ISSN 1798-4769 (print); ISSN 2053-0684 (online)
Journal of Language Teaching and Research, Vol. 17, No. 3, pp. 1019-1027, May 2026
DOI: https://doi.org/10.17507/jltr.1703.26

A DQF-MQM Evaluation of Machine Translation in English–Arabic Customs Law

Karam Damseh
School of Languages, Literacies and Translation, Universiti Sains Malaysia, Malaysia

Mozhgan Ghassemiazghandi*
School of Languages, Literacies and Translation, Universiti Sains Malaysia, Malaysia

***Abstract*—This study presents a corpus-based diagnostic evaluation of Google Translate's performance in translating the Revised Kyoto Convention (RKC) from English into Arabic. Building on the "Adequacy-Fluency Tradeoff" observed in recent literature, researchers employ an explanatory sequential triangulation design to assess whether high-resource NMT models can satisfy the strict compliance requirements of Customs law. The methodology applies the COMET metric to a stratified test suite of the RKC corpus (N = 1,000 segments) to establish a quantitative baseline. To diagnose potential metric discordance, researchers conducted a targeted human evaluation: a Likert-scale assessment of the highest-scoring segments (n = 100) to verify the "performance ceiling," followed by a granular DQF-MQM error analysis of the lowest-scoring segments (n = 200) conducted by a panel of Experts. The analysis identifies a profound discordance within the RKC corpus: while the system achieved a high mean COMET score of 87.70, the detailed analysis of the "performance floor" revealed 1,579 discrete errors. The resulting penalty score (1,292.5) exceeded the professional acceptance threshold for this text type by a factor of 5.7. These findings indicate that, within this specific legal corpus, Google Translate exhibits "specification-blindness," systematically failing to disambiguate polysemous "Terms of Art" required for international trade compliance. The research concludes that for the RKC, automated metrics do not yet serve as a reliable proxy for legal review, and that a Human-in-the-Loop framework remains an absolute prerequisite for the deployment of Google Translate in Customs administration.**

***Index Terms*—compliance hazard, Google Translate, metric discordance, neural machine translation, Revised Kyoto Convention**

I. Introduction

A. The Macro-Crisis: The "Compliance Hazard" in Automated Law

The rapid integration of Neural Machine Translation (NMT) into international regulatory frameworks has precipitated a "compliance hazard" that traditional evaluation paradigms fail to capture. As global trade volumes expand, Customs administrations face increasing pressure to overcome the "document lag" of manual translation through automation. While NMT has achieved remarkable milestones in linguistic fluency—often reaching "human parity" in general domains (Wang et al., 2023)—this surface-level success masks a fundamental "Adequacy-Fluency Tradeoff" (Shayegh et al., 2025). In high-stakes legal environments, improvements in neural smoothness often correlate with a degradation in semantic fidelity, creating a phenomenon where, as Elsayed (2025) observes, high fluency scores (99.2%) can coexist with critical terminology failures that 'jeopardize the machine translation (MT) performance' in high-stakes legal arguments.

For Customs administrations, this tradeoff is not merely a linguistic curiosity but a systemic risk. A translation system that prioritizes the natural flow of the target language over the rigid semantic boundaries of the source text can trigger international trade disputes, revenue leakage, and security vulnerabilities. As Elsayed (2025) demonstrates in her analysis of International Court of Justice (ICJ) proceedings, even minor linguistic deviations in legal arguments can 'deeply weaken the overall message' and misrepresent core legal logic.

This phenomenon represents what Al Maaytah (2025) describes as the 'double hand' of MT: while it enables swift linguistic interchange, it simultaneously endangers the preservation of linguistic integrity. In the legal domain, this risk is most acute in the failure to accurately render modality—the precise linguistic encoding of obligation, prohibition, and permission (Alkatheery, 2023). In typologically distant language pairs like English and Arabic, this 'mediated translation process' often leads to a reduction in translation fidelity that aggregate metrics are ill-equipped to detect (Al Maaytah, 2025).

B. The Metric Conflict: The Granularity Gap

This reliability crisis is exacerbated by a persistent 'granularity gap.' While automated metrics correlate well with human judgment at the aggregate system level, recent shared tasks reveal a collapse in performance at finer resolutions. Lavie et al. (2025) report that current auto-raters 'fail to localize errors,' with the best systems rarely exceeding 35% F1

* Corresponding Author. Email: mozhgan.ghassemi@gmail.com

on span-level detection. Similarly, Zerva et al. (2024) conclude that despite strong sentence-level performance in Quality Estimation, there remains 'ample room for improvement' in fine-grained error span detection. This lack of transparency is a known limitation of widely used learned metrics, which estimate quality via a single scalar score but offer little insight into translation errors or their specific severity (Guerreiro et al., 2024). As Guerreiro et al. (2024) argue, the field is currently shifting toward more granular strategies that attempt to detail and categorize errors through span detection, bridging the gap between opaque "black-box" scores and human-interpretable diagnostics.

This "metric bias" (Kovacs et al., 2024) often rewards "reward hacking" where the model produces high-probability, fluent sequences that satisfy the metric's embedding model but fail the specific utility function of legal adequacy. Consequently, aggregate scores often obscure a "performance paradox": a translation may receive a high COMET score while containing a "Critical" error that renders the entire document legally void. This discrepancy is further evidenced by recent evaluations using professional certification standards. Zou et al. (2025) found that while COMET scores for English-Arabic translations remained uniformly high (averaging 0.97), manual assessments by certified graders revealed significant distinctions in quality, with aggregate metrics often overlooking an accumulation of low-severity errors that 'materially degrades professional-level adequacy'.

