace human knowledge. These findings may help shape discussions about the proper use of AI in domains where accuracy and cultural sensitivity are critical. In brief, this study is significant because it provides a much-needed evaluation of how ChatGPT compares against human translations when dealing with these types of texts. The findings may have far-reaching ramifications for industries, researchers, and developers, providing insights into the capabilities, limitations, and ethical considerations of using ChatGPT in crucial translation fields. It will also help professionals and institutions understand when they can rely on ChatGPT and when human expertise is required.

Delimitations of the Study

This study focuses on ChatGPT translation of literary texts compared to human translation. Two texts are in English translated into Arabic and one text in Arabic translated into English. In addition, the researcher examined the accuracy of translations generated by ChatGPT compared to those produced by human translators in various contexts. Moreover, she investigated which types of literary texts have shown the greatest differences in accuracy between ChatGPT and human translation.

Literature Review

This part deals with the relevant literature related to the topic. It will display a general introduction to literary translation, Reliability, Accuracy, Artificial Intelligence, and human translation vs. Machine Translation

Literary Translation

Literary Translation plays an important role in increasing awareness and understanding among diverse cultures and nations. Literary translations in particular help these different cultures reach a compromise. The increasing interest in the literature of other languages has required more studies regarding the problems of literary translation. A literary translation must reflect the imaginative, intellectual, and intuitive writing of the author. In fact, literature is distinguished by its aesthetics, especially novels. Literary translation, which is

probably the most famous of the many types of translation, is the translation of literary texts, and the people who are responsible for translating literary works (novels, short stories, poetry, etc.) are literary translators. According to Bush (1998), literary translation is "an original subjective activity at the center of a complex network of social and cultural practices" (p. 127).

Reliability

Reliability, in the context of translation, builds upon these quality components. It refers to the consistency, trustworthiness, and dependability of the translation output (O'Brien, 2012). Translation reliability can be defined as the extent to which a translation maintains the intended meaning and communication function of the source text and depends on many fundamental components: accuracy, fluency, consistency, and cultural appropriateness. These factors influence not only the quality of the translation but also its acceptability across linguistic and cultural boundaries (House, 2015, p. 23–26). The following overview examines the scholarly discourse surrounding each of these dimensions, giving both foundational theories and contemporary empirical research.

Accuracy in Translation

Accuracy is a critical parameter in Translation Quality Assessment, concentrating on how well the translated text transmits the meaning of the original material. Safei et al. (2018) recommend analyzing the source language text's vocabulary, grammatical structure, communication scenario, and cultural context before recreating its meaning in the target language. Koby et al. (2018, p126) define a high-quality translation as one that fully transfers the message from the source language to the target text, including "denotation, connotation, nuance, and style".

Fluency in Translation

Fluency is a key indicator in Translation Quality Assessment, concentrating on the linguistic quality of the target text regardless of its link to the source. Kuwanto et al. (2024) analyze if a translation "reads smoothly and naturally" for native speakers of the target language.

Artificial Intelligence:

Artificial intelligence (AI) aims to enable computers to perform tasks that humans can (Boden, 2016, p. 1). Since the Dartmouth Conference in 1956, the field has developed significantly across various areas, culminating in the third generation. The development of AI can be categorized into three major phases: the first generation, based on symbolic and rule-based systems, the second generation, with machine learning and statistical methods, and the third generation, characterized by deep learning and generative AI technologies, which facilitate advanced language processing and intelligent applications (Nilsson, 2010; Russell & Norvig, 2016, Goodfellow et al., 2016)

Human Translation vs. Machine Translation

The comparison between human and machine translation quality has been an ongoing area of research, evolving with each new MT paradigm. Early studies often highlighted the significant gap between human capabilities and the often-flawed outputs of Rule-Based Machine Translation (RBMT) and Statistical Machine Translation (SMT) systems. With the advent of Neural Machine Translation (NMT), this gap began to narrow for certain types of text and language pairs, leading to claims that NMT had reached or even surpassed human quality in some instances. However, such claims often relied on automatic metrics or limited human evaluations, and critical examinations frequently revealed persistent shortcomings, especially concerning nuance and accuracy (Toral & Way, 2018).

Research Methodology

Research Design

The major concern of this study is investigating the reliability of CHATGPT translation in comparison with human translation. Although ChatGPT has the ability to translate various languages, this study chooses Arabic and English texts. The researcher used a mixed methods approach. To analyse the data collected from the texts, the researcher employed a Comparative Textual Analysis, a Quantitative Textual Analysis utilizing Multidimensional Quality Metrics (MQM) framework. Ranking Questionnaire was analysed using SPSS program.

Data Collection Methods:

To conduct a comparative analysis of ChatGPT translation and human translation, the researcher employed a comparative textual analysis method. The study has adopted a mixed-methods approach, focusing on key reliability features: accuracy, fluency, consistency, and cultural appropriateness. Data collection and assessment will be carried out using the following instruments:

- **Comparative Textual Analysis** The researcher applied a comparative analysis technique to collect the required qualitative data. The analysis focused on linguistic precision, style, and cultural adaptation in literary texts. It is a qualitative research technique used to examine similarities and differences across multiple data sets, texts, or cases. A quantitative textual analysis in conjunction with the Multidimensional Quality Metrics (MQM) framework has been used to compare the collected data. In the context of this study, it allows the researcher to systematically compare translations to identify patterns of variation between ChatGPT and human translation in terms of accuracy, fluency, consistency, and cultural appropriateness, which are the core reliability features.
- A ranking questionnaire was developed and administered to professional translators. Participants were asked to evaluate and rank the translations according to predefined criteria, including accuracy, fluency, consistency, and cultural appropriateness. This instrument was used to collect quantitative data for the comparative analysis.

Sampling procedures

- The study employed a purposive sampling technique to select the literary texts included in the dataset. This approach was appropriate as the study focused on evaluating translation quality through specific literary features and translation challenges rather than achieving statistical generalizability. The selected texts were chosen based on their literary significance, relevance to the research objectives, and suitability for analysis using the Multidimensional Quality Metrics (MQM) framework. The dataset consisted of **three literary texts** representing English–Arabic and Arabic–English and translation directions, including the first is an extract from Shakespeare's *King Lear* translated by Mohamed Anani, the second is Shakespeare's *Sonnet 37* translated by Bader Tawfik (1988), and the third is an extract from Al-Mutanabbi's *Waharrā Qalbāh* translated by Mahmoud Abbas Al-Aqqad. These texts were selected due to their rich linguistic, cultural, and stylistic features, enabling the examination of accuracy, fluency, cultural appropriateness, and interpretive choices. To ensure analytical reliability, the translations were evaluated using predefined MQM categories and consistent criteria, allowing systematic identification and classification of translation issues in both human and AI-generated translations

Validity and Reliability of Data Collection Methods

Validity refers to the extent to which a data collection method or instrument measures what it is intended to measure. It addresses whether the research truly captures the concept or phenomenon it purports to study. Babbie (1989) described validity as the degree to which an empirical measure adequately reflects the real meaning of the concept under consideration. For comparative textual analysis, construct validity was

established through clear operational definitions of each MQM category (e.g., accuracy, terminology, fluency), supported by illustrative examples, and confirmed through coding performed by a colleague to ensure consistent interpretation of the categories. For quantitative textual analysis content validity of the quantitative textual analysis was established by employing a customized MQM taxonomy covering key translation quality dimensions (accuracy, terminology, fluency, style, and cultural appropriateness), which was reviewed and approved by expert translators to ensure comprehensive representation of translation quality. For the ranking questionnaire, content validity was confirmed through expert review.

