3L: Language, Linguistics, Literature® The Southeast Asian Journal of English Language Studies
Vol 31(4), December 2025 http://doi.org/10.17576/3L-2025-3104-17

Can ChatGPT Translate Like a Pro? A Pilot Benchmarking Study of English–Malay Translation Quality

M. ZAIN SULAIMAN *
Universiti Kebangsaan Malaysia
zain@ukm.edu.my

INTAN SAFINAZ ZAINUDIN
Universiti Kebangsaan Malaysia

HASLINA HAROON
Universiti Sains Malaysia

ABSTRACT

Artificial intelligence (AI) tools such as ChatGPT have significantly advanced machine translation, yet their performance in low-resource language pairs, particularly English–Malay, lags behind. While existing studies have compared AI and human translation quality, most have relied on academic assessment frameworks, leaving a gap in evaluating AI translation through professional certification standards. From a professional standpoint, translation competence is most reliably assessed through formal certification frameworks that combine analytic rubrics, performance descriptors, and expert judgment. To determine whether AI systems can perform at a professional standard, they must be evaluated using the same criteria applied to human translators. This pilot study addresses that gap by benchmarking ChatGPT's English–Malay translation performance against a novice and a professional translator using the National Accreditation Authority for Translators and Interpreters (NAATI) Certified Translator examination framework. Thirteen professional raters from the Malaysian Translators Association assessed the translations based on Meaning Transfer, Textual Norms and Conventions, and Language Proficiency. Findings revealed a clear performance hierarchy—Professional Translator > ChatGPT > Novice Translator—indicating that while ChatGPT achieved near-professional competence in fluency and meaning accuracy, it remained limited in idiomatic precision and cultural adaptation. The study highlights ChatGPT's potential as an assistive tool for translation and training, while reaffirming the need for human oversight. It also validates the NAATI framework as a robust benchmark for evaluating AI translation quality. As AI models continue to evolve, future research involving larger translator samples and a wider range of language pairs is essential to evaluate ongoing progress and ensure the responsible integration of AI translation into professional practice.

Keywords: AI translation; English–Malay translation; ChatGPT; NAATI; professional translator assessment

INTRODUCTION

Artificial Intelligence (AI) has transformed translation practices through rapid advances in Neural Machine Translation (NMT) and Large Language Models (LLMs). Tools such as ChatGPT have demonstrated remarkable fluency, contextual awareness, and stylistic adaptability, narrowing the gap between human and machine translation (Belgacem et al., 2023; Chowdhury & Haque, 2023; Keshamoni, 2023; Kumar et al., 2024). Yet, despite these advancements, questions persist regarding whether AI systems can perform at a level comparable to professional translators.

Most comparative studies of AI translation have focused on major language pairs well represented in large language model (LLM) training data, such as English–Spanish, English–Chinese and English–Arabic (e.g., Johnson & Kathirvel, 2025), while low-resource combinations, such as English–Malay, which receive far less training data, remain underexplored. This research gap is significant, as Malay presents distinctive linguistic and cultural challenges that can test the limits of AI translation systems. Moreover, existing evaluations have primarily relied on academic or linguistic criteria and automated metrics (e.g., Alkhawaja, 2024), overlooking how AI translation would perform when measured against professional certification standards. To our knowledge, no published study has systematically assessed AI translation performance using an established professional assessment framework. In Malaysia, English-Malay translation serves vital roles across education, administration, trade, and intercultural communication, making it an ideal context for evaluating the performance of AI-based translation tools. A core concern in this emerging area is that AI-generated translations, while often impressively fluent, may still fall short in conveying cultural nuance, pragmatic intent, and the level of professional adequacy required in real-world practice. These shortcomings have significant implications for the responsible integration of AI into professional workflows, where accuracy, accountability, and cultural appropriateness remain paramount.

The National Accreditation Authority for Translators and Interpreters (NAATI) framework, internationally recognised for its rigour, fairness, and strong alignment with professional standards, provides a suitable benchmark for assessing translation competence beyond surface-level fluency. Addressing the current gap in AI translation research, this study benchmarks ChatGPT's English–Malay translation performance against both a novice translator and a qualified professional translator using NAATI's Certified Translator rubric. By applying a formal certification framework rather than academic assessment tools, the study offers a new methodological lens for evaluating AI translation quality. In doing so, it bridges academic inquiry and industry expectations, generating insights that are directly relevant to professional evaluation practices and the evolving role of AI in translation.

AI TRANSLATION PERFORMANCE

Recent advancements in Artificial Intelligence (AI), mostly observed through Neural Machine Translation (NMT) and Large Language Models (LLMs), have revolutionised machine translation by enhancing accuracy, fluency, and accessibility across diverse applications (Awashreh & Aboeisheh, 2025; de los Reyes Lozano & Mejías-Climent, 2023; Łukasik, 2024; Siu, 2024; Tanni, 2025). NMT has surpassed earlier Statistical Machine Translation (SMT) systems, offering improved semantic precision and better handling of idiomatic expressions, although challenges persist in capturing cultural and contextual subtleties (Alkhatnai, 2025; Ozyumenko & Larina, 2025; Tanni, 2025).

The rise of LLMs has further expanded AI translation's potential by generating context-sensitive and stylistically nuanced output (Hassani et al., 2025; Lee, 2024; Siu, 2024). These models have shown particular promise in creative translation domains such as literary and audiovisual translation, where tone, style, and voice fidelity are essential (de los Reyes Lozano & Mejías-Climent, 2023). Widely available tools like Google Translate, DeepL, and especially ChatGPT have democratised translation access, enabling multilingual communication and collaboration on an unprecedented scale (Alkhatnai, 2025; Awashreh & Aboeisheh, 2025; Moneus & Sahari, 2024). Nonetheless, idiomatic language, cultural nuance, and ethical concerns continue to challenge AI-based translation systems, reinforcing the importance of human oversight (Amaro & Zhang, 2025; Tanni, 2025).

CHATGPT AS A TRANSLATION TOOL

Among AI tools, ChatGPT has become one of the most widely used models, owing to its accessibility, versatility, and generative capacity (Rustici, 2025). Built on transformer architecture and deep learning, ChatGPT represents a major step forward in AI translation by producing contextually coherent and semantically rich outputs. Its attention mechanisms and tokenisation processes enhance contextual understanding and generative flexibility, allowing it to handle complex linguistic structures effectively (Belgacem et al., 2023; Chowdhury & Haque, 2023; Keshamoni, 2023; Kumar et al., 2024). However, despite strong grammatical and structural performance, the model still struggles with nuanced, culturally embedded, and domain-specific language, often reflecting biases within its training data (Alomari, 2024; Alshalan, 2025; Bansal et al., 2024; Qamar et al., 2024). Recent developments, particularly self-attention mechanisms and reinforcement learning from human feedback, have refined its contextual comprehension and long-range dependency management, positioning ChatGPT as a milestone in natural language understanding, though continued improvement is required to address bias and interpretive limitations (Bansal et al., 2024; Bhattacharya et al., 2024).

Empirical studies examining ChatGPT's translation performance across multiple language pairs reveal a pattern of notable progress alongside persistent limitations. Research on relatively high-resource combinations, such as English–Spanish, English–Chinese, and English–Arabic, shows that ChatGPT consistently produces coherent, fluent, and contextually appropriate translations, yet it continues to struggle with idiomatic expressions and culturally nuanced meanings (Johnson & Kathirvel, 2025). In the English–Arabic and Arabic–English directions, the model demonstrates strong semantic fidelity and handles specialised terminology reasonably well, but it frequently fails to capture tone, emotion, and other pragmatic subtleties (Alkhawaja, 2024; Al-Khresheh, 2025; Farghal & Haider, 2025). Further evidence suggests that ChatGPT tends to rely on meaning-oriented, semantic translation strategies rather than literal rendering (Darawsheh et al., 2025), a tendency that can be effective for general comprehension but may not always align with stylistic or cultural expectations.

In domain-specific contexts and particularly in lower-resource language combinations, ChatGPT's performance tends to decline, revealing more pronounced limitations at both the lexical and syntactic levels. In English–Chinese and Chinese–English legal translation, for example, the model underperforms in the English-to-Chinese direction, exposing weaknesses in its encoding capabilities (Ding, 2024). Meanwhile, in German–Slovak financial translation, ChatGPT generally maintains coherence but struggles with specialised terminology and the accurate rendering of compound nouns (Kalaš, 2025). Studies involving other relatively low-resource pairs, such as Chinese–Portuguese and Indonesian–English, report comparable outcomes: although the model can produce fluent output and demonstrates some cross-cultural adaptability, it continues to exhibit inconsistencies in lexical precision, terminological accuracy, and syntactic structure (Jiang et al., 2024; Sutrisno, 2025).

