International Journal of Translation and Interpretation Studies
ISSN: 2754-2602
DOI: 10.32996/ijtis
Journal Homepage: www.al-kindipublisher.com/index.php/ijtis

Assessing AI-Driven Translation of Arabic Legal Texts: Developing a Quality Evaluation Framework through Comparative Human–Machine Analysis

Hussein Al-Bo'ul[1], **Zakaryia Almahasees**[2], **Islam Husienat** ✉[3]

[1] *Ministry of Education, Directorate of Ajloun Education, Ajloun, Jordan*

[2] *Department of Translation, Yarmouk University, Irbid, Jordan*

[3] *Janai translation Agency, Perth, Australia*

Corresponding Author: Islam Husienat**, E-mail:** motransperth@outlook.com

ABSTRACT

The research aims to develop a comprehensive quality assessment framework for evaluating AI-assisted translation of Arabic into English in the domain of law through a comparative human-machine analysis. Building on legal translation theory, corpus-based approaches, and Multidimensional Quality Metrics (MQM), the research identifies six dimensions of assessment: legal conceptual equivalence, terminological correctness and consistency, semantic correctness, functional and pragmatic adequacy, linguistic quality, and jurisdictional and cultural adequacy. In terms of methodology, the research uses qualitative comparison of representative legal texts in Arabic and their potential problematic machine-generated translations and contextually sensitive human-made alternatives. Seven types of translation problems have been analyzed: conceptual inaccuracy, terminological inconsistency, omission/addition, legal modality, culturally-specific terms, conditional provisions, and legal register. The findings of the study show that acceptable linguistically translations can distort legal rights, obligations, and institutions. This research argues that linguistic evaluation alone cannot be used to evaluate the quality of legal translation and suggests the integration of an error taxonomy based on severity and expert legal judgment. Future research directions include empirical research using actual machine-generated translations, larger multilingual legal corpora, expert evaluations, and interrater reliability tests.

KEYWORDS

Artificial Intelligence (AI); Arabic–English Legal Translation; Translation Quality Assessment; Legal Conceptual Equivalence; Human–Machine Translation; Multidimensional Quality Metrics (MQM).

ARTICLE INFORMATION

ACCEPTED: 05 November 2026 **PUBLISHED:** 14 November 2026 **DOI:** 10.32996/ijtis.2026.6.6.4

1. Introduction

The development of artificial intelligence has significantly impacted the field of professional translation, raising serious concerns regarding linguistic proficiency, translation accuracy, and the significance of human skills in specialized communication. Neural Machine Translation and, more recently, Large Language Models have made it possible for computers to translate languages and specializations in ways that are fluent and context-appropriate. Transformers have played an important role in these advances, allowing for better linguistic relations (Vaswani et al., 2017). While these innovations have improved translation practices through increased availability and efficiency, they have raised a number of concerns as to the validity of AI-generated texts in those situations where any linguistic error may result in considerable professional, organizational, or even legal problems. One such area is legal translation, in which translation quality cannot be measured only through grammatical and lexical accuracy, but must also take into account the correct translation of legal meanings, concepts, and functions (Šarčević, 1997; Cao, 2007).

Legal translation is a particular kind of communication where language serves as a means for conveying information, as well as the instrument for creation, regulation and interpretation of legal relations. The linguistic structures found in legislative acts, terms of contracts, court decisions and regulations define the rights and obligations, exceptions and consequences of particular legal rules. As a result, small changes in modality, negation, conditionality, terminology can significantly affect the interpretation of a translated rule. For example, rendering an obligatory duty in terms of discretion may change the nature of a legal statement, and omission of an exception will turn a conditional ban into an absolute one. This makes legal translation different from the other kinds of multilingual communication and requires the translators to combine their linguistic expertise with legal knowledge and skills of context analysis. As noted by Šarčević (1997) and Prieto Ramos (2015), the quality of legal translation largely depends on

Copyright: © 2026 the Author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) 4.0 license (https://creativecommons.org/licenses/by/4.0/). Published by Al-Kindi Centre for Research and Development, London, United Kingdom.

preservation of legal meaning and ability to solve the problems of translation associated with the differences between legal systems.

The above-mentioned issues are particularly applicable in Arabic–English legal translation since they involve not only different languages but also different legal traditions, institutions, and culturally-based concepts. Arabic legal discourse is based on different aspects, such as Islamic law, civil law, national laws, and regulatory systems of different jurisdictions. Thus, legal terms which might be linguistically simple to understand might actually refer to some specific concepts that are not necessarily included in English legal lexicon. Such terms as الشخص الاعتباري (legal entity), الولي الشرعي (legally authorized guardian), المهر (deferred mahr), and الطلاق البائن (irrevocable divorce) demonstrate that translating legal terms into English means not finding their equivalents in the dictionary but understanding the legal relations, functions, and doctrine that the corresponding Arabic terms reflect. Besides, Arabic-speaking jurisdictions cannot be considered as a single unified legal system, so the meaning or usage of certain legal term might differ from one country to another. This means that conceptual equivalence of Arabic to English legal translation should be evaluated considering the specific source jurisdiction and the purpose of the target text (Cao, 2007; Prieto Ramos, 2015).

An increase in the use of AI-based translation technologies only increases the difficulties in the already challenging environment for translation activities. While the traditional approach implies conscious review of legal terminology and removal of conceptual ambiguities by professional translators, an AI technology does not give any guarantee that the process of translation will include transparent reasoning related to the legal issues involved. While it is possible that an AI system produces fluently written and stylistically convincing legal texts, the results of such work may contain inaccuracies which are difficult to detect at the surface level of linguistic assessment. Such inaccuracies might include misuse of legal terms, lack of consistency in terminology, omissions of legally relevant conditions, unnecessary additions, and incorrect translation of obligations and exceptions. In this context, the specific problem of LLM translation should be highlighted since the process of generation may allow for fluent translation of the target language text while at the same time including some information which is not justified by the source text. However, it is also inappropriate to assume that the human-made translations are free of such mistakes. Both types of translations should be assessed using certain quality criteria. The main issue here is not only the possibility of producing a fluent legal text by AI but whether this text keeps the meaning, concept relations, and communicative functions of the source text.

The issue of evaluating AI-generated legal translations brings a variety of methodological difficulties which are not related only to the problems inherent to translation generation processes. Automatic evaluation techniques, in particular reference-based metrics like BLEU, have been created to determine how close the machine translations are to the reference texts done by humans (Papineni et al., 2002). Even though automatic metrics provide some kind of standardized quantitative measures, they rely on the lexical equivalence which is unable to evaluate the whole spectrum of translation qualities. It can happen that the machine translation has many similarities with a reference translation but lacks some critical negations or changes the structure of conditionality in the text. The correct legal translation may use other wording that would make it get a lower similarity rate. Such difficulties are especially relevant in legal translation because the significance of any mistake depends on its legal consequences, not on its lexical nature.

Scholarly literature about the assessment of the quality of translations serves as a basis for both theoretical and methodological frameworks related to the aforementioned problems. According to House (2015), the focus on translation quality assessment relies on comparing systematically the sources and targets in linguistic, pragmatic, and contextual aspects. Simultaneously, the Multidimensional Quality Metrics (MQM) system offers a classification of translation errors according to such aspects as accuracy, terminology, fluency, among others (Lommel et al., 2014). Despite the fact that the above frameworks provide an important basis for assessing the quality of translations, their application to legal translations demands the incorporation of particular considerations associated with legal conceptual equivalence, appropriateness from the perspective of jurisdictions, and the consequences of incorrect choices during the translation process. According to Prieto Ramos (2015), it is essential for legal translation quality assurance to encompass process factors, specialized knowledge, and product assessment. This means that the quality of legal translations requires a combination of linguistic and non-linguistic aspects.

Corpus-based translation studies offer another methodological approach to the analysis of these questions. After Baker's (1993) pioneering use of corpus linguistics in the analysis of translation, the use of corpus-orientated methodologies has made it possible to conduct systematic analyses of the translated language, typical linguistic patterns, and discrepancies between source and target languages. In the case of legal translation, the use of corpus analysis techniques proved especially helpful in the analysis of specialized terminology, traditional phraseology, lexical variation, and textual structuring patterns. As stated by Biel (2014), the importance of using corpus analysis for the description of the features of legal translated discourse and the connection between linguistic conventions and legal communication cannot be overstated. Using corpus-based analysis techniques in the analysis of human and AI-generated translations makes it possible to detect typical terminological replacement, discrepancies in the translation of legal terms, as well as the omission or transformation of various parts of texts. However, using only corpus-based measures is not enough to assess whether the translation correctly conveys the idea behind the legal concept. A frequently used term might be conceptually incorrect while an infrequent one might better reflect the idea of the law.

