SKASE Journal of Translation and Interpretation, 2025; 18(2): 165–87
doi: 10.33542/JTI2025-S-9

Beyond automated metrics: Assessing GPT-4o and Google Translate against professional translation standards

Longhui Zou, Kent State University
Ali Saeedi, University of Illinois Urbana-Champaign
Geoffrey S. Koby, Kent State University

Abstract
This study evaluates the translation capabilities of GPT-4o, a large language model (LLM), and Google Translate, a neural machine translation (NMT) system, using the American Translators Association (ATA) certification examination framework. We assess translations in two high-resource language pairs: English-to-Chinese (eng-chi) and English-to-Arabic (eng-ara). The evaluation combines both automatic metrics using COMET and manual assessment by ATA-certified graders following the standardized ATA grading framework. Two source texts from retired ATA certification exams were translated by both systems, producing eight target texts in total. Our findings indicate varying performance across systems and language pairs, with only GPT-4o's eng-ara translations achieving superior quality for both required texts. Error analysis reveals distinct patterns between systems and language pairs: GPT-4o's eng-chi translations primarily exhibit challenges with Terminology, Literalness, and Omission, while Google Translate shows a different distribution dominated by Cohesion issues, followed by Literalness and Misunderstanding. For eng-ara translations, both systems display similar error patterns, primarily in Terminology and Literalness, suggesting consistent challenges in this language pair. While COMET scores indicate high performance across all translations, manual assessment reveals more nuanced distinctions in translation quality, particularly in handling rhetorical expressions and idiomatic language use. These findings highlight the importance of complementing automatic metrics with human assessment in translation quality evaluation. The study also suggests that translation challenges extend beyond text complexity, reflecting distinct linguistic characteristics of each language pair and varying approaches in handling these challenges by different machine translation (MT) systems.

Keywords: *Machine Translation; Large Language Models; Neural Machine Translation; Translation Quality Assessment; ATA Certification; Professional Translation Standards*

1. Introduction

Since its release by OpenAI in November 2022, ChatGPT (GPT-3.5-Turbo), a chatbot powered by the GPT (Generative Pre-trained Transformer) language model, has quickly gained popularity due to its impressive performance on a wide array of natural language processing (NLP) tasks (OpenAI 2022). These tasks encompass text generation, question answering, content creation, text summarization, sentiment analysis, acting a role to perform a task, and machine translation (MT), among others (Jiao et al. 2023b; Liu et al. 2023). Underlying ChatGPT is a multilayer Transformer model pre-trained on vast text corpora using self-supervised learning (Ouyang et al. 2022). The Transformer architecture, initially proposed for neural machine translation (NMT), comprises encoder and decoder networks to map input text

to target text output (Vaswani et al. 2017). Building upon the Transformer decoder model, OpenAI developed the GPT series starting in 2018, designed for broad language generation capabilities (Radford et al. 2019). GPT models are first pre-trained in an unsupervised manner on extensive unlabeled texts to learn textual representations, then fine-tuned on labeled data to adapt to downstream tasks (OpenAI 2019a). Their self-supervised pre-training methodology enabled strong performance across tasks without task-specific architectures or datasets (OpenAI 2019b).

ChatGPT integrates the GPT approach with reinforcement learning using human preferences to further improve response quality (Ouyang et al. 2022). By providing natural language interaction, ChatGPT has demonstrated its potential to grasp context and generate coherent, relevant responses, making it a promising tool for various industries without specialized fine-tuning, including the translation sector (Jiao et al. 2023a; Dwivedi et al. 2023). The rapid evolution of these models underscores the timeliness of this research. GPT-4, released in March 2023, introduced enhanced visual capabilities and strengthened language mastery (OpenAI 2023b). More recently, GPT-4o was released in May 2024 as part of OpenAI's multimodal model lineup. This iteration represented a significant advancement in multimodal capabilities, efficiency, and language support (OpenAI 2024). Particularly relevant to our research interests, GPT-4o offers improved handling of multiple languages with greater accuracy and fluency compared to GPT-4. OpenAI has enhanced tokenization specifically for languages that do not use a Western alphabet, such as Chinese, Arabic, Hindi, and Korean. The new tokenizer more efficiently compresses non-English text, processing prompts in those languages in a cheaper, quicker way (Craig 2024). This recent development directly impacts our investigation into its translation capabilities.

Although GPT is not specially fine-tuned for language translation tasks, its potential for accurate and efficient MT has been reported by many scholars and researchers for different language pairs. Cady et al. (2023) compared the performance of seven MT systems on a set of 3,000 bidirectional translation sentences between English and Chinese drawn from patent documents on science and technology. The systems evaluated were commercial MT services (Google Translate, DeepL, Baidu, Youdao, and Niutrans) and GPT models (GPT-3.5 and GPT-4). Based on BLEU metrics, the results indicate that GPT-3.5 and GPT-4 performed comparably to the specialized commercial MT systems, with no single system showing consistent superiority across all test cases. However, it is important to note that BLEU metrics (Papineni et al. 2002) have recognized limitations, particularly for language pairs with significant structural differences like English and Chinese, and for specialized technical text such as patents. These findings highlight a research gap regarding how newer GPT models might perform in translation tasks compared to dedicated MT systems, especially for diverse language pairs and text types that require more fine-grained evaluation beyond BLEU scores.

Primarily utilizing the BLEU score for their analysis, Jiao et al. (2023b) assessed the MT capabilities of GPT-4 against GPT-3.5, uncovering significant improvements by GPT-4 across six language pairs, including English-to-German (eng-ger), German-to-English (ger-eng), English-to-Chinese (eng-chi), Chinese-to-English (chi-eng), German-to-Chinese (ger-chi), and Romanian-to-Chinese (rom-chi)[1]. Their results also showed that GPT-4 is competitive with leading commercial systems like Google Translate, DeepL, and Tencent TranSmart, for both high-resource European languages and low-resource or distant ones. Furthermore, human evaluations of the translations in their study suggest that GPT-3.5 is more prone to producing

[1] ATA uses ISO 639-2, the three-letter bibliographic codes (ISO 1998).

errors and hallucinations, whereas GPT-4 demonstrates fewer inaccuracies, indicating that GPT-4 has evolved into an effective translation tool (Jiao et al. 2023a).

In their evaluation of GPT-4o's translation capabilities across six major languages in 2024, Shahriar et al. (2024) conducted a comprehensive analysis using 500 randomly sampled data points from each language dataset. The study utilized the OPUS dataset for Spanish, Arabic, French, Portuguese, and Russian translations, while Hindi data was sourced from the IIT Bombay English-Hindi Parallel Corpus. Their methodology employed BERT-based sentence embeddings (specifically the paraphrase-MiniLM-L6-v2 model) and cosine similarity metrics to evaluate translation quality, though this computational approach had limitations in capturing cultural and contextual nuances. Their results demonstrated varying performance across languages, with Spanish and Portuguese achieving the highest accuracy rates of 88% and 86% respectively, followed by Hindi (82%), Russian (80%), and French (75%). Notably, Arabic performed relatively poorly at 78%, which the researchers attributed to its intricate script system, complex word forms, and unique linguistic challenges that pose significant difficulties for MT. The findings suggest GPT-4o approaches the quality of dedicated translation systems, despite not being specifically optimized for translation. However, the study did not include an evaluation of eng-chi translation performance, and the researchers acknowledged limitations in their sampling approach and the absence of human evaluators for assessing linguistic subtleties. Like many MT evaluation studies, this research raises questions about the balance between computational metrics and human assessment in evaluating translation quality, particularly for morphologically rich languages like Arabic where computational metrics alone may not fully capture translation adequacy.

In their recent evaluation conducted in May 2024, Intento assessed the translation capabilities of 52 MT systems, including 24 Large Language Models (LLMs) (Intento 2024). Their analysis, spanning nine content domains and eleven language pairs, revealed that GPT-4o, DeepL, and Google Translate demonstrated superior performance with the highest proportion of translations containing no or minor issues. The evaluation employed a multi-metric approach combining Intento Language Quality Assessment (LQA), an LLM-based DQF-MQM metric, and the COMET semantic similarity framework (Rei et al. 2020). Each system was tested using approximately 500 source segments per language pair, with LLMs receiving zero-shot prompts such as "You are a professional translator. Translate this from <source language> to <target language>: <source segment>." Translations in the colloquial domain and in the eng-ara language pair presented the most major and critical issues across all systems. While GPT-4o achieved top-tier performance across most domains, analysis of general domain translation revealed that Google Translate ranked among the top performers for both eng-ara and eng-chi pairs, whereas GPT-4o was among the leaders specifically for eng-chi translation. However, the report's segment-by-segment translation approach may not adequately account for broader background and context, potentially resulting in higher rates of false positives. This methodology raises questions about how LLMs and commercial MT systems might perform in more comprehensive full-text translation scenarios where contextual understanding is essential.