When compared with traditional machine translation systems such as Google Translate, ChatGPT generally delivers more natural-sounding translations and tends to require less post-editing, particularly in Arabic–English contexts (Alafnan, 2025). Translation quality can be improved further through prompt engineering, where domain-specific instructions and contextual cues enhance accuracy and adaptability in specialised tasks (Gao et al., 2024; Peng et al., 2023). Despite these advantages, significant challenges remain. ChatGPT frequently mismanages idiomatic and culturally specific expressions, underperforms in low-resource language pairs, and occasionally produces hallucinated or factually inaccurate content (Peng et al., 2023; Saehu &

Hkikmat, 2025). These limitations highlight the need for continued human oversight to safeguard cultural and contextual appropriateness in professional translation practice (Algaraady & Mahyoob, 2025).

Consequently, scholars increasingly advocate for hybrid human–AI translation models that combine ChatGPT's fluency and speed with human translators' contextual judgment and cultural competence. Such collaborative frameworks are being refined to enhance translation quality, consistency, and domain adaptability across professional and academic settings (Alafnan, 2025; Wang et al., 2024).

TRANSLATION QUALITY ASSESSMENT METHODOLOGIES

The earliest form of Translation Quality Assessment (TQA), known as intuitive assessment, relied mainly on assessors' subjective impressions and personal experience (Eyckmans et al., 2012). Although grounded in mentalist and neo-hermeneutic traditions that value individual interpretation, this approach lacked objectivity, consistency, and replicability, limiting its usefulness for systematic evaluation (House, 2014).

To address these limitations, error analysis emerged as a more structured approach. It evaluates translations by identifying and classifying errors according to type, severity, and weight (Martínez Melis & Hurtado Albir, 2001; Waddington, 2001a). Models such as the error-to-scale transformation (EST) and the error-penalisation/points-deduction (EPPD) systems were developed to enhance consistency in scoring (Williams, 2001, 2004). However, despite its clearer structure, error analysis remains partly subjective because error categories and weightings still depend on assessor judgement (Colina, 2008; Jiménez-Crespo, 2011). It is also time-consuming and tends to focus narrowly on linguistic accuracy. Nevertheless, it marked an important shift toward more systematic, evidence-based TQA (Martínez Mateo, 2014).

In the early 21st century, corpus-based evaluation gained traction. This method uses reference corpora and concordance tools to assess how well a translation aligns with authentic language use, particularly in terms of lexical and phraseological choices (Bowker, 2000, 2001; Jiménez-Crespo, 2009, 2011). Statistical methods, such as those developed by De Sutter et al. (2017), improved the objectivity of assessments. However, corpus-based methods require substantial technical expertise, making them more suitable as a complement to—rather than a replacement for—human evaluation.

Alongside these developments, rubric-referenced (scale-based) scoring emerged as an alternative to the reductionism of error analysis (Angelelli, 2009; Colina, 2008, 2009; Waddington, 2001a, 2001b, 2003). Rubrics use structured rating scales with clearly defined performance bands, allowing assessors to evaluate not only linguistic accuracy but also functional adequacy and pragmatic effectiveness. Research shows that rubric scoring correlates strongly with error analysis and produces high reliability (Lai, 2011; Turner, Lai, & Huang, 2010). Still, some concerns remain about the psychometric soundness of rubric descriptors, which are often theoretically rather than empirically derived (Clifford, 2007; Martínez Mateo et al., 2017;). Even so, rubric scoring represents a significant methodological shift toward holistic, top-down translation assessment.

To balance analytical precision with broader evaluative insight, several advanced Translation Quality Assessment (TQA) methods have emerged. Mixed-methods scoring combines error analysis and rubric-based evaluation—often using 70/30 or 50/50 weightings—to improve both reliability and comprehensiveness (Amini, 2018; Waddington, 2001a, 2001b). Building on this, item-based assessment applies psychometric principles by treating translation errors as statistically testable items, as in the Calibration of Dichotomous Items (CDI) model (Eyckmans et al., 2009). A streamlined version, the Preselected Item Evaluation (PIE) method, focuses on key text segments but offers reduced coverage and weaker empirical validation (Eyckmans & Anckaert, 2017; Kockaert & Segers, 2017).

Finally, automated metrics such as BLEU and METEOR offer cost-effective large-scale evaluation (Chung, 2020; Yating et al., 2025). Although they correlate with human judgments overall, reliability declines at the sentence level and in capturing semantic nuance (Gladkoff & Han, 2022; Tan et al., 2015). While automation enhances efficiency, human assessment remains the gold standard, ensuring contextual, cultural, and communicative accuracy (Graham et al., 2013; Nuriev & Egorova, 2021).

NAATI PROFESSIONAL ASSESSMENT

Building on the evolution of Translation Quality Assessment (TQA) methodologies, professional certification systems must balance theoretical soundness with practical demands for validity, reliability, transparency, and fairness. In high-stakes contexts such as translator certification, assessment frameworks need to evaluate competence accurately while ensuring consistency and procedural defensibility. Among available TQA models, rubric-based analytic scoring has emerged as the most practical and defensible choice (Angelelli, 2009). Unlike error-deduction methods, which are granular and time-intensive, or corpus- and psychometric approaches, which require complex technical resources, rubric-based assessment combines structure with professional judgment. It uses explicit, criterion-referenced descriptors that support both inter-rater reliability and holistic evaluation of performance—capturing linguistic accuracy, communicative effectiveness, and functional adequacy.

This approach has gained prominence in professional certification contexts. The National Accreditation Authority for Translators and Interpreters (NAATI) in Australia provides a representative example of its effective application (NAATI, 2012, 2024). Recognised internationally for its rigorous and comprehensive certification framework (Sulaiman et al., 2024), NAATI's system exemplifies how rubric-based models can balance methodological rigour with operational practicality. Its standards-based, analytic framework, revised in 2024, evaluates translation performance holistically against qualitative criteria rather than deducting points for individual errors (NAATI, 2024). This allows assessors to measure candidates' ability to convey meaning accurately, adhere to genre conventions, and demonstrate idiomatic proficiency.

NAATI'S TQA RUBRIC

NAATI's Certified Translator rubric (see Appendix) applies to the translation of a non-specialised text task and assesses performance across three key criteria, grouped under two overarching domains: Transfer Competency and Language Competency. These criteria are coded A, C, and D.[1] And defined as follows:

[1] The previous *"Criterion B – Follow Brief"* was merged into *"Criterion C – Application of Textual Norms and Conventions"* during the 2024 revision of the rubric to streamline assessment and reduce redundancy. The letter sequence (A–D) has been retained for consistency across NAATI's credentialing suite and to maintain comparability with historical data and examiner training materials.

- Criterion A – Meaning Transfer: Evaluates the translator's ability to convey both the intent and the content of the source message accurately and consistently.
- Criterion C – Application of Textual Norms and Conventions: Assesses control of register, style, structure, and terminology appropriate to the genre, audience, and purpose of the target text.
- Criterion D – Language Proficiency Enabling Meaning Transfer: Examines written command of the target language, including grammar, syntax, lexical precision, and idiomatic usage.

Each criterion is rated on a five-band scale (Band 1 = highest, Band 5 = lowest). To meet certification standards, candidates must achieve at least Band 2 or above for Criteria A and D, and Band 3 or above for Criterion C. These thresholds represent the minimum level of professional competence expected of a certified translator in Australia (see Appendix).

Although numerous studies have evaluated AI translation performance, most have focused on high-resource language pairs that are well represented in large language model (LLM) training data. In contrast, low-resource combinations such as English–Malay, which receive far less training data, remain underexplored. Furthermore, existing evaluations have predominantly relied on academic or linguistic assessment criteria, rather than the professional standards used in certification contexts. Yet from a professional standpoint, translation competence is most reliably judged through formal certification frameworks that integrate analytic rubrics, performance descriptors, and expert human judgment. If the question is whether AI systems can operate at a professional level, they must be evaluated using the same criteria applied to human translators.

This study responds to that gap by adopting the NAATI Certified Translator (CT) assessment framework, which offers a well-recognised and robust rubric-based system for evaluating translation competence. The TQA literature demonstrates that although numerous assessment methods exist, rubric-based analytic scoring has emerged as the most practical, reliable, and defensible approach for professional certification, balancing structured criteria with expert judgement. NAATI's framework reflects these strengths: it uses clear performance descriptors, focuses on meaning transfer, textual norms, and language proficiency, and has been validated through long-standing professional practice. ChatGPT is selected as the AI system for analysis because prior research shows that it can handle complex linguistic structures, contextual nuances, and stylistic variation with a level of fluency that surpasses earlier AI models. Building on this foundation, the present study evaluates ChatGPT's English–Malay translation using NAATI's framework and compares the results with those of a qualified professional and a novice translator. By situating AI translation within an established professional standard, the study directly examines whether ChatGPT's output aligns with, approximates, or diverges from the competence expected of certified human translators—addressing a critical gap in current AI translation research.