Reliability refers to the consistency and stability of a measurement method over repeated applications or observations. A reliable instrument yields similar results under consistent conditions, indicating low random error. For example, if a respondent answers a questionnaire item one way and gives the same answer upon retesting, that indicates reliability. Reliability of the study's data collection instruments was established through systematic procedures. For both comparative and quantitative textual analyses, inter-rater reliability was assessed by having multiple coders independently annotate translations using the MQM categories, with agreement measured, it was found that more than 70% agreement among coders, while intra-rater reliability was confirmed through re-coding a subset of texts after a two-week interval, ensuring consistent and stable coding over time. The use of a standardized MQM coding manual further supported replicability and methodological consistency. For the ranking questionnaire, the reliability of the ranking questionnaire was verified by calculating the agreement among professional translators, treating them as inter-coders, to ensure consistent evaluation of translations across the criteria of accuracy, fluency, consistency, and cultural appropriateness, where it was found that the agreement among inter-coders was 80%.

Data Analysis

Creswell describes data analysis as the process of organizing, interpreting, and integrating the collected data to uncover patterns, themes, or test hypotheses. Quantitative data analysis typically involves statistical techniques, while qualitative data analysis involves thematic or content analysis. In mixed methods research, both types of analysis are used for comprehensive interpretation (Creswell & Creswell, 2018).

Literary Text 1

Below is a systematic comparative analysis of the literary prose translation (*King Lear* extract) produced by Human Translation (HT: Mohamed Anani) and ChatGPT-5 Translation, strictly aligned with your research framework, objectives, and questions.

Table 1: Human Translation and ChatGPT Translation of Text 1 (Literary Texts)

Original Text	Human Translation	ChatGPT Translation
Lear: Why, thou wert better in thy grave than to answer with thy uncovered body this extremity of the skies. Is man no more than this? Consider him well. Thou owest the worm no silk, the beast no hide, the sheep no wool, the cat no perfume. Ha! Here is three on's are sophisticated. Thou art the thing itself; unaccommodated man is no more but such a bare, forked animal as thou art. Off, off, you lending! Mohamed, A. (2003). <i>The literary Translation between Theory and Practice</i>. Egyptian International Publishing Company p.67	لير: أن تترقد في القبر خير من أن تواجه بجسدك العاري عضبة السماوات الجائعة. ألا يزيد الإنسان عن ذلك الجسد؟ تأملوه انعموا النظر! إنك لا تدين بالحرير لدود القز، ولا للحيوان بالجلود، ولا للأغنام بالصوف، ولا للغزال بالمسك. انظروا! إن ثلاثتنا مزيفون! أما أنت فالإنسان في صورته الأصلية. فالإنسان المجرد حيوان ضعيف عار يمشي على اثنتين! فلا تخلص! إن من هذه الثياب المستعارة Translated by: Mohamed Anani	لير: لأن كنت هكذا، لكان خيرا لك أن تكون في قبرك من أن تجيب غريان الجسد على قسوة السماء. أفالإنسان ليس أكثر من هذا؟ تأمله جيدا. أنت لا تدين للدود بدبياج، ولا للوحش بجلد، ولا للغنم بصوف، ولا للقطعة بعطر. ها! هاهنا ثلاثة مَن قد تكلفوا أنت هو الشيء نفسه؛ فالإنسان بلا غدة ولا لتزيين زينة ليس إلا حيوانا عاريا مشقوق الساقين كهذا الذي! أنت عليه. انزعوا، انزعوا، يا أثواب السلف Translated by: Chat Gpt5

Analysis and Discussion Text 1

In line with the topic of this study “A Comparative Analysis of the Reliability of ChatGPT and Human Translation with Reference to Literary, Legal and Political Texts”, this subsection addresses the literary genre and seeks to answer the three research questions by applying the MQM framework to a parallel human and AI translation of a Shakespearean prose passage. Reliability is operationalized through accuracy, fluency, consistency, and cultural appropriateness, using both qualitative and quantitative analysis. The quantitative textual analysis was conducted following the qualitative comparative textual analysis of the same data.

Accuracy

Accuracy is operationally defined as the degree to which the target text faithfully reproduces the source text’s semantic content, propositional meaning, and pragmatic intent, without unjustified omission, addition, or distortion. The qualitative textual analysis confirms that both translations achieve a high level of propositional accuracy, yet with measurable differences in reliability. The human translation by Mohamed Anani shows greater semantic stability and closer adherence to the source text’s philosophical structure. For example, the evaluative judgment in “*thou wert better in thy grave*” is rendered as «*أن ترقد في القبر خير من أن ...تواجه*», which faithfully reproduces the comparative force and rhetorical severity of the original. Similarly, the philosophical notion of “*unaccommodated man*” is accurately preserved in «*الإنسان المجرد*», maintaining Shakespeare’s ontological reflection on human vulnerability.

By contrast, while ChatGPT’s translation remains largely accurate, it demonstrates occasional semantic drift due to over-interpretation, such as reframing imperatives and intensifiers in ways not explicitly licensed by the source text. The most serious accuracy-related problem appears in the final clause “*Off, off, you lending!*”, which ChatGPT renders as «*يا أثواب السلف*». This represents a lexical–semantic distortion, as “*you lending*” refers to borrowed garments, not ancestry or predecessors. This error directly affects referential meaning and constitutes a reliability weakness in the AI translation. Consequently, in response to Research Question 1, the findings indicate that human translation remains more reliable in preserving core semantic and philosophical meanings in literary texts, despite occasional stylistic awkwardness.

Fluency

Fluency is a key indicator in Translation Quality Assessment, concentrating on the linguistic quality of the target text regardless of its link to the source. Kuwanto et al. (2024) stated that if a translation “reads smoothly and naturally” for native speakers of the target language. Popović defined fluency as “how much the translated text adheres to the structural system of the target language” (Popovic, 2020). This criterion differs from accuracy in that it evaluates the translation as a solo text rather than comparing it to the original content. With regard to fluency, the human translation exhibits superior stylistic smoothness and rhetorical force, conforming closely to established conventions of Arabic literary discourse. Anani’s rendering, for instance, «*أن ترقد في القبر ...خير من أن تواجه بجسدك العاري غضبة السماوات الجائحة*», effectively captures the dramatic intensity and elevated tone of Shakespeare’s original “*thou wert better in thy grave than to answer with thy uncovered body this extremity of the skies*”. The Arabic formulation is idiomatic, rhythmically balanced, and stylistically resonant, reflecting a high level of genre awareness. By contrast, although the ChatGPT translation, grammatically well-formed, shows stylistic rigidity in places, notably in «*حيوانًا عاريًا مشقوق الساقين*», which is accurate but pragmatically marked and less idiomatic in Arabic literary discourse. From an MQM perspective, this constitutes a minor fluency-related stylistic issue, rather than a semantic error.

Consistency

In terms of consistency, the human translation maintains a uniform register throughout, sustaining the tragic–philosophical tone without noticeable shifts. Conversely, ChatGPT displays register fluctuation, alternating between classical diction («*تأمله حيًا*») and semi-modern abstractions («*بلا عُدّة ولا زينة*»). This inconsistency does not distort meaning but affects textual cohesion and stylistic reliability. These findings

contribute to Research Question 3, showing that register inconsistency and stylistic rigidity are among the most frequent non-critical issues in ChatGPT's literary translation.

