3L: Language, Linguistics, Literature® The Southeast Asian Journal of English Language Studies
Vol 32(1), March 2026 http://doi.org/10.17576/3L-2026-3201-11

Assessing Legal Translations Generated by GPT-4 Turbo Using MQM: A Comparative Study

LAMA ABDULLAH ALDOSARI *
Translation Department, English Language and Sciences College,
King Saud University, Saudi Arabia
lamaabdullahmd@gmail.com

NASRIN ALTUWAIRESH
Translation Department, English Language and Sciences College,
King Saud University, Saudi Arabia

ABSTRACT

Advances in artificial intelligence (AI), particularly through GPT models, have significantly enhanced machine translation (MT) capabilities, offering more accurate and nuanced translations. This study investigates the influence of different translation prompts on the quality of legal translations generated by Generative Pretrained Transformer-4 (GPT-4) Turbo. By applying the Multidimensional Quality Metrics (MQM) framework, the research evaluates both the types and frequency of errors found in translations produced by existing and newly developed prompts. The study focuses on comparing translation commission prompts—designed to enhance context-specific texts—with existing prompts in terms of translation quality. A dataset of 14 Saudi Laws, drawn from official sources, serves as the basis for analysis, with reference translations used as benchmarks. The findings reveal that the newly developed prompts, specifically tailored for legal translation, resulted in significantly lower error rates (10-12%) compared to existing prompts, which demonstrated error rates ranging from 34% to 45%. These results underscore the transformative potential of tailored prompt engineering in achieving high-quality legal translations by reducing errors in terminology, accuracy, and style. By categorising and ranking translation errors by severity, the research highlights the impact of prompt engineering on improving legal translation performance. These findings contribute to the development of more effective MT systems, offering practical insights for refining machine translation in the legal field and beyond.

Keywords: GPT-4 Turbo; Legal Translation; Machine Translation; Multidimensional Quality Metrics; Translation Commission

INTRODUCTION

Translation is fundamental in bridging communication gaps across diverse languages, cultures, and regions. The evolution of machine translation (MT) has significantly influenced the field by facilitating the exchange of information on a broader scale (Mohamed et al., 2024). Recent advancements in artificial intelligence (AI) have further transformed MT, with large language models (LLMs) such as Generative Pretrained Transformer-4 (GPT-4) Turbo representing a major advancement in this domain. These models demonstrate enhanced capabilities in understanding and generating human language, leveraging extensive datasets and sophisticated algorithms to produce increasingly nuanced and contextually accurate translations. Nevertheless, despite these technological breakthroughs, the question of whether AI systems can achieve parity with professional human translators remains a subject of ongoing investigation (Sulaiman et al., 2025).

A key factor in enhancing the performance of LLMs is prompt engineering, which entails formulating clear and specific instructions to guide the model in producing optimal results (Wang et al., 2023). In particular, the structure of translation prompts—used to direct models like GPT-4 Turbo—is crucial in determining the quality of translations from MT systems. While there has been considerable progress in AI-driven translation technologies, the influence of different translation prompts on translation quality remains an important subject of study. This research seeks to investigate how varying prompts affect the effectiveness of LLMs, focusing on their impact on translation accuracy and overall performance.

In translation studies (TS), a major transformation occurred during the 1970s and 1980s, leading to the adoption of a functionalist approach (Holz-Mänttäri, 1984; Nord, 1997; Reiss, 1971; Reiss & Vermeer, 1984; Snell-Hornby, 1988, as cited in Munday et al., 2022). Vermeer (1989/2021) introduced the concept of translation commission, which refers to the set of instructions provided either by oneself or another party to complete a specific translation task. It involves providing clear guidelines to the translator about the purpose, tone, and audience of the translation, etc. These instructions help the translator make informed decisions that align with the translation receiver's expectations, ensuring that the final product not only conveys the correct meaning but also meets the specific needs of the target audience. Without proper guidance, key elements such as cultural nuances, specialised terminology, or the desired level of formality might be lost or misinterpreted. This commission defines essential aspects of the target text (TT), such as its intended function, the target audience, the time and place of reception, the medium, and the reason for translating (Munday et al., 2022).

Machine translation faces considerable challenges in specialised fields such as legal translation due to the complex terminology and intricate structures involved (Anesa & Kulbicki, 2022). Similarly, translation prompts proposed by researchers as a step toward improving translation quality used with GPT models for translation often fail to generate accurate, specialised translations, largely because the instructions provided are insufficient (Gao et al., 2023). Therefore, further advancements can be achieved by incorporating the translation commission approach into GPT-4 translation prompts. This approach offers two key advantages. First, GPT-4's extensive training on large datasets enables it to handle complex and specialised texts, such as legal documents, with greater precision (OpenAI, 2023). Second, the translation commission framework is widely recognised as a valuable tool for ensuring high-quality legal translations (Soriano Barabino, 2020).

Previous research has primarily focused on proposing and evaluating general and customised prompts for translation tasks across different language pairs. However, limited attention has been paid to assessing the types and frequency of errors in translations produced by both existing and new prompts using GPT-4 Turbo, as measured by the Multidimensional Quality Metrics (MQM) framework. The aim of this study was to address this gap by analysing the types and frequency of errors in translations generated by both existing and newly developed prompts with GPT-4 Turbo, utilising the MQM framework. Furthermore, the study evaluated how newly developed prompts, particularly those incorporating translation commissions, influence translation quality compared to existing prompts.

RESEARCH OBJECTIVES

The purpose of this study is to assess the types and frequency of errors in legal translations generated by GPT-4 Turbo using two categories of prompt designs: (1) prompt designs from previous studies ("existing translation prompts"), and (2) newly developed translation commission prompts ("new translation prompts"). The study applies the MQM framework, which allows systematic categorisation and severity ranking of translation errors. This approach provides a comprehensive measure of translation quality in legal contexts. MQM is particularly suitable for legal translation assessment because it not only identifies the type of errors (e.g., terminology, accuracy, style) but also ranks them by severity, enabling a nuanced assessment of translation quality.