METHODOLOGY

RESEARCH DESIGN

This study employed a comparative evaluation design to benchmark ChatGPT's English–Malay translation performance against human translators of differing experience levels, using NAATI's professional assessment framework. The research comprised four key components: the source text, translation agents, raters, and the evaluation framework. The source text was a 250-word English passage selected to meet the NAATI Certified Translator examination specifications. It represented a non-specialised text designed for a general educated readership.

Three translation agents were included: a novice translator, a professional translator, and ChatGPT. The novice translator was an undergraduate student enrolled in a university-level translation course, while the professional translator was a qualified English–Malay translator with over ten years of industry experience. ChatGPT-4o (hereafter 'ChatGPT'), the most advanced version of ChatGPT available at the time of the study, represented a state-of-the-art large language model. The inclusion of these three agents enabled a comparative analysis across a continuum of translation competence—machine, novice, and expert—providing insight into ChatGPT's performance relative to human standards. It should be noted that the study involved only one novice and one professional translator and thus does not seek to generalise results to all human translators. Rather, it serves as a pilot benchmarking exercise, using exemplar performances to explore how NAATI's professional certification framework can be applied to AI-generated translations.

Thirteen professional English–Malay translators, all registered members of the Malaysian Translators Association (MTA), participated as raters. An open call was distributed to MTA members, and thirteen volunteers with a minimum of five years of professional experience were selected. This ensured sufficient subject-matter expertise and supported reliability in judgment across raters. Translation quality was assessed using NAATI's Certified Translator Assessment Rubric, comprising three evaluation components: A: Meaning transfer, C: Application of textual norms and conventions, and D: Language proficiency enabling meaning transfer. This rubric was chosen because NAATI is widely recognised for its rigorous and comprehensive certification framework.

RESEARCH PROCEDURE

The study followed a standardised three-phase procedure: translation, rating, and data analysis.

Phase 1: Translation task

The human translators completed the translation task under controlled conditions that closely simulated the NAATI certification examination environment. They were not permitted to use external tools such as machine translation systems or online dictionaries beyond standard reference materials typically allowed in professional settings. For the ChatGPT translation, the same source text was input into ChatGPT using a structured prompt: "Act as a professional English–Malay translator. Translate the attached news article on Afghanistan's narcotic problem from English into Malay for a general Malay-speaking audience". To ensure optimal performance, the prompt was used in three independent sessions to generate three versions of the translation. The version judged to be the most natural and accurate by an English–Malay translation expert was selected for evaluation.

Phase 2: Rater briefing and assessment

Before assessing the translations, all raters attended an online briefing session on the application of the NAATI Certified Translator Assessment Rubric to ensure uniform interpretation of rating criteria. Each rater was then provided with three anonymised translations (ChatGPT, novice, and professional), coded to conceal their identities and sources. Raters were instructed to evaluate all translations independently within 48 hours of the briefing session, applying the rubric without consultation with other participants.

Phase 3: Analysis of Findings
The third phase focused on the quantitative analysis of the assessment results. Data from the thirteen raters were compiled for each translation agent—novice translator, professional translator, and ChatGPT—across the three NAATI rubric components: (A) Meaning Transfer, (C) Textual Norms and Conventions, and (D) Language Proficiency Enabling Meaning Transfer.

Descriptive statistics, including mean, variance, and standard deviation, were calculated to summarise the central tendencies and variability of scores for each component and translation agent. These measures provided insight into the overall performance pattern, distinguishing between consistent strengths and weaknesses across the three evaluative dimensions. Lower mean scores reflected higher translation quality, as Band 1 represented the highest and Band 5 the lowest performance. Inter-rater reliability was examined using Fleiss' Kappa to examine consistency among the 13 evaluators.

FINDINGS

INITIAL RESULTS

Initial results based on the full set of 13 raters indicated that the professional translator achieved a pass rate of 76.9% (10/13), ChatGPT 53.8% (7/13), and the novice translator 30.8% (4/13). Descriptive statistics showed a clear performance hierarchy, though with moderate variability in ratings (see Figure 1).

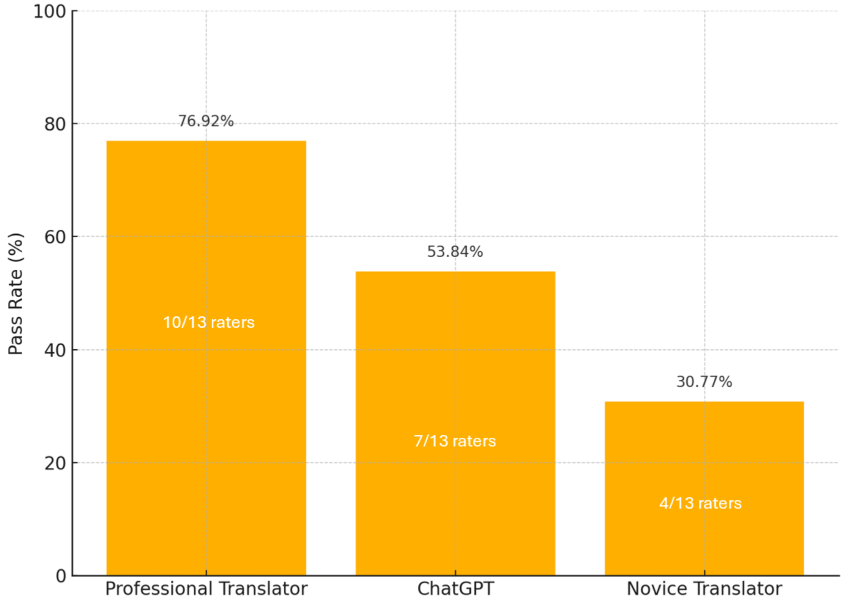


FIGURE 1. Initial results - Pass rate by translator type

Inter-rater reliability was calculated using Fleiss' Kappa (κ) for categorical Pass/Fail outcomes across 13 raters. Results indicated substantial agreement for the professional translator (κ = 0.62), moderate agreement for ChatGPT (κ = 0.47), and fair agreement for the novice translator (κ = 0.34)[2]. These findings indicate that raters exhibited substantial consistency in evaluating the professional translator's performance, whereas their assessments of ChatGPT and the novice translator were more variable.

However, closer examination revealed anomalies in the scoring patterns of three raters (R3, R8, and R10), whose evaluations deviated notably from the consensus of the other ten raters. Their ratings were inconsistent with the overall trend in which the professional translator typically outperformed both the novice translator and ChatGPT. A particularly notable anomaly was observed in the ratings of R3 and R8, both of whom assigned failing scores to the professional translator while passing the novice translator, contrary to the strong consensus among the other eleven raters (see Table 1). This pattern could not be explained by differences in the evaluators' judgment of translation quality but was most plausibly due to a misinterpretation of the band scale. Under the NAATI Certified Translator assessment framework, Band 1 represents the highest level of performance and Band 5 the lowest. However, it appears that R3 and R8 may have reversed this hierarchy, assuming Band 1 was the lowest and Band 5 the highest. This inversion was consistent across their scoring patterns, with both raters systematically assigning higher band numbers to stronger translations and lower band numbers to weaker ones.

TABLE 1. Initial results - Scores given by raters for components A, C, D and NAATI Pass Classification

Rater	**Professional Translator**				**ChatGPT-4o**				**Novice Translator**			
	A	C	D	Pass/Fail	A	C	D	Pass/Fail	A	C	D	Pass/Fail
R1	2	2	2	Pass	3	3	3	Fail	4	4	4	Fail
R2	1	2	2	Pass	1	2	2	Pass	2	3	2	Pass
R3	3	2	3	Fail	3	3	4	Fail	1	2	2	Pass
R4	2	1	2	Pass	1	1	2	Pass	2	3	3	Fail
R5	1	1	1	Pass	2	3	3	Fail	3	4	4	Fail
R6	1	1	1	Pass	1	2	1	Pass	2	3	2	Pass
R7	2	2	2	Pass	3	3	2	Fail	4	4	4	Fail
R8	4	3	4	Fail	4	3	3	Fail	2	2	2	Pass
R9	1	1	1	Pass	2	3	2	Pass	3	3	3	Fail
R10	3	4	5	Fail	1	1	2	Pass	2	4	4	Fail
R11	1	1	2	Pass	2	3	2	Pass	4	4	3	Fail
R12	1	1	1	Pass	1	2	1	Pass	3	3	2	Fail
R13	1	1	1	Pass	3	3	3	Fail	4	4	4	Fail

When their results were reinterpreted according to the correct band orientation, their evaluations aligned closely with the majority of raters, restoring the expected performance hierarchy among the three agents. This confirms that the anomaly was procedural rather than interpretive, arising from a misunderstanding of the rubric's scoring direction rather than from differences in evaluative judgment.