Cultural Appropriateness

The most significant divergence between the two translations lies in cultural mediation. The human translation demonstrates functional equivalence and adaptive cultural recontextualization, as seen in the metaphor sequence: لا تدين بالحرير لنود الفز... ولا للغزال بالمسك.

Although the substitution of "cat" with "gazelle" constitutes a **lexical inaccuracy**, the strategy itself reflects an attempt to activate culturally resonant imagery for the Arabic reader. From an MQM perspective, this is a **culture-oriented shift**, albeit imperfectly executed. In contrast, ChatGPT adheres to **formal equivalence**, rendering the sequence literally («ولا للقطعة بعطر»). While semantically accurate, this choice reduces pragmatic and aesthetic impact in Arabic, where such imagery lacks the same cultural resonance. This pattern exemplifies **cultural literalism**, a recurrent characteristic of AI-generated literary translations. Accordingly, in response to **Research Question 2**, the analysis shows that **ChatGPT captures surface-level cultural references but struggles with deeper cultural re-encoding**, whereas the human translator demonstrates greater sensitivity to target-culture expectations.

Quantitative textual Analysis

Table 2: Comparative Analysis of Human and ChatGPT Translation

MQM Category	Human Translation (HT)	ChatGPT Translation (CHT)
Accuracy	1 (16.7%)	1 (16.7%)
Fluency	0 (0%)	1 (16.7%)
Consistency	0 (0%)	2 (33.3%)
Cultural Inappropriateness	0 (0%)	1 (16.7%)

The comparative analysis reveals a marked difference between Human Translation (HT) and ChatGPT Translation (CHT). Human Translation produced only one notable issue (16.7%), whereas ChatGPT Translation produced five issues (83.3%), representing a difference of 66.6 percentage points and a fivefold higher observed error frequency in the ChatGPT output. At the category level, both translation modes recorded one accuracy-related error (16.7%), indicating comparable performance with respect to accuracy. However, ChatGPT Translation generated additional problems in fluency (16.7%), consistency (33.3%), and cultural appropriateness (16.7%), whereas no errors were identified in these categories for Human Translation. Consistency constituted the most prominent weakness in the ChatGPT output, accounting for two of the five identified notable issues. Overall, the findings suggest that Human Translation demonstrated considerably fewer notable errors than ChatGPT Translation in the evaluated data.

Quantitatively, ChatGPT exhibits a significantly higher concentration of MQM-notable issues, primarily related to register control and cultural literalism, rather than meaning loss. This distribution pattern supports findings in prior studies on neural and generative AI translation, which report that AI systems tend to prioritize semantic continuity over stylistic and cultural optimization in literary discourse (Toral & Way, 2018; Daems et al., 2017).

Quantitative Comparative Analysis of Translation Errors

Table 3 presents a quantitative comparison of the error patterns identified in the Human Translation (HT) and ChatGPT Translation (CT). Overall, ChatGPT produced a higher number of errors across most categories, indicating a greater incidence of translation problems compared with the human translation.

Table 3: Quantitative comparison of the error patterns identified in the Human Translation (HT) and ChatGPT Translation

Error Category	Human Translation	ChatGPT Translation
Omissions	0	0
Additions	1	2
Distortions	1	2
Terminology	1	3
Style / Register	0	2
Total	3	9

Table 3 reveals that the human translation recorded 3 errors, whereas ChatGPT recorded 9 errors. Thus, ChatGPT produced 6 more errors, representing a 200% higher total error count than the human translation. Put differently, ChatGPT's total number of identified errors was three times that of the human translation. The largest difference occurred in Terminology, where ChatGPT produced 3 errors compared with 1 error in the human translation. This represents a 200% increase in terminology errors. Similarly, ChatGPT recorded 2 additions and 2 distortions, compared with only 1 addition and 1 distortion in the human translation, representing a 100% increase in each category. A notable difference was also observed in Style/Register. The human translation contained no errors in this category, whereas ChatGPT produced 2 errors. This suggests that ChatGPT's translation was comparatively more vulnerable to stylistic or register-related problems. However, this difference should be interpreted cautiously because a percentage increase cannot be calculated when the baseline value is zero. In contrast, both translation approaches recorded zero omission errors, indicating that neither translation exhibited omissions according to the adopted error-identification criteria.

To sum up, the quantitative results suggest that the human translation demonstrated better performance in terms of error frequency, with only 3 recorded errors compared with 9 in ChatGPT's translation. However, because the absolute number of errors is small, these findings should be interpreted as descriptive evidence rather than evidence of statistically significant differences, unless the study includes an appropriate inferential statistical test.

Literary Text 2

The following section presents a systematic comparative analysis of Shakespeare's *Sonnet 37*, examining the Arabic translation by Bader Tawfik (1988) alongside the ChatGPT-generated translation.

Table 4: Text 2 Literary Text (*Sonnet 37* by Shakespeare) and both translations

Original Text	Human Translation	ChatGPT Translation
As a decrepit father takes delight, To see his active child do deeds of youth, So I, made lame by fortune's dearest spite, Take all my comfort of thy worth and truth; For whether beauty, birth, or wealth, Or wit, or any of these all or all, or more, SONNET 37 by: Shakespeare ST:	حين يرى مثلما يسعد الأب الذي أقعدته الشيخوخة هكذا أنا وقد صرت طفله الذي يمارس أفعال الشباب معوقاً يفعل الطبيعة الحاقدة وأجد راحتي الكبرى في منزلتك وفي حقيقة سواء كان الجمال، أو الميلاد، أو أو أي من هذه أو كلها معاً، أو ما الثروة، أو الذكاء يزيد عليها. Translated by: Bader Tawfik (1988)	ابنه النشط يؤدي كما يفرح الأب الهرم حين يرى أخذ كذلك أنا، وقد أقعدتني قسوة الحظ، أفعال الصبا، كان الجمال، أو فإن كل عزائي من يهلك وصدقك؛ أو الذكاء، أو هذه كلها، أو أكثر. النسب، أو الغنى، Translated by: Chat GPT 5

Analysis and Discussion Text 2

This section, building on the research presented in "A Comparative Analysis of the Reliability of ChatGPT and Human Translation with Reference to Literary, Legal, and Political Texts," investigates the literary

translation of Shakespeare's Sonnet 37, contrasting a human-generated translation with one produced by ChatGPT-5. Employing the MQM framework, reliability is defined by accuracy, fluency, consistency, and cultural appropriateness; both qualitative textual and quantitative content analyses are subsequently utilized.

Accuracy

Accuracy is operationally defined as the degree to which the translated text preserves the semantic, propositional, and pragmatic meaning of the source without unjustified additions or omissions. Both translations maintain high semantic fidelity to Shakespeare's original sonnet; however, differences emerge in precision and interpretive nuance.

The human translation:

«مثلما يسعد الأب الذي أقعدته الشيخوخة حين يرى طفله الذكي يمارس أفعال الشباب هكذا أنا وقد صرت معوقاً بفعل الطبيعة الحاقدة وأجد راحتي الكبرى في منزلتك وفي حقيقتك سواء كان الجمال، أو الميلاد، أو الثروة، أو الذكاء أو أي من هذه أو كلها معاً، أو ما يزيد عليها»

The translation demonstrates considerable semantic consistency. It effectively communicates the original's emotional depth and philosophical contemplation, preserving the comparative metaphor ("as a decrepit father takes delight") and the reflective quality characteristic of Shakespeare. The expression "أقعدته الشيخوخة" accurately reflects "decrepit father," thereby retaining both the physical deterioration and the emotional warmth.

