Large Language Models as Legal Translators of Arabic Legislation: Do ChatGPT and Gemini Care for Context and Terminology?

Khadija Ait ElFqih[1], **Johanna Monti**[1]
[1] UNIOR NLP Research Group, University of Naples 'L'Orientale'
{k.aitelfqih, jmonti}@unior.it

Abstract

Accurate translation of terminology and adaptation to in-context information is a pillar to high quality translation. Recently, there is a remarkable interest towards the use and the evaluation of Large Language Models (LLMs) particularly for Machine Translation tasks. Nevertheless, despite their recent advancement and ability to understand and generate human-like language, these LLMs have not been thoroughly investigated and are still far from perfect, especially in domain-specific scenarios. This is particularly evident in automatically translating legal terminology from Arabic into English and French, where, beyond the inherent complexities of legal language and specialised translations, technical limitations of LLMs further hinder accurate generation of text. In this paper, we present a preliminary evaluation of two evolving LLMs 'Generative Pre-trained Transformer' and 'Gemini' as legal translators of Arabic legislation to test their accuracy and the extent to which they care for context and terminology across two language pairs (AR→EN / AR→FR). The study targets the evaluation of Zero-Shot prompting for in-context and out-of-context scenarios of both models relying on a gold standard dataset, verified by professional translators who are also experts in the field. We evaluate the results applying the Multidimensional Quality Metrics to classify translation errors. Moreover, we also evaluate the general LLMs outputs to verify their correctness, consistency, and completeness. In general, our results show that the models are far from perfect and recalls for more fine-tuning efforts using specialised terminological data in the legal domain from Arabic into English and French.

1 Introduction

Over the last five years, there has been incremental progress in the field of Machine Translation (MT) (Koehn, 2020; Herold et al., 2022; Almahasees, 2021; Rossi & Carre, 2022) to the point where some researchers claim parity with human translation (HT) (Hassan, 2018; Thierry, 2022). Even recently, with the huge advent of Large Language Models (LLMs) such as Generative Pre-trained Transformer (GPT) (Brown et al., 2020) and Gemini (Rohan et al., 2024), it is anticipated that they will become more accurate, robust, and adaptable to a wider range of languages, and domains, making them even more valuable tools in our interconnected world (Gao & Lin, 2004; Kombrink et al., 2011; Alves et al., 2018; Chang et al., 2024). LLMs significantly demonstrate excellent performance in diverse Natural Language Processing (NLP) application tasks, including MT, due to their self-supervised and transfer learning abilities (Huang, 2024). This progress has led to focus on model scalability, efficiency, and human-like understanding and translation capabilities. However, their potential in domain-specific MT remains under-explored. In other words, the extent to which these models are consistent and accurate in the translation of technical texts and the terminology used therein is uninspected. As a matter of fact, consistent term translation is an important facet of quality assurance for specialized translation. Since terminologies are essential for communication among domain experts, term forms must be consistent and their translation must respond to the contextual requirements to maintain the integrity of the underlying conceptual system during knowledge exchange (Darwish, 2009; Sager, 1990). Nevertheless, some knowledge domains and languages still suffer from the lack of high-quality MT results due to the

Proceedings of The Second Arabic Natural Language Processing Conference, pages 111–122

multidimensional mistranslations of terminology (Mediouni, 2016; Killman, 2014; Zakraoui et al., 2021). This is the case, especially in the legal domain and the Arabic language. Consider example 1 from the Moroccan family code, taking the terms {**الفراش**}, {**الطعن**}, {**اللعان**}, {**القطع**} into consideration:

1. **AR:** يعتبر **الفراش** بشروطه حجة قاطعة على ثبوت النسب، لا يمكن **الطعن** فيه إلا من الزوج عن طريق **اللعان**، أو بواسطة خبرة تفيد **القطع**.

   EN (ChatGPT-4): the **consummation** with its conditions constitutes conclusive evidence of the lineage, which can only be **challenged** by the husband through "**li'an**" or by means of expertise proving its **falsity**.

   FR (ChatGPT-4) : The marital bed, under its conditions, is conclusive evidence of the establishment of **lineage**. It cannot be **challenged** except by the husband through **cursing (li'an)** or through expert testimony that provides **certainty**.

   EN (Gemini): the **marriage bed**, with its conditions, is considered conclusive evidence of **lineage**; it can only be **challenged** by the husband through invoking **curses**, or by expert testimony that provides **conclusive** evidence.

   FR (Gemini): la **consommation** est considérée comme une preuve concluante de la lignée, qui ne peut être **considerée** que par le mari à travers la **"li'an"** ou par des moyens d'expertise prouvant sa **fausseté**.

   HT (EN): Marriage consummation is considered a strong proof of paternity, it can be **rebutted** only by the husband through **accusation** or through a **certain** evidence.

   (HT) (FR): La **consommation du mariage** est considérée comme une preuve solide signifiant la paternité, il ne peut être **réfutée** que par le mari soit à travers **l'accusation** ou bien une **certaine** preuve.