[2] Fleiss' Kappa (κ) quantifies how consistently raters agree beyond what would be expected by chance. It ranges from:
< 0.20 → slight agreement
0.21–0.40 → fair agreement
0.41–0.60 → moderate agreement
0.61–0.80 → substantial agreement
0.81–1.00 → almost perfect agreement

Rater 10 exhibited an inconsistent scoring pattern, showing an unexpected reversal in performance hierarchy. Although both the novice and professional translators were rated below the passing threshold, the ratings suggested that the novice translator outperformed the professional translator — a result that contradicts the general consensus among the other raters. Overall, these three raters (R3, R8, and R10) were identified as outliers. As these anomalies likely reflect procedural rather than evaluative errors, their data were excluded from subsequent quantitative analyses to preserve the integrity and validity of the dataset.

OUTLIER-ADJUSTED RESULTS

After excluding the anomalous data provided by the three outliers (R3, R8, and R10) to ensure data integrity, the ratings were reanalysed to obtain a more accurate assessment. The reanalysis revealed that the professional translator achieved a perfect pass rate of 100% (10/10), while ChatGPT attained a 60% (6/10) pass rate, and the novice translator achieved 20% (2/10) (see Figure 2 and Table 2).

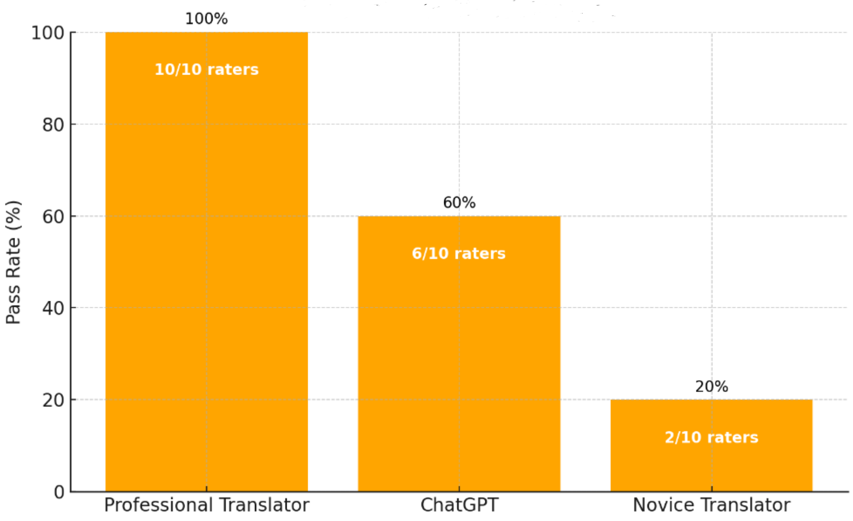


FIGURE 2. Outlier-adjusted results - Pass rate by translator type

TABLE 2. Outlier-adjusted results - Scores given by raters for components A, C, D and NAATI Pass Classification

Rater	Professional Translator				ChatGPT-4o				Novice Translator			
	A	C	D	Pass/Fail	A	C	D	Pass/Fail	A	C	D	Pass/Fail
R1	2	2	2	Pass	3	3	3	Fail	4	4	4	Fail
R2	1	2	2	Pass	1	2	2	Pass	2	3	2	Pass
R4	2	1	2	Pass	1	1	2	Pass	2	3	3	Fail
R5	1	1	1	Pass	2	3	3	Fail	3	4	4	Fail
R6	1	1	1	Pass	1	2	1	Pass	2	3	2	Pass
R7	2	2	2	Pass	3	3	2	Fail	4	4	4	Fail
R9	1	1	1	Pass	2	3	2	Pass	3	3	3	Fail
R11	1	1	2	Pass	2	3	2	Pass	4	4	3	Fail
R12	1	1	1	Pass	1	2	1	Pass	3	3	2	Fail
R13	1	1	1	Pass	3	3	3	Fail	4	4	4	Fail

Inter-rater reliability, measured using Fleiss' Kappa, showed perfect agreement among raters for the professional translator's output, indicating unanimous recognition of professional competence. The ChatGPT translation achieved a Kappa of 0.07, reflecting *slight agreement* and diverse rater opinions on its adequacy for certification, while the novice translator scored 0.21, showing *fair agreement* and more consistent recognition of its limitations. Consequently, inter-rater reliability improved substantially, indicating stronger agreement and consistency among the remaining ten raters. This adjustment ensured that the subsequent statistical analyses accurately reflected true evaluative consensus rather than procedural variation. Overall, raters demonstrated strong consensus in identifying professional-level quality, while their evaluations of the AI and novice translations were less consistent.

ANALYSIS OF MEAN, VARIANCE AND STANDARD DEVIATION (AFTER EXCLUDING OUTLIERS)

Descriptive statistics, including the mean, variance, and standard deviation, were then computed for each component and translation agent to examine performance trends and consistency. The results are summarised in Table 3 below.

TABLE 3. Mean, variance, and standard deviation of the assessment components

Component	Statistics	Professional	ChatGPT	Novice
A (Meaning Transfer)	Mean	1.30	1.90	3.10
	Variance	0.04	0.69	0.69
	Std. Deviation	0.19	0.83	0.83
C (Textual Norms & Conventions)	Mean	1.30	2.5	3.50
	Variance	0.21	0.45	0.25
	Std. Deviation	0.46	0.67	0.50
D (Language Proficiency)	Mean	1.50	2.10	3.10
	Variance	0.25	0.49	0.69
	Std. Deviation	0.50	0.70	0.83

PROFESSIONAL TRANSLATOR

The professional translator consistently received the highest ratings across all components, with mean scores ranging from 1.3 to 1.5. For Meaning Transfer (A), the mean of 1.3 and very low variance (0.04) indicate strong agreement among raters that the translation conveyed the source text's meaning accurately and naturally. In Textual Norms and Conventions (C), the mean of 1.3 and variance of 0.21 similarly show that raters perceived the translation as stylistically appropriate, well-structured, and aligned with professional norms. For Language Proficiency (D), the mean of 1.5 with a variance of 0.25 confirms the consensus that the text exhibited precise grammatical and lexical control. Overall, the low dispersion (SD ≤ 0.5) demonstrates exceptionally high inter-rater consistency, suggesting minimal subjective variation. These findings affirm the professional translation as the benchmark of quality, comfortably exceeding NAATI certification thresholds.

CHATGPT

Raters' evaluations of ChatGPT's translation positioned it between the professional and novice translators. Mean scores ranged from 1.9 to 2.5, corresponding to Band 2–3, indicating generally competent performance that meets or approaches NAATI's minimum passing criteria in all areas. Meaning Transfer (A = 1.9) received relatively strong evaluations, showing that raters agreed ChatGPT conveyed the source message accurately with minor stylistic inconsistencies. For Textual Norms (C = 2.5), higher variance (0.45) and standard deviation (0.67) suggest moderate rater disagreement, reflecting differing perceptions of its stylistic naturalness and idiomatic appropriateness. Language Proficiency (D = 2.1) yielded moderate variance (0.49) and SD (0.70), indicating more variation in raters' judgments of grammatical precision and lexical fluency. Collectively, the moderate variability in ChatGPT's scores indicates that while most raters viewed its output positively, a subset found stylistic or idiomatic issues that prevented it from reaching professional parity. Nonetheless, its consistent performance near or above the NAATI passing threshold demonstrates substantial translation competence.

NOVICE TRANSLATOR

The novice translator received the lowest mean ratings across all three components, reflecting performance below certification standards. Mean scores of 3.1–3.5 correspond to Bands 3–4, indicating partial meaning transfer and limited control of target-language conventions. Meaning Transfer (A = 3.1) and Language Proficiency (D = 3.1) share the same variance (0.69) and SD (0.83), suggesting moderate dispersion and some inconsistency among raters—possibly due to uneven quality within the translation. Textual Norms (C = 3.5) recorded the lowest variance (0.25) and SD (0.5), implying that raters were relatively consistent in their view that the translation lacked professional cohesion and appropriate register. Overall, the novice translator's performance falls below NAATI competency thresholds, reflecting developing translation skills and inconsistent control of meaning and expression.

In summary, the professional translator met or exceeded NAATI's certification thresholds across all assessment components, demonstrating consistently high-quality performance and alignment with professional standards. In comparison, ChatGPT attained passing scores in *Meaning Transfer* and *Textual Norms and Conventions* and performed at a borderline level in *Language Proficiency*, indicating near-professional accuracy overall but with occasional weaknesses in idiomatic expression and stylistic naturalness. The novice translator, on the other hand, failed to meet the minimum pass criteria across all components, reflecting developing competence yet a limited ability to manage complex textual structures and linguistic demands effectively (see Table 4).