C. The Domain Challenge: The Task-Data Mismatch in Arabic

The challenges of NMT are further amplified in the context of Arabic, a morphologically rich language that remains low-resource in specialized domains such as finance and law (Alghamdi et al., 2024). Recent scholarship attributes the reliability gap in this language pair to a fundamental "Task-Data Mismatch" (Uhlig et al., 2025). General-purpose models like Google Translate are trained on massive, interdisciplinary web corpora that diverge significantly from the high-precision requirements of legislative drafting. This challenge is rooted in the high degree of lexical ambiguity inherent in Arabic, where homonyms and polysemes frequently lead to systemic misunderstandings in traditional MT systems (Aldawsari, 2025).

As Semenov et al. (2025) clarify, while sentence-level terminology has reached near-saturation in general domains, performance degrades precipitously in document-level contexts. This contextual decay is particularly dangerous in Customs law, where terminological consistency must be maintained across complex regulatory frameworks to ensure legal certainty. Furthermore, the translation of lexical collocations remains a significant hurdle for English-Arabic MT (Sabtan et al., 2024). Recent evaluations of Google Translate in this language pair reveal an adequacy rate of only 60% for collocations, with literal translation and omission identified as the primary causes of systemic failure (Sabtan et al., 2024). Recent evaluations confirm that even state-of-the-art NMT systems struggle to match the sophisticated requirements of Arabic's gendered grammar and syntactic flexibility (Shaalan et al., 2019; Al Maaytah, 2025). Al Maaytah (2025) identifies syntactic errors (28%) and semantic mismatches (24%) as the most prevalent challenges in this language pair, suggesting that NMT tools frequently reduce the complexity of syntax or fail to uphold the specific meanings intended by human authors.

D. The Research Intervention: A Diagnostic Analysis of the State-of-the-Art

To navigate this disruption, the translation industry has shifted toward structured quality management, culminating in the publication of ISO 11669:2024 and ISO 5060:2024 (Strandvik, 2025). These standards prioritize explicit "translation specifications"—such as purpose, audience, and risk tolerance—as the non-negotiable foundation for evaluation. A critical limitation of current NMT architectures, as identified by Kayano and Sugawara (2025), is the specification gap: models operate without awareness of the "translation brief" (Skopos), defaulting instead to the most probable statistical output. In the Customs domain, where the RKC serves as the "blueprint" for Customs administration of global trade (World Customs Organization [WCO], 1999), this lack of specification awareness leads to outputs that are linguistically valid but functionally disastrous.

This study addresses these critical gaps by providing a comprehensive diagnostic evaluation of Google Translate's performance in the English-Arabic Customs domain. Researchers select Google Translate not as a competitor in a comparative analysis, but as a proxy for the current State-of-the-Art (SOTA) in widely accessible NMT. Moving beyond surface-level scoring, researchers adopt a functionalist approach grounded in the 'Five Perspectives' of translation quality (Fields et al., 2014), recently reinvigorated by Strandvik (2025) to address the specific risks of automated translation. By triangulating automated COMET metrics with a granular DQF-MQM error analysis conducted by a panel of experts, the research aims to identify the specific failure points where NMT fluency masks legal liability. In doing so, this paper provides actionable insights for the integration of NMT into professional legal workflows and establishes a quantitative baseline for the "Human-in-the-Loop" (HITL) requirements necessary for autonomous deployment in specialized Customs administration.

II. Literature Review

A. The Metric Evolution: From Lexical Overlap to Neural Semanticity

The evaluation of English-Arabic MT is currently undergoing a paradigm shift from surface-level lexical metrics to deep neural architectures. Traditional metrics like BLEU have been increasingly criticized for their inability to capture the expressive integrity and contextual nuances required for morphologically rich languages (Wang et al., 2023; Gamal

et al., 2023). This inadequacy is particularly acute in Arabic, where lexical similarity does not equate to semantic adequacy (Babaali & Salem, 2022). To address this, recent scholarship has introduced "linguistic-based" evaluation methods that incorporate POS tagging and multilingual BERT models to penalize errors in sentence position and context (Beseiso et al., 2022). However, as Al-Khalifa et al. (2024) demonstrate, even advanced neural metrics like BERTScore and COMET can exhibit substantial limitations when faced with Arabic's morphological complexity, necessitating a return to human-centric, multidimensional frameworks.

To systematically uncover the weaknesses of neural metrics, recent research has turned to "challenge sets" designed to probe specific linguistic phenomena. The Translation Adequacy Challenge Set (ACES), introduced by Moghe et al. (2024), benchmarks metrics across 68 error phenomena based on the MQM ontology. Their large-scale study of 47 metrics revealed that reference-based neural metrics still rely excessively on surface-level overlap and frequently ignore information present in the source (Moghe et al., 2024). Furthermore, Moghe et al. (2024) demonstrate that LLM-based evaluators fail to provide reliable segment-level performance, often producing peaked distributions and hallucinations that render them less effective than traditional string-overlap metrics in zero-shot settings. These findings validate the need for the granular, span-based human analysis employed in this study.