The bold terms in example 1 are domain-specific and context-dependent, so their translation require the consideration of the context, as well as of the cultural, lexical, morphological, and semantic properties of the terms in addition to their equivalences across languages and legal systems (i.e., English, and French), as the HT does. Both models, instead, produce results with high levels of multidimensional critical errors. This example highlights the fact that these models can indeed capture general language patterns, generate coherent responses, and perform excellent MT tasks in general but struggle with translating domain-specific texts and terminology due to their limited domain-knowledge and limited training data. In addressing this gap, many authors conducted a thorough evaluation of LLMs potentials in MT application including Moslem et al. (2023) who find out that while some high-resource languages such as English-to-French, English-to-Spanish and even English-to-Chinese show excellent results, other languages have lower support either because they are low-resource languages such as English-to-Kinyarwanda or because of issues in the ChatGPT-3.5 tokenizer such as English-to-Arabic. Siu (2023) survey reveals that current LLMs exhibit certain limitations in numerous tasks, notably reasoning and robustness tasks. They suggest the need for contemporary evaluation systems that ensure the accurate assessment of LLMs' inherent capabilities and limitations. Hendy et al. (2023) presents a comprehensive study of the MT capabilities of the latest ChatGPT models. Their investigation covers 18 language pairs across four different domains, enabling a broad understanding of the models' general performance. Their findings demonstrate that ChatGPT-3.5 can produce highly fluent and competitive translation outputs especially and only for the high-resource language translations. All of them and others suggest that fine-tuning is an ultimate option to improve the in-context learning capability of LLMs and improve their translation quality and adherence to domain terminology and style (Schioppa et al., 2023; Sia & Duh, 2022; Alves et al., 2023; Iyer et al., 2023; Jiao et al., 2023; Xu et al., 2023; Zhang et al., 2023; Yang et al., 2023). In this paper, we present a comprehensive evaluation and an in-depth study of the MT capabilities of ChatGPT-4 and Gemini in In-Context and Out-of-Context levels in the legal domain in two language pairs AR→EN / AR→FR using the Multidimensional Quality Metric (MQM) (Burchardt, 2013). We also evaluate the potential of ChatGPT-4 and Gemini in MT of legal texts and terminology in Arabic using a gold standard dataset developed specifically for assessing the quality and accuracy of machine-

translated legal terms from Arabic into English and French. This paper is structured as follow: section 1 presents the introduction, section 2 details the experimental set-up and methodology, section 3 discusses the evaluation and results, and section 4 presents conclusions and future work.

2 Experiment Set-up and Methodology

Despite their great ability to perform MT tasks, both ChatGPT-4, developed by OpenAI [1], and Gemini, by Google DeepMind[2], encounter several and different pitfalls in accurately translating legal terminology from AR→EN / AR→FR, due to limited exposure to specialized terminological data in the legal domain during training. This leads to challenges in contextual and cultural understanding, besides inconsistencies, and potential misinterpretations of legal concepts in Arabic. Therefore, in this paper, we present a comprehensive evaluation of both models in the translation of Arabic legislation and answer the question of whether these models care for context and terminology in two language pairs (AR→EN / AR→FR). To do this, i) we develop an Arabic legal corpus (more details about the sources, size, coverage, terminology extraction and annotation etc., are mentioned in Corpus and Terminology extraction paragraph),, ii) generate translations from ChatGPT-4 and Gemini, iii) evaluate the accuracy of both LLMs outputs based on our gold standard dataset, which to the best of our knowledge is the first formalized legal terminological resource for Arabic developed specifically for assessing the quality and accuracy of machine-translated legal terms from Arabic into English and French, and then v) apply the MQM[3] (Burchardt, 2013) framework to classify translation errors and calculate the severity levels presented by the models.

Our gold dataset comprises a collection of 1,015 Arabic legal terms, each accompanied by 1,015 sentences where these terms are used, along with their translations into English and French. To make it reliable, our reference dataset is developed in accordance with the online gateways of EU laws, including EUR-Lex[4], IATE[5], Juremy[6] to retrieve the exact equivalence of legal terms in English and French, and is then submitted to two annotators to verify and validate it. The first is a legal expert whose language skills are excellent both in the source and the target languages. He validates the translations after checking their degree of accuracy and adequacy in the target languages. The second is a native Arabic speaker with a linguistic background meticulously who annotates the Part-of-Speech tags, Geographical Usage (following the ISO 20771:2020[7] standard for Legal translation Requirements, to indicate where a given term is adapted to express a legal practice). This validation method for such sensitive dataset is important for the sake of equivalence reference to ensure an adequate and accurate analysis. This dual-annotator approach enhances the quality of the data by reducing the chances of errors and inconsistencies, and it provides a standardized point of reference for evaluating MT systems objectively and systematically in the area of legal terminology translation for Arabic.

Corpus and Terminology Extraction. We use NooJ[8], an NLP software application, to develop a corpus of Arabic legislation of different Arab countries called the ARabic Legal Corpus (ARLC) that consists of 532,688 tokens (ElFqih & Monti, 2023; ElFqih et al., 2024; Silberztein, 2015) including marriage contracts, divorce provisions, constitutions, decrees, codes, etc., (see Table 1).

Legislation	Country	Tokens
Mariage Contracts	MAR/ UAE/ KSA/ EG	1,002
Divorce Provisions	MAR/ EG/ DZ/ KSA/ UAE/TN	1,069
Family Code	MAR	20,945
Code of Penal Procedure	MAR/ DZ/ TN	201,870
Code of Obligations and Contracts	MAR	82.365
Civil Code	QA/ DZ	175,888
Code of Personal Status	TN	11,638
Constitutions	MAR	12,494
Decrees	MAR/ EG	22,698
Local and External Pilgrims Law	KSA	2,719
Total of Tokens		532,688

Table 1: Arabic Legislation, contracts, and provisions.