TABLE 4. Descriptive summary of mean performance and NAATI's pass classification

Component	NAATI Pass Threshold	Novice (mean)	Pass/Fail	ChatGPT (mean)	Pass/Fail	Professional (mean)	Pass/Fail
A (Meaning Transfer)	Band ≤ 2	3.1	Fail	1.9	Pass	1.3	Pass
C (Textual Norms & Conventions)	Band ≤ 3	3.5	Fail	2.5	Pass	1.3	Pass
D (Language Proficiency)	Band ≤ 2	3.1	Fail	2.1	Borderline	1.5	Pass

The aggregated data demonstrate a clear and statistically consistent performance hierarchy: Professional Translator > ChatGPT > Novice Translator. Across all three components, ChatGPT's mean scores were approximately 1.0–1.4 bands higher than those of the novice translator, while the professional translator outperformed ChatGPT by 0.6–1.2 bands. These mean differences reflect meaningful performance distinctions that align with qualitative expectations. The pattern of variances and standard deviations further substantiates these findings:

- The professional translator's very low variance (0.04–0.25) and SD (0.19–0.50) reflect *strong inter-rater reliability* and shared perceptions of quality.
- ChatGPT's moderate dispersion (variance 0.45–0.69) suggests some diversity in stylistic assessment, typical when raters evaluate hybrid linguistic features (e.g., mechanically fluent but less idiomatic output).
- The novice translator's higher variance (0.25–0.69) indicates lower reliability, consistent with mixed perceptions of accuracy and language control.

In statistical terms, the low within-group variability relative to between-group differences supports high discriminative validity of the NAATI rubric: raters reliably distinguished between levels of translation competence across human and AI-produced texts. These patterns also suggest that, while ChatGPT's performance overlaps partially with professional expectations, it still lacks the stylistic and idiomatic precision that characterises expert human translation.

DISCUSSION

The results of this pilot benchmarking study demonstrate a clear performance hierarchy—Professional Translator > ChatGPT > Novice Translator—which aligns with established expectations in recent empirical research on AI translation (Alafnan, 2025; Tanni, 2025). However, the fact that ChatGPT achieved a 60% pass rate and near-professional mean scores across all components indicates that large language models (LLMs) have reached a level of linguistic sophistication previously unseen in machine translation systems.

CHATGPT'S NEAR-PROFESSIONAL COMPETENCE

The reanalysed data show that ChatGPT's mean ratings (1.9–2.5) positioned its output within or near the NAATI certification thresholds across all components. This aligns with previous findings that LLM-based systems produce translations with greater fluency, coherence, and contextual awareness than earlier neural or statistical models (de los Reyes Lozano & Mejías-Climent, 2023; Hassani et al., 2025; Siu, 2024). In particular, ChatGPT's strength in Meaning Transfer (A), with a mean of 1.9, suggests that it can accurately preserve semantic fidelity and logical coherence. This supports claims by Ding (2024) and Al-Khresheh (2025) that ChatGPT effectively manages cross-linguistic meaning equivalence in both high- and low-resource languages.

Nevertheless, the moderate variance and standard deviation values for Textual Norms and Conventions (C) and Language Proficiency (D) indicate inconsistency among raters regarding stylistic and idiomatic naturalness. Such variability echoes prior studies highlighting that, while ChatGPT excels in grammatical and structural precision, it struggles with idiomatic expressions, pragmatic subtleties, and culturally embedded meanings (Alshalan, 2025; Johnson & Kathirvel, 2025). These stylistic inconsistencies point to ChatGPT's inability to fully reproduce the register, tone, and cultural alignment that human translators intuitively achieve.

PROFESSIONAL TRANSLATOR AS A BENCHMARK OF QUALITY

The professional translator's performance—mean scores between 1.3 and 1.5 with minimal variance—confirmed not only mastery of the NAATI criteria but also high inter-rater agreement. The uniformity of ratings supports the psychometric reliability of rubric-based assessment models (Lai, 2011) and validates the NAATI framework's robustness for distinguishing professional competence levels. Moreover, the professional translator's consistent superiority across all components underscores the enduring value of human expertise in achieving nuanced, idiomatic, and culturally coherent translations. In particular, the professional translator's exceptional performance in Meaning Transfer (A) and Textual Norms (C) reflects the core attributes of professional translation competence—the ability to balance fidelity with naturalness while adhering to genre and audience expectations. This level of cohesion remains beyond the reach of AI systems, which, despite syntactic precision, often lack pragmatic intuition and contextual sensitivity.

NOVICE TRANSLATOR PERFORMANCE AND LEARNING IMPLICATIONS

The novice translator's failure to meet the NAATI pass thresholds across all components highlights the developmental nature of translation competence acquisition. The relatively high variance (0.25–0.69) suggests uneven quality and a lack of consistent control over meaning transfer and stylistic adaptation. From a pedagogical standpoint, these results suggest that translation training may benefit from more explicit attention to the development of students' critical translation evaluation skills. While AI systems such as ChatGPT can produce fluent and contextually appropriate texts, it may not be advisable for student translators to rely on such tools to complete translation tasks, as doing so could limit opportunities to develop the problem-solving and decision-making abilities that underpin professional competence. At the same time, it may be valuable for students to gain experience in critically assessing AI-generated translations, recognising both their potential usefulness and their limitations. One possible approach is to introduce AI tools *after* students have produced an initial human translation, allowing them to compare their own work with AI output to explore differences in lexical choice, cohesion, and stylistic register. Such reflective comparison may support metacognitive awareness, enhance revision strategies, and strengthen students' ability to evaluate translation quality, while still preserving the role of authentic translation practice as the foundation of learning.

RELIABILITY AND VALIDITY OF HUMAN ASSESSMENT

The identification and exclusion of three anomalous raters (R3, R8, and R10) underscore the importance of methodological rigour and rater calibration in translation quality assessment. Misinterpretation of the band scale by R3 and R8—who appeared to reverse the intended hierarchy where Band 1 represents the highest performance and Band 5 the lowest—illustrates procedural issues that can affect reliability even among experienced professionals. Such anomalies were likely not evaluative in nature but procedural, arising from conditions surrounding the online rater briefing. Conducted via videoconference rather than in person, the briefing may have limited opportunities for immediate clarification and visual reinforcement of the rubric orientation. Moreover, some participants did not activate their cameras, which likely reduced engagement and attentional focus. In high-stakes professional settings such as NAATI's operational certification assessments, such misunderstandings are typically mitigated through face-to-face calibration sessions, guided practice, and examiner moderation.

Another contributing factor may lie in the design of the rubric sheet itself. If the layout or visual hierarchy does not clearly emphasise that Band 1 denotes the highest standard of performance, raters unfamiliar with the system may easily invert the scale. Enhancing the rubric's visual clarity—through bolded labels, colour coding, or explicit orientation cues—could therefore improve usability and minimise interpretive errors. Once the anomalous data were removed, inter-rater reliability increased substantially, with the remaining raters demonstrating high consistency across all assessment components. This confirms that the NAATI rubric provides both discriminative validity and procedural robustness when used under adequately controlled conditions.

IMPLICATIONS FOR AI IN PROFESSIONAL TRANSLATION

These results must not be interpreted as evidence that AI could replace human translators. Even when AI systems such as ChatGPT produce output that approximates or occasionally surpasses human performance in certain linguistic aspects, this alone is insufficient for professional equivalence. Translation is not merely a matter of linguistic accuracy; it is a profession grounded in accountability, responsibility, and ethical judgment. Human translators are bound by professional codes of conduct and bear direct responsibility for their work, dimensions that cannot be mechanised or transferred to an algorithm.

The present study does not seek to prepare AI for professional accreditation but to evaluate its performance against established certification standards. The findings indicate that while ChatGPT does not yet meet the threshold of professional competence, it demonstrates a high degree of reliability as a translation aid. Its consistent ability to produce contextually appropriate and linguistically coherent output under the NAATI rubric suggests practical value as a supportive tool, for example, in generating draft translations, maintaining terminological consistency, or assisting with quality assurance. Ultimately, AI might better be viewed as an assistive technology that complements, rather than replaces, human expertise. Its responsible integration into translation practice must reinforce the human translator's central role as the final arbiter of meaning, ethics, and professional accountability.

CONCLUSION

The results of this pilot benchmarking study, which specifically concern English–Malay translation, indicate that ChatGPT demonstrates near-professional competence in meaning transfer and fluency, though it remains limited in idiomatic precision, stylistic coherence, and cultural adaptation. The findings underscore the growing potential of large language models in translation, especially for low-resource languages such as Malay. ChatGPT's consistent performance suggests it could serve as an assistive tool for drafting and quality assurance in professional translation workflows, provided human oversight ensures contextual and ethical accuracy. The study also highlights the value of the NAATI framework as a rigorous method for assessing AI translation quality within professional standards.