B. The "Fluency Trap" in Arabic NMT: Morpho-Syntactic Divergence

A recurring theme in recent English-Arabic MT research is the "Fluency Trap"; a phenomenon where NMT systems produce grammatically plausible but semantically vacant outputs. Zakraoui et al. (2021) and Ameur et al. (2020) identify Arabic's complex syntax, diacritics, and long-distance dependencies as primary triggers for NMT failure. This is further evidenced by Nagi (2023), who found that when translating complex English relative clauses into Arabic, Google Translate produced a high volume of fluency errors (70.23%), suggesting that the system's ability to maintain 'linguistic well-formedness' collapses when faced with the recursive syntax of the target language. This divergence is often attributed to the 'Task-Data Mismatch' (Uhlig et al., 2025), as general-purpose models are typically trained on interdisciplinary web corpora that fail to capture the specific constraints of specialized domains (Ameur et al., 2020), which leads to "literalism" and "wrong equivalents" in colloquial or specialized registers (Sabtan et al., 2021).

Recent evaluations confirm that even state-of-the-art NMT systems struggle to match the sophisticated requirements of Arabic's gendered grammar and syntactic flexibility (Shaalan et al., 2019; Al Maaytah, 2025). Al Maaytah (2025) identifies syntactic errors (28%) and semantic mismatches (24%) as the most prevalent challenges in this language pair, suggesting that NMT tools frequently reduce the complexity of syntax or fail to uphold the specific meanings intended by human authors. This linguistic barrier is a persistent theme in recent surveys of the field. Idrysy et al. (2025) observe that while significant progress has been made in neural architectures, the quality of Arabic MT systems still lags behind that of some other languages. This inherent lag necessitates a more rigorous, domain-specific evaluation to identify the linguistic and technical challenges that general-purpose systems fail to resolve in specialized contexts (Idrysy et al., 2025).

C. Fine-Grained Error Analysis: The DQF-MQM Turn

To move beyond scalar scores, the field has pivoted toward fine-grained error typologies. The Harmonized DQF-MQM framework (Lommel, 2018) has emerged as the industry standard for diagnosing specific translation failures. Recent large-scale evaluations by Freitag et al. (2021) demonstrate that MQM-based human annotation provides a far more reliable signal of translation quality than scalar ratings. Furthermore, the framework's capacity to distinguish error severity allows for the identification of 'Critical' failures that pose safety risks (Lommel, 2018). In expressive domains, Fakih et al. (2024) report a 90% failure rate, with 'Critical' adequacy errors compromising the core message of literary texts. Conversely, in informative news contexts, Abdelaal and Alazzawie (2020) observe that while the majority of NMT errors are minor, the system persistently struggles with lexical omissions and semantic ambiguities arising from homonymy. This suggests that as the 'expressive load' of a text increases—which is the case in Customs Law—the severity of NMT failure escalates from cosmetic to critical.

For instance, in medical and literary contexts, semantic and syntactic errors remain the most frequent, requiring significant human intervention to reach professional standards (Almahasees et al., 2021; Fakih et al., 2024). This underscores a critical consensus: automated translation systems can provide a general understanding, but they cannot yet serve as a proxy for legal or technical review (Ali, 2020).

This shift is further driven by the "flattening" of meaning in NMT. A critical concern in modern linguistics is the tendency of NMT to "flatten" or oversimplify culturally and legally embedded meanings (Al Maaytah, 2025). While systems like Google Translate perform adequately on technical or literal prose where structures are predictable, they often destabilize when faced with context-dependent translations (Al Maaytah, 2025). This bifurcation of performance necessitates a mixed-methods approach that combines benchmark-based scoring with expert human review to provide a holistic evaluation of MT performance (Al Maaytah et al., 2025; Al Maaytah, 2025).

D. The Domain-Specific Gap: From Finance to Customs Law

While domain adaptation has shown promise in specialized fields, a significant regulatory gap remains in literature. Studies in the financial (Alghamdi et al., 2024) and industrial (Hameed et al., 2022) sectors demonstrate that fine-tuning pre-trained models on domain-specific parallel corpora can significantly outperform general-purpose engines like Google Translate. However, in the legal domain, the register-related failures of NMT remain a primary concern. Alkatheery (2023)

observes that Google Translate systematically struggles with the preservation of legal discourse features, often failing to disambiguate specialized terminology. In addition, NMT systems continue to suffer from 'literal translation practices' and inadequacies in style and register, which can lead to 'semantically hollow' outputs in specialized domains (Al Maaytah, 2025). Nevertheless, as AlAfnan (2025) argues, while AI has made substantial strides, 'human oversight' remains indispensable to ensure contextual adequacy in specialized fields like law.

Despite these findings, there is a marked scarcity of empirical research addressing the Customs domain, where the intersection of international trade law under Customs administration and Arabic regulatory syntax creates a unique set of high-liability challenges that this study aims to address.

III. Methodology

A. Research Design: Explanatory Sequential Triangulation

This study employs a mixed-methods, explanatory sequential research design to evaluate the performance of Google Translate in the specialized domain of Customs law. To mitigate the biases inherent in any single evaluation instrument (Kit & Wong, 2023), the methodology achieves a triangulation perspective by integrating automated neural metrics with granular human expert analysis. This approach is designed to capture both the "performance ceiling" (optimal output) and the "performance floor" (systemic failure points), providing a detailed account of how neural fluency interacts with legal precision.