Despite their impressive capabilities, generative AI models like ChatGPT face several inherent limitations in translation tasks. These include issues with accuracy, outdated terminology, and inappropriate linguistic patterns stemming from their training data (Liu et al. 2023; Ray 2023). Moreover, these models can perpetuate various embedded biases, including gender, cultural, religious, political, and regional language biases that exist within their training corpora (Ghosh & Caliskan 2023; Motoki et al. 2024; Babaei et al. 2024). Human oversight thus remains indispensable in professional translation contexts, particularly for high-

stakes content such as legal documents, tourism and hospitality industry content, and healthcare texts where precision is paramount (Siu 2023). The necessity for careful human review and post-editing of LLM-generated translations underscores the enduring complementary relationship between technological innovation and human expertise in professional translation workflows.

This paper addresses a significant gap in MT quality assessment. While prior studies have relied primarily on automated metrics like BLEU or limited human evaluation protocols, few have assessed translation quality within established professional standards. The American Translators Association (ATA) certification grading framework offers a comprehensive, industry-standard method for evaluating full-text translation, yet it remains underutilized in comparing LLMs and NMT systems. This study simulates the ATA certification exam using two retired ATA source texts (STs) to evaluate MT performance in English-to-Chinese (eng-chi) and English-to-Arabic (eng-ara) language pairs. The selection of Google Translate and GPT-4o as test candidates was informed by Intento's (2024) report, which identified Google Translate as the leading performer across both pairs in the general domain, and GPT-4o as a top-tier LLM system comparable to commercial MT systems across most domains and language pairs. Google Translate currently employs a Transformer-based NMT architecture, incorporating innovations in attention mechanisms and sequence-to-sequence learning. Since transitioning from statistical methods in 2016, this architecture has significantly improved translation quality across both high-resource and low-resource language pairs (Wu et al. 2016; Caswell & Liang 2020).

The choice of the ATA framework ensures consistent proficiency assessment through its systematic evaluation of accuracy, style, and language conventions (Koby & Champe 2013). This offers a detailed and relevant benchmark that more accurately reflects industry requirements than the evaluation approaches commonly used in previous studies in related fields. Through this experimental setup, we address three research questions:

1. How does the translation quality of LLM compare to dedicated NMT system when evaluated through professional translation standards represented by the ATA framework?
2. What specific error patterns emerge in both LLM and NMT translations for eng-chi and eng-ara language pairs?
3. What are the relative strengths and weaknesses of LLM versus NMT systems in performing MT?

2. Methodology

This section outlines our approach to comparing the translation quality of GPT-4o and Google Translate using the ATA framework. We begin by explaining the ATA certification process, followed by a description of our experimental design, including text selection, automatic evaluation, and manual assessment procedures. This methodology enables us to evaluate system-generated translations against professional standards, rather than relying solely on commonly used automated metrics.

2.1. The ATA certification process

The ATA certification examination requires candidates to translate two STs, each approximately 225–275 words in length. These texts are generally at university reading level but do not require highly specialized knowledge. They are designed to include manageable challenges in terminology and phraseology that skilled translators can address using comprehensive general dictionaries (Zou 2024).

Each translated text is evaluated in various aspects including errors that concern the form of the exam, meaning transfer and strategic errors, and mechanical errors. The ATA employs a points-addition scoring system where errors are categorized by severity and impact on the overall translation. Unlike point-deduction systems, ATA begins at zero (a theoretical perfect exam would have zero points) and adds points for each error, up to a maximum threshold of 18 (Zou et al. 2024). Less severe errors result in fewer points (1, 2, or 4), while more serious errors are assigned higher point penalties (8 or 16). To pass the exam, a candidate's translations must accumulate fewer than 18 error points for each text. The ATA error annotation scheme will be further illustrated in Section 2.4.

Each ATA exam (i.e. two translated texts) is evaluated by two ATA-certified translators who have been vetted and trained as graders. Each grader assesses the exam independently, and their results are then compared for consistency. If the two graders agree, the evaluation is finalized. In cases of disagreement, a third grader is brought in to review. This rigorous process ensures that ATA-certified translators meet the high standards required for professional translation, demonstrating their ability to handle complex and diverse content. The selectivity of this process is reflected in the current pass rate of less than 20%, with fewer than 2,000 certified translators worldwide (American Translators Association n.d.).

2.2. Selection of source texts

For this study, two English STs were selected from previous ATA certification examinations with general topics. These exams were intended and considered a general professional-level assessment for translators of a certain language combination (for instance, eng-chi or eng-ara) in the US (Koby & Champe 2013). Each text is around 250 words long and contains about ten segments. As shown in Table 1, their readability index scores (Flesch-Kincaid Grade Level) are similar, and thus comparable.

Table 1: Description of the two STs

ST	Topic	Readability Index	Word count	Segment count
1	Philanthropy	12.2	245	12
2	Social Media	12.2	274	15
Total			519	27

Both GPT-4o and Google Translate were used to generate raw translations of the two STs into simplified Chinese and Arabic. For GPT-4o, we employed a zero-shot prompt: "You are a professional translator. Please translate the following text from English to Chinese/Arabic: <source text>." Each system produced four translations: two for Chinese (GPTchi1, GPTchi2, GNMTchi1, GNMTchi2) and two for Arabic (GPTara1, GPTara2, GNMTara1, GNMTara2). The evaluation involved 27 target segments for each language pair.

2.3. Automatic assessment

For automatic assessment, this study employs the COMET framework (Rei et al. 2020), which has demonstrated stronger correlation with human judgment compared to traditional lexical-based metrics like BLEU (Papineni et al. 2002) and CHRF (Popović 2015). Recognizing that the quality of reference translations significantly impacts the reliability of automatic metrics (Freitag et al. 2020), we established a two-step reference preparation process. For each language pair (eng-chi and eng-ara), two ATA-certified professional translators were engaged in the process. One translator translated the texts from scratch, while the second translator validated and verified its quality. This meticulous approach ensures high-quality reference translations for reliable automatic assessment.

2.4. Manual assessment

The manual evaluation was conducted following the ATA certification grading framework (Koby 2015). This framework employs a points-addition scoring system where candidates must demonstrate proficiency across two texts. For certification, translations must receive fewer than 18 error points per text, with lower scores indicating better performance.

The evaluation of translated texts follows a comprehensive framework that assesses both error types and their severity. Translation errors fall into three main categories under the ATA error taxonomy. Form-related errors involve technical aspects of the submission, such as unfinished translations (UNF), illegible content (ILL), or instances of indecision where multiple translation options are provided (IND).

Meaning transfer or strategic errors negatively affect the clarity and utility of the target text. These encompass errors such as addition (A) and omission (O) of content, ambiguity (AMB) and cohesion (COH) issues, faithfulness (F) deviations from ST meaning, faux amis (false friend) (FA), overly literal translations (L), misunderstanding of ST (MU), terminology or word choice (T) issues, and text type errors (TT) including register (R) and style deviations, incorrect verb tense (VT) that alters meaning, as well as other meaning transfer issues (OTH-MT).

Mechanical errors negatively impact the overall quality of the target text. These include grammar (G) issues, which can be divided into two subcategories: syntax (phrase, clause, or sentence structure) (SYN) and word form or part of speech (WF/PS). Mechanical errors also encompass punctuation (P) mistakes and spelling or character (SP/CH) errors. Spelling or character errors can be further categorized into issues with diacritical marks/accents (D) and capitalization (C). Additionally, mechanical errors include usage (U) mistakes and other mechanical issues (OTH-ME).

The ATA scoring system assigns severity levels to translation errors, with corresponding point deductions of 1, 2, 4, 8, or 16 points. This approach reflects a graded scale to some extent similar to MQM (Multidimensional Quality Metrics), which categorizes errors by type and severity, typically labeled as minor, major, or critical, and assigns penalty weights accordingly (Lommel et al. 2014). In the ATA framework, mechanical errors (e.g. spelling, punctuation) are capped at a maximum of 4 points per instance, aligning with MQM's practice of treating such issues as lower severity. Spelling and character errors typically incur 1-point deductions, with a maximum of 2 points. Additionally, graders may award up to three quality points for exceptional translation choices, which are subtracted from the total error points, a feature not in standard MQM scoring but present in the ATA framework to recognize excellence. For

example, if a text contains four COH-2 (cohesion) errors at 2 points each, one L-2 (literalness) error at 2 points, and one COH-4 error at 4 points, the total error point for the text is (4 + 1) × 2 + 1 × 4 = 14 points. Since this falls below the 18-point threshold, the translation would receive a passing grade from that grader.