Even though there have been significant achievements in legal translation scholarship, AI translation and multidimensional quality evaluation of translation, there is a crucial methodological question to be considered – how to incorporate all the aforementioned perspectives into one process that will make it possible to conduct a systematic evaluation of Arabic–English legal translation. First, general frameworks on the quality of translations offer important classification of linguistic and semantic mistakes to be found,

while corpus-based approaches give quantitative information about translation practices. Furthermore, theories of legal translation emphasize the necessity to pay attention to such aspects as conceptual equivalence, function, and jurisdiction. Nevertheless, all these approaches should be applied in practice, which means that it is necessary to develop an approach in which linguistic differences can be differentiated from the mistakes that change the meaning of translation in terms of rights, obligations, and legal institutions. This issue becomes particularly important in case of comparison of translations done by humans and machines, as the difference in fluency, terminology, and legality may not correspond to the same patterns.

A further point relates to the handling of error severity. In current practices of translation evaluation, error frequency or quality score is often indicated. The quantity of errors, however, does not indicate their legal significance. There can be many errors in terms of grammar and style that do not play a big role in understanding a legal document but one incorrectly translated negation, conditional clause or modal verb can completely change the meaning of the provision. Thus, the framework for assessing legal translation needs to take into account both the number of errors and the possible effects they cause. For this, severity criteria need to be developed which will include the role of the translated legal expression, the degree of semantic deformation and the likelihood that such a translation will give a misleading meaning. This will be consistent with the multidimensionality of MQM criteria (Lommel et al., 2014), but its application for the analysis of legal translation will make it more versatile. Moreover, such an approach will be needed to solve a specific problem of AI-assisted translation – to determine whether a smooth translation contains hidden mistakes which should be corrected by experts to make the document legally correct.

Against this backdrop, the proposed study will aim to develop a multidimensional framework for assessing the quality of English translations of Arabic legal texts generated via artificial intelligence technology, with reference to a human-machine comparative assessment. In a corpus-based approach, the research is expected to conduct an analysis of a large number of Arabic legislative, contractual, and regulatory texts translated into English by professional human translators and chosen AI models. The framework to be developed will include quantitative analysis of the corpus, categorization of translation mistakes, assessment of legal conceptual equivalence, and severity-oriented evaluation of significant inaccuracies with regard to law. The framework will consist of six main dimensions: legal conceptual equivalence, terminological precision and consistency, semantic correctness, functional and pragmatic adequacy, linguistic quality, and jurisdictional and cultural adequacy.

The study focuses on answering three main research questions, which are: (1) How does machine-based translation ensure conceptual fidelity and terminology consistency in translating Arabic legal documents into English compared to professional human translation? (2) What kinds of linguistic, terminological, semantic, and doctrinal mistakes appear in machine-based translations of Arabic legal discourse? And (3) In what ways can corpus-based comparative analysis be utilized in developing a standardized quality evaluation framework for machine-assisted translation between Arabic and English legal texts? Such questions lead to a linear progression that starts with a comparative analysis and leads to error recognition and framework formulation. This research also includes examples of legal translation to show how the proposed categories are used in dealing with conceptual equivalence, legal modality, additions and omissions, culturally-specific terminology, conditionals, and legal register problems. The provided examples play an interpretive role rather than empirical one.

It makes its contribution in three main ways. From the theoretical perspective, the study makes a contribution to the ongoing discussions about the quality of legal translation by considering not only linguistic accuracy but also conceptual equivalence and legal adequacy of AI-based translations. From the methodological point of view, it offers an integrated approach to evaluating specialized translations that integrates the corpus linguistics metrics with legal analysis and severity of errors classification to fill the gap of existing methodologies that evaluate the quality of translation either from the linguistic perspective or based on the reference corpora. Finally, from the practical perspective, the study is relevant for quality assurance processes of legal translators, translation agencies, legal institutions, and translation educators who aim to introduce AI into their translation process.

2. Literature Review

2.1 Legal Translation, Conceptual Equivalence, and Legal-System Differences

Legal translation has always been regarded as a kind of specialized communication process, whereby linguistic meaning is inherently associated with legal systems and contexts of operation of the texts. The theoretical basis was established by Šarčević (1997), who showed that legal translation is a specific form of communication that involves functional and conceptual correspondence rather than mechanical translation of linguistic forms from one language into another. Cao (2007) has also emphasized that legal translation involves communication of language, law and culture, and in many cases the translator acts as a mediator who has to reconcile discrepancies between institutions that do not have any equivalents in the source language system. The approaches are united in their criticism of linguistic equivalence as the only possible measure of legal translation quality. However, there is an important methodological problem that emerges from these theories – conceptual equivalence is situation-bound and can never be simply translated into linguistic equivalences of legal terms. It becomes especially relevant in case of evaluation of AI-generated legal translations.

In light of this theoretical framework, Prieto Ramos (2015) proposes an integrated model for legal translation quality assurance that connects translation competence, the process of decision making, and evaluation of translated output. This author extends previous research on equivalence by emphasizing the methodological aspects of identifying and solving legal translation issues. The interaction between these theories is especially important from the standpoint of using AI technology for translation quality assessment. Although Šarčević (1997) and Cao (2007) define the conceptual and practical demands of legal translation, Prieto Ramos (2015) provides a perspective of quality assurance which makes these demands operational. However, none of these studies,

separately, provides the criteria for assessing the performance of AI models of today in comparison with human professionals in legal translation between Arabic and English languages. Therefore, the current research takes advantage of the common focus on legal meaning of these theories, trying to translate it into assessable quality criteria.

2.2 Arabic–English Legal Translation and Jurisdictional Specificity

The Arabic-English language combination represents specific difficulties in the process of legal translation due to the nature of concepts which have to be transferred from one legal tradition to another, each having its own historical, ideological and institutional background. Arabic law may contain elements of Islamic jurisprudence, civil law systems, national legislation and administrative practice, while the English law may involve the terms and notions typical of common law institutions. In this respect, the theoretical approaches proposed by Cao (2007) and Šarčević (1997) are especially valuable since they help understand why legal equivalence is impossible without taking into consideration the jurisdiction of the law. Terms related to guardianship, legal personality, contract law, personal status and procedure of judicial proceedings may require different English equivalents depending on the specific legal system and the aim of the translation.

Another implication from the reviewed literature is that the Arabic legal translation cannot be measured using a uniform terminology list or an uncontextualized set of sentence pairs. The meaning of laws differs from one Arabic-speaking jurisdiction to another. While a certain English term will work well in a particular jurisdiction, the same term can create confusion in another. In the context of AI system evaluation, the fact that the translation has produced the commonly used English equivalent does not prove that the term is fit for use in the legal provision in question. As such, the literature recommends a jurisdiction-specific evaluation method, which takes into account source text information, genre, target readership, and linguistic quality of the translation. However, how well different AI systems deal with the problem remains a question to be answered empirically.

2.3 Corpus-Based Translation Studies and Legal Terminological Consistency

The corpus-based approaches to translation studies provide a methodological framework for studying linguistic repetition and systemic differences between translations. Baker (1993) illustrated how corpus linguistics could revolutionize translation studies with an empirical study of the language of translation. This methodology shifted attention from singular examples of translation to repetitions observed in a larger corpus. As far as the legal translation is concerned, Biel (2014) applied the method of corpus study to the properties of legal texts translated into other languages, thus emphasizing the importance of the phraseological nature of the translations and the link between legal translations and linguistic norms of the genre. Baker and Biel are, therefore, complementary: the former offers a methodological framework, whereas the latter applies it to the special case of legal translation.