RESEARCH QUESTIONS

To achieve the research objectives, the study aims to answer the following two research questions:

1. How do existing and new translation prompts in GPT-4 Turbo compare in terms of the types and frequency of errors encountered in legal translations, as measured by the MQM framework?
2. To what extent do new translation prompts improve legal translation quality, as evaluated using the MQM framework, compared to existing translation prompts in GPT-4 Turbo?

LITERATURE REVIEW

This study aims to evaluate the types and frequency of errors found in legal translations generated by both existing and new translation prompts using GPT-4 Turbo, assessed through the MQM framework. In addition, the research investigated the impact of the new translation prompts on translation quality compared to the existing ones, utilising the MQM framework for evaluation. This section addresses two key concepts that underpin the study. The first part explores MT, highlighting the importance of prompts in optimising GPT models and how customised prompts can lead to improved translations. The second part focuses on translation quality assessment (TQA), including human quality evaluation through the MQM framework.

MACHINE TRANSLATION

Advancements in AI and natural language processing (NLP) have significantly influenced MT, which refers to the automatic translation between natural languages (Lopez, 2008). Traditional systems like Google Translate use neural machine translation, relying on large multilingual datasets, while GPT models leverage deep learning techniques to better capture context and deliver more accurate translations (Islam et al., 2021; Liu et al., 2021). The use of MT has expanded across various fields, including legal translation (Anesa & Kulbicki, 2022), where researchers have highlighted its potential benefits (Dunđer, 2020; Knap-Dlouhá, 2022). However, despite these advancements, there is still a need to enhance the quality and precision of MT to fully meet the functional requirements of legal texts, underscoring the importance of refining MT for legal language through prompt engineering.

Recent studies have emphasised the growing use of GPT models such as GPT-3 and GPT-3.5 for translating specialised texts between Arabic and English (Banat & Adla, 2023; Khoshafah, 2023). These studies suggest that while GPT models generally produce understandable translations, they often struggle with handling nuances, cultural context, and complex content. Nevertheless, the primary focus of these studies has been on evaluating overall translation quality rather than the specific impact of translation prompts. This differs from earlier research, which primarily assessed NLP models' performance on general-purpose text, often leading to lower quality translations (Hendy et al., 2023; Jiao et al., 2023). Despite the advancements in translation quality, the role of specific translation prompts in enhancing the accuracy and handling of specialised content remains relatively unexplored. This suggests a need for further research into how tailored prompts can improve the performance of GPT models, particularly in complex and culturally nuanced translations.

IMPORTANCE OF PROMPTS

A prompt is a text input provided to an LLM to direct its output. For instance, a translation prompt might be as simple as: 'Translate from Arabic to English' (Zhang et al., 2023). Recent research has underscored the importance of prompts in enhancing the performance of GPT models. Jiao et al. (2023) highlighted that the style and structure of prompts can greatly affect translation quality. Similarly, Ekin's (2023) study revealed how prompt design can substantially influence ChatGPT's performance across various tasks, including translation.

The emergence of GPT models has brought about challenges in prompt engineering, as LLMs do not interpret prompts the way humans do (Lu et al., 2021; Webson & Pavlick, 2021). As a result, many studies have focused on the careful selection and engineering of prompts to enhance the translation quality produced by GPT models. Researchers have investigated various strategies for optimising prompt design, including adjusting the length, specificity, and phrasing of prompts to better guide the model's output. These studies have shown that even small modifications in the wording of prompts can lead to significant improvements in translation accuracy, fluency, and contextual understanding (Agrawal et al., 2022; Brown et al., 2020; Gao et al., 2023; Hendy et al., 2023; Jiao et al., 2023; Shin et al., 2020; White et al., 2023; Yamada, 2023).

CUSTOMISED PROMPTS

Customised prompts have become essential for improving translation quality in GPT models. Jiao et al. (2023) proposed a general prompt: "This is a [SL] to [TL] translation task, please provide the translation for these sentences: [ST]," which they found to generally yield satisfactory translations with minimal performance variation. However, Gao et al. (2023) critiqued Jiao et al.'s (2023) prompt for lacking supplementary information that could guide GPT models to produce superior translations. To address this, Gao and colleagues suggested a more customised prompt: "Please provide the [TL] translation of the [SL] sentences taken from the [Domain]. [ST]," which incorporates domain-specific details and part-of-speech tagging. This tailored approach resulted in better translations, especially for specialised texts like news and e-commerce.

Despite these advancements, early studies have not yet fully integrated translation theories into prompt design. Building on this, Yamada (2023) proposed a more detailed prompt: "Translate the following [SL] sentences into [TL]. Please fulfil the following conditions when translating. Purpose of the translation: Target audience: Dynamic equivalence is a strategy for translating from

the perspective of equalising the reader's response to [SL] and the [TT]. [ST]." Yamada incorporated Nida's (1977) equivalence theory, particularly dynamic equivalence, into the prompts and evaluated them using human metrics such as the MQM and Dynamic Quality Framework (DQF). While this approach showed promise and led to improved translation quality, it may have overlooked other essential factors necessary for consistently achieving high-quality translations. This underscores the ongoing exploration within translation studies (TS) into how translation theories can be better applied in practice.

TRANSLATION QUALITY ASSESSMENT

Translation quality assessment (TQA) is the process of evaluating texts to pinpoint issues and determine their compliance with professional standards (Mossop, 2001). Recent developments in translation studies have led to a focus on assessing functional equivalence rather than relying solely on subjective evaluations, as noted by House (2001). Nonetheless, scholars and practitioners widely recognise the absence of a singular objective method for measuring translation quality (Drugan, 2013). The assessment of translation quality involves key methodologies, including human evaluation, which is discussed below.