However, the results are specific to the English–Malay translation direction and cannot be generalised to other language pairs, which may yield different outcomes depending on data richness and linguistic complexity. As a pilot study, its scope was limited to one novice and one professional translator, providing an exploratory snapshot rather than a generalisable conclusion. Given the rapid evolution of AI models, most recently with the release of ChatGPT-5, translation performance may already have improved significantly, warranting further investigation. Future research might, therefore, better include larger translator samples, analyse multiple text genres, and explore different translation directions, encompassing both high and low-resource languages. Comparative studies across evolving LLMs and prompting strategies may deepen understanding of AI translation competence and support the responsible integration of AI tools into professional translation practice and training contexts.

ACKNOWLEDGEMENT

This work is supported by Geran Kecil Penyelidikan (GKP) FSSK, Universiti Kebangsaan Malaysia [grant number SK-2024-046] and Dana Inisiatif Belanjawan 2025: AI Enhancement [grant number WARISAN-2025-008].

REFERENCES

Alafnan, M. A. (2025). Large language models as computational linguistics tools: A comparative analysis of ChatGPT and Google Machine Translations. *Journal of Artificial Intelligence and Technology*, *5*, 20-32. https://doi.org/10.37965/jait.2024.0549

Algaraady, J., & Mahyoob, M. (2025). Exploring ChatGPT's potential for augmenting post-editing in machine translation across multiple domains: Challenges and opportunities. *Frontiers in Artificial Intelligence.* 8,1526293. https://doi.org/10.3389/frai.2025.1526293

Alkhatnai, M. (2025). The role of artificial intelligence tools in mediating Sino-Arab cultural exchanges through intercultural translation. *Babel, 71*(6), 740-769. https://doi.org/10.1075/babel.25101.alk

Alkhawaja, L. (2024). Unveiling the new frontier: ChatGPT-3 powered translation for Arabic-English language pairs. *Theory and Practice in Language Studies*, *14*(2), 347-357. https://doi.org/10.17507/tpls.1402.05

Al-Khresheh, M. (2025). A back translation analysis of AI-generated Arabic-English texts using ChatGPT: Exploring accuracy and meaning retention. *Dragoman*, *17*, 97-117. https://doi.org/10.63132/ati.2025.abackt.95444806

Alomari, E. A. (2024). Unlocking the potential: A comprehensive systematic review of ChatGPT in natural language processing tasks. *Computer Modeling in Engineering & Sciences*, *141*(1), 43-85. https://doi.org/10.32604/cmes.2024.052256

Alshalan, A. (2025). Bridging the divide: Saussurean Structure and Derridean complexity in ChatGPT's meaning-making. *Arab World English Journal (AWEJ) Special Issue on Artificial Intelligence*, 81-95. https://dx.doi.org/10.24093/awej/AI.5

Amaro, V., & Zhang, X. (2025). Intercultural interfaces: Artificial intelligence and its challenges of cultural sensitivity. *E-Revista de Estudos Interculturais.13,* 1-27.

Amini, M. (2018). How to evaluate the TEFL students' translations: through analytic, holistic or combined method? *Language Testing in Asia, 8*(10), 1-8. https://doi.org/10.1186/s40468-018-0063-6

Angelelli, C. V. (2009). Using a rubric to assess translation ability: Defining the construct. In C. V. Angelelli, & H. E. Jacobson (Eds.), *Testing and assessment in translation and interpreting studies: A call for dialogue between research and practice* (pp. 13-47). John Benjamins.

Awashreh, R., & Aboeisheh, A. (2025). The collaborative future of translation between human-AI partnerships. In M. H. Al Aqad (Ed.), *Role of AI in translation and interpretation* (pp. 205-236). IGI Global.

Bansal, G., Chamola, V., Hussain, A., Guizani, M., & Niyato, D. (2024). Transforming conversations with AI – A comprehensive study of ChatGPT. *Cognitive Computation*, *16*(5), 2487-2510. https://doi.org/10.1007/s12559-023-10236-2

Belgacem, A., Bradai, A., & Beghdad-Bey, K. (2023, October 23-26). *ChatGPT backend: A comprehensive analysis* [Paper presentation]. International Symposium on Networks, Computers and Communications (ISNCC 2023), Doha, Qatar

Bhattacharya, P., Prasad, V. K., Verma, A., & Dhiman, G. (2024). Demystifying ChatGPT: An in-depth survey of OpenAI's robust large language models. *Archives of Computational Methods in Engineering*, *8*, 4557-4660. https://doi.org/10.1007/s11831-024-10115-5

Bowker, L. (2000). A corpus-based approach to evaluating student translations. *The Translator, 6*(2), 183-210. https://doi.org/10.1080/13556509.2000.10799065

Bowker, L. (2001). Towards a methodology for a corpus-based approach to translation evaluation. *Meta, 46*(2), 345-364. https://doi.org/10.7202/002135ar

Chowdhury, M. N.-U.-R., & Haque, A. (2023, June 23-25). *ChatGPT: Its applications and limitations* [Paper presentation]. 3rd International Conference on Intelligent Technologies (CONIT 2023), Hubli, India.

Chung, H.-Y. (2020). Automatic evaluation of human translation: BLEU vs. METEOR. *Lebende Sprachen*, *65*(1), 181-205. https://doi.org/10.1515/les-2020-0009

Clifford, A. (2007). Grading scientific translation: What's a new teacher to do? *Meta, 52*(2), 376-389. https://doi.org/10.7202/016083ar

Colina, S. (2008). Translation quality evaluation: Some empirical evidence for a functionalist approach. *The Translator, 14*(1), 97-134. https://doi.org/10.1080/13556509.2008.10799251

Colina, S. (2009). Further evidence for a functionalist approach to translation quality evaluation. *Target, 21*(2), 215-244. https://doi.org/10.1075/target.21.2.02col

Darawsheh, K., Hamamra, B., & Abuarrah, S. (2025). Addressing untranslatability: ChatGPT's compensatory strategies for translating verbified proper nouns in English–Arabic contexts. *Traduction et Langues, 24*(1), 298-326.

de los Reyes Lozano, J., & Mejías-Climent, L. (2023). Beyond the black mirror effect: the impact of machine translation in the audiovisual translation environment. *Linguistica Antverpiensia, New Series – Themes in Translation Studies, 20,* 1-19. https://doi.org/10.52034/lans-tts.v22i.790

De Sutter, G., Cappelle, B., De Clercq, O., Loock, R., & Plevoets, K. (2017). Towards a corpus-based, statistical approach to translation quality: Measuring and visualising linguistic deviance in student translations. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, *16*, 25-39. https://doi.org/10.52034/lanstts.v16i0.440

Ding, L. (2024). A comparative study on the quality of English-Chinese translation of legal texts between ChatGPT and neural machine translation systems. *Theory and Practice in Language Studies*, *14*(9), 2823-2833, https://doi.org/10.17507/tpls.1409.18

Eyckmans, J., & Anckaert, P. (2017). Item-based assessment of translation competence: Chimera of objectivity versus prospect of reliable measurement. *Linguistica Antverpiensia, New Series – Themes in Translation Studies, 16,* 40-56. https://doi.org/10.52034/lanstts.v16i0.436

Eyckmans, J., Anckaert, P., & Segers, W. (2009). The perks of norm-referenced translation evaluation. In C. V. Angelelli, & H. E. Jacobson (Eds.), *Testing and assessment in translation and interpreting studies: A call for dialogue between research and practice* (pp. 73-93). John Benjamins.

Eyckmans, J., Segers, W., & Anckaert, P. (2012). Translation assessment methodology and the prospects of European collaboration. In D. Tsagari, & I. Csépes (Eds.), *Collaboration in language testing and assessment* (pp. 171-184). Peter Lang.

Farghal, M., & Haider, A. (2025). A Cogno-Prosodic Approach to Translating Arabic Poetry into English: Human vs. Machine. *3L: Language, Linguistics, Literature*®, *31*(1), 255-271.

Gao, Y., Wang, R., & Hou, F. (2024, December 3-6). *How to design translation prompts for ChatGPT: An empirical study* [Paper presentation]. 6th ACM International Conference on Multimedia in Asia Workshops, Auckland, New Zealand.

Gladkoff, S., & Han, L. (2022, June 20-25). *HOPE: A task-oriented and human-centric evaluation framework using professional post-editing towards more effective MT evaluation* [Paper presentation]. 13th Language Resources and Evaluation Conference (LREC 2022), Marseille, France.