[1] https://openai.com/index/chatgpt/
[2] https://deepmind.google/technologies/gemini/
[3] https://themqm.org/
[4] https://eur-lex.europa.eu/browse/eurovoc.html?locale=en
[5] https://iate.europa.eu/home
[6] https://www.juremy.com/about/
[7] https://www.iso.org/obp/ui/en/#!iso:std:69032:en
[8] https://nooj.univ-fcomte.fr/

The data used for corpus development is collected from the official website of Ministries of Justice, Official Gazettes, and others were provided by lawyers (contracts and provisions). Secondly, we launch a pre-processing operation on the corpus where we eliminate common typographical errors such as confusion between Alif (ا) and Hamza (ء) or the substitution of (ة) and (ه) at the end of the word, the false writing of (همزة/Hamza), the addition or omission of a character in a word, and any additional space that might be found between terms, etc. Then, we annotate it in ElFqih et al., (2024) using NooJ linguistic resources for Arabic (Mesfar, 2008). Building upon the initial step, we then extract Arabic legal terms and the sentences in which they occur using NooJ syntactic grammars (ElFqih et al., 2024).

Prompting. To start with the ChatGPT-4 and Gemini assessment for the automatic translation of legal terminology from Arabic, we utilize a Python script that iterates through our data to generate prompts. It implements a series of prompts for Out-of-Context and In-Context translation evaluation in two language pairs AR→EN / AR→FR.

1. `prompt1 = 'Write the corresponding English/French term for "{Arabic legal term}". Write only the English/French translation without providing comments.' 'Write the translations following the order provided.'`
2. `Prompt2 = 'Write the corresponding English/French sentence for "{Arabic legal sentence}". Write only the English/French translation without providing comments.' 'Write the translations following the order provided.'`

Prompts 1 and 2 generate from the LLMs the translation for Out-of-Context and In-Context legal terms and sentences into English and then into French, each separately respecting the order provided. Prompt 1 is instructed to provide translations without providing additional context or comments and Prompt 2 is directed to provide no additional comment. The generated output, subject for evaluation, is refined and classified as follow:

- Out-of-Context and In-Context translation results of ChatGPT-4 from AR→EN / AR→FR,
- Out-of-Context and In-Context translation results of Gemini from AR→EN / AR→FR.

3 Evaluation and Results

	ChatGPT-4			
	Out-of-Context AR→EN	**In-Context AR→EN**	**Out-of-Context AR→FR**	**In-Context AR→FR**
False	54%	98%	60%	98%
	Gemini			
False	57%	97%	61%	99%

Table 2: Comparison of ChatGPT-4 and Gemini's MT of Legal Terms from AR→EN / AR→FR in Out-of-Context and In-Context Scenarios, Based on Accuracy Criterion.

Our evaluation consists of three fundamental phases. The first is to assess the obtained results of both models, GPT-4 and Gemini, based on the accuracy criterion against the gold dataset. The second is the evaluation of the results applying the MQM (Burchardt, 2013), which is a framework for analytic Translation Quality Evaluation (TQE)[9] that can be applied to evaluate machine-translated legal terms for Arabic by ChatGPT-4 and Gemini against our gold dataset. The third consists of evaluating the general LLMs outputs to verify their correctness, consistency, and completeness in the MT of legal terminology for Arabic.

Table 2 presents machine-translated legal terms results of both models from AR→EN / AR→FR based on accuracy criterion. In the out-of-context evaluation, ChatGPT-4 yielded 54% false results when translating from Arabic into English, whereas it produced 60% false results in Arabic into French. Conversely, Gemini generated 57% false results for Arabic into English, and 61% false results for Arabic into French. Notably, both systems demonstrated a consistent trend of higher false counts than true counts across both language pairs, highlighting the challenge of accurately translating legal terminology without contextual cues. While ChatGPT-4 generally exhibited slightly fewer false results than Gemini in Arabic into English translations, Gemini showcase marginally better performance in Arabic into French, indicating nuanced differences in its capabilities across different language pairs. However, in the In-context assessments, both ChatGPT-4 and Gemini

[9] https://themqm.org/introduction-to-tqe/

faced significant challenges, with ChatGPT-4 scoring 98% false results in Arabic into English, and 98% false results in Arabic into French. Gemini performed similarly, registering 97% false results in Arabic into English, and 99% false results in Arabic into French. Despite slight variations in the true counts, both systems demonstrate a notable decrease in performance when translating legal terminology within specific contextual frameworks. While the accuracy measure is important, these first findings do not, however, capture the nuances of the translation quality beyond word-for-word accuracy i.e., the Arabic linguistic structures and cultural nuances Arabic legal terminology denotes. They also do not provide an accountable decision on the overall quality of the models.

