

Evaluating Google's new TranslateGemma for English–Arabic machine translation: A corpus-based assessment using United Nations documents

Yassine El Rhaffouli ^①, Hicham Boughaba ^②

^{①②} King Fahd School of Translation – Abdelmalek Esaaddi University, Tangier, Morocco

 yassine.elrhaffouli@etu.uae.ac.ma

Abstract

This study presents a systematic evaluation of Google's newly released model TranslateGemma-4B for English-Arabic machine translation using official United Nations parallel documents. Despite the recent proliferation of large language models claiming multilingual competence, empirical assessments of translation quality for morphologically rich languages which are languages that exhibit many grammatical cases and inflected word forms such as Arabic remain limited, particularly when evaluated against professionally approved institutional reference translations. We assessed TranslateGemma in a zero-shot configuration (i.e., the prompt used contained only the task objective and no examples or demonstrations) on 10,000 sentence pairs drawn from the UN English–Arabic Parallel Corpus. The model was executed on Google Colab due to computational requirements exceeding local hardware capacity. Translation outputs were evaluated using five complementary automatic metrics: BLEU (6.95), chrF++ (33.28), METEOR (20.90), BERTScore (74.21), and COMET (71.60). Additionally, we implemented diagnostic heuristics to detect omissions, hallucinations (i.e., inventing irrelevant texts), digit mismatches, and terminology inconsistencies. Results indicate that while TranslateGemma achieves moderate semantic similarity scores on neural metrics, it exhibits substantial deficiencies in surface-form accuracy and institutional terminology consistency. Sentence length ratio analysis revealed systematic under-translation patterns, with 16.86% of outputs flagged for potential omission. Digit mismatch errors affected 24.48% of the corpus, raising concerns for high-stakes translation contexts. Terminology consistency analysis using an expanded UN glossary indicated that 7.95% of sentences containing frozen institutional terms failed to preserve standard Arabic equivalents. These findings demonstrate that TranslateGemma, while showing promise in capturing broad semantic adequacy, requires significant refinement before deployment in institutional translation workflows where precision and terminological fidelity are paramount.

Keywords

machine translation, English to Arabic translation, neural machine translation, evaluation metrics, United Nations corpus

How to cite this article in APA format:

El Rhaffouli, Y., & Boughaba, H. (2026). Evaluating Google's new TranslateGemma for English–Arabic machine translation: A corpus-based assessment using United Nations documents. *Linguistics Initiative*, 6(1), 155–171. <https://doi.org/10.53696/27753719.61411>

Introduction

Machine translation has undergone rapid transformation with the advent of neural architectures and, more recently, large language models. In fact, commercial systems such as Google Translate and Microsoft Translator have achieved substantial improvements in translation quality for high-resource language pairs. Nonetheless, the evaluation of newly released open-source models on morphologically complex and/or under-resourced languages remains an urgent research priority. Particularly speaking, Arabic is characterized by its rich inflectional morphology, non-concatenative word formation, and substantial dialectal variation. These linguistic properties pose persistent challenges for automated translation systems (Alkhatib & Shaalan, 2018; Ameer et al., 2020; Zakraoui et al., 2021). Additionally, the linguistic distance between Arabic and English necessitates rigorous evaluation frameworks that extend beyond conventional string-matching metrics (Shquier & Sembok, 2008; Alqudsi et al., 2012).

In January 2026, Google released TranslateGemma, and it is a family of open-source translation models designed to provide efficient multilingual translation across diverse hardware environments (Vilar & Black, 2026). TranslateGemma represents a significant development in accessible machine translation technology as it offers model variants ranging from 4 billion to 27 billion parameters. The 4B instruction-tuned variant (i.e., designed to follow specific translation instructions/prompts rather than generate general text) evaluated in this study is positioned as suitable for resource-constrained environments such as consumer-grade laptops and desktops. However, Google's technical documentation reports competitive performance on the WMT24++ benchmark across 55 languages, yet independent academic evaluation of TranslateGemma's performance on domain-specific, professionally translated corpora remains absent from the published literature due to its recency.

Recent scholarship has documented persistent difficulties in English–Arabic machine translation across multiple dimensions. Al-Khalifa et al. (2024) demonstrated that pretrained language models frequently produce grammatical errors in Arabic output, particularly in subject-verb gender agreement and functional morpheme omission. Abdelaal and Al Sawi (2025) identified systematic challenges in translating legal terminology, and they noted that both neural machine translation and large language model systems struggle with culturally embedded legal concepts and jurisdiction-specific vocabulary. Similarly, Altakhaine et al. (2024) found that AI-generated translations of legal documents into Arabic exhibited lower accuracy than human translations across all assessed criteria that included terminology precision and adherence to legal discourse conventions. These findings underscore the importance of evaluation that both takes domains into consideration and uses professionally validated reference translations.