To assess the translation quality under professional certification standards, we conducted an experiment simulating the ATA certification examination process. Both GPT-4o and Google Translate were treated as certification candidates for eng-chi and eng-ara language pairs. The machine-generated translations were evaluated by ATA-certified graders following the standardized ATA grading framework. To ensure unbiased assessment, the graders were not informed that they were evaluating MTs, and they were requested to provide brief written feedback following each evaluation.

3. Results

3.1. Automatic evaluation results for the two systems

Based on COMET scores for the same set of STs (27 segments), both GPT-4o ($p = 0.0000224$) and Google Translate ($p = 0.000037$) perform significantly better in the eng-ara language pair compared to eng-chi. These p-values were calculated using a paired sample t-test, confirming that the differences are statistically significant ($p < 0.05$). As shown in Table 2, the average COMET scores for eng-ara are 0.9665 for GPT-4o and 0.9724 for Google Translate, whereas for eng-chi, they are 0.8832 and 0.8878, respectively. While Google Translate shows slightly higher COMET scores than GPT-4o across both language pairs, these differences are not statistically significant at the segment level.

Table 2: Descriptive statistics of COMET results

System & Language Pair	Mean	STD	Min	Max	Overall Score
GPT-4o (eng-chi)	0.8832	0.1140	0.5777	1.0000	0.8832
Google (eng-chi)	0.8878	0.1006	0.7060	1.0000	0.8878
GPT-4o (eng-ara)	0.9665	0.0447	0.8642	1.0000	0.9665
Google (eng-ara)	0.9724	0.0413	0.8475	1.0000	0.9724

As shown in Figure 1, both GPT-4o and Google Translate demonstrate strong performance across both language pairs, with COMET scores frequently exceeding 0.95. The eng-ara translations show notably consistent performance, with both systems maintaining high scores around 0.98-1.0 across most segments. In contrast, the eng-chi translations exhibit more variability. We observe pronounced performance drops in several segments where both systems show parallel decreases in performance, such as Segments 9 (with scores around 0.71), 11 (around 0.74), and 19 (around 0.75). These segments include figurative and context-dependent expressions that are especially challenging for MT.

For instance, Segment 11, "Rather than simply write checks for existing institutions, these 'philanthrocapitalists', as they are often called, aggressively seek to shape their operations", contains the quoted term *philanthrocapitalists*, which carries a sarcastic or critical tone in its original context. It also includes idiomatic expressions like *write checks* and *shape their operations*, which require contextual understanding to avoid awkward or overly literal

translations. This pattern suggests that content demanding pragmatic or conceptual interpretation poses greater challenges in the eng-chi pair, regardless of the system used. These results from automatic assessment provide initial insights, but a comprehensive analysis integrating both automatic and manual assessment findings, along with detailed error analysis of challenging segments, will be presented in the discussion (Section 4).

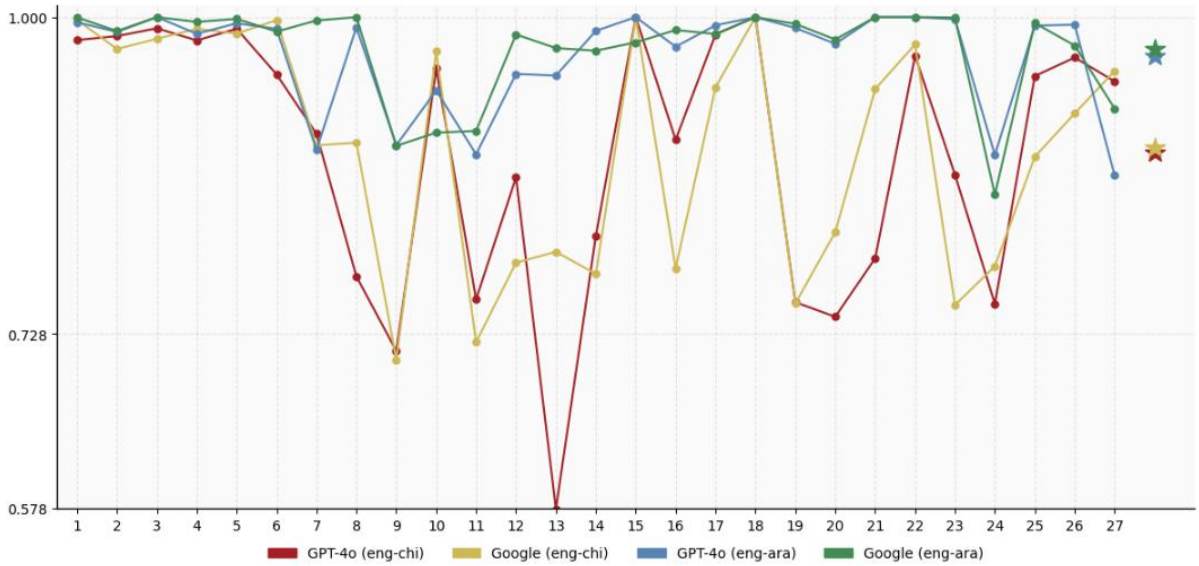

Figure 1: Segment-wise COMET scores across MT systems and language pairs[2]

For eng-ara translations, Google Translate (green line) shows marginally higher COMET scores compared to GPT-4o (blue line) across most segments. This performance gap becomes more pronounced in specific segments, particularly in Segments 7 and 27, where Google Translate demonstrates consistently more stable performance across challenging segments. Statistical analysis confirms that the difference between Google Translate and GPT-4o is not statistically significant for the eng-ara pair, despite Google's generally higher scores.

For eng-chi translations, Google Translate (yellow line) and GPT-4o (red line) exhibit more varied performance patterns compared to the eng-ara pair. While Google Translate generally achieves higher scores in several segments, its superiority is less consistent. Both systems experience significant performance fluctuations across segments. The performance gap favors GPT-4o in segments like 16 and 23, where Google Translate's scores drop to approximately 0.77 while GPT-4o maintains higher scores around 0.88. Conversely, in segments such as 13 and 21, the gap shifts in favor of Google Translate, which performs significantly better than GPT-4o. This variability suggests that Google Translate and GPT-4o may handle different types of translation challenges with varying degrees of resilience in the eng-chi pair. Statistical testing confirms that the difference between Google Translate and GPT-4o is not statistically significant for the eng-chi pair, despite these visible segment-level variations.

[2] Note: Asterisk markers (★) indicate overall average scores for each system-language pair.

3.2. Manual evaluation results for the two systems

As shown in Table 3, the results of the manual evaluation based on the ATA framework reveal varying levels of translation capabilities across systems and language pairs[3]. Among all four system-language pair combinations tested in this experiment, only GPT-4o's eng-ara translations met the certification threshold, with both texts receiving acceptable scores from both graders. The remaining three combinations failed to meet ATA requirements for different reasons. GPT-4o's eng-chi translations performed adequately on Text 1 but substantially exceeded the error threshold on Text 2. Google Translate's eng-chi translations showed inconsistent quality on Text 1, with graders disagreeing on whether it met the threshold, while both found Text 2 acceptable. For eng-ara, Google Translate performed well on Text 1 but had mixed results on Text 2, with one grader finding it exceeded the threshold. These results indicate that while certain language-specific MT systems have made significant progress, the majority of MT outputs examined do not yet meet the professional translation standards required by ATA.

Table 3: Manual evaluation results across systems and language pairs

System & Language Pair	**Text 1 Performance (% of threshold)**		**Text 2 Performance (% of threshold)**		**Certification Threshold Met?**
	Grader 1	Grader 2	Grader 1	Grader 2	
GPT-4o eng-chi	22%	83%	122%	189%	No
Google eng-chi	67%	128%	11%	78%	No
GPT-4o eng-ara	61%	72%	61%	39%	Yes
Google eng-ara	33%	44%	72%	133%	No

Note: Values represent error points as a percentage of the ATA certification threshold (18 points). Percentages below 100% indicate performance within acceptable limits; values above 100% indicate performance exceeding error threshold limits.

The interrater agreement analysis reveals a 75% overall agreement rate, with graders concurring on six out of eight text evaluations. The magnitude of score differences varies between language pairs. For eng-ara translations, graders showed stronger scoring consistency, particularly in their evaluations of GPT-4o's outputs. In contrast, eng-chi evaluations exhibited larger scoring variations between graders, despite achieving the same 75% agreement rate on overall text quality judgments as the eng-ara evaluations.