HUMAN QUALITY ASSESSMENT

Unlike automated methods for evaluating MT quality, human evaluation relies on expert assessors to evaluate the quality of translations (Chatzikoumi, 2020). Several types of human evaluation methods exist, including Direct Assessment (DA) introduced by Graham et al. (2013), DQF established by TAUS in 2011, MQM proposed in 2013 and updated in 2022 by the Quality Translation Launch Pad project, and the Scalar Quality Metric (SQM) suggested by Freitag et al. (2021). Collectively, these human evaluation methods contribute to a more thorough understanding of translation quality, addressing the limitations of automated evaluations by incorporating expert insights and contextual awareness.

The MQM framework is a leading approach for human evaluation, offering a comprehensive framework for assessing translation quality within the realm of MT (Mariana et al., 2015). The effectiveness of MQM lies in its ability to capture the multifaceted nature of translation quality by dividing the evaluation into eight dimensions of error typology, including fluency, accuracy, style, and terminology (Burchardt et al., 2021). This structure enables an assessor to pinpoint specific areas needing improvement (Lommel, 2018).

The MQM framework is highly versatile, applicable in numerous fields such as psychology, politics, medicine, and science, as demonstrated by Al-Khalifa et al. (2024). Vilar et al. (2023) further showcased its use in areas like biography, business, and news, emphasising its effectiveness in evaluating MT quality across diverse contexts. In legal translation, MQM has emerged as a key resource; Sosoni et al. (2022) applied it to analyse legal translations, revealing critical insights into errors in normative laws. This analysis not only enhanced the understanding of the legal translation process but also identified specific areas for improvement, ultimately contributing to higher translation quality in this specialised sector.

While previous research has primarily focused on developing and evaluating general and customised prompts for translation tasks across different language pairs, there has been limited attention given to analysing the types and frequency of errors in translations produced by both existing and new prompts using GPT-4 Turbo, as measured by the Multidimensional Quality

Metrics (MQM) framework. This study aims to address this gap by examining the error types and frequencies in translations generated by both established and newly developed prompts with GPT-4 Turbo, utilising the MQM framework. Furthermore, the study assessed how the newly designed prompts, particularly those that incorporate translation commissions, influence translation quality compared to existing prompts.

METHODOLOGY

The study is a qualitative study which aims to assess the types and frequency of errors encountered in translations generated by existing and new translation prompts using GPT-4 Turbo, as measured by the MQM framework. The study further aims to assess the impact of newly developed translation prompts (i.e., translation commission prompts) on the quality of translations, compared to existing prompts, using the MQM framework for evaluation, specifically, in terms of accuracy, fluency, and terminology.

To improve the quality of legal translations and bridge the gap between theory and practice, the integration of the translation commission framework has been proposed as a potential enhancement for GPT-4 Turbo translations. By incorporating specific instructions into the prompts, this framework aims to refine the translation process. The first prompt, which was referred to as translation commission prompt 1 (TCP1) in this study, includes: "This is a translation task from [SL] to [TL]. Please adhere to the requirements specified by the translation commission. Translation purpose: Translation Receiver: Time and place of translation reception: Medium: Motive: [ST]." The second prompt, referred to as translation commission prompt 2 (TCP2), instructs: "As a translator, your task is to translate from [SL] to [TL], making the content accessible to [target audience] for [purpose]. The translation was used at [time and place of reception], requiring a [translation strategy] that maintains [function]. The tone should be [tone], and the translation will be available in [medium]. [ST]."

For the purpose of this study, a corpus of fourteen publicly available Saudi Laws originally written in Arabic and translated into English, along with the English translations generated by GPT-4 Turbo by using existing translation prompts and new translation prompts, was compiled and analysed. The data that was utilised in this study comprised the Saudi Laws extracted from two Government websites: the National Centre for Archives and Records (Royal Court, n.d.-a) and the Bureau of Experts at the Council of Ministers (Royal Court, n.d.-b). The Laws were originally written and governed in Arabic and have been translated into English by the Official Translation Department at the Bureau of Experts at the Council of Ministers. Nevertheless, it is essential to clarify that the English versions of the Saudi Laws are intended for guidance purposes only. In all circumstances, the Arabic versions remain the authoritative text governing the laws in Saudi Arabia.

The laws included in this study were chosen based on two main criteria: ease of access and the availability of authentic STs paired with their professionally translated TTs. This selection helps to avoid the pitfalls of using unnatural STs, which can lead to inaccurate evaluations. These officially published English translations acted as reference texts for assessing the quality of machine translations produced by GPT-4 Turbo, using the MQM framework for evaluation.

The choice to centre the study on the legal domain is justified by the extensive prevalence of legal texts and research highlighting the challenges MT faces in accurately translating the complex language used in legal documents (Alkatheery, 2023; DeMattee et al., 2022). Hence, the primary objective of the study is to bridge the gap in leveraging MT to improve the quality of legal

translation, particularly by assessing the types and frequency of errors encountered in translations generated by existing and new translation prompts using GPT-4 Turbo, as measured by the MQM framework. The study further aims to assess the impact of newly developed translation prompts (i.e., translation commission prompts) on the quality of translations, compared to existing prompts, using the MQM framework for evaluation. The research employed existing translation prompts proposed by Jiao et al. (2023), who suggested the general prompt (GP), Gao et al. (2023), who proposed the domain-specific prompt (DP), and Yamada (2023), who introduced the target audience, purpose, and dynamic equivalence prompt (TAP) to instruct GPT-4 Turbo. Moreover, TCP1 and TCP2, specifically designed to enhance the MT quality of legal texts, were also employed. Then, the study assessed the translation quality of each prompt by identifying the errors encountered by them according to terminology, accuracy, and fluency.

The English translations produced by GPT-4 Turbo using existing translation prompts and new translation prompts of the Arabic Saudi Laws were assessed using the MQM framework. Each detected error was ranked according to its severity to provide a detailed assessment of translation quality. Upon identifying an error, its type was categorised, and its severity was evaluated based on three main levels: minor, major, and critical. Minor errors are those that have a minimal impact on the overall meaning or usability of the translation, such as minor grammatical issues or slight stylistic deviations. Major errors significantly affect the translation's clarity or accuracy, including substantial grammatical mistakes, incorrect terminology, or omissions that could cause confusion. Critical errors severely undermine the translation's integrity, leading to potential misunderstandings or misinterpretations, such as severe mistranslations or errors that distort the original message (Kocmi & Federmann, 2023). The severity assessment was conducted with reference to the official translation, which served as a benchmark for evaluating the accuracy and reliability of the generated translations. This approach enabled a comprehensive analysis of how different translation prompts influence the quality of translations generated by GPT-4 Turbo.