The institutional translation domain presents distinct challenges that differentiate it from general-purpose machine translation evaluation. Ahmed et al. (2025) analyzed figurative language in United Nations General Assembly speeches. They demonstrated that metaphorical expressions and culturally specific rhetorical devices are frequently neutralized or distorted in machine translation output. Further, Bechara et al. (2024) highlighted methodological challenges in constructing multilingual evaluation corpora from UN documents. These studies collectively emphasize the necessity of rigorous, corpus-based evaluation frameworks that account for institutional discourse characteristics.

Despite growing interest in large language model performance for machine translation, several critical gaps persist in the literature. First, the majority of published evaluations have focused on general-domain texts or news corpora. This gap leaves institutional and legal translation underexplored (Zakraoui et al., 2020). Second, evaluation studies frequently rely exclusively on BLEU scores, which have been demonstrated to capture inadequately morphological variation and semantic adequacy in Arabic (Bouamor et al., 2014; Guzmán et al., 2016). Third, few studies

systematically assess error patterns such as omission, hallucination, and terminology inconsistency, despite their critical importance in high-stakes translation contexts (Diab, 2021; Nassar, 2025). Finally, independent academic assessment of newly released open-source models remains limited because most evaluation efforts are concentrated on proprietary commercial systems.

This study addresses these gaps through a comprehensive evaluation of the newly released TranslateGemma-4B on English–Arabic translation using the United Nations Parallel Corpus. Our research is guided by three primary objectives. First, we assess translation quality using a multi-metric framework that incorporates both surface-form measures (BLEU, chrF++, METEOR) and neural semantic metrics (BERTScore, COMET). Second, we implement diagnostic heuristics to identify systematic error patterns, including omissions, hallucinations, factual distortions, and terminology inconsistencies. Third, we examine the model's performance in preserving institutionally frozen terminology specific to United Nations documents, in order to provide insight into its suitability for high-stakes institutional translation workflows.

The research questions guiding this investigation are as follows: (1) What levels of translation quality does TranslateGemma achieve when evaluated against official UN Arabic reference translations using established automatic metrics? (2) What systematic error patterns emerge in TranslateGemma's English–Arabic translation output, particularly regarding omissions, hallucinations, and terminology consistency? (3) To what extent does the new model preserve institutionally standardized terminology when translating formal legal-administrative discourse? By addressing these questions through corpus-based evaluation, this study contributes empirical evidence regarding the capabilities and limitations of TranslateGemma for Arabic machine translation, with implications for both academic research and practical deployment in institutional translation contexts.

Method

This study adopted an empirical, corpus-based evaluation design to assess the performance of TranslateGemma-4B on English–Arabic institutional translation. The evaluation framework integrated automatic translation quality assessment metrics with diagnostic error analysis. The study used a zero-shot translation prompt without relying on prompt engineering or fine tuning. This methodological approach reflects realistic deployment scenarios in which large language models are employed without extensive task-specific customization.

Computational Infrastructure

All translation experiments were conducted using Google Colab, which is a cloud-based Jupyter notebook environment providing recent Python installation, relevant libraries, and GPU acceleration. This computational infrastructure was selected due to local hardware limitations. Specifically, the main reason for such a choice was the absence of sufficient GPU memory and system RAM required to load and execute a 4-billion-parameter model efficiently. Google Colab enabled stable inference with controlled resource allocation in order to secure reproducibility across the evaluation corpus. Model access and authentication were handled through the Hugging Face ecosystem using secure access tokens, and the TranslateGemma model was loaded via the transformers pipeline interface.

Translation Model

The experiments employed TranslateGemma-4B, which is an instruction-tuned translation model released by Google in January 2026. Translation was performed in a zero-prompt

configuration. This means that no additional contextual instructions, examples, or task-specific constraints were provided beyond basic source-target language specification. This design choice enabled unbiased evaluation of the model's intrinsic translation capabilities without confounding effects from prompt engineering. To ensure deterministic outputs and eliminate variation, sampling was disabled during generation, and a fixed maximum token length of 128 tokens was applied to balance translation completeness with computational efficiency. The 128 tokens limit means that the model only processes around 90 to 100 words in a single interaction, and the choice of this length limit ensured that the model did not exceed its computational limits and at the same time eliminates any incomplete outputs. The model was executed using bfloat16 precision on GPU hardware, with batch processing of 16 sentences to optimize quantity.