Therefore, we apply the MQM (Burchardt, 2013) framework. It includes a set of error types and a scoring model. Error types are organized in a

		ChatGPT-4				Gemini			
Error Types	**Error sub-types**	**Out-of-Context AR→EN**	**In-of-Context AR→EN**	**Out-of-Context AR→FR**	**In-of-Context AR→FR**	**Out-of-context AR→EN**	**In-context AR→EN**	**Out-of-context AR→FR**	**In-of-Context AR→FR**
Terminology	**Inconsistent with terminology resource**	1000	1006	998	1001	998	1009	862	1000
	Inconsistent use of terminology								
	Multiple terms for concept in source (multiple terms for concept)								
	Wrong term								
Style	**Language register**	1008	1005	768	1000	980	1003	950	989
	Awkward style								
Linguistic convention	**Duplication**	645	760	765	864	500	800	750	987
	Grammar								
	Word form								
	Agreement								
	Word order								
	Function words								
	Grammatical register								
	Transliteration								
	Coherence								
Accuracy	**Undertranslation**	986	999	921	1001	909	1005	901	999
	Overtranslation								
	Addition								
	Omission								
	Untranslated								
	Incomplete procedure								
	Mistranslation								
	Ambiguous target content								
	False friend								
	Completeness								
	Incomplete List (incomplete-list)								
	Incomplete procedure (incomplete -procedure)								
Audience appropriateness	**Culture-specific reference**	1002	1010	1002	1002	1006	1011	963	1001
	Locale-specific content								
	Legal requirements								
	End-user suitability								

Table 3: Errors count and classification of machine-translated legal terms by ChatGPT-4 and Gemini for AR→EN / AR→FR in Out-of-Context and In-Context scenarios.

hierarchical system[10] under seven high-level core error dimensions, together with subordinate error types and their associated severity levels (neutral, minor, major, and critical). The scoring model[11] features a system of weights and parameters assigned to the error types and severity levels, as well as a scoring formula used to calculate a numerical score that represents the quality of the evaluated translation i.e., Absolute Penalty Total according to agreed-upon specifications and the Error Count. To achieve the MQM scoring stage, it is fundamental to first annotate the errors applying the MQM error typology framework. Table 3 details per each LLM i) six high level dimensions namely terminology, style, linguistic convention, accuracy and audience appropriateness, ii) the errors subtypes iii) and the number per error type for the two language pairs AR→EN / AR→FR in Out-of-Context and In-Context scenarios.

Subsequently, to be more precise towards the evaluation and the final judgement of the MT results produced by the models, we calculate the final quality of the translation results using the MQM scoring model as shown Tables 4, 5, 6, 7, 8, 9, 10 & 11 in Appendix A in by assigning scores to:

- **The Error Severity Levels** (neutral, minor, major, and critical) that we define depending on the extent to which the error poses a risk to the quality of the translation,
- **The Severity Penalty Multiplier (SPM)** score which reflects the increased risk and impact between the Error Severity Levels (for example, in our case study, we give 1 to neutral, 1 to minor, 5 to major, and 25 to critical),
- **Error Type Weight (ETW)** that reflects the importance of certain error types that should be given more prominence than others. For example, we give 5 ETW score for terminology, accuracy, and audience appropriateness and only 3 to style and linguistic convention because they represent the highest error rates.
- **Error Type Penalty Total (ETPT)** is the sum of penalty points calculated for the individual error types annotated. The error count for a specific error type and severity level is multiplied by the respective SPM and ETW to obtain the ETPT. For example, the ETPT for terminology in Table 4 is determined as follows: 0x1(x5)=0; 6x1(x5)=30; 10x5(x5)=250; 979x25(x5)=122.375 then, (0+30+250+122.375)x5=613.275 therefore, 613.275 is the ETPT for terminology errors in the Out-of-Context from Arabic into English scored by ChatGPT-4.
- **Errors Count (EC)** is the number of errors set for each dimension and its types which we classify according to the level of severity across the error severity level.
- **Absolute Penalty Total (APT)** is considered the most important value used for quality score calculation, and the one that we consider to decide over the translation quality and compare the models' quality and performance on MT of legal terms in and out of context from AR→EN / AR→FR. APT is the sum of all ETPTs.

The evaluation comparing MT application of ChatGPT-4 and Gemini in the AR→EN / AR→FR language pairs using the MQM scoring model provides valuable insights into their performance, both in and out of context (see Tables 4, 5, 6, 7, 8, 9, 10 & 11 in Appendix A). In the Out-of-Context scenario for AR→EN translation, ChatGPT-4 exhibited a higher absolute penalty total compared to Gemini (see Table 4 & 8). This suggests that ChatGPT-4 struggled more with maintaining accuracy and coherence when translating legal terms without the surrounding context.

We assign 5 ETW for terminology (ranging from 0 neutral, 6 minor, 10 Major, and 979 critical terms), accuracy (ranging from 0 neutral, 23 minor, 35 Major, and 916 critical terms), and audience appropriateness (ranging from 0 neutral, 23 minor, 11 Major, and 959 critical terms), as ChatGPT-4 shows an inconsistent use of terminology, the use of wrong terms and false friends, untranslated entries, lack of legal requirements, and culture-specific references etc., which all lead to present wrong concepts in the target language with a total number of 1000 errors out of 1015 term. For example, the English translation of the term **{متعة} 'compensation'** is left untranslated by ChatGPT-4 and results instead in the following comment **'nan**

[10] Available here: https://themqm.org/error-types-2/typology/

[11] Detailed process available here: https://themqm.org/error-types-2/detailed-process/

(there doesn't seem to be an English translation for this term)', and **'البناء' 'marriage consummation'** is translated as **'construction'.** The high critical severity rate, considering the number, dimensions, and types of errors, underscores the importance of addressing these translation issues, especially in legal contexts where precision is paramount.