The scores of eng-chi translations reveal intriguing performance patterns across both systems. GPT-4o exhibited considerable inconsistency in translation quality, with Text 1 performing within acceptable limits (22% and 83% of the threshold from Graders 1 and 2, respectively), while Text 2 significantly exceeded the error threshold (122% and 189%). This demonstrates that GPT-4o's error points more than doubled from Text 1 to Text 2, with Grader 2's assessment showing it exceeded the certification threshold by 89% on Text 2. Google Translate displayed a similarly variable performance, but with an inverse pattern. It struggled with Text 1 (67% and 128% of threshold) but performed better on Text 2 (11% and 78% of threshold). This means Google's performance on Text 2 was substantially better, with

[3] Due to ATA certification protocols, this paper presents only relative scoring and comparative analyses. Complete grading data is available to researchers upon request, subject to confidentiality agreements.

Grader 1 annotating only 11% of the error threshold, a 56-percentage improvement over its Text 1 performance.

While the performance differences between systems are substantial in absolute terms, the small sample size in this study, limited to only two STs, limits the reliability of statistical significance testing. Broader evaluation with more texts would be necessary to establish statistically significant differences. Although neither system met the certification threshold for Chinese translation, the inverse performance patterns, where GPT-4o performing better on Text 1 but worse on Text 2, and Google Translate showing the opposite, suggest that translation quality in eng-chi pair is influenced by factors beyond surface-level text complexity. Furthermore, the contrasting patterns on the same texts indicate that each system encounters distinct challenges in eng-chi translation, aligning with findings from the automatic evaluation metrics in Section 3.1. These results point to the need for more fine-grained error analysis, as ST readability does not necessarily correlate with translation difficulty across different systems.

In eng-ara translations, GPT-4o exhibited consistent performance across STs of similar readability, effectively translating both Text 1 and Text 2 with relatively low error points (61–72% and 39–61% of threshold, respectively). Google Translate, despite effective translation for Text 1 (33–44% of threshold), showed significant performance deterioration in Text 2 (72–133% of threshold). This disparity between systems' performance also presents an interesting contrast with the automatic evaluation results. While COMET scores indicated uniformly high quality for both systems, with overall averages around 0.97, the manual assessment revealed more significant distinctions in translation quality. This discrepancy between automatic and human evaluation aligns with findings from Freitag et al. (2021), who demonstrated that professional translators with access to full document context identify substantially different quality rankings compared to those established through automatic metrics. Similarly, Läubli et al. (2020) found that claims of human-machine parity in translation were often based on evaluation designs that failed to capture the types of errors professional translators readily identify, particularly when evaluating full documents rather than isolated sentences.

3.3. Relationship between automatic and human evaluation

We examined the relationship between COMET scores and manual error annotations for each segment. As shown in Figure 2, there is an overall negative correlation between COMET scores and the average human grader scores by segment. This is expected since the grader scores represent accumulated error points and COMET scores represent translation quality automatically predicted. The strongest alignment between COMET scores and human grader scores appears to be GPT-4o eng-chi among the four combinations ($r=-0.44$, $p=0.022<0.05$), but with moderate correlation. Other three combinations, however, show both weak agreement between COMET and human evaluation and also lack statistical significance. COMET scores align more closely with human ratings for Chinese (especially GPT-generated translations) than Arabic.

Segments with multiple errors or high-severity errors, such as segment 13, which includes errors like SYN-4 and SYN-2 annotated by Grader 1, and L-8 by Grader 2, generally received lower COMET scores (e.g. 0.58). However, the analysis also revealed instances where segments flagged with many error annotations by ATA graders still received quite high COMET scores. For example, segment 23 received the largest average accumulated error points in the dataset (T-2 and MU-4 by Grader 1, and MU-16 by Grader 2) despite its high COMET score of 0.86. This pattern supports findings from previous research (Freitag et al. 2021;

Läubli et al. 2020), which demonstrate that automatic evaluation metrics, while useful, often fail to capture certain types of translation errors that professional translators deem significant. These results reinforce the growing consensus that automatic metrics should be complemented by human evaluation in professional translation workflows. Furthermore, a closer look at error annotations across both graders revealed distinct error patterns unique to each combination of MT system and language pair, underscoring the importance of detailed, system-specific analysis in MT evaluation.

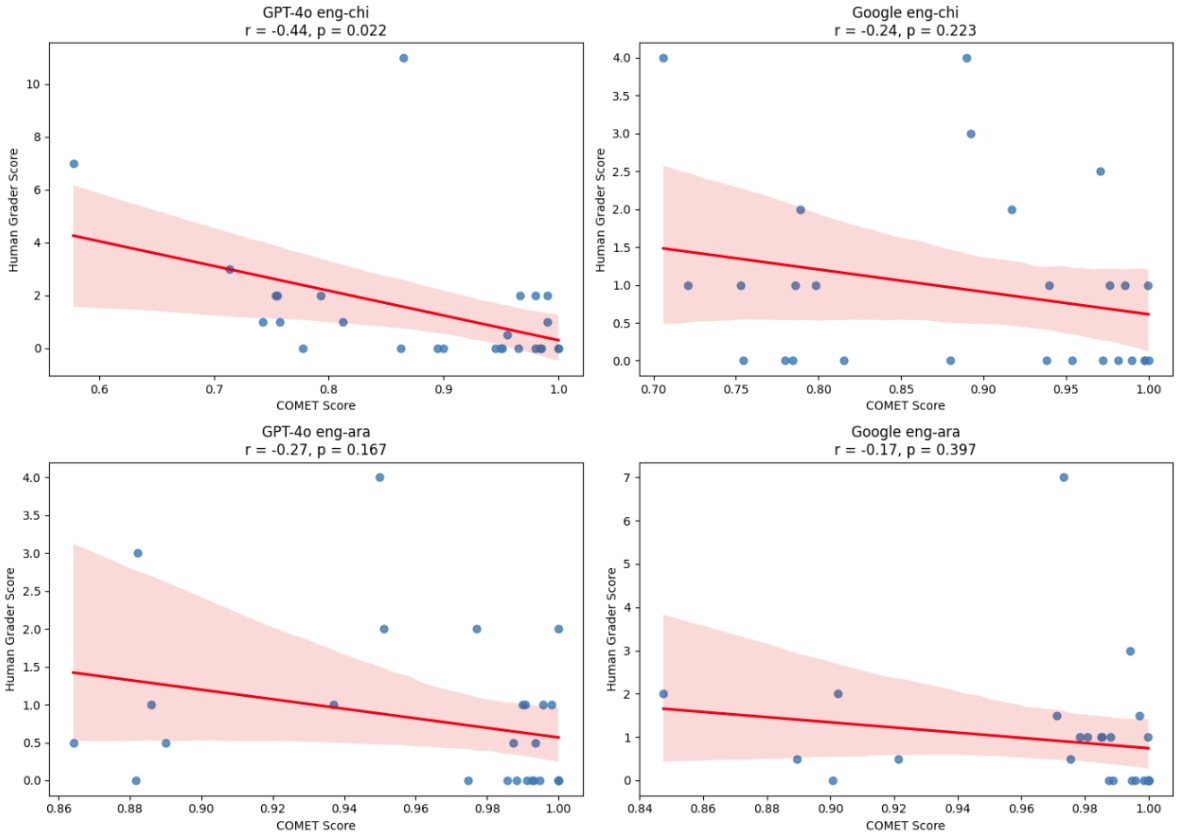

Figure 2: Correlation between COMET scores and human evaluation across language pairs and systems

3.4. Predominant error patterns

In general, many human candidates do not fail the ATA certification exam because of one large error, but rather due to an accumulation of many smaller errors that lead to an overall flawed translation. This may be true of MT as well. Our analysis reveals distinctive error patterns across different translation systems and language pairs. These patterns provide valuable insights into areas where NMT and LLM systems continue to face challenges despite recent advances.

3.4.1. Comparative analysis of error types across systems and language pairs

To facilitate direct comparison across systems and language pairs, we present a comparative analysis of error distributions in Figure 3. This visualization demonstrates that Terminology (T) and Literalness (L) errors constitute the most significant portion of translation errors across all system-language pairs, with particularly high frequencies in Arabic translations. GPT-4o (eng-

ara) shows the highest percentage of Terminology errors at 40.91%, closely followed by Google (eng-ara) at 37.04%.

Language-specific patterns emerge clearly in the data. Arabic translations (regardless of system) show concentrated error distributions primarily in Terminology and Literalness, together accounting for 72.7% of GPT-4o's errors and 74% of Google's errors. Chinese translations, in contrast, exhibit more diverse error profiles spread across multiple categories.