B. Corpus Selection and Stratification: The "Test Suite" Approach

The primary instrument for this study is the RKC, the foundational international legal instrument governing cross-border trade under Customs administration (WCO, 1999). Standard randomized test sets frequently mask translation deficiencies by averaging performance across disparate sentence types (Mukherjee & Shrivastava, 2023). To rigorously assess the system's competence in Customs law—a domain characterized by the complex syntactic structures found in 'Judgments' and 'Technical-writing'—researchers adopted the 'Test Suite' methodology. Following the protocol in Mukherjee and Shrivastava (2023), which prioritizes clustering by linguistic features such as Sentence Length (SL), researchers stratified the 1,000-segment parallel English-Arabic corpus (31,655 words) into three complexity tiers. This stratification also addresses the 'complexity degradation' phenomenon often observed in NMT (Koehn & Knowles, 2017; Zakraoui et al., 2021):

Short: ≤20 words (n=198 segments, 19.6%)
Medium: 21-40 words (n=498 segments, 49.4%)
Long: >40 words (n=313 segments, 31.0%)

By isolating these categories, researchers move beyond monolithic scoring to perform a targeted analysis of how syntactic density impacts semantic coherence (Singh et al., 2025).

C. The Expert Panel: Ensuring Ecological Validity

To ensure high ecological validity (Läubli et al., 2018), the evaluation was conducted by a panel of three expert annotators with deep domain-specific expertise. The panel included two Senior Translators concurrently serving as Customs Officers (10+ years of experience) and one Senior Linguist specializing in Arabic (7+ years of experience). This panel represents the "professional gold standard," ensuring that the evaluation of "Critical" errors is grounded in actual regulatory risk rather than abstract linguistic preference.

Inter-Rater Reliability (IRR): To validate the consistency of the human judgments in rating the Adequacy and Fluency of the top-tier segments (n=100), IRR was calculated using Krippendorff's Alpha (α) for ordinal data. The analysis yielded moderate to substantial agreement for Adequacy (α=0.51-0.58) and Fluency (α=0.53-0.58), confirming that the ratings reflect a shared professional consensus rather than individual annotator bias.

D. Evaluation Framework: Decoupling Fluency From Adequacy

The study utilizes a three-tiered evaluation framework designed to prevent metric masking:

Automated Metric (COMET): Researchers deployed the Cross-lingual Optimized Metric for Evaluation of Translation (COMET) via the MAchine Translation Evaluation Online (MATEO) (https://mateo.ivdnt.org/) platform (Vanroy et al., 2023). Unlike n-gram metrics, COMET utilizes cross-lingual embeddings to assess semantic similarity, a feature recently validated as a reliable proxy for professional judgment in multilingual settings (Uhlig et al., 2025).

Human Quality Rating (Likert Scale): A subset of high-ranked segments (n=100) was subjected to a five-point Likert scale for Adequacy and Fluency. Following Shayegh et al. (2025), these dimensions were analyzed independently to detect metric discordance.

Fine-Grained Error Analysis (DQF-MQM): The 200 lowest-ranked segments underwent an in-depth analysis using the Harmonized DQF-MQM Error Typology (Lommel, 2018). This choice is validated by Elsayed (2025), who successfully employed the DQF-MQM framework to diagnose critical terminology errors in English-Arabic legal transcripts, identifying it as a 'comprehensive and widely accepted' method for pinpointing the root causes of translation failure. As well as the work by Junczys-Dowmunt (2025) on GEMBA-MQM V2 demonstrates that MQM remains the most robust instrument for capturing fine-grained, severity-weighted terminology violations in high-stakes contexts.

E. Quality Assessment Scoring

To provide a definitive verdict on usability, a quantitative Quality Assessment was conducted using the standard TAUS DQF scoring algorithm. This methodology assigns weighted penalty points to errors based on their severity: 0.5 points for Minor errors, 1.0 point for Major errors, and 1.5 points for Critical errors. Consistent with industry benchmarks for "Business-to-Business" content, the maximum allowable penalty threshold was set at 50 points per 1,000 words. Any score exceeding this limit indicates that the translation quality is too low to be considered professionally viable without extensive re-working.

IV. Results and Discussion

A. The Performance Ceiling: Quantitative and Human Validation

The initial phase of the evaluation established the "performance ceiling" of the NMT system. Quantitatively, Google Translate demonstrated robust linguistic well-formedness across the full corpus (N=1,000), achieving a mean COMET score of 87.70 (SD=5.45; 95% CI [87.4, 88.0]). This score significantly exceeds the typical baseline for raw MT output, suggesting a high degree of semantic proximity to the professional human reference.

Moghe et al. (2024) demonstrate that reference-based neural metrics frequently over-rely on surface-level lexical overlap, leading to high scores even when the translation contains hallucinations or ignores the source context. To validate that the results indicate genuine quality rather than lexical artifacts, researchers conducted a human expert analysis of the top-tier segments (n=100). The results confirmed the system's capacity for high-performance output:

Adequacy: 97.0% of segments were rated as preserving "All" (68.3%) or "Much" (28.7%) of the source meaning.

Fluency: 96.7% of segments were rated as "Flawless" (78.0%) or "Good" (18.7%).