Conversely, in the In-Context scenario (Table 5 & 9) for AR→EN both models, despite the slight variations in favor of ChatGPT-4, face a notable failure in performance when contextual information was provided. This suggests that the models performance did not benefit from context. If we consider example 2 from the Moroccan Family Code[12] and pay careful attention to the translation of the terms {**مستحقات**}, {**الصداق المؤخر**}, {**العدة**}, {**المتعة**}:

2. **AR:** إن تشمل **مستحقات** الزوجة: **الصداق المؤخر** وجد، ونفقة **العدة**، و**المتعة** التي يراعى في تقديرها فترة الزواج والوضعية المالية للزوج.

   EN (ChatGPT-4): the wife's **dues** include: the **deferred dower**, if any, the maintenance of the **period of waiting**, and the **pleasure** that is taken into account in its estimation of the period of marriage and the financial status of the husband.

   EN (Gemini): The wife's **entitlements** include: **delayed dowry** if applicable, post-marriage support, and **temporary marriage**, which takes into account the duration of the marriage, the husband's financial situation.

   HT: The wife's **entitlements** include: the **rest of the dowry**, if any, the **waiting period** expenses, and the **compensation** that is estimated according to the marriage duration and the financial situation of the husband.

we find out that the two models fail to accurately translate the legal terms highlighted and capture their context. The translations demonstrate several pitfalls including the models' inconsistent use of terminology, the use of wrong terms, word order that is not compliant with the target language norms. They also fail render how **'compensation'** is estimated which is a critical and important information, part of acknowledging the husband duties towards the wife after divorce. The models results also lack legal requirements and undertranslate or mistranslate legal terms in a legal system where religion, culture, and law meet and mutually influence each other. The critical severity rate and the EC remaining high in both models indicate persistent translation issues that need attention.

In the In-Context AR→FR pair, both ChatGPT-4 and Gemini demonstrate a high critical number of errors than the major and minor ones compared to the Out-of-Context (see Table 7 & 11). In Out-of-Context (see Table 6 & 10), they score a higher number of minor and major errors than the critical ones. In other words, ChatGPT-4 approximately generates around 28% and Gemini 44% of correct terms in Out-of-Context, and only 7% (ChatGPT-4) and 15% (Gemini) of correct phrases in in In-Context Scenario with, however, higher ETW score in both scenarios for both models i.e., 5 for terminology, 3 for style, 3 for linguistic conventions, 5 for accuracy, and 5 for audience appropriateness. This implies that the two models struggle with leveraging and disambiguating legal phrases that solely rely on contextual and cultural knowledge, leading to inaccuracy, transliterations, non-compliant results with legal requirements, and violation of culture-specific references etc.

Our terminological dataset consists of judicial documents (i.e., contracts, provisions, codes, decrees, etc.) of different Arab countries (Morocco, Algeria, Tunisia, United Arab Emirates, Saudi Arabia, Egypt). Therefore, the use of distinct legal terminology to convey similar legal practices in different countries can significantly impact the outcomes of MT for Arabic. Due to variations in legal systems, cultural nuances, idiomatic expressions, linguistic variations, and the specific precision required in legal language, ChatGPT-4 and Gemini may struggle to accurately capture the intended meanings. This could lead to mistranslations, misinterpretations, and errors that have potentially serious legal consequences. For example, the term **{مأذون}** is used mostly in Qatar and Egypt. It is used to refer to the person certified by the judge to perform certain legal formalities, especially to draw up or certify marriage contracts, deeds, and other documents for use in other jurisdictions[13]. ChatGPT-4, however, translates it as **'authorized'** into English and **'autorisé'** into French. Whereas Gemini, as well, translates it as 'authorized' into English and **'autorisé'** into

[12] https://shorturl.at/EG567

[13] https://www.almaany.com/ar/dict/ar-ar/

French. Therefore, we notice that both models not only transform the grammatical category of the term from a noun, which represents a person into an adjective, but they also misinterpret the intended legal practice in the target legal systems. Hence, in France, the equivalence of {مأذون} is **'maire'** (i.e., the person who chairs the municipal council[14]), he/she is the one who oversees approving and drawing up marriage contracts. Whereas in England the person in charge of approving and celebrating the marriage requests is called the **'superintendent registrar[15]'** of the district. This unveils that these models are not trained on a diverse and comprehensive dataset that covers a wide range of legal terminologies from different countries. In other words, these models need to be equipped with region-specific legal dictionaries and context-aware algorithms that consider the nuances of each country's legal language. Additionally, leveraging parallel legal texts in different terms can help train the models to better handle these variations.

4 Conclusion and Future Work

The results show that Gemini surpasses ChatGPT-4 in achieving less critical errors in Out-Of-Context scenario from AR→EN, conversely, in In-Context scenario, ChatGPT-4 demonstrates less critical results, but more minor and major errors compared to Gemini with only slight differences regarding the critical error severity count. In AR→FR pair, ChatGPT-4 and Gemini achieve better results in Out-of-Context scenarios but face significant challenges with accurately capturing the nuances and complexities of legal terminology when the context is provided. This is because the complexity of legal language and the terminology used to express certain concepts that are culturally and religiously bound present critical barriers. Secondly, these models are pre-trained on vast amount of text from diverse sources and on a large amount of unlabeled corpora, and they might not have access to sufficient amounts of domain-specific texts for every possible domain. In this regard, we believe that efforts should be invested on enhancing context sensitivity of LLMs' MT tasks by i) developing models capable of handling longer text sequences, such as Transformer-XL (Dai et al., 2019), which can capture broader dependencies, ii) incorporating hierarchical modeling to process text at multiple levels (sentence, paragraph, document) to ensure better context continuity, iii) training with datasets that preserve discourse structure and using dynamic context integration techniques, like Retrieval-Augmented Generation (RAG) (Gao et al., 2023) which combines generative models with retrieval mechanisms to dynamically incorporate relevant context from external sources during the generation process. Indeed, MQM offers an insightful and comprehensive approach to evaluating machine-translated legal terms for Arabic by ChatGPT-4 and Gemini. This framework enables a granular assessment of the accuracy state of the models better than any automatic metric would do, by considering various dimensions such as terminology, style, linguistic convention, accuracy, and audience appropriateness, and their respective sub-dimensions. This detailed analysis provides us with actionable feedback for improvement as it presents detailed insights that go beyond automatic metrics.