Corpus Selection and Preparation

The evaluation corpus was drawn from the United Nations Parallel Corpus, officially published on the United Nations Department for General Assembly and Conference Management website (Ziems et al., 2016). This corpus was selected for three primary methodological advantages. First, UN documents exhibit formal, legal language and are of high-stakes institutional discourse, which is known to pose substantial challenges for neural and large language model-based translation systems. Second, the corpus includes professionally produced and institutionally validated Arabic translations that have undergone rigorous quality assurance processes. These reference translations served as reliable standards for evaluation. Third, source and target texts are sentence-aligned with high precision, and, thus, it ensures exact correspondence between English inputs and Arabic reference translations.

From the full UN English–Arabic parallel corpus, a subset of 10,000 sentence pairs was extracted for evaluation. This sample size was selected to balance computational feasibility with statistical robustness. By adopting this design, the study enabled both large-scale quantitative evaluation and targeted qualitative error analysis. The extraction procedure prioritized sentence pairs with complete alignment, excluding document headers, metadata fragments, and malformed entries. Each data point in the final evaluation set consisted of one English source sentence and its corresponding human-produced official Arabic translation. Before moving forward with the study, we ensured that the directionality of the corpus was correct by verifying sentence alignment. Finally, the extracted dataset was stored in CSV format with UTF-8 encoding to preserve Arabic script integrity across platforms and processing stages.

Translation Procedure

The selected English sentences were translated into Arabic using TranslateGemma in batch mode to improve computational efficiency and reduce inference latency. Sentences were processed in fixed-size batches of 16, with generation parameters optimized for institutional-length sentences. A maximum token limit of 128 was imposed to accommodate typical UN sentence structures and to prevent excessive computation time as well. To ensure data integrity during extended inference sessions, intermediate results were periodically checkpointed to Google Drive at intervals of 500 sentences. Upon completion of the full corpus translation, outputs were consolidated into a single results file containing the English source sentence, the human Arabic reference translation, and the TranslateGemma-generated Arabic rendition.

Output Normalization and Cleaning

Raw model outputs occasionally contained prompt-related artifacts, metadata markers, and mixed language content resulting from the instruction-tuned model architecture. A robust post processing procedure was therefore applied to extract exclusively the Arabic translation content. This cleaning protocol involved several sequential steps. First, prompt-related markers

such as 'Translation:', 'Arabic:', and their Arabic equivalents were detected and removed using regular expression patterns. Second, non-Arabic characters were eliminated while preserving Arabic Unicode ranges, numerals, and standard punctuation. Third, the first contiguous block of Arabic script was extracted as the final translation output, and any subsequent explanatory text and/or alternative suggestions were discarded. The cleaned outputs were validated to ensure non-empty content and absence of English remnants. Sentences failing this validation process were excluded from subsequent analysis. This normalization step ensured that all downstream analyses targeted actual model translation output rather than formatting artifacts or metadata.

Evaluation Framework

The evaluation framework comprised two complementary components: automatic reference-based metrics and diagnostic error analysis. Automatic metrics were selected to capture multiple dimensions of translation quality, including surface-form accuracy, character-level similarity, lexical-semantic matching, and neural semantic adequacy. This multi-metric approach addresses documented limitations of relying exclusively on BLEU scores for Arabic machine translation evaluation (Bouamor et al., 2014; Guzmán et al., 2016).

Surface-form and character-level metrics were computed using SacreBLEU. Corpus-level BLEU measured n-gram overlap between TranslateGemma output and UN reference translations so as to provide conventional baseline comparability. The chrF++ metric was included as a character n-gram F-score that captures subword-level similarity, which is particularly suitable for Arabic due to its rich inflectional system. Character-level metrics mitigate vocabulary sparsity issues inherent in word-level matching and demonstrate greater robustness to inflectional variation and orthographic normalization (Popović, 2015).

Semantic similarity was assessed using BERTScore, which leverages contextualized embeddings from multilingual BERT models to evaluate meaning preservation beyond exact lexical overlap (Zhang et al., 2020). BERTScore computation employed a multilingual model with Arabic language support, executed in batches of 64 sentences with GPU acceleration. Sentence-level F1 scores were aggregated to produce a corpus-level metric reflecting overall semantic adequacy. This metric is particularly valuable for evaluating modern neural machine translation systems that may produce fluent paraphrases rather than literal translations.

Neural quality estimation was performed using COMET, which is a reference-based metric trained to correlate with human judgment of translation quality (Rei et al., 2020). COMET evaluation incorporated source sentences, system renditions, and reference translations, and it utilized GPU acceleration to process the full corpus. The COMET metric was employed due to its demonstrated strong correlation with professional human assessments across multiple language pairs. COMET served as the primary indicator of semantic adequacy and contextual appropriateness in this study.