SAMPLE

The study relied on a dataset comprising 14 publicly available legislative documents, namely, Saudi Laws. Table 1 below provides a comprehensive overview of the datasets, including their corresponding sizes.

TABLE 1. List of the Saudi Laws included in the Data

No.	Title of the Law	Sentence Count	Word Count
1	Succession Commission Law	71	2,321
2	Law of the Flag	67	1,309
3	Law of the Council of Ministers	91	1,079
4	Basic Law of Governance	195	1,764
5	Shura Council Law	69	1,075
6	Law of Provinces	112	2,514
7	Law of Civil Procedure	675	14,033
8	Public Prosecution Law	96	1,857
9	Law of Criminal Procedure	516	11,098
10	Law of Procedure	177	4,009
11	Law of the General Commission for Guardianship	164	3,123
12	Law of the Judiciary	261	4,480
13	Tourism Law	97	2,722
14	Patent Law of the Cooperation Council	143	3,876
Total		1,971	55,260

SELECTION OF TRANSLATION PROMPTS

During the generation of MTs, the existing prompts from Gao et al. (2023), Jiao et al. (2023), and Yamada (2023), along with newly suggested prompts TCP1 and TCP2, were used. The resulting English translations were compiled into separate Excel spreadsheets for easy data manipulation and analysis. This iterative process produced 70 Excel files, representing five machine legal translation iterations for each of the fourteen laws examined. These prompts are carefully chosen to provide GPT-4 Turbo with different inputs, enabling it to generate high-quality legal translations. Their use contributed to the translation of the dataset, demonstrating GPT-4 Turbo's ability to produce accurate and contextually relevant translations from various sources.

The large volume of datasets was integrated into an Excel file to handle them efficiently, and a custom programming script was developed using the OpenAI Application Programming Interface (API) key and the Python programming language to automate the MT process by applying the existing translation prompts and the proposed prompts to the datasets, streamlining and speeding up the translation process. These scripts incorporated the translation prompts in multiple iterations, leading to the independent translation of the dataset. In this process, each prompt was used to generate translations for the dataset separately.

Tables 2 and 3 below present examples of legal MT output generated by both existing and new translation prompts, accompanied by the corresponding STs and reference translations. These tables offer a comprehensive comparison by juxtaposing the STs and reference translations with their corresponding MTs. This presentation allows for a thorough analysis of the translation quality and effectiveness of different prompts in the legal domain.

TABLE 2. Example of MT Outputs Taken from the Tourism Law, presenting GP, DP, TAP, TCP1, and TCP2

Text Type	Text
Source	دون إخلال بالأنظمة واللوائح ذات العلاقة، يحظر الإساءة إلى سمعة السياحة في المملكة والتعدي على الوجهات السياحية والمقوّمات السياحية، أو إلحاق الضرر بها، أو القيام بأي فعلٍ من شأنه الإضرار بقيمتها أو أهميتها السياحية.
Reference	Without prejudice to relevant laws and regulations, it is prohibited to undermine the reputation of tourism in the Kingdom, encroach on or cause damage to tourist destinations and tourist resources, or take any action that would harm their touristic value or importance.
GP	Without prejudice to the related laws and regulations, it is prohibited to harm the reputation of tourism in the Kingdom, to infringe upon tourist destinations and tourism assets, to damage them, or to engage in any act that would harm their value or their tourism importance.
DP	Without prejudice to the related regulations and bylaws, it is prohibited to harm the reputation of tourism in the Kingdom, infringe upon tourist destinations and their components, cause damage to them, or engage in any act that could harm their value or their significance for tourism.
TAP	Without prejudice to the related laws and regulations, it is prohibited to harm the reputation of tourism in the Kingdom, to infringe upon tourist destinations and their assets, to cause damage to them, or to engage in any act that could harm their value or their significance for tourism.
TCP1	Without prejudice to relevant laws and regulations, it is prohibited to disparage the reputation of tourism in the Kingdom, encroach upon or cause damage to tourist destinations and resources, or take any action that could harm their touristic value or importance.
TCP2	Without prejudice to related laws and regulations, it is prohibited to undermine the reputation of tourism in the Kingdom, trespass on or cause damage to tourist destinations and resources, or take any action that would detract from their touristic value or importance.

TABLE 3. Example of MT Outputs Taken from the Shura Council Law presenting GP, DP, TAP, TCP1, and TCP2

Text Type	Text
Source	لا يكون اجتماع مجلس الشورى نظاميا إلا إذا حضر الاجتماع ثلثا أعضائه على الأقل، بمن فيهم الرئيس أو من ينوب عنه، ولا تكون القرارات نظامية إلا إذا وافقت عليها أغلبية المجلس.
Reference	Meetings of the Shura Council shall not be valid unless attended by at least two-thirds of its members, including the Chairman or his designee. Resolutions shall not be valid unless approved by the majority of Council members.
GP	The Shura Council meeting is not considered official unless at least two-thirds of its members are present. This includes the President or his deputy, and decisions are not considered official unless approved by the majority of the Council.
DP	The meeting of the Shura Council is not considered valid unless at least two-thirds of its members are present. It includes the President or his deputy, and decisions shall not be considered valid unless approved by the majority of the Council.
TAP	A meeting of the Shura Council is not considered official unless at least two-thirds of its members are present. This includes the President or his deputy. Decisions are not considered official unless approved by a majority of the Council.
TCP1	The meeting of the Shura Council shall not be deemed valid unless attended by at least two-thirds of its members, including the Chairman or his deputy. Resolutions shall lack validity unless approved by the majority of Council members.
TCP2	Meetings held by the Shura Council shall not be valid unless attended by no less than two-thirds of its members, including the Chairman or his deputy. Resolutions shall not have legal standing unless they receive the approval of the majority of Council members.