System-specific patterns are evident in Chinese translations, with GPT-4o showing more Terminology (29.17%), Literalness (20.83%), and Omission (16.67%) errors out of the total 24 identified errors, while Google demonstrates significantly more Cohesion issues (36.84%) out of the total 19 identified errors. This notable divergence in Cohesion errors is unique to Google (eng-chi) translations, while being completely absent in both Arabic translation systems.

GPT-4o exhibits a more diverse error profile than Google across both languages, spreading across multiple error categories. This pattern suggests that different neural architectures may produce distinctly different error patterns even when processing the same STs.

The data further reveals specialized error tendencies. Word Form (WF) errors appear exclusively in GPT-4o (eng-ara) translations (9.09%), while Faithfulness (F) errors are unique to GPT-4o (eng-chi) translations (4.17%). Grammar (G) errors are only found in GPT-4o's Arabic translations, and Style (ST) errors only in Google's Arabic translations, highlighting how error patterns can be both system-dependent and language-dependent.

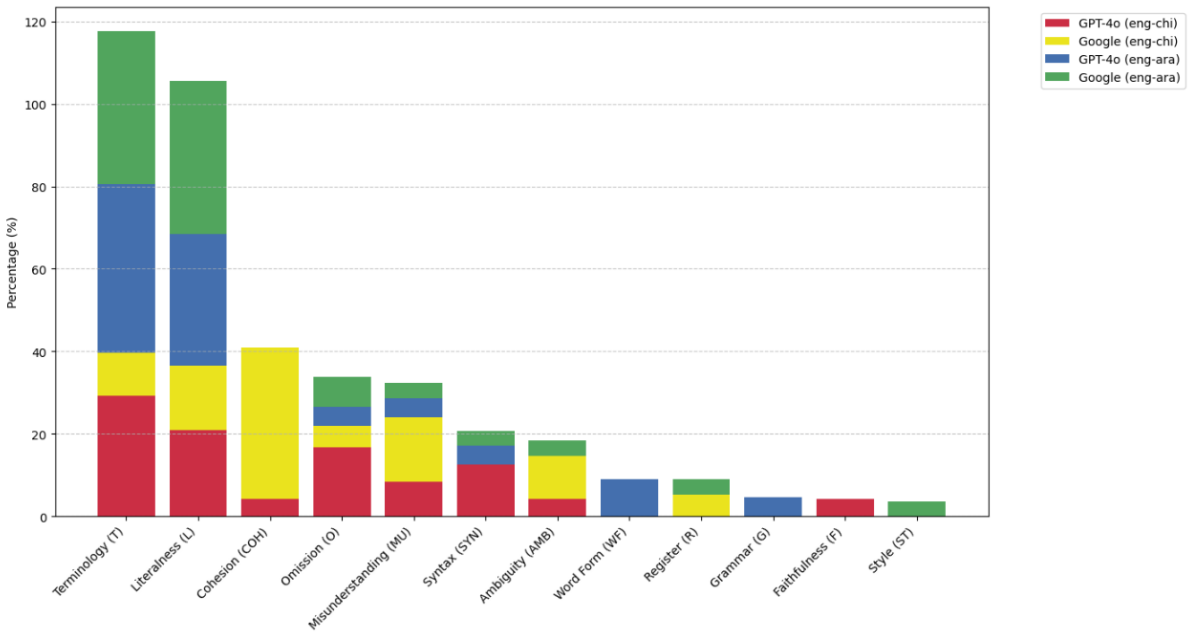

Figure 3: Composition of error types across systems and language pairs

3.4.2. Severity analysis across systems and language pairs

We also compared the severity levels of translation errors across different systems (GPT-4o and Google) and language pairs (eng-chi and eng-ara), categorized by error type. As illustrated in Figure 4, major to critical severity errors (Levels 8–16) are found exclusively in GPT-4o (eng-chi), specifically within the Misunderstanding (MU) category. This category has a notably

high average severity level of 10.00, substantially higher than any other error type across both systems.

GPT-4o's Chinese output also shows a high average severity of 4.00 in the Literalness (L) category. According to the ATA grading framework, a misunderstanding error results from a mistranslated word, idiom, or misinterpreted sentence structure, while a literalness error occurs when a translation sticks too closely to the ST, producing awkward or incorrect phrasing (American Translators Association 2022). Both are classified as transfer errors, which negatively impact the accurate conveyance of meaning. These findings point to serious issues with semantic transfer in GPT-4o's eng-chi translations, suggesting the model faces significant challenges in preserving the ST's meaning in this language pair.

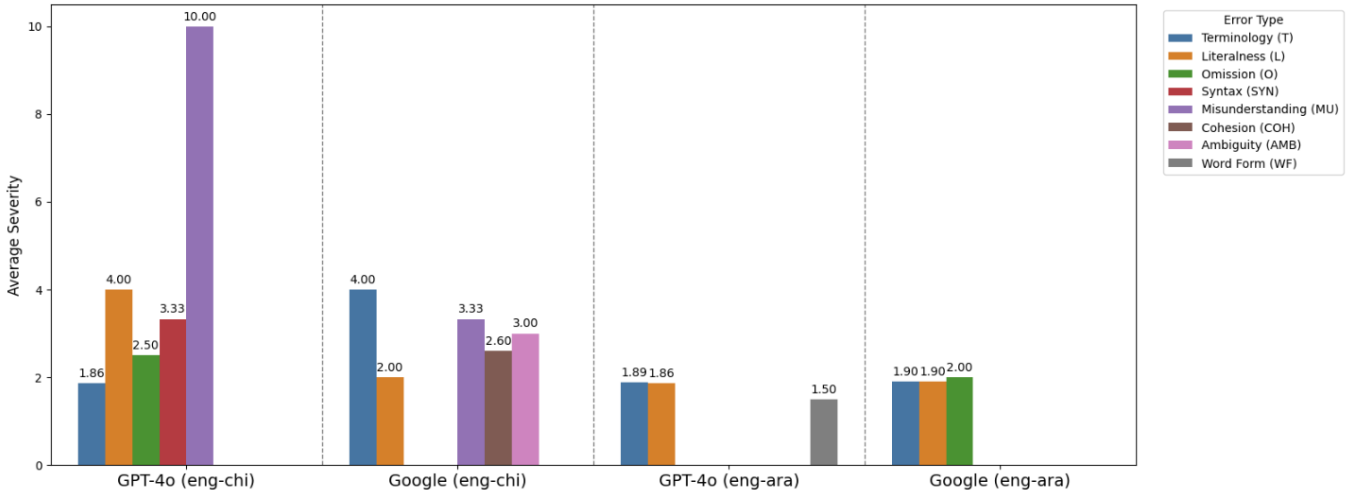

Figure 4: Average severity level by system-language pair and error type

Google's eng-chi translations show a more balanced distribution of error severity across categories. The most notable issue appears in the Terminology (T) category, with an average severity of 4.00, followed by Misunderstanding (MU) at 3.33 and Ambiguity (AMB) at 3.00, indicating moderate to major concerns. While these are also classified as meaning transfer errors, Google does not display extreme outliers in any single category, unlike GPT-4o. The terminology issues in Google's output tend to occur at the word or phrase level, suggesting localized problems rather than broader semantic breakdowns.

Minor to moderate errors (Levels 1–2) are the most common in Arabic translations across both systems. For GPT-4o, 95.45% of errors fall within this range, while Google shows a full 100% of its errors at these levels. Compared to eng-chi translations, eng-ara outputs exhibit consistently lower severity scores, suggesting that this language pair may be handled more effectively overall. However, as previously noted, ten 2-point errors still total 20 points, enough to fail an exam. This suggests that a high volume of low-severity errors can still significantly impact overall quality.

While Terminology and Literalness errors are common across systems and language pairs, their severity varies considerably. For GPT-4o's eng-chi translations, Terminology errors are relatively minor to moderate (1.86), whereas Literalness errors are significantly more severe, averaging 4.00 (major). Conversely, Google's eng-chi output shows the reverse trend. Terminology errors are major (4.00), while Literalness errors are moderate (2.00). In the case

of eng-ara, both GPT-4o and Google exhibit minor to moderate severity in these categories, suggesting that these issues are generally less problematic in this language pair.

Beyond these shared error types, distinct patterns emerge for specific system-language pairs. For instance, Syntax errors appear mostly in GPT-4o's eng-chi output, with an average severity of 3.33 (moderate to major), while Word Form errors are unique to GPT-4o's eng-ara translations at a lower average severity of 1.50 (minor to moderate). Meanwhile, Cohesion (2.60) and Ambiguity (3.00) errors predominantly occur in Google's eng-chi translations, indicating moderate to major impact. These variations highlight how different systems handle linguistic challenges in distinct ways. We will provide qualitative examples evaluating both error type and severity to better understand each system's translation quality and limitations in the discussion section.