This data confirms the existence of a 'fluency trap'; a state where, as Shayegh et al. (2025) demonstrate, metrics lean so heavily toward fluency that they fail to penalize catastrophic semantic errors. At its best, the system produces output that is statistically indistinguishable from human translation, creating a false sense of security regarding its reliability in complex contexts.

B. The Performance Floor: Fine-Grained Error Analysis

The illusion of reliability collapses when the analysis shifts to the "performance floor." The detailed analysis of the lowest-ranked segments (n=200) using the DQF-MQM framework revealed a systemic breakdown in translation integrity. The expert panel identified a total of 1,579 discrete error instances within this subset, resulting in an average Error Density of 7.895 errors per failing segment.

C. The Severity Profile and Penalty Score

The application of the TAUS DQF scoring model resulted in a cumulative penalty score of 1,292.5. Given the sample size (~4,500 words), the maximum allowable penalty threshold for professional "Business-to-Business" acceptance is 225 points. The NMT output exceeded this safety threshold by a factor of 5.7, rendering it professionally unusable without extensive post-editing.

The severity distribution (Table 1) reveals a high-risk profile:

Table 1
Distribution of Error Severity by High-Level Category

High-level Error Type	Critical	Major	Minor	Frequency (n)	Percentage (%)
Adequacy	83	278	163	524	33.20%
Terminology	80	217	161	458	29.00%
Fluency	3	149	188	340	21.60%
Style	0	30	227	257	16.20%
TOTAL	166	674	739	1,579	
Percentage (%)	10.51%	42.69%	46.80%		100.0%

Crucially, the risk is not evenly distributed. 98.2% of all Critical errors (163/166) occurred within the Adequacy and Terminology categories. This confirms that while the system rarely fails grammatically (Fluency/Style), it systematically fails to convey the correct legal meaning. This discrepancy aligns with the "Metric Bias" identified by Kovacs et al. (2024), where models optimize for high-probability fluent sequences at the expense of the specific utility function of legal adequacy.

D. Qualitative Analysis: Mechanisms of Legal Failure

The diagnostic analysis identified three specific failure modes where the NMT system's lack of specification awareness—a 'specification gap' identified by Kayano and Sugawara (2025)—created direct legal liability.

(a). Critical Over-Translation (Hallucination)

The most dangerous failure mode identified was the addition of restrictive qualifiers not present in the source text.

- Source Text: "...processing of goods for home use..."
- NMT Output: ...al-istikhdam al-manzili (Literally: "residential/household use").

- Legal Implication: In Customs law, "Home Use" is a Term of Art referring to goods cleared for free circulation in the domestic market (whether industrial, commercial, or personal). By adding the semantic constraint "household" (manzili), the NMT system effectively rewrote the regulation to exclude industrial importers. This hallucination fundamentally misrepresents the scope of the law, creating a false legal condition that could trigger trade disputes. This mirrors findings by Zou et al. (2025), who noted that NMT systems often stick too closely to literal meanings, failing to recognize specialized domain constraints.

(b). Critical Under-Translation (Terminological Dilution)

The system frequently defaulted to general-domain definitions for polysemous terms, stripping them of their specific regulatory force.

- Source Text: "Issue of stores for consumption."
- NMT Output: Isdar makhazin lil-istihlak (Literally: "Issuing warehouses for consumption").
- Legal Implication: In the context of the RKC, "stores" refers to vessel supplies (fuel, food). The NMT system failed to suppress the high-probability association of "stores" with "shops/warehouses." The resulting translation is semantically nonsensical ("consuming a warehouse"), rendering the provision functionally inoperative. This failure provides a textbook example of distributional interference (Uhlig et al., 2025), where the model defaults to the most probable general definition rather than the domain-specific definition required by the Skopos. This mirrors the 'contextual blindness' observed by Elsayed (2025), where Google Translate failed to recognize 'The Wall' as a legal entity in ICJ texts, instead rendering it as a person with an opinion. In both cases, the NMT system prioritized the most probable general meaning over the specific jurisdictional definition.

(c). Critical Omission (Jurisdictional Authority)

The analysis revealed instances where the system omitted binding provisos, altering the constitutional balance of the text.

- Source Text: "A Customs or Economic Union... shall, for the matters within its competence, exercise in its own name the rights..."
- NMT Output: (Omitted the phrase "for the matters within its competence").
- Legal Implication: This omission transforms a conditional delegation of power into a blanket authorization. It implies that a Customs Union has unlimited authority to act on behalf of member states, overriding national sovereignty in areas (like criminal enforcement) where competence has not been transferred. While the Arabic output was grammatically fluent, it was legally unconstitutional. where terminological consistency must be maintained across complex regulatory frameworks to ensure the precise expression of legal modality; the linguistic encoding of obligation and permission (Alkatheery, 2023). As Elsayed (2025) argues, the failure of NMT to preserve these 'deontic' features in international legal discourse can 'deeply weaken the overall message' and lead to the misinterpretation of core regulatory logic.

E. The Complexity Penalty: Syntactic Degradation

Finally, the stratified analysis confirms that syntactic complexity is a primary trigger for these failures. While "Long" sentences (>40 words) constituted only 19.6% of the original corpus, they accounted for 33.0% of the total error sample.