In contrast, ChatGPT's translation:

«كما يفرح الأب الهرم حين يرى ابنه النشيط يؤدي أفعال الصبا، كذلك أنا، وقد أقعدتني قسوة الحظ، أخذ كل عزائي من فضلك وصدقك؛ فإن كان الجمال، أو النسب، أو الغنى، أو الذكاء، أو هذه كلها، أو أكثر»

The translation is largely accurate in its propositional meaning, though it exhibits a slight semantic shift. The phrase "أقعدتني قسوة الحظ" ("crippled by the cruelty of fate") amplifies the original tone somewhat when compared to "fortune's dearest spite," thereby introducing a moralized emotional weight that is not present in the more restrained Shakespearean expression. This exemplifies interpretive expansion, a minor divergence within MQM's accuracy dimension.

In general, both translations effectively preserve meaning; however, the human translation demonstrates greater philosophical and emotional accuracy, remaining closer to Shakespeare's tone of dignified melancholy.

Overall, ChatGPT achieves **propositional accuracy**, yet occasionally compresses Shakespeare's rhetorical unfolding for brevity. At the same time, Anani's version sustains both propositional and structural fidelity, demonstrating slightly higher reliability under the MQM *accuracy* metric. The findings indicate that human translation is more dependable in preserving the semantic depth and emotional equilibrium of literary texts. In contrast, ChatGPT, despite its accuracy, tends to exhibit emotional or stylistic overextension.

Fluency

Fluency refers to the **naturalness and readability** of the target text in its own linguistic system, independent of the source.

The human translation exhibits a heightened degree of fluency and a natural cadence characteristic of Arabic literary tradition. The expression «وأجد راحتي الكبرى في منزلتك وفي حقيقتك» demonstrates idiomatic accuracy, conforming to the stylistic expectations of elevated Arabic prose while preserving the poetic gravitas of Shakespeare's original work. Lexical selections such as «الطبيعة الحاقدة» elicit a classical register, thereby aligning with established Arabic literary conventions.

Conversely, ChatGPT's version, although grammatically sound and stylistically polished, occasionally manifests formulaic constructions and syntactic constraints. ChatGPT's rendering is grammatically correct but stylistically flatter. Expressions like «أخذ كل عزائي من فضلك وصدقك» are semantically valid but syntactically ordinary, lacking the rhetorical sonority expected in literary Arabic. The lexical choice of «فضلك» (your virtue)

and «صدقك» (your truthfulness) simplifies the complex moral-philosophical nuance of “worth and truth,” which in Shakespeare connotes intrinsic value and integrity rather than social politeness or honesty. Thus, in *fluency* terms, ChatGPT demonstrates coherence but limited stylistic sophistication. Anani’s translation exhibits superior idiomatic fluency and literary tone alignment, affirming the human translator’s greater sensitivity to rhetorical appropriateness. Furthermore, the parallel structure of the clauses exhibits a more mechanical quality, thereby diminishing emotional impact.

Within the framework of MQM, the observed problem constitutes a minor reduction in fluency, primarily of a stylistic nature rather than a grammatical one. While the AI-generated translation is deemed to be “readable,” it lacks the aesthetic embellishments characteristic of Arabic literature produced by human authors.

Concerning fluency, human translation once more demonstrates a more organic and rhetorically balanced presentation, thereby exhibiting a closer alignment with the literary conventions of the target language.

Consistency

Translation consistency necessitates the preservation of a uniform register, tone, and lexical coherence throughout the text. The human translation successfully maintains a consistent philosophical and contemplative tone, thereby adhering to the elegiac mood of the original sonnet. The register is consistently elevated, and lexical fields, including those pertaining to “virtue,” “truth,” and “worth,” are treated with coherence. Conversely, ChatGPT’s rendition exhibits occasional fluctuations between modern phrasing (“أبهرم”) and classical poetic diction (“أقعدتني قسوة الحظ”), resulting in mild register inconsistency. Although the meaning remains intact, this tonal variation diminishes the translation’s stylistic unity.

From the perspective of the MQM framework, this inconsistency constitutes a non-critical cohesion issue; however, it does impact textual polish and reader immersion.

Cultural Appropriateness

Cultural appropriateness necessitates a translator’s capacity to transpose imagery and conceptual frameworks into the target culture, all while preserving the original’s aesthetic and functional integrity. Within this domain, human translation demonstrates superiority. It deftly recontextualizes Shakespearean imagery within an Arabic poetic structure, avoiding any distortion of meaning. The phrase «وأجد راحتي الكبرى في منزلتك» exemplifies an understanding of Arabic rhetorical preferences, specifically the use of parallel noun structures, thereby augmenting the text’s cultural resonance. Conversely, the ChatGPT translation prioritizes formal equivalence to the source text, yet it exhibits limited cultural adaptation. Its expressions are largely literal, emphasizing semantic correspondence over cultural resonance. For example, “أفعال الصبا” is grammatically correct but lacks the connotative depth that could be achieved through culturally nuanced metaphorical phrasing, such as “مظاهر الشباب” or “روح الشباب.”

Consequently, although the AI-generated translation demonstrates sufficient cultural neutrality, it falls short of the complete poetic re-encoding anticipated in Arabic literary translation. This disparity highlights that human translation more successfully incorporates the cultural and poetic expectations of the target language, whereas ChatGPT’s adherence to literalism constrains its aesthetic as well as cultural richness.

In overall, the comparative examination of Sonnet 37 reveals that ChatGPT’s translation demonstrates considerable structural and semantic fidelity, thereby illustrating notable advancements in AI’s capacity for literary translation. Nevertheless, human translation is deemed more dependable when assessed comprehensively via the framework, particularly excelling in semantic precision, cultural interpretation, and idiomatic expression. Consequently, while AI translation instruments like ChatGPT can furnish accurate initial drafts, the human element, encompassing interpretive insight and cultural sensitivity, remains essential in literary translation, especially for texts distinguished by philosophical depth and emotional intricacy, such as Shakespeare’s sonnets.

Quantitative textual Analysis

Table 5: Error categorization in both translations

Reliability Category	Human Translation (HT)	ChatGPT Translation (CHT)
Accuracy	16.7%	16.7%
Fluency	0%	33.3%
Consistency	0%	16.7%
Cultural Inappropriateness	0%	16.7%
Total MQM-notable issues	16.7%	83.3%

The comparative analysis of reliability aspects reveals a marked difference between Human Translation (HT) and ChatGPT Translation (CHT). Human Translation recorded only one MQM-notable issue (16.7%), classified as an accuracy-related interpretive addition, whereas ChatGPT Translation recorded five notable issues (83.3%). Although both translation conditions exhibited one accuracy-related issue (16.7%), ChatGPT Translation generated additional problems in fluency, consistency, and cultural appropriateness. Fluency constituted the most prominent weakness in the ChatGPT output, with two stylistic-rigidity errors (33.3%), followed by one consistency error associated with register shift (16.7%) and one cultural-appropriateness error associated with literalism (16.7%). In contrast, Human Translation recorded no errors in these three dimensions. Overall, the findings indicate that Human Translation demonstrated substantially fewer MQM-notable issues and greater consistency across the evaluated quality dimensions, whereas the ChatGPT translation showed a broader range of stylistic, register-related, and culturally sensitive weaknesses.