In summary, although GPT-4o's performance in translating eng-ara met the ATA certification threshold in this experiment, it still has several areas for improvement. The system struggles with contextual awareness, often producing literal translations that fail to convey the intended meaning. Terminological accuracy is another challenge, which was mistranslated due to a lack of context sensitivity. The system also needs to handle idiomatic expressions more effectively to ensure cultural and linguistic appropriateness. Grammatical errors such as incorrect forms and inconsistent phrasing further undermine the translation quality. To address these issues, GPT-4o could benefit from enhanced context modeling, better handling of idiomatic language, and training on more diverse datasets that emphasize cultural nuances and grammatical accuracy.

4. Discussion

Our analysis in Section 3 revealed distinct performance patterns across language pairs. For eng-chi translations, both automatic and manual evaluations showed that Google Translate and GPT-4o excel at different segments, with each system demonstrating unique strengths. For eng-ara translations, both systems performed at a consistently higher level overall, though they still encountered some common challenges with certain segments. This section presents a qualitative analysis of representative examples that illustrate these patterns.

4.1. Complementary strengths in eng-chi translation across systems

The segment-level analysis of COMET results in Section 3.1. (Figure 1) showed that GPT-4o and Google Translate excel at translating different segments in Chinese. This pattern was reinforced by our manual evaluation in Section 3.2., which found contrasting performance patterns between GPT-4o and Google Translate on the same texts. Our error analysis in Section 3.4. further revealed different error profiles, with GPT-4o producing primarily Terminology (29.17%), Literalness (20.83%), and Omission (16.67%) errors, while Google Translate exhibited substantially more Cohesion issues (36.84%). These quantitative findings are illustrated through the following representative examples.

4.1.1. Example illustrating GPT-4o's advantage

Example (1) in Table 4 (Segment 7) illustrates how the two systems handle the translation of the same source sentence "Every year, an estimated $40 billion is diverted from the public treasury through charitable donations." with different word choices.

GPT-4o rendered this as "每年，估计有 400 亿美元通过慈善捐赠从公共财政中转移出去。" ('Every year, an estimated $40 billion is **transferred** through charitable donations from public treasury.'). In contrast, Google Translate produced "每年，估计有 400 亿美元通过慈善捐款从国库中挪用。" ('Every year, an estimated $40 billion is **misappropriated** through charitable donations from national treasure.').

Although COMET scores showed only a marginal difference (0.90 for GPT-4o vs. 0.89 for Google Translate), human evaluation identified a more pronounced disparity in translation quality. The ATA graders identified no errors in GPT-4o's translation, whereas both graders marked Google Translate's version with a T-4 error (terminology error of severity level 4). The critical issue was Google's word choice of 挪用 ('misappropriated') for *diverted*, which introduces an unwarranted negative connotation of financial misconduct that is not present in the ST.

Table 4: Segment-level outputs and ATA grader feedback for Example (1) in the eng-chi pair

ST	Every year, an estimated $40 billion is **diverted** from the public treasury through charitable donations.			
System	**TT**	**COMET**	**Grader 1**	**Grader 2**
GPT-4o (back-translation)	Every year, an estimated $40 billion is **transferred** through charitable donations from public treasury.	0.90		
Google (back-translation)	Every year, an estimated $40 billion is **misappropriated** through charitable donations from national treasure.	0.89	T-4 *Diverted* was not translated correctly in this context.	T-4 被转移出去 ('is transferred out')

4.1.2. Example illustrating Google Translate's advantage

Example (2) in Table 5 (Segment 23) examines the two systems' different approaches to translating a rhetorically complex sentence: "The ignorant comments about historically complex subjects are laughable at best and frightening at worst."

GPT-4o translated this as "关于历史上复杂问题的无知评论至多令人发笑，至少令人害怕。" ('Ignorant comments about historically complex problems are at most hilarious and at least terrifying.'). Google Translate produced "关于历史复杂主题的无知评论充其量是可

笑的，最坏的情况是令人恐惧的。” (‘Ignorant comments about historically complex subjects are laughable at best and terrifying under the worst situation’).

Interestingly, while GPT-4o achieved a higher COMET score (0.86 compared to Google Translate’s 0.75), the human evaluation revealed significant issues in GPT-4o’s translation. For this example, GPT-4o received the largest accumulative error points from the two graders combined in our dataset. The ATA graders identified multiple serious errors in GPT-4o’s translated version. Grader 1 noted a moderate T-2 error (terminology error of severity level 2) for mistranslating *subjects* as *problems*. Grader 1 also identified a major MU-4 error (misunderstanding error of severity level 4) for reversing the meaning of *at best* and *at worst* in the translation. Grader 2 assessed the entire sentence as having a critical misunderstanding error (MU-16), suggesting instead “轻则令人发笑，重则令人恐惧” (‘At the mildest, it’s hilarious; at the worst, it’s terrifying’).

In contrast, Google Translate’s version received only one moderate error mark: a cohesion error of severity level 2 (COH-2) from one grader, who suggested using “往好里说，往坏里说” (‘say it at best, say it at worst’) to improve the flow and naturalness of the Chinese expression.

Table 5: Segment-level outputs and ATA grader feedback for Example (2) in the eng-chi pair

ST	The ignorant comments about historically complex **subjects** are laughable **at best** and frightening **at worst**.			
System	**TT**	**COMET**	**Grader 1**	**Grader 2**
GPT-4o (back-translation)	Ignorant comments about historically complex **problems** are **at most** hilarious and **at least** terrifying.	0.86	T-2 *Topic* was translated as *problem.* MU-4 *At best* and *at worst* were reversed.	MU-16 轻则令人发笑，重则令人恐惧。 (‘At the mildest, it’s hilarious; at the worst, it’s terrifying.’)
Google (back-translation)	Ignorant comments about historically complex subjects are laughable **at best** and terrifying **under the worst situation**.	0.75		COH-2 往好里说，往坏里说 (‘Say it at best, say it at worst’)

The two examples above also resonate with the findings from Section 3.3, that automatic metrics might not always align with human assessment. For Example (1), despite GPT-4o and Google Translate’s similarly high COMET scores (0.90 vs. 0.89), human evaluators identified major terminology errors in Google’s translation that are contextually wrong and significantly impact the meaning transfer from the ST to the TT. For Example (2), although GPT-4o received a much higher COMET score than Google (0.86 vs 0.75), human evaluators identified moderate terminology errors and critical misunderstanding errors that

substantially impacted translation quality. These examples demonstrate that the discrepancy between automatic and human evaluation was particularly evident in segments containing contextually sensitive terminology or nuanced rhetorical expressions and idiomatic language use in the eng-chi language pair.

4.2. Common challenges in eng-ara translation across systems

The segment-level COMET analysis in Section 3.1. (Figure 1) showed that GPT-4o and Google Translate deliver consistently high scores in Arabic, frequently exceeding 0.95, with Google maintaining a marginal lead on particularly challenging segments such as 7 and 27. However, manual evaluation in Section 3.2. painted a more sophisticated picture: only GPT-4o's Arabic output met the ATA certification threshold, whereas Google Translate fell short on Text 2 despite strong automatic scores. Error profiling in Section 3.4. revealed that both systems share a similar error landscape dominated by Terminology and Literalness issues, accounting for 72.7% of GPT-4o's and 74% of Google's errors. However, they diverge at the margins: Word-Form and Grammar errors appear exclusively in GPT-4o's output, while Style errors occur only in Google Translate's. Many Arabic errors are minor to moderate in severity (≤ 2 points), so even high COMET scores can conceal the cumulative impact of many low-level issues. These quantitative findings set the stage for the representative Arabic examples that follow.

4.2.1. Example (3): Idiomatic and terminological pitfalls in Segment 11

The source sentence under review is the rhetorically loaded statement: "Rather than simply **write checks** for existing institutions, these 'philanthrocapitalists', as they are often called, **aggressively** seek to shape their operations."

GPT-4o translated this as "بدلاً من مجرد كتابة الشيكات للمؤسسات القائمة، يسعى هؤلاء "الرأسماليون الخيريون ..."بشكل عدواني إلى تشكيل عمليات تلك المؤسسات." ('Instead of merely writing checks to existing institutions, these *philanthro-capitalists*... in a hostile manner strive to shape the operations of those institutions.'). Google Translate produced "وبدلاً من مجرد كتابة الشيكات للمؤسسات القائمة، يسعى هؤلاء "الرأسماليون الخيريون ..."إلى تشكيل عملياتهم بقوة." ('And instead of merely writing checks to existing institutions, these *philanthro-capitalists*... seek to shape their operations forcefully').