This complexity degradation manifests as a loss of contextual integrity. As sentence length increases, the NMT attention mechanism struggles to maintain long-distance dependencies, leading to the decoupling of conditional clauses (e.g., "provided that...") from their subjects. This data validates the hypothesis by Al Maaytah (2025) that current NMT architectures, despite their fluency, lack the cognitive depth required to parse the recursive syntax characteristic of international legal conventions.

V. Conclusion

This study provides an in-depth diagnostic of Google Translate's performance in the English-Arabic Customs domain, uncovering a profound safety gap that aggregate metrics fail to register. While the system demonstrates a high "performance ceiling" in general contexts, its "performance floor" in complex regulatory segments is characterized by systemic failures in adequacy and terminology. The research concludes that NMT, in its current state, cannot function as an autonomous agent for the translation of international Customs conventions. The 5.7x failure gap identified in this research provides a quantitative baseline for the compliance hazard inherent in unmonitored NMT deployment.

The results underscore a persistent performance paradox: high fluency scores actively camouflage critical adequacy failures. This "metric discordance" suggests that current evaluation paradigms—and the NMT objectives themselves—cannot reliably resolve the tradeoff in favor of strict legal precision. The findings confirm that while NMT has mastered the form of legal Arabic, it remains blind to the function of legal obligation.

To address these limitations, future research must focus on the development of Grounded Metrics such as COMET-poly-ic (Züfle et al., 2025), which retrieve human-scored legal segments to calibrate automated scores. Additionally, the use of granular post-edits as 'preference data' for fine-tuning models via Direct Preference Optimization (DPO) (Uhlig et al., 2025) offers a promising pathway toward aligning NMT systems with the rigorous demands of international law.

Furthermore, integrating human error markings directly into the inference process, as demonstrated by Berger et al. (2024), could enable real-time self-correction for critical legal terminology. Future research should specifically evaluate the efficacy of xCOMET-XXL (Guerreiro et al., 2024) in the legal domain. By utilizing a metric that integrates fine-grained error detection directly into the scoring process, Customs administrations may be able to automate the identification of localized critical errors and hallucinations, thereby reducing the cognitive load on human posteditors while maintaining regulatory safety.

REFERENCES

[1] Abdelaal, N. M., & Alazzawie, A. (2020). Machine translation: the case of Arabic-English translation of news texts. *Theory and Practice in Language Studies*, *10*(4), 408–418. https://doi.org/10.17507/tpls.1004.09

[2] AlAfnan, M. A. (2025). Artificial Intelligence and Language: Bridging Arabic and English with Technology. *Journal of Ecohumanism*, *3*(8). https://doi.org/10.62754/joe.v3i8.4961

[3] Aldawsari, H. A. H. (2025). Evaluating the Performance of Large Language Models on Arabic Lexical Ambiguities: A Comparative Study with Traditional Machine Translation Systems. *World Journal of English Language*, *15*(3), 354. https://doi.org/10.5430/wjel.v15n3p354

[4] Alghamdi, E. A., Zakraoui, J., & Abanmy, F. A. (2024). Domain adaptation for Arabic machine Translation: Financial Texts as a case study. *Applied Sciences*, *14*(16), 7088. https://doi.org/10.3390/app14167088

[5] Ali, M. A. (2020). Quality and Machine Translation: An Evaluation of Online Machine Translation of English into Arabic Texts. *Open Journal of Modern Linguistics*, *10*(05), 524–548. https://doi.org/10.4236/ojml.2020.105030

[6] Alkatheery, E. R. (2023). Google Translate Errors in Legal Texts: Machine Translation Quality Assessment. *Arab World English Journal for Translation and Literary Studies*, *7*(1), 208–219. https://doi.org/10.24093/awejtls/vol7no1.16

[7] Al-Khalifa, H., Al-Khalefah, K., & Haroon, H. (2024). Error analysis of Pretrained Language Models (PLMS) in English-to-Arabic machine translation. *Human-Centric Intelligent Systems*, *4*(2), 206–219. https://doi.org/10.1007/s44230-024-00061-7

[8] Al Maaytah, S. A. A. (2025). Evaluating Three neural machine translation platforms for English-Arabic translation: A Comparative study of linguistic accuracy and Cultural fidelity. *World Journal of English Language*, *16*(2), 1. https://doi.org/10.5430/wjel.v16n2p1

[9] Al Maaytah, S. A., Aalzobidy, S. A., & Alwidyan, M. F. (2025). Using machine translation English - Arabic procedures and challenges - A systematic review. *Power System Technology*, *49*(1), 588–607. Retrieved February 18, 2026, from https://powertechjournal.com/index.php/journal/article/view/1582

[10] 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

[11] Ameur, M. S. H., Meziane, F., & Guessoum, A. (2020). Arabic Machine Translation: A survey of the latest trends and challenges. *Computer Science Review*, *38*, 100305. https://doi.org/10.1016/j.cosrev.2020.100305

[12] Babaali, B., & Salem, M. (2022). Arabic machine Translation: A panoramic survey. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4312742

[13] Berger, N., Riezler, S., Exel, M., & Huck, M. (2024). Prompting Large Language Models with Human Error Markings for Self-Correcting Machine Translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation* (Vol. 1, pp. 636–646). European Association for Machine Translation (EAMT). Retrieved February 18, 2026, from https://aclanthology.org/2024.eamt-1.54/