Nonetheless, the extant literature highlights a drawback of applying a corpus-based approach to legal translation quality assessment. It is possible to detect recurring equivalences and variation using frequency and concordance analysis. However, these types of analysis cannot, by themselves, assess the accuracy of the legal meaning rendered in the target text. In other words, an inaccurate legal term may recur in the text, whereas the correct term may recur infrequently owing to its jurisdictional specificity. Hence, there needs to be a clear distinction made between terminology consistency and accuracy. This is particularly relevant when comparing human and machine translation because different systems may display varying levels of terminological consistency while being equally conceptually accurate. The current research project will address this problem through the combination of corpus-based metrics and expert legal evaluation, allowing quantitative data to be interpreted from the perspective of concepts and functions conveyed in the source texts in Arabic.

2.4 Neural Machine Translation and the Emergence of Large Language Models

The advent of neural machine translation was indicative of a significant shift from statistical approaches to frameworks able to handle complex linguistic relationships. The Transformer framework by Vaswani et al. (2017) represented a key milestone in this trend and became the basis for further developments in the field of neural machine translation and large language models. Recently, Hendy et al. (2023) have conducted a thorough evaluation of translation with GPT models for 18 language directions including translation quality, prompting strategies, domain adaptation, and document translation. The findings reveal that GPT models can deliver competitive performance in high-resource language settings, but their performance in low-resource languages is more limited. At the same time, there might be benefits from using hybrid approaches that involve the integration of GPT models with other systems for translation. Altogether, it could be concluded that translation performance depends on linguistic resource availability, model properties, and other factors rather than being inherent to any particular artificial intelligence framework.

The implications of the above findings for legal translation need careful consideration. Hendy et al. (2023) provide empirical proof concerning the performance of general machine translation but cannot say anything about whether one can expect a similar level of performance in the case of jurisdiction-specific Arabic legal documents. Legal translation requires certain limitations that go beyond the linguistic and situational factors considered in the benchmarks for general translation. A translation may turn out to be accurate but choose a wrong legal equivalent or change the nature of the contractual obligation. Thus, proof of good general translation performance does not equal legal translation accuracy. This difference means that the comparison of human translators, NMTs, and LLMs must be carried out using identical texts and identical instructions and criteria of translation quality that take into account the legal consequences of mistakes in translation. The current study abides by this rule by testing the accuracy of the translated concept and its legal viability along with linguistic quality.

2.5 AI-Based Translation Quality Evaluation and the Limits of Automatic Metrics

Evaluation of the machine translation process has traditionally been performed by means of automatic measures used to assess similarity between machine-generated translations and their human reference equivalents. Papineni et al. (2002) suggested the

application of BLEU as a tool for automated evaluation that is based on n-gram precision and can serve as an effective way of comparing various translation engines. However, there are specific issues related to the application of lexical reference-based metrics in the domain of specialized legal texts. Even if the translation preserves most of the text's content, an exception, a negation, or even just a modal verb can be missing or misused which may result in severe consequences regardless of the limited lexical dissimilarity between the translations. In the meantime, legally correct translations may still differ from references because of using conventional alternatives to certain constructions.

There have been some attempts to test the capabilities of large language models (LLMs) as translators' performance evaluators. For instance, Kocmi and Federmann (2023) proposed the model GEMBA, which is based on GPT and evaluates translation quality by means of using or not using reference translations. As a result of the research, the authors conclude about the high correlation between the system-level agreement of the results and human evaluations of the quality in the MQM framework in the language pairs evaluated within the WMT22 metrics task. The importance of this finding consists in the fact that, probably, generative AI can help not only in translation but in its quality assessment as well. Nevertheless, the described research results refer to certain language pairs and evaluation settings and do not prove the same reliability of AI translation quality evaluation in case of Arabic–English legal translation and detection of legally important errors. Therefore, using the LLM as both translator and evaluator poses the problem of the lack of independence of the results and replicability of similar mistakes.

2.6 Translation Quality Assessment and the Need for Legal Error Severity

Quality assessment of translation research provides a useful basis for the establishment of procedures designed for the identification and assessment of translation errors. In this regard, the importance of taking into account contextual and functional aspects in assessing translation quality is emphasized by House (2015). In turn, Lommel et al. (2014) propose Multidimensional Quality Metrics (MQM) as an approach allowing for classification of errors according to their dimensions, such as accuracy, terminology and fluency. While these two approaches provide different angles on the subject under discussion, with House focusing on the context of translation and MQM providing a methodology for the classification of errors, it should be noted that legal translation requires paying particular attention to the implications of the translation errors. Indeed, a grammar mistake might decrease readability without changing the legal meaning of the text, while a mistake related to negation, liability or contractual conditions might affect the legal relationship reflected in the source text.

One of the most substantive contributions to resolving the problem is made by Prieto Ramos (2015), who proposes a holistic model for the quality assurance in legal translation. The model incorporates legal, contextual, macrotextual and microtextual dimensions with the translation process itself as well as its evaluation. Importantly, the model presents legal translation adequacy as a special type of decision making and highlights the professional competence of those who conduct this evaluation. The relevance of such an approach in case of evaluation of AI translation results is obvious insofar as it questions the idea that linguistic appropriateness alone is enough for translating a document adequately without regard for the relevant legal context. The comparison of MQM with the model proposed by Prieto Ramos suggests the possibility of merging structured types of errors and their severity assessment in light of legal considerations.

2.7 Empirical Studies on AI-Driven Arabic Legal Translation

Current empirical research has shed light on some of the advantages and disadvantages of AI translation in relation to Arabic legal translation. Ait ElFqih and Monti (2024) tested the abilities of GPT-4 and Gemini to translate legislative texts from Arabic into English and French while paying attention to the issue of accuracy of terms and the influence of context. The research used zero-shot prompting in both contextual and non-contextual settings, used the dataset approved by professionals as well as MQM error categorization. It has shown some flaws in the translation of specialized terms and consistency issues, despite the general multilingual capacities of the models. This research is relevant due to its focus on Arabic legal translation as well as its methodological approach based on combining domain-specific reference data with error classification. Nevertheless, being focused only on the particular models and legislative texts, it may benefit from further research which would consider other legal genres and human benchmarks.

Other research has further pursued this question via a comparison of human and AI legal translations. For example, Altakhaineh et al. (2025) evaluated a corpus containing around 20,000 words of legal documents translated into Arabic by humans and by ChatGPT. The assessment criteria were legal terminology, clarity, and conformity with the Arabic legal system. The research has revealed the superiority of human translators with regard to the studied criteria, although the possible efficiency of AI in simpler tasks was noted. In a similar manner, Abdelaal and Al Sawi (2025) compared human translations with those produced by ChatGPT and Google Translate, using 16 legal texts as a corpus. The research showed that the human translation had some advantages in terms of terminology and style. However, despite the usefulness of both studies, their results have to be viewed in context of the language directions, sizes of corpora, and evaluation techniques used. It is especially important not to generalize the findings from the English to Arabic translation to Arabic to English one due to the unique difficulties related to the direction.

New studies have found that the performance of the AI may be influenced by variables such as the type of legal text being translated, cultural complexity of the text being translated, and the criteria of evaluation used. Merabet and Roser Nebot (2026) performed a comparative evaluation of the AI models: ChatGPT, DeepL, and Google Translate on Algerian Supreme Court decisions using accuracy, fluency, clarity, and cultural appropriateness as the criteria. While the results confirmed the strong performance of ChatGPT in terms of translating cultural nuances of the selected texts, it reinforced the continued relevance of human knowledge. The findings do not conflict with the studies showing the superiority of human performance but rather imply that depending on

the systems, texts, and criteria being compared, different outcomes will emerge. Overall, the results of these studies suggest a developing empirical consensus that AI can be used for legal translation but there are variations in the performance.

2.8 Human Expertise, Evaluation Reliability, and Hybrid Translation Workflows

The significance of human expertise in the process of evaluating machine translations has become more pronounced as the fluency of such systems improved. Freitag et al. (2021) carried out a large-scale study of human evaluation of machine translations based on professional translators, document-level context, and MQM-based error annotations. The results have shown that the methodology and competence of the evaluator can have a major effect on the ranking of systems, as the expert-based assessment can yield different results than that carried out by crowd workers. This paper offers a valuable methodological insight for the research of legal translation: the reliability of the comparison does not depend only on the translation products themselves but also on the competence of evaluators and criteria used to evaluate them. This is especially true for legal translation because detecting the conceptual errors can require the understanding of comparative law and the purpose of the document.