Graham, Y., Baldwin, T., Moffat, A., & Zobel, J. (2013, December 4-6). *Crowd-sourcing of human judgments of machine translation fluency* [Paper presentation]. Australasian Language Technology Association Workshop 2013 (ALTA 2013), Brisbane, Australia.

Hassani, G., Malekshahi, M., & Davari, H. (2025). AI-powered transcreation in global marketing: Insights from Iran. *ELOPE: English Language Overseas Perspectives and Enquiries*, *22*(1), 203-221. https://doi.org/10.4312/elope.22.1.203-221

House, J. (2014). Translation quality assessment: Past and present. In J. House (Ed.), *Translation: A multidisciplinary approach* (pp. 241-264). Palgrave Macmillan.

Jiang, L., Jiang, Y., & Han, L. (2024). The potential of ChatGPT in translation evaluation: A case study of the Chinese-Portuguese machine translation. *Cadernos de Tradução, 44*(1), 1-22. https://doi.org/10.5007/2175-7968.2024.e98613

Jiménez-Crespo, M. A. (2009). The evaluation of pragmatic and functionalist aspects in localization: Towards a holistic approach to quality assurance. *The Journal of Internationalization and Localization, 1*, 60–93. https://doi.org/10.1075/jial.1.03jim

Jiménez-Crespo, M. A. (2011). A corpus-based error typology: Towards a more objective approach to measuring quality in localization. *Perspectives, 19*(4), 315–338. https://doi.org/10.1080/0907676X.2011.615409

Johnson, J. W., & Kathirvel, K. (2025). Evaluating the effectiveness of ChatGPT in language translation and cross-lingual communication. In Bui Thanh Hung, M. Sekar, Ayhan ESI, R. Senthil Kumar (Eds.), *Applications of Mathematics in Science and Technology* (pp. 434-439). CRC Press.

Kalaš, F. (2025). Evaluation of German–Slovak AI translation of stock market news. *Linguistische Treffen in Wrocław*, *27*(1), 117-129. https://doi.org/10.23817/lingtreff.27-7

Keshamoni, K. (2023). ChatGPT: An Advanceds Natural Language Processing System for Conversational AI Applications—A Comprehensive Review and Comparative Analysis with Other Chatbots and NLP Models. In: Tuba, M., Akashe, S., Joshi, A. (Eds.) *ICT Systems and Sustainability. ICT4SD 2023. Lecture Notes in Networks and Systems*, vol 765 (pp. 447 – 455). Springer, Singapore. https://doi.org/10.1007/978-981-99-5652-4_40

Kockaert, H. J., & Segers, W. (2017). Evaluation of legal translations: PIE method (Preselected Items Evaluation). *The Journal of Specialised Translation*, *27*, 148-163. https://doi.org/10.26034/cm.jostrans.2017.263

Kumar, H., Damle, M., Natraj, N. A., & Lapina, M. (2024, December 3-5). *AI-driven natural language processing: ChatGPT's potential and future advancements in generative AI* [Paper presentation]. 6th International Symposium on Advanced Electrical and Communication Technologies (ISAECT), Alkhobar, Saudi Arabia

Lai, T. (2011). Reliability and validity of a scale-based assessment for translation tests. *Meta, 56*(3), 713-722. https://doi.org/10.7202/1008341ar

Lee, T. K. (2024). Artificial intelligence and posthumanist translation: ChatGPT versus the translator. *Applied Linguistics Review*,*15*(6), 2351-2372. https://doi.org/10.1515/applirev-2023-0122.

Łukasik, M. (2024). The future of the translation profession in the era of artificial intelligence: Survey results from Polish translators, translation trainers, and students of translation. *Lublin Studies in Modern Languages and Literature*, *48*(3), 25-39. http://dx.doi.org/10.17951/lsmll.2024.48.3.25-39

Martínez Mateo, R. (2014). A deeper look into metrics for translation quality assessment (TQA): A case study. *Miscelánea: A Journal of English and American Studies*, *49*, 73-93. https://doi.org/10.26754/ojs_misc/mj.20148792

Martínez Mateo, R., Montero Martínez, S., & Moya Guijarro, A. J. M. (2017). The Modular Assessment Pack: A new approach to translation quality assessment at the Directorate General for Translation. *Perspectives, 25*(1), 18-48. https://doi.org/10.1080/0907676X.2016.1167923

Martínez Melis, N., & Hurtado Albir, A. (2001). Assessment in translation studies: Research needs. *Meta, 46*(2), 272-287. https://doi.org/10.7202/003624ar

Moneus, A. M., & Sahari, Y. (2024). Artificial intelligence and human translation: A contrastive study based on legal texts. *Heliyon, 10*(6), e28106. https://doi,org/10.1016/j.heliyon.2024.e28106

NAATI. (2024, February). *Certified translator assessment rubrics (Final version 2.0).* https://www.naati.com.au/wp-content/uploads/2023/07/Certified-Translator-Assessment-Rubrics.pdf

NAATI (National Accreditation Authority for Translators and Interpreters). (2012). *Improvements to NAATI testing: Development of a conceptual overview for a new model for NAATI standards, testing and assessment* [Report]. NAATI.

Nuriev, V. A., & Egorova, A. Y. (2021). Methods of quality estimation for machine translation: State-of-the-art. *Informatika i Ee Primeneniya [Informatics and its Applications], 15* (2), 104-111. https://doi.org/10.14357/19922264210215

Ozyumenko, V. I., & Larina, T. V. (2025). Artificial intelligence in translation: Advantages and limitations. *Science Journal of Volgograd State University, 24*(1), 117-130. https://doi.org/10.15688/jvolsu2.2025.1.10

Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., Ouyang, Y., & Tao, D. (2023, December 6-10). *Towards making the most of ChatGPT for machine translation* [Paper presentation]. Findings of the Association for Computational Linguistics: EMNLP 2023 Conference.

Qamar, M. T., Yasmeen, J., Pathak, S. K., & Rangarajan, M. (2024). Big claims, low outcomes: Fact checking ChatGPT's efficacy in handling linguistic creativity and ambiguity. *Cogent Arts & Humanities, 11*(1), 1-22. https://doi.org/10.1080/23311983.2024.2353984

Rustici, C. (2025, August 20). *What are the top AI chatbots? Data-driven insights from the AI "Big Bang" study. DirectIndustry e-Magazine.* https://emag.directindustry.com/2025/08/20/best-ai-chatbots-data-insights-study/

Saehu, A., & Hkikmat, M. M. (2025). The quality and accuracy of AI-generated translation in translating communication-based topics: Bringing translation quality assessments into practices. In M. H. Al Aqad (Ed.), *Role of AI in translation and interpretation* (pp. 237-266). IGI Global. https://doi.org/10.4018/979-8-3373-0060-3.ch009

Siu, S. C. (2024). Revolutionising translation with AI: Unravelling neural machine translation and generative pre-trained large language models. In Y. Peng, H. Huang, & D. Li (Eds.), *New advances in translation technology: Applications and pedagogy* (pp. 29-54). Springer. https://doi.org/10.1007/978-981-97-2958-6_3

Sulaiman, M. Z., Zainudin, I. S., & Haroon, H. (2024). Pemprofesionalan amalan terjemahan dan kejurubahasaan di Malaysia: Satu tinjauan awal (*The professionalisation of translation and interpreting practice in Malaysia: A preliminary study). GEMA Online Journal of Language Studies, 24*(4), 387-409. http://doi.org/10.17576/gema-2024-2404-21

Sutrisno, A. (2025). Inter-sentential translation and language perspective in Neural Machine Translation: Insights from ChatGPT as a transformer-based model. *Asia Pacific Translation and Intercultural Studies*, *12*(1), 81-94. https://doi.org/10.1080/23306343.2025.2485609

Tan, L., Dehdari, J., & van Genabith, J. (2015, October 16). *An awkward disparity between BLEU/RIBES scores and human judgements in machine translation* [Paper presentation]. 2nd Workshop on Asian Translation (WAT 2015), Kyoto, Japan.

Tanni, S. A. (2025). The role of artificial intelligence in translation sites. In B. S. Awwad (Ed.), *Sustainability in light of governance and artificial intelligence applications* (pp. 157-178). Emerald. https://doi.org/10.1108/9781837081981

Turner, B., Lai, M., & Huang, N. (2010). Error deduction and descriptors: A comparison of two methods of translation test assessment. *Translation & Interpreting, 2*(1), 11–23.

Waddington, C. (2001b). Different methods of evaluating student translations: The question of validity. *Meta, 46*(2), 311-325. https://doi.org/10.7202/004583ar

Waddington, C. (2001a). Should translations be assessed holistically or through error analysis? *HERMES - Journal of Language and Communication in Business*, *14*(26), 15-37. https://doi.org/10.7146/hjlcb.v14i26.25637

Waddington, C. (2003). A positive approach to the assessment of translation errors. In M. M. Ricardo (Ed.), *Actas del I Congreso Internacional de la Asociación Ibérica de Estudios de Traducción e Interpretación* (pp. 409-426). AIETI.