METEOR was included as an additional lexical-semantic metric that evaluates translation quality through exact matches, stemmed matches, and synonym matches (Banerjee & Lavie, 2005). METEOR provides an intermediate perspective between surface overlap and neural semantic similarity, with sensitivity to lexical variation and partial morphological correspondence. Although less prominent than neural metrics in contemporary evaluation practice, METEOR offers historical continuity with earlier Arabic machine translation evaluation literature (Hadla et al., 2015).

Diagnostic Error Analysis

Beyond aggregate quality metrics, we implemented diagnostic heuristics to detect systematic translation pathologies. These analyses targeted four categories of errors particularly consequential in institutional translation contexts: omissions, hallucinations, factual distortions, and terminology inconsistencies. Still, these diagnostics do not constitute comprehensive semantic evaluation, but they provide scalable indicators for structural translation failures warranting qualitative inspection.

Digit mismatch detection served as a high-precision proxy for factual inconsistencies. For each sentence pair, numeric expressions were extracted from both English source and Arabic translation using regular expression patterns. Any discrepancy between the two sets was flagged as a potential factual distortion. This diagnostic is particularly critical for UN documents, where numerical information such as resolution numbers, dates, article citations, and statistical data must be preserved with absolute accuracy.

Length-based heuristics identified potential omissions and hallucinations through sentence length ratio analysis. The ratio between Arabic word count and English word count was computed for each sentence pair. Following empirically motivated thresholds established in machine translation evaluation literature, ratios below 0.6 were flagged as potential omissions indicating severe content compression, while ratios above 1.4 were flagged as potential hallucinations indicating excessive content expansion. These thresholds were calibrated to balance sensitivity and specificity, but still nonetheless acknowledged that some length variation is inherent to cross-linguistic translation.

Incomplete sentence detection identified translations exhibiting premature termination. Sentences were flagged as incomplete when the English source terminated with sentence final punctuation (period, question mark, or exclamation point) but the Arabic output did not. This diagnostic captures a known failure mode in neural and large language model-based generation, particularly when operating under strict token limits or experiencing decoding errors.

Terminology consistency evaluation assessed preservation of institutionally standardized UN vocabulary. An expanded UN glossary was manually curated, encompassing principal organs (General Assembly, Security Council, Economic and Social Council, Secretariat, International Court of Justice), legal instruments (Charter of the United Nations, Resolution), governance concepts (Member States, Human Rights, Sovereignty, International Law), and institutional roles (Secretary-General, Special Representative). For each English source sentence containing a glossary term, the Arabic output was inspected for the presence of approved Arabic equivalents. Sentences were flagged for terminology inconsistency when standard equivalents were absent, so that it could indicate potential substitution with non-standard variants or paraphrases.

Named entity information was extracted from English source sentences using spaCy's English language model to support contextual interpretation of error patterns. While not directly scored, entity counts (organizations, locations, persons, dates) were retained to facilitate subsequent qualitative analysis, particularly for sentences flagged by multiple diagnostic criteria. All diagnostic analyses were implemented in Python using regular expressions, natural language processing libraries, and batch processing to ensure scalability across the 10,000-sentence evaluation corpus.

Results

The evaluation of TranslateGemma-4B on English–Arabic translation of United Nations documents yielded a multi-dimensional assessment of translation quality and error patterns. After preprocessing and output normalization, the final evaluation set comprised 8,357 valid sentence pairs with complete source texts, reference translations, and non-empty system outputs. Sentences containing missing values or failed translations were excluded to ensure metric reliability and cross-method comparability.

Automatic Metric Performance

Table 1. Scores of the Automatic Translation Quality Assessment Metrics

Metric	Score
BLEU	6.95
chrF++	33.88
BERTScore (F1)	74.21
COMET Score	71.60
METEOR	20.90

Surface-form and character-level metrics revealed substantial challenges in exact lexical correspondence. The corpus-level BLEU score of 6.95 indicates minimal n-gram overlap between TranslateGemma outputs and official UN reference translations. This score falls considerably below benchmarks reported for commercial neural machine translation systems on Arabic institutional texts, which typically achieve BLEU scores in the range of 25 to 40 on comparable corpora (Zakraoui et al., 2020). The chrF++ score of 33.28 demonstrated improved performance at the character level. This score reflects partial morphological correspondence and subword similarity despite limited word-level matching. The substantial gap between BLEU and chrF++ scores suggests that indeed TranslateGemma captures some morphological patterns, yet it frequently produces lexical variants that diverge from standardized UN terminology.