DATA ANALYSIS

The methodology employed in this study consists of conducting a human evaluation using the MQM framework. The evaluation of translation quality in this study was conducted using the MQM framework to ensure a comprehensive and systematic analysis. The evaluation followed a structured approach where each error was first identified, categorised by type (e.g., accuracy, fluency, etc.), and then ranked according to its severity. We compared the machine-generated translations to the official reference translation to ensure consistency and reliability in the assessment process. This evaluation process allowed for a detailed analysis of the impact that different translation prompts have on the overall quality of the translations produced by GPT-4 Turbo.

The evaluation process was structured to ensure consistency and accuracy at each stage. Initially, every error in the translated output was meticulously identified and mapped to one of the primary MQM categories, namely accuracy, fluency, and terminology, with a further breakdown into subcategories when applicable. For example, accuracy errors were sub-classified into omissions, additions, or distortions, while fluency errors included grammatical inconsistencies, awkward phrasing, and stylistic deviations. Each identified error was then assessed for its severity—minor, major, or critical—depending on the extent to which it impacted the meaning or usability of the translation.

FINDINGS AND DISCUSSION

According to MQM, the errors detected in this paper in the GPT-4 Turbo output can be classified into three categories: (1) terminology, (2) accuracy, and (3) style. Terminology errors arise when terms do not align with standard terminology or when target terms do not accurately reflect their source text equivalents. Accuracy errors occur when the target text fails to convey the intended meaning of the source, often due to distortions, omissions, or additions to the message. Style errors

involve grammatically correct text that, nonetheless, deviates from organisational style guides or uses an inappropriate tone. Table 4 below shows the percentages and frequencies of each error for each prompt.

TABLE 4. Percentages and Frequencies of Each Error for Each Prompt

Prompt	MQM Error Rate	Terminology	Accuracy	Style
GP	45%	**18%**: *"Financial data"* for *" بياناتها المالية"* (should be "financial statements").	**16%**: *"Up to 5 million Saudi Riyals"* for *" لا تتجاوز خمسة ملايين"*(misses the definitive maximum limitation).	**11%**: *"Contracts must include all agreed terms"* for *" العقود المبرمة بين الأطراف...الشروط المتفق عليها"* (informal tone).
DP	38%	**15%**: *"International Financial Standards"* for *" المعايير الدولية لإعداد التقارير المالية"* (should be IFRS).	**14%**: *"Will be fined 5 million Riyals"* for *" تُفرض غرامة لا تتجاوز"*(misleading fixed interpretation).	**9%**: *"Contracts signed between parties must have clarity"* for *" يجب أن تتسم العقود بالوضوح والدقة"* (omitted precision).
TAP	34%	**14%**: *"Firms"* for *"الشركات"*(should use "companies" for formality).	**13%**: *"Environmental guidelines* "for *" الأنظمة البيئية"* (inappropriate term with less legal weight).	**7%**: *"All agreements must cover every condition agreed upon"* for *" العقود...على جميع الشروط"* (lacks specificity).
TCP1	12%	5%: *"Foreign companies must submit financial statements in accordance with IFRS"*(accurate).	7%: *"May be imposed on companies violating environmental regulations"*(captures legal nuance).	None observed.
TCP2	10%	None observed.	4%: *"A fine not exceeding five million Saudi Riyals"* (accurately reflects *" لا تتجاوز خمسة ملايين"*)	6%: *"Contracts concluded between the parties shall be characterised by clarity and precision..."* (formal tone).

The Jiao et al. (2023) general prompt exhibited notable deficiencies across all evaluated categories, underscoring its limited applicability to the specialised domain of legal translation. With a total error rate of 45%, the distribution was as follows: terminology (18%), accuracy (16%), and style (11%). A prominent example of a terminological error involved the phrase " يجب على الشركات الأجنبية تقديم بياناتها المالية وفقًا للمعايير الدولية لإعداد التقارير المالية (IFRS)" translated as "Foreign companies must submit their financial data according to international financial reporting standards (IFRS)." Here, the use of "financial data" for "بياناتها المالية" lacks the precision required to denote "financial statements," which is a term of art in accounting and legal contexts. Such terminological inaccuracies undermine the legal and regulatory validity of the translation. Accuracy errors were similarly prevalent, as seen in the translation of " تُفرض غرامة لا تتجاوز خمسة ملايين ريال سعودي على الشركات المخالفة للأنظمة البيئية" as "A fine of up to 5 million Saudi Riyals may be imposed on companies violating environmental regulations." The phrase "up to" failed to accurately reflect the legal limitation inherent in "لا تتجاوز," which specifies a definitive maximum amount. Additionally, stylistic shortcomings were evident, particularly in the translation of " يجب أن تتسم العقود المبرمة بين الأطراف بالوضوح والدقة وأن تشتمل على جميع الشروط المتفق عليها" as "Contracts between parties must be clear, precise, and include all agreed-upon terms." The lack of formality and insufficient emphasis on legal rigour highlight the inadequacy of this prompt for highly specialised texts such as Saudi legal documents.