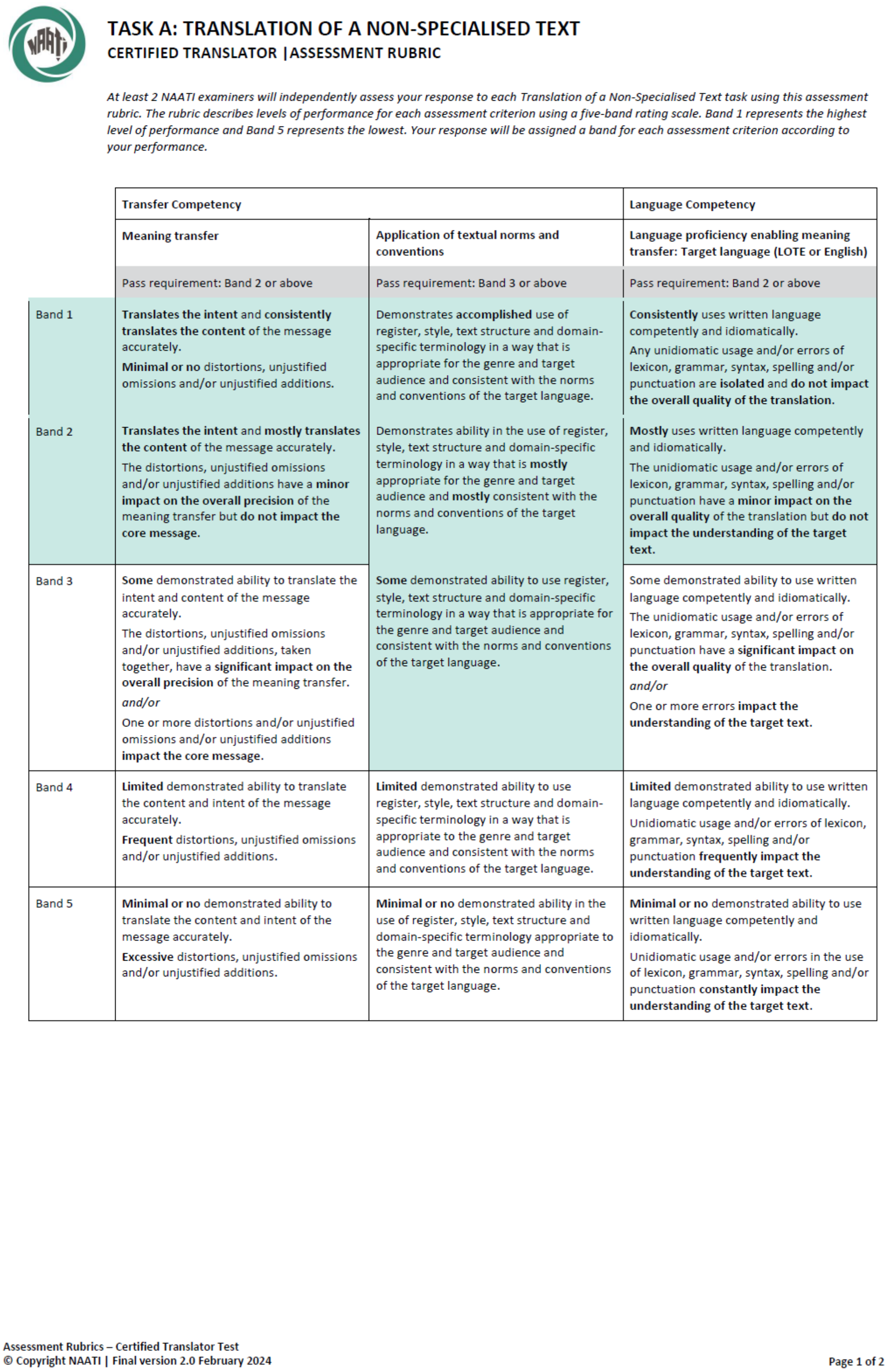
Wang, Y., Zhang, J., Shi, T., Deng, D., Tian, Y., & Matsumoto, T. (2024). Recent advances in interactive machine translation with large language models. *IEEE Access*, *12*, 179353-179382. https://doi.org/10.1109/ACCESS.2024.3487352

Williams, M. (2001). *The application of argumentation theory to translation quality assessment. Meta, 46*(2), 327-344. https://doi.org/10.7202/004605ar

Williams, M. (2004). *Translation quality assessment: An argumentation-centred approach.* University of Ottawa Press.

Yating, L., Afzaal, M., Shanshan, X., & El-Dakhs, D. A. S. (2025). TQFLL: a novel unified analytics framework for translation quality framework for large language model and human translation of allusions in multilingual corpora. *Automatika*, *66*(1), 91-102. https://doi.org/10.1080/00051144.2024.2447652

APPENDIX



TASK A: TRANSLATION OF A NON-SPECIALISED TEXT

CERTIFIED TRANSLATOR |ASSESSMENT RUBRIC

At least 2 NAATI examiners will independently assess your response to each Translation of a Non-Specialised Text task using this assessment rubric. The rubric describes levels of performance for each assessment criterion using a five-band rating scale. Band 1 represents the highest level of performance and Band 5 represents the lowest. Your response will be assigned a band for each assessment criterion according to your performance.

	Transfer Competency		**Language Competency**
	Meaning transfer	**Application of textual norms and conventions**	**Language proficiency enabling meaning transfer: Target language (LOTE or English)**
	Pass requirement: Band 2 or above	Pass requirement: Band 3 or above	Pass requirement: Band 2 or above
Band 1	**Translates the intent** and **consistently translates the content** of the message accurately. **Minimal or no** distortions, unjustified omissions and/or unjustified additions.	Demonstrates **accomplished** use of register, style, text structure and domain-specific terminology in a way that is appropriate for the genre and target audience and consistent with the norms and conventions of the target language.	**Consistently** uses written language competently and idiomatically. Any unidiomatic usage and/or errors of lexicon, grammar, syntax, spelling and/or punctuation are **isolated** and **do not impact the overall quality of the translation.**
Band 2	**Translates the intent** and **mostly translates the content** of the message accurately. The distortions, unjustified omissions and/or unjustified additions have a **minor impact on the overall precision** of the meaning transfer but **do not impact the core message.**	Demonstrates ability in the use of register, style, text structure and domain-specific terminology in a way that is **mostly** appropriate for the genre and target audience and **mostly** consistent with the norms and conventions of the target language.	**Mostly** uses written language competently and idiomatically. The unidiomatic usage and/or errors of lexicon, grammar, syntax, spelling and/or punctuation have a **minor impact on the overall quality** of the translation but **do not impact the understanding of the target text.**
Band 3	**Some** demonstrated ability to translate the intent and content of the message accurately. The distortions, unjustified omissions and/or unjustified additions, taken together, have a **significant impact on the overall precision** of the meaning transfer. *and/or* One or more distortions and/or unjustified omissions and/or unjustified additions **impact the core message.**	**Some** demonstrated ability to use register, style, text structure and domain-specific terminology in a way that is appropriate for the genre and target audience and consistent with the norms and conventions of the target language.	Some demonstrated ability to use written language competently and idiomatically. The unidiomatic usage and/or errors of lexicon, grammar, syntax, spelling and/or punctuation have a **significant impact on the overall quality** of the translation. *and/or* One or more errors **impact the understanding of the target text.**
Band 4	**Limited** demonstrated ability to translate the content and intent of the message accurately. **Frequent** distortions, unjustified omissions and/or unjustified additions.	**Limited** demonstrated ability to use register, style, text structure and domain-specific terminology in a way that is appropriate to the genre and target audience and consistent with the norms and conventions of the target language.	**Limited** demonstrated ability to use written language competently and idiomatically. Unidiomatic usage and/or errors of lexicon, grammar, syntax, spelling and/or punctuation **frequently impact the understanding of the target text.**
Band 5	**Minimal or no** demonstrated ability to translate the content and intent of the message accurately. **Excessive** distortions, unjustified omissions and/or unjustified additions.	**Minimal or no** demonstrated ability in the use of register, style, text structure and domain-specific terminology appropriate to the genre and target audience and consistent with the norms and conventions of the target language.	**Minimal or no** demonstrated ability to use written language competently and idiomatically. Unidiomatic usage and/or errors in the use of lexicon, grammar, syntax, spelling and/or punctuation **constantly impact the understanding of the target text.**

Assessment Rubrics – Certified Translator Test
© Copyright NAATI | Final version 2.0 February 2024

Page 1 of 2