In the future, we will consider fine-tuning these models with extra context to improve their in-context learning abilities, as well as the translation quality and adherence to the domain terminology and style. We might also consider the use of dynamic context integration techniques like RAG to enhance the models ability to handle complex and context-dependent translations by providing them with access to a broader and more relevant set of contextual information during the translation process. Additionally, we will extend this evaluation framework to additional domains and language pairs and include more models to not only capture the understanding of the models capabilities and limitations but to also systematically improve the MQM dimensions so that to ensure it remains a relevant, accurate, and actionable metric for such models.

[14] EESC/COR-FR, d'après le Conseil des communes et régions d'Europe (CCRE), «Gouvernements locaux et régionaux en Europe — Structures et compétences» (2016) (3.5.2022), page 26

[15] Term reference: https://www.citizensadvice.org.uk/family/living-togethermarriage-and-civil-partnership/getting-married/

References

Almahasees, Z. (2021). Analysing English-Arabic Machine Translation: Google Translate, Microsoft Translator and Sakhr. Routledge.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N.M., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., García, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A.M., Pillai, T.S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Díaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K.S., Eck, D., Dean, J., Petrov, S., & Fiedel, N. (2022). PaLM: Scaling Language Modeling with Pathways. J. Mach. Learn. Res., 24, 240:1-240:113.

Alves, D., Guerreiro, N., Alves, J., Pombal, J., Rei, R., de Souza, J., Colombo, P., and Martins, A. (2023). Steering Large Language Models for Machine Translation with Fine tuning and In-Context Learning. In Bouamor, H., Pino, J., and Bali, K., editors, Findings of the Association for Computational Linguistics: EMNLP 2023, pages 11127–11148, Singapore. Association for Computational Linguistics.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., 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.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language Models are Few-Shot Learners. ArXiv, abs/2005.14165.

Burchardt, A. (2013). Multidimensional quality metrics: a flexible system for assessing translation quality. In Proceedings of Translating and the Computer 35.

Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., Yang, L., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P.S., Yang, Q., & Xie, X. (2023). A Survey on Evaluation of Large Language Models. ACM Transactions on Intelligent Systems and Technology, 15, 1 - 45.

Darwish, A. (2009). Terminology and translation: A phonological-semantic approach terminology. Writescope Publishers.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J.G., Le, Q.V., & Salakhutdinov, R. (2019). Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. Annual Meeting of the Association for Computational Linguistics.

ElFqih, K.A., Di Buono, M.P., Monti, J. (2024). Towards a Linguistic Annotation of Arabic Legal Texts: A Multilingual Electronic Dictionary for Arabic. In: Bartulović, A., Mijić, L., Silberztein, M. (eds) Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities. NooJ 2023. Communications in Computer and Information Science, vol 1816. Springer, Cham. https://doi.org/10.1007/978-3-031-56646-2_5 .

ElFqih, K. A., & Monti, J. (2023, September). On the Evaluation of Terminology Translation Errors in NMT and PB-SMT In the Legal Domain: A Study on the Translation of Arabic Legal Documents into English and French. In Proceedings of the Workshop on Computational Terminology in NLP and Translation Studies (ConTeNTS) Incorporating the 16th Workshop on Building and Using Comparable Corpora (BUCC) (pp. 26-35).

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., & Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. ArXiv, abs/2312.10997.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. (2023). Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805,2023.

Winata, G.I., Madotto, A., Lin, Z., Liu, R., Yosinski, J., & Fung, P. (2021). Language Models are Few-shot Multilingual Learners. ArXiv, abs/2109.07684.

Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., ... & Mirjalili, S. (2023). A survey on large language models: Applications, challenges, limitations, and practical usage. Authorea Preprints.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W.D., Li, M., Liu, S., Liu, T., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., & Zhou, M. (2018). Achieving Human Parity on Automatic Chinese to English News Translation. ArXiv, abs/1803.05567.

Hendy, A., Abdelrehim, M.G., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y., Afify, M., & Awadalla, H.H. (2023). How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. ArXiv, abs/2302.09210.

Herold, C., Rosendahl, J., Vanvinckenroye, J., & Ney, H. (2022). Detecting various types of noise for neural machine translation. Findings of the Association for Computational Linguistics: ACL 2022. http://dx.doi.org/10.18653/v1/2022.findings acl.200

Huang, Y. (2024). Leveraging Large Language Models for Enhanced NLP Task Performance through Knowledge Distillation and Optimized Training Strategies. arXiv preprint arXiv:2402.09282.

Iyer, V., Chen, P., and Birch, A. (2023). Towards Effective Disambiguation for Machine Translation with Large Language Models. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, Proceedings of the Eighth Conference on Machine Translation, pages482–495, Singapore. Association for Computational Linguistics.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. North American Chapter of the Association for Computational Linguistics.