The Gao et al. (2023) domain-specific prompt demonstrated moderate improvement, yet the error rate remained significant at 38%, distributed across terminology (15%), accuracy (14%), and style (9%). While the domain specificity of this prompt enabled marginally more appropriate word choices, substantial issues persisted. For instance, the translation of " يجب على الشركات الأجنبية تقديم بياناتها المالية وفقًا للمعايير الدولية لإعداد التقارير المالية (IFRS)" as "Foreign corporations are required to submit financial statements based on the International Financial Standards (IFRS)"* introduced a terminological error by rendering "المعايير الدولية لإعداد التقارير المالية" inaccurately as "International Financial Standards." This deviation from the internationally recognised "International Financial Reporting Standards" represents a critical oversight in the legal and financial context. Furthermore, accuracy issues arose in the translation of *" تُفرض غرامة لا تتجاوز خمسة ملايين ريال سعودي على الشركات المخالفة للأنظمة البيئية" as "Companies violating environmental laws will be fined five million Saudi Riyals." The fixed interpretation of the fine amount misconstrues the flexibility and conditionality expressed in "لا تتجاوز," which could mislead readers regarding the law's actual stipulations. Stylistic errors were also prominent, as in *"Contracts signed between parties must have clarity and include all agreed terms" for " يجب أن تتسم العقود المبرمة بين الأطراف بالوضوح والدقة وأن تشتمل على جميع الشروط المتفق عليها." The informal phrasing and omission of "precision" detracted from the professional tone essential in legal discourse, further indicating the prompt's inadequacy.

The Yamada (2023) target audience and purpose-oriented prompt exhibited a comparatively lower error rate of 34%, divided as follows: terminology (14%), accuracy (13%), and style (7%). However, it also struggled to meet the precision and formal tone required for translating Saudi Laws. For example, " يجب على الشركات الأجنبية تقديم بياناتها المالية وفقًا للمعايير الدولية لإعداد التقارير المالية (IFRS)" was rendered as "Foreign firms should provide their financial reports in compliance with IFRS standards." The phrase "firms" introduced a less formal tone, and the redundancy of "IFRS standards" undermined the translation's stylistic and terminological appropriateness. Similarly, in the translation of *" تُفرض غرامة لا تتجاوز خمسة ملايين ريال سعودي على الشركات المخالفة للأنظمة البيئية" as "A maximum fine of 5 million Riyals will be applied to companies breaching environmental guidelines," the term "guidelines" lacked the legal weight of "الأنظمة," thereby diminishing the translation's accuracy. Additionally, style errors, such as in *"All agreements must be written clearly, precisely, and cover every condition agreed upon by the parties" for "يجب أن تتسم العقود المبرمة بين الأطراف بالوضوح والدقة وأن تشتمل على جميع الشروط المتفق عليها," revealed inconsistencies in maintaining the requisite formal tone and specificity.

In contrast, the new prompts introduced by the researchers demonstrated markedly superior performance, achieving significantly lower error rates. TCP1, which was optimised for terminological consistency, recorded a total error rate of 12%, with terminology errors (5%) and accuracy errors (7%). For instance, " يجب على الشركات الأجنبية تقديم بياناتها المالية وفقًا للمعايير الدولية لإعداد التقارير المالية (IFRS)" was translated as "Foreign companies must submit their financial statements in accordance with the International Financial Reporting Standards (IFRS)." This translation exemplified precise terminology while preserving the formality required in legal texts. TCP2, designed to ensure stylistic and contextual fidelity, performed even better, with an error rate of 10%, comprising accuracy errors (4%) and style errors (6%). An example is " تُفرض غرامة لا تتجاوز خمسة ملايين ريال سعودي على الشركات المخالفة للأنظمة البيئية," translated as "A fine not exceeding five million Saudi Riyals may be imposed on companies violating environmental regulations."* This translation successfully conveyed the legal nuance of "لا تتجاوز" and maintained the formal tone necessary for the target audience. Stylistic issues were minimal, as illustrated in *"Contracts concluded between the parties shall be characterised by clarity and precision and shall include all

agreed-upon terms," accurately translated from "يجب أن تتسم العقود المبرمة بين الأطراف بالوضوح والدقة وأن تشتمل على جميع الشروط المتفق عليها."

These findings underscore the transformative potential of tailored prompt engineering in achieving high-quality legal translations. The reduced error rates of the new prompts—by more than half compared to existing prompts—highlight their capacity to meet the unique demands of translating Saudi legal documents. They indicate that the structured design of the translation commission prompts significantly reduced terminology and accuracy errors, demonstrating the effectiveness of prompt engineering in legal MT. These results also demonstrate the importance of developing translation prompts that are both domain-specific and aligned with the linguistic and stylistic conventions of the target text, ultimately enhancing the reliability of machine-assisted legal translation.

These findings underscore the transformative potential of tailored prompt engineering in achieving high-quality legal translations, particularly in the context of specialised domains such as Saudi legal texts. By leveraging domain-specific insights and considering the unique linguistic and stylistic nuances of the legal context, tailored prompts can significantly enhance the output quality of machine translation systems. The substantial reduction in error rates observed in the new prompts—more than a 50% improvement over existing prompts—demonstrates their effectiveness in addressing the challenges inherent in translating legal documents, which often contain complex terminologies, precise legal expressions, and culturally specific references.

The exceptional performance of the new prompts highlights their capacity to meet the unique demands of translating Saudi Laws. Saudi legal texts are characterised by a formal, highly structured style and the use of technical legal terminology that requires a nuanced understanding of both the source and target legal systems. The high error rates found in the existing prompts (ranging from 34% to 45%) clearly indicate that generic translation models or those designed for broader domains struggle to capture the intricacies of legal language. In contrast, the new prompts, specifically designed with the legal domain in mind, addressed these issues more effectively, reducing error rates to just 10-12%. This improvement suggests that tailored prompt engineering is critical in fine-tuning machine translation outputs to meet the expectations and requirements of highly specialised fields like legal translation.

Moreover, these results highlight the importance of developing translation prompts that are both domain-specific and aligned with the linguistic and stylistic conventions of the target text. Legal translations require more than just linguistic accuracy; they must reflect the formal, authoritative tone and the precise legal constructs inherent in the source material. By tailoring prompts to align with these specific needs, machine translation systems can produce outputs that not only preserve the meaning of the original text but also maintain the legal integrity and professionalism required in the target document. The new prompts demonstrated a significant improvement in terminological accuracy, fluency, and style, ensuring that the translations were not only linguistically accurate but also legally sound and contextually appropriate.