Expert evaluation is also emphasized in more recent researches proposing cooperation of humans and machines in AI-powered legal translations. Both Altakhaineh et al. (2025) and Abdelaal and Al Sawi (2025) mention some possible advantages of integration of automated translation with human expertise, especially with regard to human revision of the initial draft produced by AI systems and requiring terminological and contextual editing. Nevertheless, one cannot simply assume that collaboration of AI systems and professional translators will ensure high-quality result since post-editing quality is directly dependent on the expertise level of a person who performs revision, time allowed for the process, transparency of the context, and detection of the error hidden behind fluent language. Therefore, any future comparison of human, machine, and post-edited machine translation should be conducted with consideration of these three distinct options.

2.9 Critical Synthesis and Research Gap

The above-reviewed literature identifies three connected trends. The first is that studies in legal translation prove that translation quality should be based on conceptual equivalence, functionality of translation in the context of law, and jurisdiction-specific interpretation rather than mere linguistic match (Šarčević, 1997; Cao, 2007; Prieto Ramos, 2015). The second trend is that the progress in NMT and LLM technologies has enhanced the fluency and flexibility of machine translation but still, its performance varies depending on language pairs, domain and contextual factors (Hendy et al., 2023; Jiao et al., 2023). The third trend is that the studies on translation quality evaluation show the importance of error analysis, contextual information and professional expertise for obtaining valid evaluation (Lommel et al., 2014; Freitag et al., 2021; Kocmi & Federmann, 2023). These trends offer complementary perspectives to evaluate machine-translated legal texts but at the same time reveal the complexity of their integration.

The most relevant Arabic legal translation studies provide support to the significance of the proposed integration while showing empirically that the domain is not unexplored anymore. Indeed, Ait ElFqih and Monti (2024) have already looked into contextual and terminological performance in Arabic legislative translation through LLMs and MQM. Altakhaineh et al. (2025) and Abdelaal and Al Sawi (2025) have compared human and AI performance in the legal context, whereas Merabet and Roser Nebot (2026) have compared several machine translation systems on the basis of Arabic judicial texts. A wider literature review by Bajčić and Golenko (2024) also pointed out the necessity for empirical research on interdisciplinary nature of the problem under question. Therefore, the research gap must not be described as an absence of comparative studies and approaches to the classification of errors altogether. The research gap consists in the necessity of validation of an integrated framework including legal conceptual equivalence, terminological consistency, semantic accuracy, functional adequacy, and jurisdictional appropriateness.

The suggested research fills this gap by merging corpus-based comparative analysis with a multidimensional and severity-sensitive legal translation evaluation framework. The potential contribution of the research is distinguished from research focusing on comparing translation accuracy and fluency in general because the aim is to provide a means of defining the legal importance of translation errors and distinguish linguistic mistakes from those that influence the rights, duties, conditions and institutions. This framework allows expert validation, inter-rater reliability estimation, and comparative analysis of translation outputs both generated by humans and machines. Integrating the theory of legal translation and the machine translation evaluation into one methodology will lead to a more transparent and replicable approach to assessing the quality of Arabic–English legal translations. However, its contribution will become clear after empirical implementation and verification whether the suggested dimensions and severity criteria are applicable and informative.

3. Methodology

The current research utilizes a comparative mixed-methods research methodology which consists of a combination of quantitative corpus-based and qualitative legal conceptual evaluation of the quality of AI-generated English translations of Arabic legal texts against the quality of translations produced by professional human translators. The quantitative part looks into various quantifiable linguistic and translational factors such as terminology consistency, lexical density, and the number and distribution of translation errors. The qualitative part considers concepts equivalence, legal doctrine accuracy, and legal implications of certain translation decisions. Methodological underpinnings of the research include descriptive translation studies, corpus-based translation studies (Baker, 1993; Biel, 2014), and traditional models for assessing quality of legal translations (Šarčević, 1997; Prieto Ramos, 2015). In addition, the research employs multidimensional approach to classifying translation errors typical of Multidimensional Quality Metrics (Lommel et al., 2014).

3.1 Research Sample

The research sample consists of a purposively selected corpus of Arabic legal texts that fall within three main categories of legal discourse: legislative texts, commercial contracts, and regulatory texts. The corpus includes roughly 5,000-10,000 source words in Arabic, varying in terms of complexity and special terminology. These genres are chosen to guarantee variety of legal discourses, especially in terms of legal rights, obligations, prohibitions, conditions, and procedures. Each Arabic source text is translated into English by selected AI translation software programs and large language models, as well as by professional human translators specializing in Arabic-English legal translation. Therefore, the research dataset includes three parts, each consisting of one text: Arabic source texts, English AI-translated texts, and English professional human-translated texts. The selection of the text is guided by certain criteria concerning the genre of the legal text, jurisdiction of the text, terminological complexity, and the form of the text to ensure comparability among the different translation products. The same source text is selected for both conditions of translation.

3.2 Research Instruments

In this study, three research tools are used to measure linguistic accuracy, terminological consistency, and legal adequacy of the translations. First, frequency lists, keyword distribution, collocations, and lexico-grammatical measures, including lexical density, are produced using corpus analysis software, such as Sketch Engine, thus allowing to make systematic comparisons of linguistic parameters of the AI-generated and human translations. Second, a taxonomy of translation errors based on the Multidimensional Quality Metrics framework (Lommel et al., 2014) identifies six types of translation problems: terminological inconsistency, conceptual distortion, omission, addition, ambiguity, and stylistic deviation. Every detected problem is additionally rated according to its severity and legal significance. Third, a legal adequacy evaluation grid evaluates the presence and degree of transfer of legal concepts, legal doctrines, terminology, and legal relations present in the source text. Evaluations of legal adequacy are made by experts trained in the field of legal translation and comparative law following the criteria described in Prieto Ramos (2015) on legal translation quality assurance. However, automatic translation evaluation metrics can be used as additional indicators of translation adequacy.

3.3 Research Procedure

The research procedure is comprised of five interrelated steps, which should guarantee systematic data collection, translation comparability and quality assessment reliability. At the first step, Arabic legal texts have been gathered, sorted according to their genre and jurisdiction and prepared for corpus analysis through text cleansing, formatting and segmentation. At the second step, the selected AI systems perform translation of the source texts into English according to standardized translation instructions and documented settings of the systems. The third step implies performing independent human translation of the same Arabic texts into English under similar task instructions and intended target audience and legal context without seeing the output of the AI translation. The fourth step includes alignment of the Arabic source texts and AI and human translations according to sentences and legal clauses. Finally, the Arabic source texts and their AI and human translations have been independently assessed by experts through a structured error taxonomy and legal adequacy grid, during which translation errors will be identified, their severity will be rated, and legal meaning will be evaluated. When two independent raters assess the same categorical data, Cohen's kappa have been used to estimate inter-rater reliability. Any disagreement will be discussed in order to reach consensus.

3.4 Data Analysis

The data analysis includes both quantitative corpus-based linguistic measures and qualitative conceptual assessment of the translations in accordance with the laws and principles in order to reveal possible distinctions between the AI-based translation and professional human translation. The quantitative analysis includes the measures of terminological consistency ratio, normalized error frequency, lexical density, and cohesion, and the analysis is performed using the proper inferential statistics. Error frequency is evaluated in connection with its level of significance and distinguishing insignificant language-related deviations from the errors which can lead to the misinterpretation of the text. The qualitative analysis includes the investigation of the selected legal terms and phrases which are characterized by the conceptual complexity, jurisdictional peculiarities, contractual and conditional relations, and culturally specific legal concept. These examples are investigated using the source-target conceptual mapping and are used to reveal the ability of the translation to reproduce the intended meaning and doctrine of the source texts.

4. Findings and Analysis

This section analyzes the quality of AI-produced English translations of Arabic legal texts in terms of seven crucial criteria: conceptual equivalence, terminological accuracy, omissions/ additions, legal modality, cultural concepts, semantic accuracy, and legal style/register. Legal Translation Quality Assessment (as described in Prieto Ramos, 2015) and Multidimensional Error Classification (as defined by Lommel et al., 2014) are taken as the bases for the analysis, which includes concrete examples that highlight translation errors and analyze their possible impact on legal meaning/interpretation. The presented examples serve to illustrate how the evaluation system can be applied, not as a report of empirical data on AI systems.