Quantitatively, the ChatGPT translation demonstrates a significantly greater occurrence of MQM-identified problems. The most prominent of these are register inconsistency and literal cultural transference, both of which impact stylistic cohesion rather than the fundamental meaning. The accuracy deviation, exemplified by the intensified “أفعدنتني قسوة الحظ” for “fortune’s dearest spite,” represents a semantic amplification that diverges from the pragmatic tone of Shakespeare’s original text.

This pattern aligns with observations in contemporary neural and generative translation research, which highlights AI translation’s proficiency in syntactic-semantic mapping while simultaneously revealing a consistent deficiency in pragmatic style control and target-culture reconstitution (Toral & Way, 2018; Daems et al., 2017).

Table 6: Error categorization in both translations

Error Category	Human Translation	ChatGPT Translation
Omissions	0	0
Additions	0	1
Distortions	1	3
Terminology	0	2
Style / Register	0	0

Table 6 shows that the Human Translation (HT) contains 1 identifiable error, compared with 6 errors in the ChatGPT Translation (CHT). The HT records only one distortion, while no omissions, additions, terminology, or style/register errors are identified. In contrast, CHT contains 3 distortions, 2 terminology errors, and 1 addition, with no omissions or style/register problems. Thus, the most prominent error category in the ChatGPT version is distortion (3 instances), followed by terminology (2 instances) and addition (1 instance). The absence of omission and style/register errors in both translations suggests that both versions successfully retained the source-text content and maintained an appropriate literary register. However, the substantially higher frequency of distortions and terminology errors in CHT indicates greater semantic and lexical deviation from

the source text in this particular translation. Overall, the frequency data show that HT performs more reliably than CHT for this text, with only 1 error compared with 6 in ChatGPT, suggesting that the human translator achieved closer semantic and terminological correspondence, whereas ChatGPT introduced more interpretive and lexical deviations. These findings also demonstrate why the reliability of AI translation cannot be generalized across texts: the relative performance of HT and CHT varies according to the linguistic, literary, and cultural characteristics of the source text.

Literary Text 3: Poetry written by (Al-Mutanabbi's *Waharrā Qalbāh*)

The following section presents a systematic comparative analysis of Al-Mutanabbi's *Waharrā Qalbāh*, examining the English translation by Mahmoud Abbas Alaqaad (n.d) alongside the ChatGPT-generated translation.

Table 7: Text 2 Literary Text

Original Text	Human Translation	ChatGPT Translation
واحر قلباه ممن قلبه ومن بجسمي وحالي عنده مالي اكتم حيا قد يرى جسدي وتدعي حب سيف الدولة الامم ان كان يجمعنا حب فليت انا بقدر الحب قد زرتة وسيف الهند وقد نظرت اليه والسيوف دم فكان احسن خلق الله كلهم وكان احسن ما في الاحسن فوت العدو الذي يمتنه ظفر في طيه اسف في طيه واحر قلباه: أبو الطيب المتنبي	*My heart is aflame, burning with love for you While your heart is frigid-cold toward me You think so lightly of me, treating me with such indifference My soul is sickened, my body debilitated *Why should I conceal a love that has consumed my body? When pretenders fake the love of Saif-Al-Dawlah *If what brings us together is our common love for you I wish we would meritoriously share your bounties Each according to the love he harbors for you *I have been in your presence while rapiers sheathed And gazed at you when swords blood-stained And found you to be the handsomest of God's creation Superb in manners, incomparable in character *When you charge an enemy and he escapes You are the winner just the same Albeit disappointing, such escape is yet a blessing Translated by Mahmoud Abbas Alaqaad	*Alas, how my heart burns, while his is cold—and to him my body and my state seem naught but illness. *Why should I hide a love that has worn my body thin, while nations lay claim to loving Sayf al-Dawla? *If what unites us is love for his radiant brow, would that we divided that love according to our shares! *I visited him while the Indian swords were sheathed; I gazed upon him while the swords were dripping blood. *He was the fairest of all God's creation, and of all that is fairest, his character was the fairest. *The escape of the foe you set your face toward is victory—in it lies their regret, and within it lie your boons. Translate by: ChatGPT-5

Results and Discussion

Accuracy

The human translation (HT) demonstrates greater semantic accuracy and preserves the rhetorical nuances, metaphors, irony, and paradoxes of Al-Mutanabbi's poetry more effectively. ChatGPT generally conveys the core meaning accurately but occasionally introduces explanatory elements, over-interprets certain expressions, or simplifies complex poetic meanings, resulting in minor semantic shifts.

Fluency

HT shows superior fluency, idiomaticity, and stylistic naturalness in English. Its poetic diction and expressive phrasing contribute to a smoother and more rhetorically effective translation. Although ChatGPT's

translation is grammatically accurate, some expressions are overly literal, archaic, or rigid, which reduces its naturalness and stylistic flexibility.

Consistency

HT maintains a consistently elevated, formal, and poetic register that corresponds well with the tone of the original poem. ChatGPT displays some fluctuations between classical, archaic, and explanatory styles. These variations do not substantially affect meaning but somewhat weaken the stylistic cohesion of the translation.

Cultural Appropriateness:

HT demonstrates stronger cultural mediation by adapting culturally marked expressions and martial imagery to make them more accessible to English-speaking readers while retaining their intended effect. ChatGPT tends to favor literal or formal equivalence, preserving surface-level cultural references but sometimes failing to reproduce their deeper cultural and pragmatic resonance.

Overall Comparative Analysis of AI and HT

Overall, the comparative analysis indicates that the human translation performs better than ChatGPT across the four MQM dimensions examined. HT demonstrates greater accuracy in conveying semantic and rhetorical nuances, higher fluency and stylistic naturalness, stronger consistency of register, and greater sensitivity to cultural and literary context. ChatGPT, while generally accurate and grammatically well-formed, shows limitations in handling poetic ambiguity, stylistic subtlety, culturally embedded expressions, and register consistency. These **findings suggest that** AI translation can effectively convey the general meaning of literary texts but still requires human expertise, particularly when translating poetry that involves metaphor, irony, cultural references, and highly nuanced literary language.

Quantitative Comparative Analysis

Table 8: Comparative MQM-Based Reliability of the Human and ChatGPT Translations

Reliability Category	Human Translation (HT)	ChatGPT Translation (CHT)
Accuracy	66.7%	33.3%
Fluency	0%	16.7%
Consistency	16.7%	0%
Cultural Inappropriateness	16.7%	0%

Table 8 shows the MQM-based analysis reveals a noticeable difference in the reliability profiles of the Human Translation (HT) and ChatGPT Translation (CHT). Accuracy constitutes the largest proportion of MQM-notable issues in both translations, accounting for 66.7% of the issues identified in the human translation, compared with 33.3% in the ChatGPT translation. This substantial difference suggests that the human translation exhibits a greater proportion of problems involving semantic interpretation, lexical choice, and departures from the source meaning. In particular, the human translator's use of interpretive additions and semantic substitutions appears to contribute to the higher accuracy-related error rate.

By contrast, the **33.3%** accuracy rate for ChatGPT indicates that its translation contains substantially fewer accuracy-related problems than the human version. This relatively lower percentage suggests that ChatGPT generally maintains a closer relationship with the semantic content of the Arabic source. Its preservation of expressions such as "Sayf al-Dawla," "Indian swords," and "his radiant brow" contributes to greater retention of the source text's historical and poetic imagery.