[14] Beseiso, M., Tripathi, S., Al-Shboul, B., & Aljadid, R. (2022). Semantics based English-Arabic machine translation evaluation. *Indonesian Journal of Electrical Engineering and Computer Science*, *27*(1), 189–197. https://doi.org/10.11591/ijeecs.v27.i1.pp189-197

[15] Elsayed, A. S. O. (2025). When machines meet gavel: a case study of the English–Arabic machine translation of the Egyptian arguments before the International Court of Justice (2024). *Language and Semiotic Studies*, *11*(4), 661–690. https://doi.org/10.1515/lass-2025-0054

[16] Fakih, A., Ghassemiazghandi, M., Fakih, A. H., & Singh, M. K. M. (2024). Evaluation of Instagram's neural machine translation for literary texts: An MQM-Based Analysis. *GEMA Online Journal of Language Studies*, *24*(1), 213–233. https://doi.org/10.17576/gema-2024-2401-13

[17] Fields, P., Hague, D. R., Koby, G. S., Lommel, A., & Melby, A. (2014). What is quality? A management discipline and the translation industry get acquainted. *Tradumàtica Tecnologies De La Traducció*, *12*, 404–412. https://doi.org/10.5565/rev/tradumatica.75

[18] 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

[19] Gamal, D., Alfonse, M., Jiménez-Zafra, S. M., & Aref, M. (2023). Case study of improving English-Arabic translation using the Transformer Model. *International Journal of Intelligent Computing and Information Sciences*, *23*(2), 105–115. https://doi.org/10.21608/ijicis.2023.210435.1270

[20] Guerreiro, N. M., Rei, R., Van Stigt, D., Coheur, L., Colombo, P., & Martins, A. F. T. (2024). xCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection. *Transactions of the Association for Computational Linguistics*, *12*, 979–995. https://doi.org/10.1162/tacl_a_00683

[21] Hameed, D. A., Faisal, T. A., Abbas, A. K., Ali, H. A., & Hasan, G. T. (2022). DIA-English-Arabic neural machine translation domain: sulfur industry. *Indonesian Journal of Electrical Engineering and Computer Science*, *27*(3), 1619–1624. https://doi.org/10.11591/ijeecs.v27.i3.pp1619-1624

[22] Idrysy, F. Z. E., Hourri, S., Miqdadi, I. E., Hayati, A., Namir, Y., Ncir, B., & Kharroubi, J. (2025). Unlocking the language barrier: A Journey through Arabic machine translation. *Multimedia Tools and Applications*, *84*(14), 14071–14104. https://doi.org/10.1007/s11042-024-19551-8