4.1 Conceptual Equivalence in Legal Translation

Table 1. Conceptual Equivalence in Legal Translation

No.	Arabic source text	Potentially problematic translation	Context-sensitive translation	Error type
1	يتولى الولي الشرعي إبرام العقد وفقاً لأحكام القانون.	The legitimate parent shall conclude the contract according to the law.	The legally authorized guardian shall conclude the contract in accordance with the law.	Conceptual distortion
2	يتمتع الشخص الاعتباري بالشخصية القانونية المستقلة.	The moral person enjoys independent legal personality.	The legal entity has a separate legal personality.	Incorrect legal equivalence
3	تسقط الدعوى بالتقادم وفقاً للمدة المحددة قانوناً.	The lawsuit falls because of aging according to the legal period.	The claim becomes time-barred upon expiry of the statutory limitation period.	Misinterpretation
4	لا يجوز التصرف في الأموال المحجوزة إلا بإذن من المحكمة المختصة.	Reserved money cannot be used without court permission.	Assets subject to judicial attachment may not be disposed of without authorization from the competent court.	Conceptual ambiguity

Examples from Table 2 clearly show how crucial is the issue of conceptual equivalence in Arabic–English legal translation when ,legal expressions involve extended meanings other than the lexical meaning of the expression itself. For example, in Example 1 involves conceptual equivalence because the Arabic expression الولي الشرعي using "legitimate parent" as the translation of refers to "the authorized legal guardian of the person rather than his/her natural parent." In Example 2, using "moral person" to fails to deliver the idea of an already established concept in law that refers to the "legal entity الشخص الاعتباري translate the term misses التقادم endowed with legal personality." In Example 3, using the meaning of "aging" for the translation of the expression as الأموال المحجوزة the idea of statutory limitation and how it influences the claim. And finally, in Example 4, the translation of reserved money" does not deliver the legal meaning of the expression because it is about assets that are placed under judicial" attachment and thus cannot be freely used. Thus, the examples support the need for legal concepts and their functions to be .translated rather than simply their counterparts (Šarčević, 1997; Cao, 2007)

Considering the criteria of conceptual accuracy suggested in the Legal Translation Quality Evaluation Framework, all these cases exemplify the ways in which the following criteria can be used for assessing the level of conceptual accuracy: preservation of legal meaning, use of adequate legal terminology, and preservation of the legal effect conveyed by the source text. Conceptual equivalence is suggested to have the weight of 30% in accordance with the framework as it is considered to be the main criterion for evaluating consequences of translation errors. In accordance with the approach, the translations are evaluated based on whether they convey the correct legal concept, fit the legal context and convey the rights, obligations, or limitations set out in are worth special attention since الأموال المحجوزة and التقادم the provision. For example, translation errors that are related to incorrect equivalents can distort the legal procedure and the status of the asset which is restricted by law. As it is mentioned above, in order to differentiate between minor and serious consequences of terminological inaccuracy, it is necessary to follow the methodology of evaluating the quality of legal translations provided by Prieto Ramos (2015) and Lommel et al. (2014). The examples provided allow to understand how this can be done in practice; however, the criteria suggested by the framework .should be validated empirically

4.2 Terminological Accuracy and Consistency

Table 2. Terminological Accuracy and Consistency

No.	Arabic source text	Potentially problematic translation	Context-sensitive translation	Error type
1	يجوز للطرف المتضرر فسخ العقد إذا أخل الطرف الآخر بالتزاماته الجوهرية.	The injured party may cancel the agreement if the other party fails to meet its essential commitments.	The aggrieved party may terminate the contract if the other party breaches its material contractual obligations.	Terminological imprecision

No.	Arabic source text	Potentially problematic translation	Context-sensitive translation	Error type
2	يلتزم المدعى عليه بتقديم لائحة جوابية خلال المدة القانونية.	The accused must submit an answer paper within the legal period.	The defendant shall file a statement of defence within the statutory period.	Procedural terminology
3	يكون الحكم القضائي قابلاً للطعن أمام المحكمة المختصة.	The judicial decision can be attacked before the relevant court.	The judgment may be challenged before the competent court through the applicable appeal procedure.	Literal terminology
4	يحق للدائن المطالبة بالتعويض عن الضرر الناشئ عن الإخلال بالعقد.	The lender can ask for compensation for damage caused by breaking the agreement.	The creditor is entitled to claim damages arising from a breach of contract.	Terminological inaccuracy

The cases provided in Table 3 are indicative of the importance of terminological accuracy in retaining the legal meaning and as "cancel the agreement" and فسخ العقد procedural implications of Arabic texts. In the case of Example 1, the translation of .as "fails to meet essential commitments" reduces the accuracy of specialized contractual terms الإخلال بالتزامات جوهرية Terminate the contract" and "breaches its material contractual obligations" are context-specific alternatives that better express" is highly contingent upon the فسخ the specific contractual remedy and the type of breach; however, the accurate translation of particular legal system in which the concept is being used. The procedural inaccuracy in Example 2 is even more apparent, as in as "the accused" brings criminal law terminology into the provision that seems to deal المدعى عليه that case, the translation of as "attack" in Example 3, which fails to convey the الطعن with a civil process. The same applies to the literal translation of specialized term representing a procedure of contesting a judicial decision. Finally, in Example 4, "lender" is an inaccurate .because of the restricted interpretation of "creditor," whose right is not necessarily based on lending الدائن translation of

In the Legal Translation Quality Evaluation Framework suggested by the author, terminological accuracy and consistency are seen as a separate evaluation dimension which has a tentative weight of 20%. The dimension is aimed at the assessment of how translations use appropriate legal equivalents, keep terminological consistency within related provisions and keep the difference between the legal concepts and procedural positions. The examples provided demonstrate that the errors have to be distinguished both according to their terminological and legal consequences. In particular, the replacement of "defendant" by "accused" could lead to misunderstanding of the procedural situation, whereas the replacement of "creditor" with "lender" could result in the wrong restriction of the category of people who are entitled to compensation for damage. In this way, the proposed framework involves corpus-based terminology analysis together with an expert evaluation of the acceptability of terminological alternatives and inconsistencies. On the one hand, corpus tools could help to see what translation choices were made, on the other hand, legal specialists could evaluate the choices in terms of appropriateness in the particular jurisdiction. The approach adopted here aligns with Biel's (2014) emphasis on the legal phraseology and Prieto Ramos's (2015) combined model of legal translation quality assurance.

4.3 Omission and Addition in Legal Translation

Table 3. Omission and Addition in Legal Translation

No.	Arabic source text	Potentially problematic translation	Context-sensitive translation	Error type
1	لا يجوز إنهاء العقد قبل انتهاء مدته ما لم يتفق الطرفان كتابةً على خلاف ذلك.	The contract may not be terminated before its expiration.	The contract may not be terminated before the expiry of its term unless the parties agree otherwise in writing.	Omitted exception
2	يلتزم المستأجر بدفع الأجرة خلال الأيام الخمسة الأولى من كل شهر.	The tenant shall pay the rent every month.	The tenant shall pay the rent within the first five days of each month.	Omitted time restriction

No.	Arabic source text	Potentially problematic translation	Context-sensitive translation	Error type
3	يجوز للمحكمة تأجيل الجلسة بناءً على طلب أحد الأطراف ولأسباب مبررة.	The court may postpone the hearing at the request of either party.	The court may adjourn the hearing at the request of either party for justified reasons.	Omitted condition
4	يتحمل المقاول مسؤولية الأضرار الناتجة عن الإهمال.	The contractor shall be liable for all damages, including those caused by negligence or unforeseen events.	The contractor shall be liable for damages resulting from negligence.	Unjustified addition

As the cases in Table 4 show, omissions and additions can jeopardize the semantic accuracy and legal consequences of the translated provision. Thus, in Example 1, the omission of ما لم يتفق الطرفان كتابةً على خلاف ذلك ("unless the parties agree otherwise in writing") deletes an important exception from the contract, turning a conditional ban into a seemingly absolute one. In a similar way, the omission of a certain deadline for payments is seen in Example 2 when translating خلال الأيام الخمسة الأولى من كل شهر as "every month" thus deleting the criterion regarding the time within which the tenant is required to make payments. In Example 3, the omission of ولأسباب مبررة ("for justified reasons") eliminates the condition limiting the judge's right to suspend the procedure. On the contrary, in Example 4, additional information, not present in the Arabic source text, extends the contractor's responsibilities to events he could not foresee. Thus, omissions and additions are not only linguistic inaccuracies but might influence the conditions, limitations and obligations imposed by legal texts. This statement is supported by the ideas expressed by Šarčević (1997) and House (2015).