The results for fluency show an inverse pattern. The human translation records 0% fluency-related issues, whereas ChatGPT records 16.7%. This indicates that, although ChatGPT demonstrates relatively strong semantic fidelity, it is not equally successful in producing consistently natural English. The awkwardness of the final couplet, particularly the formulation concerning "the foe you set your face toward," illustrates how a

relatively close rendering of the Arabic structure can produce less idiomatic English. The 0% fluency-related issues in the human translation therefore suggests a clear strength of the professional human version in terms of naturalness and readability.

With respect to **consistency**, the human translation accounts for **16.7%**, while ChatGPT records **0%**. The presence of consistency-related issues in the human version can be associated with its greater use of interpretive reformulation and expansion. The translator occasionally moves away from the structural and rhetorical patterns of the Arabic original in order to produce a more aesthetically accessible English rendering. ChatGPT's 0% suggests greater consistency in maintaining the relationships and referents established in the source text.

A similar pattern occurs in cultural appropriateness. The human translation records 16.7%, whereas ChatGPT records 0%. This difference indicates that ChatGPT more consistently retains culturally significant elements of the original rather than replacing or expanding them through culturally specific interpretations. The human translation's rendering of *يقدر الحب نقّسّم* as “share your bounties,” for example, introduces a semantic association with patronage or material benefits that is not explicitly present in the Arabic wording. Consequently, the human translation's higher percentage in this category reflects a greater degree of culturally interpretive intervention.

Overall, the distribution of MQM-notable issues suggests that ChatGPT demonstrates greater overall translation reliability in this particular literary text, particularly in relation to accuracy (33.3% versus 66.7%), consistency (0% versus 16.7%), and cultural appropriateness (0% versus 16.7%). However, the human translation demonstrates a clear advantage in fluency, with 0% of identified issues compared with 16.7% for ChatGPT. The findings therefore point to a distinction between semantic/source-text fidelity and stylistic naturalness: ChatGPT appears more reliable in preserving source-text meaning and imagery, whereas the human translation produces more fluent and aesthetically natural English but does so through greater interpretive intervention.

Thus, for this text, the lower proportion of accuracy-related issues in ChatGPT (33.3%) compared with the human translation (66.7%) provides the strongest evidence of its higher MQM-based reliability, while the 0% fluency-related issues in the human translation demonstrates its comparative strength in literary English expression. The results should, however, be interpreted as text-specific findings rather than evidence that ChatGPT is categorically more reliable than human translation, since literary translation quality can vary considerably according to genre, translator expertise, textual difficulty, and the particular characteristics of the source text.

Table 9: Error categorization in both translations

Translation Problem	Human Translation (HT)	ChatGPT Translation (CHT)
Omissions	1	0
Additions	3	1
Distortions	4	2
Terminology	1	0
Style / Register	0	1
Total	9	4

Based on the frequency in Table 10, the Human Translation (HT) contains 9 identifiable translation problems, compared with 4 in the ChatGPT Translation (CHT). In the HT, the most frequent problems are distortions (4) and additions (3), followed by omissions (1) and terminology problems (1), while no notable style/register problem was identified. This pattern indicates that the human version relies more heavily on interpretive reformulation and expansion, which occasionally moves away from the semantic content of the Arabic original. In contrast, CHT contains 2 distortions, 1 addition, and 1 style/register problem, with no

omissions or terminology problems. Thus, the frequency data show that CHT has five fewer identifiable problems than HT (4 versus 9) and particularly fewer distortions and additions. This suggests greater source-text fidelity in the ChatGPT version, whereas HT demonstrates greater interpretive freedom. However, the presence of one style/register problem in CHT indicates that its closer adherence to the source sometimes produces less idiomatic or less aesthetically natural English. In MQM terms, therefore, the frequency pattern points to higher overall reliability for CHT in terms of minimizing source-text deviations, while HT's principal strength lies in its fluent and literary presentation. MQM specifically treats additions, omissions, and mistranslations/distortions as accuracy-related problems and distinguishes style as a separate quality dimension.

Questionnaire Analysis

The semi-structured questionnaire was analysed quantitatively and qualitatively. Six evaluators participated in assessing the ChatGPT translation and human translations. The assessors were given 3 translations (Two English texts and one Arabic text). The texts were translated by human and Chatgpt. The assessors were asked to rate their level of acceptance on a scale from **0 to 10**, as well as to provide feedback on translation fluency, cultural appropriateness, creativity, and consistently.

Comparison between Human Translation and ChatGPT Translation

Then a comparative analysis was made by the evaluators' response statistics using SPSS program, where mean scores and standard deviations were calculated, to compare the acceptance levels of human and ChatGPT translations. The results of the descriptive analysis is shown in Table 11.

Table 11: Human Translation Mean Scores and Standard Deviations

Text	Mean	SD
Text 1	8.58	0.49
Text 2	8.67	0.96
Text 3	8.00	0.91
Overall mean	8.42	

Table 12 shows the Evaluators responses mean scores and standard deviations regarding **ChatGPT Translation**.

Table 12: ChatGPT Translation Mean Scores and Standard Deviations

Text	Mean	SD
Text 1	8.08	1.08
Text 2	7.67	1.11
Text 3	7.92	1.25
Overall mean	7.89	

Quantitative Comparison between Human Translation and ChatGPT Translation

A quantitative comparison was conducted using the evaluators' mean scores and standard deviations to determine the relative acceptance levels of the Human Translation (HT) and ChatGPT Translation (CHT) across the three literary texts. The comparison indicates that HT received higher mean acceptance scores than CHT in all three texts. For Text 1, HT obtained a mean score of 8.58 (SD = 0.49), compared with 8.08 (SD = 1.08) for CHT, representing a mean difference of 0.50 points in favor of HT. For Text 2, the difference was more pronounced, with HT achieving 8.67 (SD = 0.96) compared with 7.67 (SD = 1.11) for CHT, producing a 1.00-point difference. For Text 3, HT again obtained a slightly higher mean score of 8.00 (SD = 0.91), compared with 7.92 (SD = 1.25) for CHT, resulting in a relatively small 0.08-point difference. Overall, the

average mean score across the three texts was 8.42 for HT, compared with 7.89 for CHT, giving an overall mean difference of 0.53 points in favor of the human translation. The standard deviations also reveal that the evaluators' ratings were generally more consistent for HT, particularly in Text 1 (SD = 0.49), whereas CHT showed greater variability across all three texts, with SDs ranging from 1.08 to 1.25. The largest variability occurred in the ChatGPT evaluation of Text 3 (SD = 1.25). These results indicate that, descriptively, human translation was more highly accepted and more consistently evaluated than ChatGPT translation across the three texts. However, because the results presented here are descriptive statistics only, the mean differences should not be interpreted as statistically significant without an appropriate inferential test, such as a paired-samples t-test, assuming that the same evaluators rated both versions.