Ultimately, the results underscore the potential for enhanced reliability of machine-assisted legal translation when tailored prompts are utilised. The findings suggest that, for machine translation systems to be truly effective in legal contexts, it is not enough to rely on general models or generic prompts. Instead, focused efforts should be directed toward developing models that account for the specificities of the legal domain, including its terminology, syntax, and style. This not only ensures a higher degree of accuracy in translation but also builds trust in the use of machine translation in high-stakes, legally sensitive contexts, where precision and clarity are paramount.

Furthermore, the improvement in translation quality observed with the new prompts opens the door for more widespread use of machine translation in legal practice, potentially reducing translation costs, enhancing workflow efficiency, and allowing legal professionals to access and interact with foreign legal documents with greater ease. This development could also pave the way for future research in other specialised domains, highlighting the broader applicability of tailored prompt engineering in achieving high-quality machine translation across various fields. As machine learning and AI continue to evolve, this approach could become a cornerstone for improving the reliability and efficiency of automated translation in both the legal field and beyond.

CONCLUSION

This research investigated the influence of translation prompts, including existing translation prompts and newly developed ones incorporating the translation commission approach, on the quality of GPT-4 Turbo-generated legal translations. Employing the MQM framework, the study revealed distinct variations in the types and frequency of errors associated with different prompts, particularly in the categories of terminology, accuracy, and style. Findings indicate that translation prompts tailored with the translation commission approach (TCP1 and TCP2) outperformed existing translation prompts (GP, DP, and TAP), achieving significantly lower error rates and exhibiting enhanced terminological precision, contextual accuracy, and stylistic appropriateness. This study underscores the importance of incorporating the translation commission framework into prompt design, as it aligns machine-generated translations more closely with the specific functional and audience-oriented needs of the legal domain.

Despite its contributions, this research has several limitations. The study focused exclusively on legal translations of Saudi Laws, which may limit the generalisability of its findings to other legal systems or specialised fields. Moreover, the study concentrated solely on the Arabic-to-English language pair, leaving the applicability of its conclusions to other language combinations unexplored. Additionally, the dataset, while extensive, was constrained to 14 legislative documents. Furthermore, the reliance on MQM as the sole evaluation metric, although comprehensive, might not fully capture the subjective nuances of translation quality.

Future research could expand on this study by exploring the impact of translation commission prompts across various legal systems and diverse genres of specialised texts. Comparative studies involving other LLMs and different MT evaluation frameworks could provide deeper insights into prompt optimisation strategies. Moreover, integrating user feedback into the evaluation process could enhance the understanding of translation quality from the perspective of end-users, particularly in professional and regulatory contexts. Finally, exploring the use of adaptive prompts that dynamically adjust based on real-time analysis of source text complexity could pave the way for more intuitive and effective MT systems. This study demonstrates the transformative potential of integrating established translation theories with advanced AI technologies, offering invaluable directions for enhancing MT quality and advancing the intersection of the two fields.

ACKNOWLEDGEMENTS

This research received a grant no. (454/2024) from the Arab Observatory for Translation (an affiliate of ALECSO), which is supported by the Literature, Publishing & Translation Commission in Saudi Arabia.

REFERENCES

Agrawal, S., Zhou, C., Lewis, M., Zettlemoyer, L., & Ghazvininejad, M. (2022). *In-context examples selection for machine translation*. ArXiv. https://doi.org/10.48550/arXiv.2212.02437

Alkatheery, E. R. (2023). Google translate errors in legal texts: Machine translation quality assessment. *Arab World English Journal for Translation & Literary Studies*, *7*(1), 208-219. http://dx.doi.org/10.24093/awejtls/vol7no1.16

Al-Khalifa, H., Al-Khalefah, K., & Haroon, H. (2024). Error analysis of pretrained language models (PLMs) in English-to-Arabic machine translation. *Human-Centric Intelligent Systems*, *4*(2), 206-219.

Anesa, P., & Kulbicki, L. (2022). The impact of digitalization on legal communication: Introduction. *The International Journal of Law, Language & Discourse*, *10*(2), 5-8. https://doi.org/10.56498/1022022408

Banat, M., & Adla, Y. A. (2023). Exploring the effectiveness of GPT-3 in translating specialized religious text from Arabic to English: A comparative study with human translation. *Journal of Translation and Language Studies*, *4*(2), 1-23. https://doi.org/10.48185/jtls.v4i2.762

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). *Language models are few-shot learners*. ArXiv. https://doi.org/10.48550/arXiv.2005.14165

Burchardt, A., Lommel, A., & Macketanz, V. (2021). A new deal for translation quality. *Universal access in the information society*, *20*(4), 701-715. https://doi.org/10.1007/s10209-020-00736-5

Chatzikoumi, E. (2020). How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, *26*(2), 137-161. https://doi.org/10.1017/S1351324919000469

DeMattee, A. J., Gertler, N., Shibaike, T., & Bloodgood, E. A. (2022). Supplemental information for overcoming the laws-in-translation problem: Comparing techniques to translate legal texts. *Qualitative and Multi-Method Research, 20*(2), 13-21, https://doi.org/10.31219/osf.io/jc5p9

Drugan, J. (2013). *Quality in professional translation.* Bloomsbury.

Dunđer, I. (2020). Machine translation system for the industry domain and Croatian language. *Journal of Information and Organizational Sciences*, *44*(1), 33-50. https://doi.org/10.31341/jios.44.1.2

Ekin, S. (2023). *Prompt engineering for ChatGPT: A quick guide to techniques, tips, and best practices.* TechRxiv. https://doi.org/10.36227/techrxiv.22683919.v2

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, *9*, 1460-1474.

Gao, Y., Wang, R., & Hou, F. (2023). *How to design translation prompts: An empirical study*. ArXiv. https://doi.org/10.48550/arXiv.2304.02182

Graham, Y., Baldwin, T., Moffat, A., & Zobel, J. (2013). Continuous measurement scales in human evaluation of machine translation. *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse* (pp. 33-31). Association for Computational Linguistics.