Both GPT-4o and Google Translate render the Arabic idiom *write checks* literally and struggle to convey the pragmatic nuance of *aggressively*. Table 6 juxtaposes the outputs and the ATA-grader feedback. Both systems mishandle two pragmatically charged elements. First, the idiom ***write checks*** is calqued as *كتابة الشيكات* ('writing checks'), a literal phrase that Arabic readers interpret either literally or as awkward bureaucratic jargon, thus obscuring its idiomatic meaning of *passively donating funds*. Second, GPT-4o's *عدواني* ('hostile,' 'belligerent') and Google's *بقوة* ('forcefully,' 'strongly') each skew the connotation of ***aggressively***: the former suggests hostility, the latter mere strength, whereas the source implies *proactive, hands-on involvement*. A more idiomatic rendering would paraphrase the segment as *بدلاً من الاكتفاء بالتبرعات المالية ... يسعون بنشاط حثيث إلى توجيه عمل تلك المؤسسات* ('instead of merely relying on monetary donations … they actively strive to steer the work of those institutions'). Although COMET scores for eng-ara average near 0.97, the graders' L-2 and T-2 annotations show that calqued idioms and imprecise lexical choices still undermine professional quality and require post-editing despite the ostensibly high automatic scores.

Table 6: Segment-level outputs and ATA grader feedback for Example (3) in the eng- ara pair

ST	Rather than simply **write checks** for existing institutions, these "philanthrocapitalists", as they are often called, **aggressively** seek to shape their operations.			
System	**TT**	**COMET**	**Grader 1**	**Grader 2**
GPT-4o (back-translation)	Instead of merely **writing checks** to existing institutions, these 'philanthrocapitalists '... **in a hostile manner** strive to shape the operations of those institutions.	0.88	L-2 كتابة الشيكات ('writing checks') – literal rendering; pragmatic force obscured. T-2 عدواني ('in a hostile manner') – conveys hostility rather than energetic engagement.	T-2 عدواني ('in a hostile manner') – the translation meant an active or impulsive way
Google (back-translation)	And instead of merely **writing checks** to existing institutions, these 'philanthrocapitalists '... seek to shape their operations **forcefully**.	0.90	L-2 كتابة الشيكات ('writing checks') – same literalism.	L-2 بقوة ('forcefully') – misses sense of *proactively / assertively*.

4.2.2. Example (4): Collocations, omissions & connotation drift in Segment 13

The source sentence for Example (4) is "Moving away from a thoughtfully researched narrative to conform to the technological limits of media platforms takes us away from **scholarship** into **sound bites** and **statements of fact without context**."

GPT-4o translated this as "الابتعاد عن السرد المدروس بعناية للتماشي مع القيود التكنولوجية لمنصات الإعلام يبعدنا عن البحث الأكاديمي إلى مقاطع مختصرة وتصريحات واقعية خالية من السياق" ('Moving away from the carefully studied narrative to keep pace with media-platform technological constraints distances us from academic research toward concise clips and realistic statements devoid of context.'). Google Translate produced "إن الابتعاد عن السرد المدروس بعناية من أجل التوافق مع الحدود التكنولوجية لمنصات الإعلام يأخذنا بعيدًا عن الدراسات العلمية إلى مقاطع صوتية وبيانات واقعية بدون سياق." ('To move away from the carefully studied narrative to align with the technological limits of media platforms takes us far from scientific studies to audio clips and realistic data without context.').

Both systems falter on three fronts: over-specific renderings of ***scholarship*****,** incomplete treatment of ***sound bites*****,** and misconstrual of ***statements of fact*****.** Table 7 contrasts the Arabic outputs with the ATA graders' notes to show how these low-severity errors accumulate.

Table 7: Segment-level outputs and ATA grader feedback for Example (4) in the eng-ara pair.

ST	Moving away from a thoughtfully researched narrative **to conform to** the technological limits of media platforms takes us away from **scholarship** into **sound bites** and **statements of fact without context**.			
System	**TT**	**COMET**	**Grader 1**	**Grader 2**
GPT-4o (back-translation)	Moving away from the carefully studied narrative to keep pace with media-platform technological constraints distances us from **academic research** toward **concise clips** and **realistic statements** devoid of context.	0.95	T-2 البحث الأكاديمي (‘academic research’) – overly narrow. O-2 مقاطع مختصرة (‘concise clips’) – omits the sound element. T-2 تصريحات واقعية (‘realistic statements’) – misrepresents *facts*.	T-2 تصريحات واقعية (‘realistic statements’) – the translation does not convey the intended meaning.
Google (back-translation)	To move away from the carefully studied narrative **to align** with the technological limits of media platforms takes us far from **scientific studies** to **audio clips** and **realistic data** without context.	0.97	T-2 الدراسات العلمية (‘scientific studies’) – overly specific. O-2 مقاطع صوتية (‘audio clips’) – captures sound but omits brevity. T-2 بيانات واقعية (‘realistic data’) – misconstrues *statements of fact*. T-2 التوافق (‘to align’) – inaccurate for *to conform.*	T-2 التوافق (‘to align’) – failed to convey the full meaning. MU-2 مقاطع صوتية (‘audio clips’) – what is meant is brief or concise excerpts. T-2 بيانات واقعية (‘realistic data’) – not the original meaning

This segment reveals three recurring Arabic MT vulnerabilities. First, terminological over-specification (T-2): both systems translate *scholarship* as *البحث الأكاديمي* ('academic research') or *الدراسات العلمية* ('scientific studies'), terms that confine the broad notion of scholarship to academic or scientific research; graders recommend the wider phrase *المنحى العلمي* ('scholarly approach'). Second, partial omission (O-2)**:** GPT-4o's *مقاطع مختصرة* ('concise clips') deletes the *sound* element, whereas Google's *مقاطع صوتية* ('audio clips') retains *sound* but ignores brevity—neither fully captures *sound bite*. Third, connotation drift in fact-rendering (T-2): both systems choose *تصريحات/بيانات واقعية* ('realistic statements/data'), where *واقعية* signals *realistic* rather than *factual*, and *بيانات* might suggest either date or formal communiqués. The graders propose *تقديم حقائق* ('presentation of facts') to restore the intended sense of context-free fact statements. Collectively, these issues, narrow term mapping, headword omission, and connotative mismatch, reinforce the broader trend noted in Section 3.4.: Arabic MT output is dominated by low-severity Terminology, Literalness, and Omission errors that automatic metrics overlook, yet their accumulation materially degrades professional-level adequacy.

5. Conclusion

This study presents an evaluation of MT quality using the ATA certification framework, revealing several important findings about the current capabilities and limitations of both LLM-based and NMT-based translation systems. While neural-based automatic metrics like COMET have shown stronger correlation with human judgment compared to traditional lexical-based metrics, our analysis reveals that such automatic metrics may overestimate MT quality and fail to capture critical errors that professional translators identify, particularly in the evaluation of nuanced rhetorical expressions and idiomatic language use in translations. These findings underscore the continued importance of incorporating human assessment into professional translation workflows.

Among the four system-language pair combinations evaluated, GPT-4o demonstrated superior performance in eng-ara translation for both required texts in this experiment. The remaining three combinations - GPT-4o's eng-chi, Google Translate's eng-chi, and Google Translate's eng-ara translations - all accumulated high error points from graders for certain texts, indicating that they do not yet achieve the translation standards required for professional translators.

Our error analysis uncovered distinct patterns between systems and language pairs. For eng-chi translations, GPT-4o and Google Translate exhibited notably different error distributions. GPT-4o's errors concentrated in three main categories: Terminology, Literalness, and Omission, while Google Translate showed a different pattern dominated by Cohesion errors, followed by equal proportions of Literalness and Misunderstanding errors. In contrast, eng-ara translations from both systems displayed remarkably similar error patterns, primarily characterized by Terminology and Literalness errors. These variations suggest that translation challenges extend beyond mere text complexity, potentially reflecting distinct linguistic characteristics inherent to each language pair and the different approaches employed by each system in handling these challenges.

While this study offers valuable insights into the current state of machine translation, it has certain limitations. Due to funding constraints, our manual assessment using ATA-certified graders was limited to two texts for each language pair and MT system. Future research would

benefit from expanding the dataset and incorporating more extensive automatic assessment metrics to validate these findings across a broader range of texts and contexts.

Looking forward, several promising avenues exist for improving LLM translation performance. The implementation of retrieval-augmented generation (RAG) and advanced prompt engineering techniques could enhance translation quality by allowing for better tone control, bias mitigation, and integration of domain-specific terminology. Additionally, the potential for fine-tuning LLMs using existing translation memories could further improve their performance on specific translation tasks.