Recommendations

Based on the findings of the present study, the following recommendations are proposed:

1. Utilize ChatGPT as an assistant tool, not substitute of human translators. ChatGPT can assist with draft translation, providing synonyms, exploring terminology, and conducting literature searching on translation-relevant topic, but the final translations of the works should be carried out by qualified human literary translators. Demand professional human review and post-editing.
2. All human translation of literary works by means of ChatGPT require professional review by qualified translators or literary-language professionals prior to publishing or application in academic setting.
3. Refine prompts designs in human translation by ChatGPT. Translators and students should be trained with the use of the effective prompts that clearly state target audiences, genre, culture context and intended style and the intended purpose. Researchers should compare different prompts, as well as different AI versions and to systemically experiment the variety of prompting strategies; and how prompts can affect translation style and translation results. Set forth ethical standards and codes of practice for AI in translation.
4. Further studies on human and AI-assisted literary translation should carry out further investigations into larger, more diverse collections of literatures, in other pair of languages (English -Arabic direction), including genres such as poetry, novels, dramas, short fiction and literature originating from diverse cultures. Large samples of literary professionals and a variety of literary experts are required, combined with appropriate tests of inferential statistics comparisons should be conducted between various AI applications and prompt types, also explore effectiveness of human post-edition on generated translations of literary works.

Conclusions

This study compared the reliability of ChatGPT-generated and human translations for three literary works in English and Arabic. Applying methods of comparative textual analysis, a modified framework of Multidimensional Quality Metrics (MQM), and feedback from six assessors, this study took into account accuracy, fluency, consistency, cultural applicability, stylistics and preservation of meaning. Findings indicate that both translational approaches yielded generally acceptable output, but each approached had its advantages.

The MQM-based examination revealed that ChatGPT generated fewer source text related deviations within the specific literary texts examined, especially with regards to omissions, additions and certain distortions. This observation aligns with current scholarship which establishes that ChatGPT can indeed produce fluent and semantically appropriate literary translations and-when well-prompted-perform strongly on measures such as semantic accuracy, fluency and cultural adequacy (Wang, 2025). Similarly, Manapbayeva et al. (2024) in a comparison of ChatGPT-4 against professional translations of a Russian classic literary masterpiece report that generative AI is capable of playing an important role in the process of literary translation while asserting the continuing value of human professional interpreter on the aspects of style, culture and poetic meanings.

Nevertheless, the evaluator-based analysis provided herein consistently favour the human translations which overall received higher acceptance rates and received relatively more consistent ratings compared to ChatGPT. With an average acceptance score of 8.42 and 7.89 for the human and ChatGPT translation, the experts' ratings show that human translators may be more skilled at maintaining the literary tone, rhetorical features, nuances, idiomatic usage, culturally referenced expressions as illustrated by the qualitative notes.

This is supported by an extensive study which shows that published human translations outperformed LLM produced translations and concluded that LLM text quality is likely to be perceived as less linguistically and aesthetically diverse in translation studies. Similar study on EN-AR translation in multi -generics revealed that human translation excelled in aspects like culture and context, whereas ChatGPT perform well with descriptive prompts. This suggests that machine-translation performance is modulated by prompts and criteria, and surface level fluency cannot be used as proxy for literary quality. The current data provide further justification for adopting a broader approach to evaluate literary translators than pure error count; as a large-scale analysis concluded that using MQM alone cannot sufficiently capture the quality criteria in translation studies.

The combined evidence therefore suggests that there is a clear distinction between the fidelity to source text and the artistic quality of a translation; whilst ChatGPT has excellent fidelity in terms of the propositional contents, human translator may more readily achieve success in capturing aesthetic, stylistic, emotional and culture dimensions of literary discourse. Therefore, we strongly support the Adoption of human-centric human-machine collaboration model for literary translation ChatGPT can support the various stages of the translation process-including for brainstorming, ideation, translation generation, alternative wording solutions, terminology clarification, text simplification and even providing post-editing suggestions.

However, human translators should be exclusively responsible of editing/and post editing and translation validation to conform. There is a growing body of evidence in literature that indicates users appreciate fluency but are aware of machine generated limitation regarding cultural, stylistic nuances (e.g., Abdelhalim et al., 2025). Hence the relationship between ChatGPT, as an AI assistive tool and the human translator is complementary, not substitutive. Human are solely responsible for meaning-making and quality Assurance. These conclusions need to be approached with cautions considering the limitation of the present study as follow: limited corpus size of: Three texts, the sample population consisted of small scale of evaluators (six assessors), the contextual peculiarities of literary translation itself, also the study was solely based on the use of the Descriptive Statistics, rather than inferential statistics. As the difference between mean rating for Human and ChatGPT translation has no yet being statistically proved.

Further research should have to extend more the variety and more literary works within broader corpora, employ a variety of different reader group evaluators, compare other major language pair as well as genres, use different kind of prompting strategies, and investigate more thoroughly impact of post- editing on the output quality. More attention should be paid to the development and usage of literary specific model or alternative frameworks in that takes into consideration quality aspects as style, imagery, metaphor, rhythm and aesthetic.

References

- [1] Abdelhalim, S. M., Alsahil, A. A., & Alsuhaibani, Z. A. (2025). Artificial intelligence tools and literary translation: A comparative investigation of ChatGPT and Google Translate from novice and advanced EFL student translators' perspectives. *Cogent Arts & Humanities*, 12(1), 2508031. <https://doi.org/10.1080/23311983.2025.2508031>
- [2] Al Rousan, R., Jaradat, R., & Malkawi, M. (2025). ChatGPT translation vs. human translation: An examination of a literary text. *Cogent Social Sciences*, 11(1), 2472916. <https://doi.org/10.1080/23311886.2025.2472916>
- [3] Babbie, E. R. (1989). *The practice of social research* (5th ed.). Wadsworth Publishing Company.

- [4] Boden, M. A. (2016). *AI: Its nature and future*. Oxford University Press.
- [5] Bush, P. (1998). Literary translation. In M. Baker (Ed.), *Routledge encyclopedia of translation studies* (pp. 127–130). Routledge.
- [6] Catford, J. C. (1965). *A linguistic theory of translation: An essay in applied linguistics*. Oxford University Press.
- [7] Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.
- [8] Daems, J. (2022). Dutch literary translators' use and perceived usefulness of technology: The role of awareness and attitude. In J. L. Hadley, K. Taivalkoski-Shilov, C. Teixeira, & A. Toral (Eds.), *Using technologies for creative-text translation* (pp. 40–65). Routledge. <https://doi.org/10.4324/9781003094159-3>
- [9] Ghazala, H. (1995). *Translation as problems and solutions: A coursebook for university students and trainee translators*. ELGA.
- [10] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- [11] House, J. (2015). *Translation quality assessment: Past and present*. Routledge.
- [12] Koby, G. S., & Lacruz, I. (2018). The thorny problem of translation and interpreting quality. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, 16. <https://doi.org/10.52034/lanstts.v16i0.487>
- [13] Kuwanto, G., Urua, E., Amuok, P., Muhammad, S. H., Aremu, A., Otiende, V., Nanyanga, L., Nyoike, T. W., Akpan, A. D., Udouboh, N., Archibong, I., Moses, I., Ige, I. A., Ajibade, B. A., Awokoya, O., Abdulmumin, I., Aliyu, S. M., Iro, R. N., Ahmad, I. S., Smith, D., Michaels, P., Adelani, D. I., Wijaya, D. T., & Andy, A. U. (2024). Mitigating translationese in low-resource languages: The storyboard approach. *arXiv*. <https://arxiv.org/abs/2407.10152>
- [14] Manapbayeva, Z., Zaurbekova, G., Ayazbekova, K., Kazezova, A., & Pirmanova, K. (2024). AI in literary translation: ChatGPT-4 vs. professional human translation of Abai's poem "Spring." *Procedia Computer Science*, 251, 526–531. <https://doi.org/10.1016/j.procs.2024.11.143>
- [15] Masoud, M. A. (n.d.). *My heart is aflame: Abou-t-Tayyib al-Mutanabbi*. PoemHunter
- [16] Mohamed, A. (2003). *The literary translation between theory and practice*. Egyptian International Publishing Company.
- [17] Newmark, P. (1988). *A textbook of translation*. Prentice Hall.
- [18] Nida, E. A. (1964). *Toward a science of translating: With special reference to principles and procedures involved in Bible translating*. E. J. Brill.
- [19] Nilsson, N. J. (2010). *The quest for artificial intelligence: A history of ideas and achievements*. Cambridge University Press.
- [20] O'Brien, S. (2012). Translation as human–computer interaction. *Translation Spaces*, 1(1), 101–122. <https://doi.org/10.1075/ts.1.05obr>
- [21] Popović, M. (2020). Relations between comprehensibility and adequacy errors in machine translation output. In R. Fernández & T. Linzen (Eds.), *Proceedings of the 24th Conference on Computational Natural Language Learning* (pp. 256–264). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.conll-1.19>
- [22] Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3rd ed.). Prentice Hall.