[23] Junczys-Dowmunt, M. (2025). GEMBA V2: Ten Judgments Are Better Than One. In *Proceedings of the Tenth Conference on Machine Translation* (pp. 926–933). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.wmt-1.67
[24] Kayano, Y., & Sugawara, S. (2025). Specification-Aware Machine Translation and Evaluation for Purpose Alignment. In *Proceedings of the Tenth Conference on Machine Translation* (pp. 113–141). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.wmt-1.7
[25] Kit, C., & Wong, B. T. M. (2023). Evaluation in Machine Translation and Computer-Aided Translation. In *Routledge Encyclopedia of Translation Technology* (2nd edition, pp. 219–244). Routledge. https://doi.org/10.4324/9781003168348-13
[26] Koehn, P., & Knowles, R. (2017). Six challenges for neural machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation* (pp. 28–39). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-3204
[27] Kovacs, G., Deutsch, D., & Freitag, M. (2024). Mitigating Metric Bias in Minimum Bayes Risk Decoding. In *Proceedings of the Ninth Conference on Machine Translation* (pp. 1063–1094). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.109
[28] Läubli, S., Sennrich, R., & Volk, M. (2018). Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4791–4796). Association for Computational Linguistics. https://doi.org/10.18653/v1/d18-1512
[29] Lavie, A., Hanneman, G., Agrawal, S., Kanojia, D., Lo, C., Zouhar, V., Blain, F., Zerva, C., Avramidis, E., Deoghare, S., Sindhujan, A., Wang, J., Adelani, D. I., Thompson, B., Kocmi, T., Freitag, M., & Deutsch, D. (2025). Findings of the WMT25 Shared Task on Automated Translation Evaluation Systems: Linguistic Diversity is Challenging and References Still Help. In *Proceedings of the Tenth Conference on Machine Translation* (pp. 436–483). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.wmt-1.24
[30] Lommel, A. (2018). Metrics for Translation Quality Assessment: A case for Standardising error typologies. In *Machine translation* (pp. 109–127). Springer. https://doi.org/10.1007/978-3-319-91241-7_6
[31] Moghe, N., Fazla, A., Amrhein, C., Kocmi, T., Steedman, M., Birch, A., Sennrich, R., & Guillou, L. (2024). Machine Translation Meta Evaluation through Translation Accuracy Challenge Sets. *Computational Linguistics*, *51*(1), 73–137. https://doi.org/10.1162/coli_a_00537
[32] Mukherjee, A., & Shrivastava, M. (2023). IIIT HYD's Submission for WMT23 Test-suite Task. In *Proceedings of the Eighth Conference on Machine Translation* (pp. 246–251). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.24
[33] Nagi, K. A. (2023). Arabic and English relative clauses and machine translation challenges. *Mağallaẗ Al-dirāsāt Al-iğtimāʿiyyaẗ, 29*(3), 145–165. https://doi.org/10.20428/jss.v29i3.2180
[34] Sabtan, Y. M. N., Hussein, M. S. M., Ethelb, H., & Omar, A. (2021). An evaluation of the accuracy of the machine translation systems of social media language. *International Journal of Advanced Computer Science and Applications, 12*(7). https://doi.org/10.14569/ijacsa.2021.0120746
[35] Sabtan, Y. M. N., Omar, A., & Hamouda, W. I. (2024). Exploring the Role of Machine Translation in Translating English Collocations into Arabic: Insights from Student Translators. *World Journal of English Language*, *14*(2), 74. https://doi.org/10.5430/wjel.v14n2p74
[36] Semenov, K., Huang, X., Zouhar, V., Berger, N., Zhu, D., Oncevay, A., & Chen, P. (2025). Findings of the WMT25 Terminology Translation Task: Terminology is Useful Especially for Good MTs. In *Proceedings of the Tenth Conference on Machine Translation* (pp. 554–576). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.wmt-1.30
[37] Shaalan, K., Siddiqui, S., Alkhatib, M., & Monem, A. A. (2019). Challenges in Arabic natural language processing. In *Computational Linguistics, Speech and Image Processing for Arabic Language* (pp. 59–83). World Scientific. https://doi.org/10.1142/9789813229396_0003
[38] Shayegh, B., Peter, J., Vilar, D., Domhan, T., Juraska, J., Freitag, M., & Mou, L. (2025). Feeding Two Birds or Favoring One? Adequacy–Fluency Tradeoffs in Evaluation and Meta-Evaluation of Machine Translation. In *Proceedings of the Tenth Conference on Machine Translation* (pp. 269–285). https://doi.org/10.18653/v1/2025.wmt-1.16
[39] Singh, K. B., Kumar, D., & Ekbal, A. (2025). Evaluation of LLM for English to Hindi Legal Domain Machine Translation Systems. In *Proceedings of the Tenth Conference on Machine Translation* (pp. 823–833). https://doi.org/10.18653/v1/2025.wmt-1.57
[40] Strandvik, I. (2025). Translation quality and the role of specifications – How standards can help the translation sector today. *Across Languages and Cultures*, *26*(S), 5–24. https://doi.org/10.1556/084.2025.01057
[41] Uhlig, K., Wuebker, J., Reinauer, R., & Denero, J. (2025). Cross-lingual Human-Preference Alignment for Neural Machine Translation with Direct Quality Optimization. In *Proceedings of the Tenth Conference on Machine Translation* (pp. 31–51). https://doi.org/10.18653/v1/2025.wmt-1.2
[42] Vanroy, B., Tezcan, A., & Macken, L. (2023). MATEO: MT evaluation Online. In *The 24th Annual Conference of the European Association for Machine Translation (EAMT 2023)* (pp. 499–500). European Association for Machine Translation. Retrieved February 18, 2026, from https://aclanthology.org/2023.eamt-1.52/
[43] Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., & Tu, Z. (2023). Document-Level Machine Translation with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2023.emnlp-main.1036
[44] World Customs Organization. (1999). *International convention on the simplification and harmonization of customs procedures (as amended)*. Brussels: World Customs Organization. Retrieved February 18, 2026, from https://www.wcoomd.org/-/media/wco/public/global/pdf/topics/facilitation/instruments-and-tools/conventions/kyoto-convention/revised-kyoto-convention/body_gen-annex-and-specific-annexes.pdf?la=en
[45] 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
[46] Zerva, C., Blain, F., De Souza, J. G. C., Kanojia, D., Deoghare, S., Guerreiro, N. M., Attanasio, G., Rei, R., Orasan, C., Negri, M., Turchi, M., Chatterjee, R., Bhattacharyya, P., Freitag, M., & Martins, A. (2024). Findings of the Quality Estimation Shared

Task at WMT 2024: Are LLMs Closing the Gap in QE? In *Proceedings of the Ninth Conference on Machine Translation* (pp. 82–109). https://doi.org/10.18653/v1/2024.wmt-1.3

[47] Zou, L., Saeedi, A., & Koby, G. S. (2025). Beyond automated metrics: Assessing GPT-4o and Google Translate against professional translation standards. *SKASE Journal of Translation and Interpretation*, *18*(2). https://doi.org/10.33542/jti2025-s-9

[48] Züfle, M., Zouhar, V., Dinh, T. A., Polo, F. M., Niehues, J., & Sachan, M. (2025). COMET-poly: Machine Translation Metric Grounded in Other Candidates. In *Proceedings of the Tenth Conference on Machine Translation* (pp. 887–904). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.wmt-1.63

Karam Damseh is a Ph.D. candidate at the School of Languages, Literacies, and Translation, Universiti Sains Malaysia. His research interests focus on Machine Translation Evaluation. With more than seven years of experience in Customs administration, his research investigates the intersection of computational linguistics and legal compliance, aiming to mitigate liability risks in automated translation.

Mozhgan Ghassemiazghandi, Ph.D., is a Senior Lecturer at the School of Languages, Literacies and Translation, Universiti Sains Malaysia. Her research interests focus on Translation Technology, Machine Translation, and Audiovisual Translation. In addition to her academic work, she is a professional translator and subtitler with more than a decade of industry experience, combining theoretical insight with practical expertise in translation practice and technology.