In the suggested Legal Translation Quality Evaluation Framework, omissions and additions will be mainly estimated in terms of semantic accuracy, the weight of which is set as 20% provisionally, but also regarding their impact on legal conceptual equivalence. These types of mistakes will be estimated based on three factors: their importance, influence on legal interpretation of the provision, and possible implications for rights and obligations interpretation. Thus, omission of an exception or a deadline in the contract could be classified as major mistake as it influences the applicability of a legal obligation; addition of the liability for unforeseeable circumstances by the contracting party could be considered as critical mistake due to the great extension of responsibilities. Relying on the severity-oriented criteria of the Multidimensional Quality Metrics framework (Lommel et al., 2014) and legal quality assessment approach of Prieto Ramos (2015), the suggested approach pays more attention to the legal implications of translation errors rather than their frequency. Therefore, the suggested analysis shows the necessity to consider systematic identification of mistakes along with their assessment by an expert lawyer.

4.4 Legal Modality and Obligation

Table 4. Legal Modality and Obligation

No.	Arabic source text	Potentially problematic translation	Context-sensitive translation	Error type
1	يجب على الشركة إخطار الجهة المختصة خلال ثلاثين يوماً من تاريخ التغيير.	The company may notify the competent authority within thirty days of the change.	The company must notify the competent authority within thirty days from the date of the change.	Duty changed to permission
2	يجوز للمحكمة أن تأمر بإجراء خبرة فنية عند الضرورة.	The court must order a technical expert examination when necessary.	The court may order a technical expert examination when necessary.	Discretion changed to duty
3	يحظر على الموظف إفشاء المعلومات السرية المتعلقة بعمله.	The employee should avoid revealing confidential work information.	The employee is prohibited from disclosing confidential information relating to their work.	Prohibition weakened
4	لا يلتزم البائع بتسليم البضاعة قبل استلام الدفعة الأولى.	The seller must deliver the goods before receiving the first payment.	The seller is not obliged to deliver the goods before receiving the first payment.	Obligation reversed

The examples provided in Table 5 clearly demonstrate the significance of legal modality in maintaining the normativity of Arabic legal texts. In Example 1, the translation of يجب as "may" renders a mandatory notification into an optional one, thereby misleading about the legal obligations of the company. On the other hand, in Example 2, the translation of يجوز as "must" turns judicial discretion into a mandatory step, changing thus the power vested in the court. In Example 3, the translation of يحظر should avoid" makes the legal prohibition more like a recommendation by weakening its obligatory nature. Most importantly, in" changes a legal exception of the seller from his duty to deliver into a لا يلتزم البائع Example 4, the omission of negation in positive obligation. These examples show that modal verbs are not mere grammatical phenomena; rather, they are important devices used in legal texts to create obligations, permission, prohibition, and discretionary power. The above discussion lends credence to Šarčević's (1997) claim regarding the preservation of the normativity of the law and Cao's (2007) importance of .accurate communication of legally operative meanings

In the context of the suggested Legal Translation Quality Evaluation Framework, modality errors are mainly considered as part of semantic accuracy (20%) and legal conceptual equivalence (30%) due to their ability to change the legal implications of the translation. In the framework, modal elements are evaluated based on three aspects: the preservation of normativity, accuracy of the relationship of obligation/permission, and legal implications that could arise due to the semantic shift. Thus, changing "must" for "may" might be regarded as a major/critical mistake in case the latter makes a mandatory provision of law voluntary, while changing a negative obligation might be a crucial mistake since it significantly distorts the legal relationship of the parties to the contract. Based on the severity-oriented classification principles of Lommel et al. (2014) and legal translation quality assurance model by Prieto Ramos (2015), the suggested framework gives preference to the legal importance of the modal errors over their linguistic frequency. Consequently, the examples illustrate the need to evaluate the legal implications of AI-generated translations via expert evaluation of deontic meaning.

5. Culturally Specific Legal Concepts

Table 5. Culturally Specific Legal Concepts

No.	Arabic source text	Potentially problematic translation	Context-sensitive translation	Error type
1	تستحق الزوجة المهر المؤجل وفقاً للشروط المنصوص عليها في عقد الزواج.	The wife is entitled to the postponed dowry according to the marriage agreement.	The wife is entitled to the deferred mahr (dower) in accordance with the conditions stipulated in the marriage contract.	Cultural imprecision
2	تبدأ عدة المطلقة من تاريخ وقوع الطلاق وفقاً لأحكام القانون.	The divorced woman's counting begins from the divorce date.	The divorced woman's ʿiddah (legally prescribed waiting period) begins on the date the divorce takes effect, in accordance with the law.	Islamic legal term
3	لا يجوز للمرأة الزواج خلال فترة العدة المقررة شرعاً.	A woman cannot marry during the religious counting period.	A woman may not marry during the waiting period (ʿiddah) prescribed under the applicable Islamic legal rules.	Cultural equivalence
4	يترتب على الطلاق البائن انحلال الرابطة الزوجية وفقاً لأحكام الشريعة.	Final divorce causes the marriage relationship to disappear according to religious rules.	An irrevocable divorce (ṭalāq bāʾin) dissolves the marital bond in accordance with the applicable provisions of Islamic law.	Doctrinal imprecision

The cases described in Table 5 illustrate the problems of translating culturally rooted legal terms from Arabic to English in cases ,where the legal terms belong to Islamic jurisprudence and do not have direct equivalents in other legal systems. For example as "postponed dowry" does not capture the special legal sense of deferred mahr, which is an economic المهر المؤجل translating right created through the marriage contract. Example 2 demonstrates how "the divorced woman's counting" is the literal which does not reflect the special doctrinal sense of the legally prescribed waiting period of the عدة المطلقة translation of divorced woman. The same goes for the phrase "religious counting period," which does not convey the legal limitations imposed as "final divorce" ignores the doctrinal distinction الطلاق البائن on the ʿiddah, especially regarding remarriage. Translation of

between revocable and irrevocable divorce in Islamic family law. All of these cases prove that culturally rooted legal terminology should be translated through conceptual interpretation. The combination of transliteration and explanation of the legal term (such as ʿiddah – legally prescribed waiting period; and ṭalāq bāʾin - irrevocable divorce) can contribute to the preservation of doctrinal specificity and readability for English-speaking audience. This conceptual interpretation of legal terminology is supported by Cao (2007) analysis of legal system differences and Šarčević (1997) conceptual and functional equivalence in legal .translation

In the Legal Translation Quality Evaluation Framework, culturally specific legal terms are primarily evaluated within the parameters of jurisdictional and cultural adequacy (5%), as well as legal conceptual equivalence (30%). Specifically, the following three interrelated criteria are considered in order to evaluate these concepts: doctrinal meaning preservation, appropriateness of needs to be المهر المؤجل, the translation strategy, and compatibility of the expression with the legal jurisdiction. For instance should be translated by taking into الطلاق البائن and العدة translated considering its contractual and financial nature, whereas ,account legal consequences associated with the status of marital status and divorce in the source-language context. In this way the suggested evaluation framework considers which translation strategy (transliteration, functional equivalence, or explanatory translation) would provide the most suitable equivalent of the source term for the target audience. Similar to Prieto Ramos's approach to legal translation quality assurance, an expert evaluation is required to ensure that the equivalents selected do (2015) not mislead the reader about the source legal institution. However, it is important to note that the appropriateness of the translations is dependent on the particular jurisdiction and purpose of communication because different countries can have various provisions in Islamic family law. Therefore, the examples presented above show the significance of cultural and doctrinal .analysis within AI-assisted legal translation assessment