- [23] Sadiq, S. (2025). Evaluating English-Arabic translation: Human translators vs. Google Translate and ChatGPT. *Journal of Languages and Translation*, 12(1), 67–95. https://jltmin.journals.ekb.eg/article_423147_3899121746ed81d7400d63401751e3f5.pdf
- [24] Safei, N. H., & Salija, K. (2018). The naturalness and accuracy of English short story translation into Indonesian. *ELT Worldwide: Journal of English Language Teaching*, 5(1), 44–53. <https://doi.org/10.26858/eltww.v5i1.5770>
- [25] Shakespeare, W. (1988). *Shakespeare's complete sonnets* (B. Tawfiq, Trans.). Akhbar Al-Youm.
- [26] Toral, A., & Way, A. (2018). What level of quality can neural machine translation attain on literary text? In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), *Translation quality assessment: From principles to practice* (pp. 263–287). Springer. https://doi.org/10.1007/978-3-319-91241-7_12
- [27] Wang, Q. (2025). Evaluating Uighur literary translation: A comparative study of ChatGPT, Google Translate, and Bing Translator. *PLOS ONE*, 20(10), e0335261. <https://doi.org/10.1371/journal.pone.0335261>
- [28] Zhang, R., Zhao, W., & Eger, S. (2025). How good are LLMs for literary translation, really? Literary translation evaluation with humans and LLMs. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 10961–10988). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.naacl-long.548>

مقالة بحثية

تحليل مقارن لموثوقية الترجمة باستخدام تشات جي بي تي والترجمة البشرية، مع الإشارة إلى النصوص الأدبية

رانيا شكري قاسم شاهر^{1*}، و عادل عبدالخالق عبدالكريم دول²

¹ قسم الترجمة، كلية اللغات والترجمة، جامعة عدن، اليمن

² قسم اللغة الانجليزية، كلية اللغات والترجمة، جامعة عدن، اليمن؛ البريد الإلكتروني: adel.dulul@gmail.com

* الباحث الممثل: رانيا شكري قاسم شاهر، البريد الإلكتروني: shahrramy29@gmail.com

استلم في: 10 أغسطس 2026 / قبل في: 06 سبتمبر 2026 / نشر في: 30 سبتمبر 2026

المُلخَص

هدفت الدراسة الموسومة بـ «تحليل مقارن لموثوقية الترجمة باستخدام تشات جي بي تي والترجمة البشرية، مع الإشارة إلى النصوص الأدبية» إلى تقويم مدى موثوقية كلٍّ من الترجمة المدعومة بتشات جي بي تي والترجمة البشرية في التعامل مع النصوص الأدبية. كما سعت إلى الكشف عن فاعلية كلتا الطريقتين من حيث الدقة، والطلاقة، والاتساق، والملاءمة الثقافية، ومستوى القبول العام. اعتمدت الدراسة المنهج الوصفي ذي المدخل المختلط، الذي يجمع بين التحليل الكمي والتحليل الكيفي. وتكوّنت مادة الدراسة من ثلاثة نصوص أدبية، تُرجمت النصوص الأصلية باستخدام تشات جي بي تي، ثم قورنت الترجمات الناتجة بالترجمات البشرية المناظرة لها. ولتحليل أخطاء الترجمة وتصنيفها، اعتمدت الدراسة إطار مقاييس الجودة متعددة الأبعاد المعدّل، في حين قُيِّمت درجة قبول الترجمات من خلال استبانة تقويم قائمة على الترتيب. كما استُخدمت الإحصاءات الوصفية للمقارنة بين مستوى قبول الترجمتين البشرية والمدعومة بتشات جي بي تي. أما التحليل الكيفي، فقد استند إلى الملاحظات والتقويمات التي قدمها ستة محكمين. وأظهرت النتائج أن كلاً من الترجمة البشرية والترجمة باستخدام تشات جي بي تي حققتا مستوى مقبولاً من الجودة في معظم الحالات. ومع ذلك، أظهرت الترجمة البشرية مستوى أعلى وأكثر اتساقاً من القبول، فضلاً عن درجة أكبر من الاتفاق بين المحكمين. فقد بلغ المتوسط الحسابي لدرجات الترجمة البشرية 8.42، مقابل 7.89 للترجمة باستخدام تشات جي بي تي، مما يشير إلى تفوق الترجمة البشرية في التعامل مع النصوص الأدبية، ولا سيما في الحفاظ على الخصائص الأسلوبية، والصور والأدوات البلاغية، والفروق الدلالية والانفعالية الدقيقة، والإيحاءات الثقافية المتضمنة في النص الأصلي. وعلى الرغم من أن تشات جي بي تي أظهر قدرة ملحوظة على أداء مهام الترجمة الأدبية، فقد كشفت النتائج عن بعض أوجه القصور في معالجة الدلالات الأسلوبية والثقافية الدقيقة. وتخلص الدراسة إلى أن تشات جي بي تي يمتلك إمكانيات وأداة بوصفه أداة مساندة في الترجمة المهنية، إلا أن هذه الإمكانيات لا تجعله بديلاً عن المترجم البشري في المهام التي تتطلب الإبداع، والوعي بالسياق الثقافي، والحساسية الأدبية والأسلوبية، والتعامل مع الدلالات الضمنية. وتُسهم نتائج الدراسة في دعم الاتجاه البحثي المتنامي نحو توظيف الذكاء الاصطناعي في الترجمة، كما تقدم دلالات تطبيقية يمكن أن تفيد المترجمين، ومدرّسي الترجمة، والباحثين في دراسات الترجمة.

الكلمات المفتاحية: أدوات الذكاء الاصطناعي؛ الذكاء الاصطناعي؛ تشات جي بي تي؛ الترجمة البشرية؛ موثوقية الترجمة؛ جودة الترجمة؛ مقاييس الجودة متعددة الأبعاد.

How to cite this article:

Shaher, R. S. Q., and A. Dulul, A. A., (2026). **A COMPARATIVE ANALYSIS OF THE RELIABILITY OF CHATGPT AND HUMAN TRANSLATION WITH REFERENCE TO LITERARY TEXTS.** *Electronic Journal of University of Aden for Humanity and Social Sciences*, 7(3), p630-650. <https://doi.org/10.47372/ejua-hs.2026.3.569>



Copyright © 2026 by the Author(s). Licensee EJUA, Aden, Yemen. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-NC 4.0) license.