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023). *How good are GPT models at machine translation? A comprehensive evaluation.* ArXiv. https://doi.org/10.48550/arXiv.2302.09210

House, J. (2001). Translation quality assessment: Linguistic description versus social evaluation. *Meta*, *46*(2), 243-257. https://doi.org/10.7202/003141ar

Islam, M. A., Anik, M. S. H., & Islam, A. B. M. A. A. (2021). Towards achieving a delicate blending between rule-based translator and neural machine translator. *Neural Computing Applications, 33*(18), 12141–12167. https://doi.org/10.1007/s00521-021-05895-x

Jiao, W., Wang, W., Huang, J. T., Wang, X., & Tu, Z. (2023). *Is ChatGPT a good translator? A preliminary study.* ArXiv. https://doi.org/10.48550/arXiv.2301.08745

Khoshafah, F. (2023). *ChatGPT for Arabic-English translation: Evaluating the accuracy*. Research Square. https://doi.org/10.21203/rs.3.rs-2814154/v1

Knap-Dlouhá, P. (2022). Machine translation: A possible solution to the law? *Brünner Beiträge zur Germanistik und Nordistik*, *1*(36), 35-46. https://doi.org/10.5817/BBGN2022-1-4

Kocmi, T., & Federmann, C. (2023). *GEMBA-MQM: Detecting translation quality error spans with GPT-4.* ArXiv. https://doi.org/10.48550/arXiv.2310.13988

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., & Chen, W. (2021). *What makes good in-context examples for GPT-3?* ArXiv. https://doi.org/10.48550/arXiv.2101.06804

Lommel, A. (2018). Metrics for translation quality assessment: A case for standardising error typologies. In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), *Translation quality assessment: From principle to practice* (pp. 109-127). Springer. https://doi.org/10.1007/978-3-319-91241-7_6

Lopez, A. (2008). Statistical machine translation. *ACM Computing Surveys (CSUR)*, *40*(3), 1–49.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., & Stenetorp, P. (2021). *Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity.* ArXiv. https://doi.org/10.48550/arXiv.2104.08786

Mariana, V. R., Cox, T., & Melby, A. (2015). The Multidimensional Quality Metric (MQM) framework: A new framework for translation quality assessment. *The Journal of Specialised Translation,* (23), 137-161.

Mohamed, Y. A., Khanan, A., Bashir, M., Mohamed, A. H. H., Adiel, M. A., & Elsadig, M. A. (2024). The impact of artificial intelligence on language translation: a review. *IEEE Access*, *12*, 25553-25579.

Mossop, B. (2001). *Revising and editing for translators*. St Jerome.

Munday, J., Pinto, S. R., & Blakesley, J. (2022). *Introducing translation studies: Theories and applications*. Routledge. https://doi.org/10.4324/9780429352461

Nida, E. A. (1977). The nature of dynamic equivalence in translating. *Babel: International Journal of Translation*, (13), 99-103.

OpenAI. (2023). *GPT-4* (Jun 13 version) [Large language model]. https://openai.com/research/gpt-4

Royal Court. (n.d.-a). *Laws and Regulations.* National Centre for Archives and Records. https://ncar.gov.sa/rules-regulations

Royal Court. (n.d.-b). *Saudi Laws.* The Bureau of Experts at the Council of Ministers. https://laws.boe.gov.sa/BoeLaws/Laws/Folders/2

Shin, T., Razeghi, Y., Logan, R. L., IV., Wallace, E., & Singh, S. (2020). *Autoprompt: Eliciting knowledge from language models with automatically generated prompts.* ArXiv. https://doi.org/10.48550/arXiv.2010.15980

Soriano Barabino, G. (2020). Cultural, textual and linguistic aspects of legal translation: A model of text analysis for training legal translators. *International Journal of Legal Discourse*, *5*(2), 285-300. https://doi.org/10.1515/ijld-2020-2037

Sosoni, V., O'Shea, J., & Stasimioti, M. (2022). Translating law: A comparison of human and post-edited translations from Greek to English. *Journal of Language and Law*, *78*, 92-120. https://doi.org/10.2436/rld.i78.2022.3704

Sulaiman, M. Z., Zainudin, I. S., & Haroon, H. (2025). Can ChatGPT translate like a pro? A pilot benchmarking study of English–Malay translation quality. *3L: Language, Linguistics, Literature®, 31*(4), 259–278. https://doi.org/10.17576/3L-2025-3104-17

Vermeer, H. (2021). Skopos and commission in translational action (A. Chesterman, Trans.). In L. Venuti (Ed.), *The translation studies reader* (pp. 219-230). Routledge. (Original work published 1989). https://doi.org/10.4324/9780429280641

Vilar, D., Freitag, M., Cherry, C., Luo, J., Ratnakar, V., & Foster, G. (2023). Prompting palm for translation: Assessing strategies and performance. ArXiv. https://doi.org/10.48550/arXiv.2211.09102

Wang, J., Liu, Z., Zhao, L., Wu, Z., Ma, C., Yu, S., Dai, H., Yang, Q., Liu, Y., Zhang, S., Shi, E., Pan, Y., Zhang, T., Zhu, D., Li, X., Jiang, X., Ge, B., Yuan, Y., Shen, D., … Zhang, S. (2023). *Review of large vision models and visual prompt engineering.* ArXiv. https://doi.org/10.48550/arXiv.2307.00855

Webson, A., & Pavlick, E. (2021). *Do prompt-based models really understand the meaning of their prompts?* ArXiv. https://doi.org/10.48550/arXiv.2109.01247

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A prompt pattern catalog to enhance prompt engineering with ChatGPT.* ArXiv. https://doi.org/10.48550/arXiv.2302.11382

Yamada, M. (2023). *Optimising machine translation through prompt engineering: An investigation into ChatGPT's customizability.* ArXiv. https://doi.org/10.48550/arXiv.2308.01391

Zhang, B., Haddow, B., & Birch, A. (2023). *Prompting large language model for machine translation: A case study.* ArXiv. https://doi.org/10.48550/arXiv.2301.07069