Gao, J., & Lin, C. (2004). Introduction to the special issue on statistical language modeling. ACM Trans. Asian Lang. Inf. Process., 3, 87-93.

Jiao, W., Huang, J.-T., Wang, W., Wang, X., Shi, S., and Tu, Z. (2023). ParroT: Translating During Chat Using Large Language Models. arXiv preprint arXiv:2304.02426 [cs.CL].

Killman, J. (2014). Vocabulary accuracy of statistical machine translation in the legal context. Proceedings of the 11th Conference of the Association for Machine Translation in the Americas, 85–98.

Koehn, P. (2020). Neural machine translation. Cambridge University Press.

Lai, Jinqi & Gan, Wensheng & Wu, Jiayang & Qi, Zhenlian & Yu, Philip. (2023). Large Language Models in Law: A Survey. 10.13140/RG.2.2.15031.09124.

Mediouni, M. (2016). Towards a functional approach to Arabic–english legal translation: The role of comparable/parallel texts. In M. Taibi (Ed.), New Insights into Arabic Translation and Interpreting (pp. 115–160). Multilingual Matters. http://dx.doi.org/10.21832/9781783095254-008

Mesfar, S. (2008). Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard (Doctoral dissertation, Besançon).

Moslem, Y., Haque, R., & Way, A. (2023). Fine-tuning large language models for adaptive machine translation. arXiv preprint arXiv:2312.12740.

Moslem, Y., Haque, R., Kelleher, J. D., & Way, A. (2023). Adaptive machine translation with large language models. arXiv preprint arXiv:2301.13294.

Rossi, C., & Carre, A. (2022). Machine translation for everyone: Empowering users in the age of artificial intelligence. Language Science Press Berlin, 18, 51. https://doi.org/10.5281/zenodo.6653406

Sager, J. C. (1990). Practical course in terminology processing. John Benjamins Publishing.

Schioppa, A., Garcia, X., and Firat, O. (2023). Cross-Lingual Supervision improves Large Language Models Pre-training. arXiv preprint arXiv:2305.11778 [cs.CL].

Sia, S. and Duh, K. (2022). Prefix Embeddings for In-context Machine Translation. In Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track), pages 45–57, Orlando, USA. Association for Machine Translation in the Americas.

Silberztein, M.: La formalisation des langues : l'approche de NooJ. ISTE Editions, London (2015).

Siu, S. C. (2023). Chatgpt and GPT-4 for professional translators: Exploring the potential of large language models in translation. Available at SSRN 4448091.

Kombrink, S., Mikolov, T., Karafiát, M., & Burget, L. (2011). Recurrent neural network-based language modeling in meeting recognition. Interspeech.

Sun, Z. (2023). A short survey of viewing large language models in legal aspect. arXiv preprint arXiv:2303.09136.

Thierry, P. (2022). On" Human Parity" and" Super Human Performance" in Machine Translation Evaluation. Language Resource and Evaluation Conference.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., 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.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language Models are Few-Shot Learners. ArXiv, abs/2005.14165.

Xu, H., Kim, Y. J., Sharaf, A., and Awadalla, H. H. (2023). A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models. arXiv preprintarXiv:2309.11674 [cs.CL].

Yang, A., Xiao, B., Wang, B., Zhang, B., Bian, C., Yin, C., Lv, C., Pan, D., Wang, D., Yan, D., Yang, F.,

Deng, F., Wang, F., Liu, F., Ai, G., Dong, G., Zhao, H., Xu, H., Sun, H., Zhang, H., Liu, H., Ji, J., Xie, J., Dai, J., Fang, K., Su, L., Song, L., Liu, L., Ru, L., Ma, L., Wang, M., Liu, M., Lin, M., Nie, N., Guo, P., Sun, R., Zhang, T., Li, T., Li, T., Cheng, W., Chen, W., Zeng, X., Wang, X., Chen, X., Men, X., Yu, X., Pan, X., Shen, Y., Wang, Y., Li, Y., Jiang, Y., Gao, Y., Zhang, Y., Zhou, Z., and Wu, Z. (2023). Baichuan2: Open Large-scale Language Models. arXiv preprint arXiv:2309.10305 [cs.CL].

Zakraoui, J., Saleh, M., Al-Maadeed, S., & Alja'am, J. M. (2021). Arabic machine translation: A survey with challenges and future directions. IEEE Access, 9,161445–161468. https://doi.org/10.1109/access.2021.3132488

Zhang, X., Rajabi, N., Duh, K., and Koehn, P. (2023). Machine Translation with Large Language Models: Prompting, Few-shot Learning, and Fine-tuning with QLoRA. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, Proceedings of the Eighth Conference on Machine Translation, pages 468–481, Singapore. Association for Computational Linguistics

A MQM Scorecards

	ChatGPT-4					
	Out-of-Context AR→EN					
Error Severity Multiplier	**Neutral**	**Minor**	**Major**	**Critical**	**Error Type Penalty**	
Severity Penalty Multiplier	1	1	5	25	**Error Type Weight**	**Error Type Penalty Totals**
Error Types Dimension	**Error Count**					
Terminology	0	6	10	979	**5**	613,275
Style	0	15	40	950	**3**	213,750
Linguistic Convention	0	90	55	430	**3**	96,750
Accuracy	0	23	35	916	**5**	572,500
Audience Appropriateness	0	23	11	959	**5**	601,325
			Absolute Penalty Total			2,097,600