These findings contribute to our understanding of the strengths and limitations of current MT systems while highlighting areas for future development. The distinct error patterns observed between systems and language pairs warrant further investigation, particularly in understanding how linguistic characteristics and system architectures interact to influence translation quality. As MT technology continues to evolve, maintaining a balanced approach that combines automatic metrics with professional human assessment remains crucial for ensuring translation quality in professional contexts.

References

American Translators Association. n.d. About the ATA Certification Exam. (https://www.atanet.org/certification/about-the-ata-certification-exam/) (Accessed 2024-01-16.)

American Translators Association. 2022. Explanation of Error Categories. (https://www.atanet.org/certification/how-the-exam-is-graded/error-categories/) (Accessed 2025-09-13.)

Cady, Larry P. & Tsou, Benjamin K. & Lee, John S. Y. 2023. Comparing Chinese-English MT performance involving ChatGPT and MT providers and the efficacy of AI-mediated post-editing. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*. 205–216. (https://aclanthology.org/2023.mtsummit-users.20/) (Accessed 2025-09-09.)

Freitag, Markus & Foster, George & Grangier, David & Ratnakar, Vikas & Tan, Qijun & Macherey, Wolfgang. 2021. Experts, errors, and context. A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics* 9: 1460–74.

Google AI Blog. 2020. Recent advances in Google Translate. (https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html) (Accessed 2025-09-09.)

Intento. 2024. The state of machine translation 2024. (https://inten.to/machine-translation-report-2024/) (Accessed 2024-06-31.)

International Organization for Standardization. 1998. *ISO 639-2:1998 Codes for the representation of names of languages — Part 2: Alpha-3 code.* (https://www.iso.org/standard/4767.html) (Accessed 2024-01-16.)

Jiao, Wenxiang & Wang, Wenxuan & Huang, Jen-tse & Wang, Xing & Shi, Shuming & Tu, Zhaopeng. 2023a . Is ChatGPT a good translator? Yes with GPT-4 as the engine. *arXiv*. *arXiv:2301.08745.* (http://arxiv.org/abs/2301.08745) (Accessed 2024-01-16.)

Jiao, Wenxiang & Wang, Wenxuan & Huang, Jen-tse & Wang, Xing & Shi, Shuming & Tu, Zhaopeng. 2023b. Is ChatGPT a good translator? A preliminary study. *arXiv preprint. arXiv:2301.08745* 1(10). (https://wxjiao.github.io/downloads/tech_chatgpt_arxiv.pdf) (Accessed 2024-01-16.)

Koby, Geoffrey S. 2015. The ATA flowchart and framework as a differentiated error-marking scale in translation teaching. In Cui, Ying & Zhao, Wei (eds.), *Handbook of research on teaching methods in language translation and interpretation*. IGI Global. 220–53.

Koby, Geoffrey S. & Champe, Gertrud G. 2013. Welcome to the real world. Professional-level translator certification. *Translation & Interpreting. The International Journal of Translation and Interpreting Research 5*(1): 156–73.

Läubli, Samuel & Castilho, Sheila & Neubig, Graham & Sennrich, Rico & Shen, Qinlan & Toral, Antonio. 2020. A set of recommendations for assessing human–machine parity in language translation. *Journal of artificial intelligence research* 67: 653–72.

Liu, Yiheng & Han, Tianle & Ma, Siyuan & Zhang, Jiayue & Yang, Yuanyuan & Tian, Jiaming & He, Hao & Li, Antong & He, Mengshen & Liu, Zhengliang & Wu, Zihao & Zhu, Dajiang & Li, Xiang & Qiang, Ning &Shen, Dingang &Liu, Tianming &Ge, Bao. 2023. Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiology* 1(2): 100017.

Lommel, Arle & Uszkoreit, Hans & Burchardt, Aljoscha. 2014. Multidimensional quality metrics (MQM). A framework for declaring and describing translation quality metrics. *Tradumàtica* 12: 0455–463.

OpenAI. 2019a. *Better language models and their implications.* (https://openai.com/research/better-languagemodels) (Accessed 2024-01-31.)

OpenAI. 2019b. *GPT-2: 1.5B release*. (https://openai.com/research/gpt-2-1-5b-release) (Accessed 2024-01-31.)

OpenAI. 2022. *Introducing ChatGPT*. (https://openai.com/blog/chatgpt.) (Accessed 2024-01-31.)

OpenAI. 2023a. *GPT-4*. (https://openai.com/research/gpt-4) (Accessed 2024-01-31.)

OpenAI. 2023b. *What is the difference between the GPT-4 models?* (https://help.openai.com/en/articles/7127966-what-is-the-difference-between-the-gpt-4-models) (Accessed 2024-01-31.)

OpenAI. 2024. *Fine-tuning*. (https://platform.openai.com/docs/guides/fine-tuning) (Accessed 2024-01-31.)

Ouyang, Long & Wu, Jeffrey & Jiang, Xu & Almeida, Diogo & Wainwright, Carroll & Mishkin, Pamela & Zhang, Chong & Agarwal, Sandhini & Slama, Katarina & Ray, Alex & Schulman, John & Hilton, Jacob & Kelton, Fraser & Miller, Luke & Simens, Maddie & Askell, Amanda & Welinder, Peter & Christiano, Paul F. & Leike, Jan & Lowe, Ryan. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35: 27730–27744.

Papineni, Kishore & Roukos, Salim & Ward, Todd & Zhu, Wei-Jing. 2002. Bleu. A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL).* Philadelphia, USA. 311–18. (https://aclanthology.org/P02-1040.pdf) (Accessed 2024-01-31.)

Popović, Maja. 2015. chrF. Character n-gram F-score for automatic MT evaluation. In *Proceedings of the tenth workshop on statistical machine translation.* 392–95.

Ray, Partha Pratim. 2023. ChatGPT. A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems* 3: 121–54.

Rei, Ricardo & Stewart, Craig & Farinha, Ana C. & Lavie, Alon. 2020. COMET. A Neural Framework for MT Evaluation. *arXiv. arXiv:2009.09025*. (http://arxiv.org/abs/2009.09025) (Accessed 2024-01-31.)

Shahriar, Sakib & Lund, Brady D. & Mannuru, Nishith Reddy & Arshad, Muhammad Arbab & Hayawi, Kadhim & Bevara, Ravi Varma Kumar & Mannuru, Aashrith & Batool, Laiba. 2024. Putting GPT-4o to the sword. A comprehensive evaluation of language, vision, speech, and multimodal proficiency. *Applied Sciences* 14(17): 7782.

Siu, Sai Cheong 2023. ChatGPT and GPT-4 for professional translators. Exploring the potential of large language models in translation (preprint). Available at *SSRN 4448091* (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4448091) (Accessed 2024-01-31.)

Wu, Yonghui & Schuster, Mike & Chen, Zhifeng & Le, Quoc V. & Norouzi, Mohammad & Macherey, Wolfgang & Krikun, Maxim & Cao, Yuan & Gao, Qin & Macherey, Klaus & Klingner, Jeff & Shah, Apurva & Johnson, Melvin & Liu, Xiaobing & Kaiser, Łukasz & Gouws, Stephan & Kato, Yoshikiyo & Kudo, Taku & Kazawa, Hideto & Stevens, Keith & Kurian, George & Patil, Nishant & Wang, Wei & Young, Cliff & Smith, Jason & Riesa, Jason & Rudnick, Alex & Vinyals, Oriol & Corrado, Greg & Hughes, Macduff & Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv preprint. arXiv:1609.08144. (https://arxiv.org/abs/1609.08144) (Accessed 2024-01-31.)

Zou, Longhui. 2024. Cognitive Processes in Human-ChatGPT Interaction during Machine Translation Post-editing. Kent, Ohio, USA: Kent State University. (Doctoral dissertation.) (http://rave.ohiolink.edu/etdc/view?acc_num=kent1731969433680887) (Accessed 2025-09-13.)

Zou, Longhui & Saeedi, Ali & Koby, Geoffrey S. 2024. The comparison of quality. Neural Machine Translation versus Large Language Models-Powered Translation in the American Translators Association Exam. *Translation in Transition 2024*. Batumi, Georgia. doi: 10.13140/RG.2.2.25890.90562.

Longhui Zou
Kent State University, USA
e-mail: lzou4@kent.edu

Ali Saeedi
University of Illinois Urbana-Champaign, USA
e-mail: saeedi2@illinois.edu

Geoffrey S. Koby
Kent State University, USA
e-mail: gkoby@kent.edu