5.1 Semantic Accuracy in Conditional Provisions

Table 7. Semantic Accuracy in Conditional Provisions

No.	Arabic source text	Potentially problematic translation	Context-sensitive translation	Error type
1	لا يكون المدين مسؤولاً عن التأخير إذا أثبت أن التأخير ناتج عن قوة قاهرة.	The debtor is responsible for the delay if it results from force majeure.	The debtor shall not be liable for the delay if the debtor establishes that it resulted from force majeure.	Liability reversed
2	يسقط الحق في المطالبة بالتعويض إذا لم تُرفع الدعوى خلال ثلاث سنوات.	The right to compensation ends after three years.	The right to claim compensation is extinguished if legal proceedings are not initiated within three years.	Condition omitted
3	لا يُعفى المتعاقد من المسؤولية إلا إذا أثبت عدم ارتكابه أي خطأ.	The contracting party is exempt from liability if an error occurs.	The contracting party shall not be exempt from liability unless it proves that it committed no fault.	Exemption distorted
4	يحق للمشتري فسخ العقد إذا تأخر البائع في التسليم لأكثر من ثلاثين يوماً.	The buyer may terminate the contract if the seller delays delivery for thirty days.	The buyer may terminate the contract if the seller delays delivery by more than thirty days.	Numerical threshold

The examples provided in Table 7 show that semantic accuracy is critical when it comes to the translation of conditional provisions of laws, especially where the understanding of rights, liabilities, and remedies hinges upon maintaining logical links leads to the لا يكون المدين مسؤولاً between different clauses. For example, Example 1 shows that failing to include negation in change of the meaning of this provision as the debtor's liability becomes positive while ignoring the need for proving the يسقط الحق في المطالبة بالتعويض إذا لم تُرفع الدعوى خلال presence of force majeure. In Example 2, it can be seen that translating as "the right to compensation ends after three years" means that one omits the condition regarding the initiation of ثلاث سنوات where the incorrect ,إلا إذا ... لا legal proceedings. Likewise, Example 3 shows the distortion of conditional exemption stated in translation interprets the right of not being responsible as dependent on an error in place of lack of fault. Example 4 shows the ,difference made by substituting "more than thirty days" for "thirty days." These examples demonstrate that conditional clauses negation, and numbers are all parts of the legal meaning that should be preserved in translation as the misinterpretation of any

of them changes the meaning of the provision. This statement is consistent with Šarčević's (1997) focus on the legal effect .preservation as well as House's (2015) functional approach to translation quality assessment

In the context of the Legal Translation Quality Evaluation Framework proposed here, conditional inaccuracies are evaluated predominantly through semantic accuracy (20%) and legal conceptual equivalence (30%) due to their possible effects on the interpretation of legal rights and obligations. In particular, the framework evaluates such inaccuracies based on three interrelated criteria: preservation of conditional relations, precision of negation and logical restrictions, and conservation of numerically significant thresholds from a legal perspective. For example, the reversal of liability in the force majeure clause or the distortion of the condition of exemption could be a critical error, as in this case, the allocation of legal responsibilities will be substantially distorted. Likewise, the deletion of a condition from a statute or the distortion of a deadline in the contract may have an effect on the situation in which legal rights become applicable and extinguishable. Based on the severity-oriented approach of the Multidimensional Quality Metrics framework (Lommel et al., 2014) and Prieto Ramos's (2015) framework of legal translation quality assurance, the Legal Translation Quality Evaluation Framework developed in the current study prioritizes the legal effects of semantic deviations as opposed to the linguistic ones. As a result, the following examples show that the integration of systematic error classification and conditional meaning evaluation by an expert can help to detect potentially significant inaccuracies in AI-assisted legal translation between Arabic and English languages.

4.7 Functional Adequacy and Legal Register

Table 7. Functional Adequacy and Legal Register

No.	Arabic source text	Potentially problematic translation	Context-sensitive translation	Error type
1	يصدر الوزير التعليمات اللازمة لتنفيذ أحكام هذا النظام.	The minister gives the necessary instructions to carry out this system.	The Minister shall issue the instructions necessary for implementing the provisions of these Regulations.	Register deviation
2	يعمل بهذا القانون من تاريخ نشره في الجريدة الرسمية.	This law works from the date it is published in the official newspaper.	This Law shall enter into force on the date of its publication in the Official Gazette.	Legislative phraseology
3	مع عدم الإخلال بأي عقوبة أشد ينص عليها قانون آخر، يعاقب كل من يخالف أحكام هذا القانون.	Without harming any stronger punishment in another law, anyone who breaks this law will be punished.	Without prejudice to any more severe penalty prescribed by another law, any person who violates the provisions of this Law shall be subject to punishment.	Literal phrasing
4	تعتبر هذه الاتفاقية ملزمة للطرفين اعتباراً من تاريخ توقيعها.	This agreement is considered compulsory for both sides from the day they sign it.	This Agreement shall be binding upon both parties as of the date of its signing.	Contractual register

The above-mentioned examples given in Table 7 clearly show the significance of functional adequacy and the use of adequate ,legal register to guarantee that the translated legal text will fulfill its intended communicational and institutional functions. First into "the minister gives the necessary instructions to carry يصدر الوزير التعليمات اللازمة لتنفيذ أحكام هذا النظام the translation of يصدر الوزير التعليمات اللازمة لتنفيذ أحكام هذا النظام out this system" uses informal language. A better alternative is to translate into "The Minister shall issue the instructions necessary for implementing the provisions of these Regulations." It should be has to be identified based on the particular jurisdiction. Second, the translation of النظام noted, however, that the equivalent of into "this law works" does not convey the intended message as well as "shall enter into force" that is a يعمل بهذا القانون will produce unnatural legal phrasing, while "without مع عدم الإخلال conventional legal term. Third, a literal translation of prejudice to" retains the relationship between the provision and the more severe penalties which may have been provided by into "compulsory for both sides" does not convey the conventional legal ملزمة للطرفين another section. Finally, the translation of phrase "binding upon both parties." Overall, the quality of legal translation requires not only semantic accuracy, but also genre-specific linguistic features and conventions of translated texts. This interpretation is compatible with the functional approach proposed by House (2015) as well as with Biel's (2014) focus on the textual and phraseological characteristics of translated legal .discourse

Within the framework of the suggested framework of Legal Translation Quality Evaluation, the dimension of functional and pragmatic adequacy is a separate criterion with a tentative weight of 15%, while linguistic quality is another criterion with a tentative weight of 10%. Functional and pragmatic adequacy evaluates if the translated provisions keep their intended functions, if they follow the legal writing conventions, and if the terminology and phraseology used are appropriate for the respective genre. In this case, the suggested framework differentiates stylistic deviations that only have an impact on the readability of the document and functional errors that can have an effect on its legal interpretation. For example, substituting "shall enter into force" with "works" can be regarded as deviation from the phraseology characteristic of legislative documents and can be associated with the inability to understand the function of the provision, whereas incorrect translation of "without prejudice to" can affect the connection between the overlapping legal provisions. Based on the legal translation quality assurance approach of Prieto Ramos (2015) and the multidimensional error classification principles by Lommel et al. (2014), the proposed framework takes into account functional evaluation and severity evaluation and allows determining whether the deviation of register has an impact on the appropriateness of the text or on its legal meaning.

6. Discussion

The analysis identifies the limitations of judging the adequacy of AI-driven translation of legal documents exclusively based on the linguistic fluency and word-for-word equivalency. In view of the seven different categories of errors, it is shown that while the translations might be considered satisfactory on the surface, they could still lack legal relationships between the elements that are found in the source documents in Arabic. Such conclusions coincide with those made by Šarčević (1997) and Cao (2007) who consider the legal translation as the process of cross-cultural communication in which the equivalence is achieved not through linguistic similarities but on the basis of the legal function of the documents. They also coincide with the research findings of Ait ElFqih and Monti (2024) who discovered problems associated with terminology and contextual interpretation of the large language model-generated translations of the Arabic legislation. Moreover, such findings are consistent with the advantages of professional human translation described by Altakhaineh et al. (2025) and Abdelaal and Al Sawi (2025) in the context of legal terminology and correct interpretation of the context. While the mentioned studies focus on the advantages and disadvantages of the process of translation, the present analysis provides a demonstration of how the different types of errors can affect each other within the provision.