Table 4: MQM Evaluation Scorecard of ChatGPT-4 performance in Out-of-Context AR→EN

	ChatGPT-4					
	In-Context AR→FR					
Error Severity Multiplier	**Neutral**	**Minor**	**Major**	**Critical**	**Error Type Penalty**	
Severity Penalty Multiplier	1	1	5	25	**Error Type Weight**	**Error Type Penalty Totals**
Error Types Dimension	**Error Count**					
Terminology	0	8	35	958	**5**	603,325
Style	0	12	33	955	**3**	216,468
Linguistic Convention	0	11	66	787	**3**	180,144
Accuracy	0	8	23	970	**5**	609,325
Audience Appropriateness	0	5	27	970	**5**	609,750
			Absolute Penalty Total			2,219.012

Table 5: MQM Evaluation Scorecard of ChatGPT-4 performance in In-Context AR→FR.

	Gemini					
	Out-of-Context AR→EN					
Error Severity Multiplier	Neutral	Minor	Major	Critical	Error Type Penalty	
Severity Penalty Multiplier	1	1	5	25	**Error Type Weight**	**Error Type Penalty Totals**
Error Types Dimension	**Error Count**					
Terminology	0	2	23	973	5	611,050
Style	0	11	33	936	3	212,184
Linguistic Convention	0	10	44	446	3	102,420
Accuracy	0	5	25	879	5	552,625
Audience Appropriateness	0	14	21	971	5	609,850
				Absolute Penalty Total		2,088.129

Table 6: MQM Evaluation Scorecard of ChatGPT-4 performance in Out-of-Context AR→EN.

	Gemini					
	In-Context AR→EN					
Error Severity Multiplier	**Neutral**	**Minor**	**Major**	**Critical**	**Error Type Penalty**	
Severity Penalty Multiplier	1	1	5	25	**Error Type Weight**	**Error Type Penalty Totals**
Error Types Dimension	**Error Count**					
Terminology	0	2	55	941	**5**	595,050
Style	0	5	13	962	**3**	217,080
Linguistic Convention	0	2	11	487	**3**	110,088
Accuracy	0	8	54	847	**5**	536,325
Audience Appropriateness	0	6	55	945	**5**	597,650
				Absolute Penalty Total		2,056.193

Table 7: MQM Evaluation Scorecard of Gemini performance in In-Context AR→EN.

	Gemini					
	Out-of-Context AR→FR					
Error Severity Multiplier	**Neutral**	**Minor**	**Major**	**Critical**	**Error Type Penalty**	
Severity Penalty Multiplier	1	1	5	25	**Error Type Weight**	**Error Type Penalty Totals**
Error Types Dimension	**Error Count**					
Terminology	0	2	55	805	**5**	510,050
Style	0	12	21	917	**3**	207,348
Linguistic Convention	0	15	32	703	**3**	159,750
Accuracy	0	11	39	851	**5**	537,025
Audience Appropriateness	0	12	42	909	**5**	573,675
				Absolute Penalty Total		1,987.848

Table 8: MQM Evaluation Scorecard of Gemini performance in Out-of-Context AR→FR.

	ChatGPT-4					
	In-Context AR→EN					
Error Severity Multiplier	**Neutral**	**Minor**	**Major**	**Critical**	**Error Type Penalty**	
Severity Penalty Multiplier	1	1	5	25	**Error Type Weight**	**Error Type Penalty Totals**
Error Types Dimension	**Error Count**					
Terminology	0	25	60	915	**5**	580,000
Style	0	16	58	913	**3**	208,179
Linguistic Convention	0	54	56	650	**3**	149,256
Accuracy	0	17	52	930	**5**	587,875
Audience Appropriateness	0	13	55	934	**5**	590,950
			Absolute Penalty Total			2,116.26

Table 9: MQM Evaluation Scorecard of ChatGPT-4 performance in In -Context AR→EN.

	ChatGPT-4					
	Out-of-Context AR→FR					
Error Severity Multiplier	**Neutral**	**Minor**	**Major**	**Critical**	**Error Type Penalty**	
Severity Penalty Multiplier	1	1	5	25	**Error Type Weight**	**Error Type Penalty Totals**
Error Types Dimension	**Error Count**					
Terminology	0	20	65	913	**5**	579,250
Style	0	55	65	648	**3**	149,220
Linguistic Convention	0	54	67	644	**3**	148,806
Accuracy	0	11	23	887	**5**	557,525
Audience Appropriateness	0	9	22	971	**5**	609,850
			Absolute Penalty Total			2,044.651

Table 10: MQM Evaluation Scorecard of ChatGPT-4 performance in Out-of-Context AR→FR.

	Gemini					
	In-Context AR→FR					
Error Severity Multiplier	**Neutral**	**Minor**	**Major**	**Critical**	**Error Type Penalty**	
Severity Penalty Multiplier	1	1	5	25	**Error Type Weight**	**Error Type Penalty Totals**
Error Types Dimension	**Error Count**					
Terminology	0	3	30	967	**5**	608,200
Style	0	9	22	958	**3**	216,621
Linguistic Convention	0	5	32	950	**3**	215,235
Accuracy	0	2	15	982	**5**	615,675
Audience Appropriateness	0	3	44	954	**5**	120,425
			Absolute Penalty Total			1,776.156

Table 11: MQM Evaluation Scorecard of Gemini performance in Out-of-Context AR→FR.