The evaluation model attempts to solve the problem by combining corpus-based research with quality assessment of legal relevance and categorization of the severity of errors. Although Baker (1993) and Biel (2014) explain the importance of using corpora to study repetitive language use patterns, the present framework maintains that frequency and consistency cannot ensure legal adequacy of a translation. In particular, it is important to distinguish the present evaluation approach from the conclusions made by Hendy et al. (2023) and Kocmi and Federmann (2023) about the abilities of advanced language models in translating texts and evaluating the quality of the translation as those do not prove their ability to identify legal mistakes characteristic of a certain jurisdiction. By combining the concept of equivalence, terminology, semantics, adequacy, and culture, the evaluation framework broadens the multi-dimensional criteria suggested by Lommel et al. (2014) and the legal quality assurance approach used by Prieto Ramos (2015). The primary methodological innovation of the framework is the separation of an error occurrence and its legal consequences. It means that a single error related to liability or obligations can have more importance for assessment than a few errors related to style. However, since the examples used are not empirical translation examples, the analysis shows that the framework is conceptually applicable rather than empirically efficient. Therefore, validation using real translation samples by humans and AI, and inter-rater reliability measures is needed to estimate the efficiency of the framework.

7. Conclusion

This research paper develops a multi-dimensional model for evaluating AI-enhanced Arabic-English legal translation based on legal translation theory, corpus linguistics, and quality assessment. The main idea is that the quality of translation in legal settings must be evaluated in terms of its preservation of the operational meaning of the legal text rather than in terms of linguistic acceptability alone. The study emphasizes the significance of looking into the legal notions, procedures, conditions, and cultural expressions in one comprehensive assessment model. Herein lies the importance of theoretical groundwork laid in this paper for analyzing problems of translation.

The current research provides a multi-dimensional model of the analysis of AI-assisted translation from Arabic to English in the context of legal texts, combining legal translation theories, corpus methods, and the process of quality evaluation. Its major thesis consists of the necessity of evaluating translation quality from the perspective of maintaining legally active meaning, not just linguistic acceptability. The research draws attention to the need of paying attention to legal notions, legal procedures, conditional relations, and culturally-bound language elements when evaluating translation quality. In this way, the research contributes to the theoretical base of analyzing translation issues.

It is important to mention certain limitations of the current study that will define directions for future research. First of all, the cases of analysis have been selected to serve as examples for the proposed categorization framework, rather than to provide an evaluation of the real-life performance of any particular AI system or prove the superiority of human translation. Second, the assessment criteria weights that have been proposed need to be validated in order to ensure the reliability of the framework. The future study needs to apply the framework to analyze a corpus of Arabic and English legal documents, compare the performance of professional human translators with different AI tools, and ensure that expert assessments are consistent. The study of human

post-editing, terminology management at the document level, and the relation between the seriousness of the error and legal risk will add to the practical importance of the framework.

Funding Statement

This research received grant no. (881) from the Arab Observatory for Translation (an affiliate of ALECSO), which is supported by the Literature, Publishing & Translation Commission in Saudi Arabia.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Abdelaal, N., & Al Sawi, I. (2025). A comparative evaluation of machine translation vs. human translation for legal texts: A case study of translations between English to Arabic. *Comparative Legilinguistics*, *63*, 186–223. https://doi.org/10.14746/cl.2025.63.1

[2] Ait ElFqih, K., & Monti, J. (2024). *Large language models as legal translators of Arabic legislation: Do ChatGPT and Gemini care for context and terminology? In Proceedings of the Second Arabic Natural Language Processing Conference (pp. 111–122)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.arabicnlp-1.10

[3] Al-Btoush, G., & Almahasees, Z. (2024). A linguistic investigation for the image translation powered by AI tool: A case study of Google Lens. *Pakistan Journal of Life and Social Sciences*, *22*(2), 100650–100669. https://doi.org/10.57239/PJLSS-2024-22.2.00805

[4] Almahasees, Z. (2017). Assessing the translation of Google and Microsoft Bing in translating political texts from Arabic into English. *International Journal of Languages, Literature and Linguistics*, *3*(1), 1–4. https://doi.org/10.18178/IJLLL.2017.3.1.100

[5] Almahasees, Z. (2020). *Diachronic evaluation of Google Translate, Microsoft Translator and Sakhr in English-Arabic translation [Doctoral dissertation, The University of Western Australia]*. https://doi.org/10.26182/5e2a3b3d81f7a

[6] Almahasees, Z., Gwasmeh, M., Dagamseh, M., Haider, A. S., Alrhamneh, F. A. A., & Ghnaim, A. A. (2026). *A multidimensional evaluation of AI-generated Arabic translations of English social media posts*. Research Journal in Advanced Humanities. https://doi.org/10.58256/0naqf817

[7] Almahasees, Z., & Husienat, I. (2024). A comparative analysis of terminological inconsistency in scientific translation from English into Arabic across different medical fields. *Training, Language and Culture*, *8*(3), 25–40. https://doi.org/10.22363/2521-442X-2024-8-3-25-40

[8] Almahasees, Z., Meqdadi, S., & Albudairi, Y. (2021). Evaluation of Google Translate in rendering English COVID-19 texts into Arabic. *Journal of Language and Linguistic Studies*, *17*(4), 2065–2080. https://doi.org/10.52462/jlls.149

[9] Almahasees, Z., & Husienat, I. (2026). AI-powered voice translation in medical settings: An evaluation of English–Arabic translation accuracy. *International Journal of Linguistics, Literature and Translation*, *9*(9), 203–214.

[10] Almahasees, Z. M. (2018). Assessment of Google and Microsoft Bing translation of journalistic texts. *International Journal of Languages, Literature and Linguistics*, *4*(3), 231–235.

[11] Al-Maaytah, M., & Almahasees, Z. (2024). A linguistic investigation for a case study of ChatGPT and Google Translate in rendering special needs texts from English into Arabic: A synchronic case study. *Pakistan Journal of Life and Social Sciences*, *22*(2), 100744–100758. https://doi.org/10.57239/PJLSS-2024-22.2.00812

[12] Altakhaineh, A. R. M., Alghathian, G. A., & Jarrah, M. M. (2025). A comparative study of accuracy in human vs. AI translation of legal documents into Arabic. *International Journal of Language & Law*, *14*, 63–80. https://doi.org/10.14762/jll.2025.063

[13] Bajčić, M., & Golenko, D. (2024). Applying large language models in legal translation: The state-of-the-art. *International Journal of Language & Law*, *13*, 171–196. https://doi.org/10.14762/jll.2024.171

[14] Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233–250). John Benjamins.

[15] Biel, Ł. (2014). Lost in the Eurofog: The textual fit of translated law. Peter Lang.

[16] Cao, D. (2007). Translating law. Multilingual Matters.

[17] 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. https://doi.org/10.1162/tacl_a_00437

[18] 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 [Preprint]*. arXiv. https://doi.org/10.48550/arXiv.2302.09210

[19] House, J. (2015). *Translation quality assessment: Past and present*. Routledge.

[20] Jiao, W., Wang, W., Huang, J.-t., Wang, X., Shi, S., & Tu, Z. (2023). *Is ChatGPT a good translator? Yes with GPT-4 as the engine [Preprint]*. arXiv. https://doi.org/10.48550/arXiv.2301.08745

[21] Kocmi, T., & Federmann, C. (2023). *Large language models are state-of-the-art evaluators of translation quality*. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation. European Association for Machine Translation. https://aclanthology.org/2023.eamt-1.19/

[22] Lommel, A., Burchardt, A., & Uszkoreit, H. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: Tecnologies de la Traducció*, *12*, 455–463.

[23] Merabet, M. H., & Roser Nebot, N. (2026). Translating legal texts: A four-criteria evaluation of ChatGPT, DeepL, and Google Translate. *Comparative Legilinguistics*, *67*. https://doi.org/10.14746/cl.2026.67.5

[24] Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *BLEU: A method for automatic evaluation of machine translation*. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 311–318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

[25] Prieto Ramos, F. (2015). Quality assurance in legal translation: Evaluating process, competence and product in the pursuit of adequacy. *International Journal for the Semiotics of Law*, *28*, 11–30. https://doi.org/10.1007/s11196-014-9390-9

[26] Šarčević, S. (1997). New approach to legal translation. Kluwer Law International.

[27] Vswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *In Advances in Neural Information Processing Systems*, *30*.