The METEOR score of 20.90 positioned between surface-form BLEU and character-level chrF++, and it indicates moderate success in synonym matching and stemming-based correspondence. This intermediate performance suggests that TranslateGemma occasionally selects lexical alternatives that preserve semantic content although it diverges from official terminology. However, the METEOR score remains substantially lower than values typically associated with adequate translation quality for institutional purposes, where scores above 40 are generally considered acceptable.

Neural semantic metrics presented a contrasting pattern. The BERTScore F1 of 74.21 indicated strong semantic similarity at the contextual embedding level; this score suggests that TranslateGemma outputs preserve broad meaning despite lexical divergence. This substantial BERTScore reflects the model's capacity to capture semantic adequacy through paraphrase and lexical variation, even when producing non-literal translations. Similarly, the COMET score of 71.60 demonstrated that neural quality estimation models perceive TranslateGemma outputs as semantically adequate relative to reference translations. The convergence of BERTScore and COMET scores above 70 suggests consistent semantic fidelity at the sentence level and indicates that meaning preservation occurs despite surface-form inaccuracies.

The divergence between surface-form metrics and neural semantic metrics is characteristic of modern large language model translation systems, which prioritize fluency and semantic

adequacy over literal correspondence. However, this pattern raises concerns for institutional translation contexts where terminological precision and standardized lexical choices are mandatory rather than optional. The low BLEU score coupled with high semantic similarity scores indicates that TranslateGemma frequently generates semantically plausible Arabic text that diverges from institutionally approved translations in lexical selection and phrasal structure.

Diagnostic Error Analysis

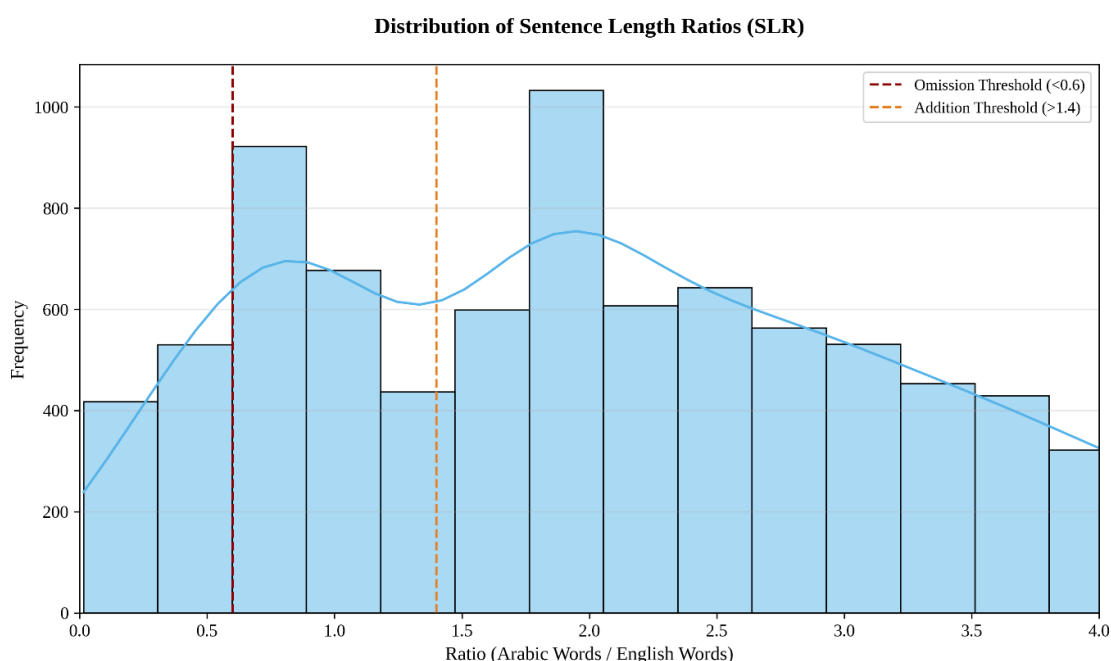


Figure 1. Distribution of Sentence Length Ratios

Sentence length ratio analysis revealed systematic structural imbalances in TranslateGemma outputs. The distribution of length ratios exhibited negative skew. In this regard, a substantial proportion of translations demonstrated compression relative to source sentences. Of the 8,357 evaluated sentences, around 1000 translations fell below the 0.6 threshold. This value indicates potential omission or severe content reduction. Conversely, over 400 translations exceeded the 1.4 threshold, and this finding suggests potential hallucination or excessive expansion. The majority of translations (72.0%) fell within the standard transfer range, which indicates proportional length correspondence.

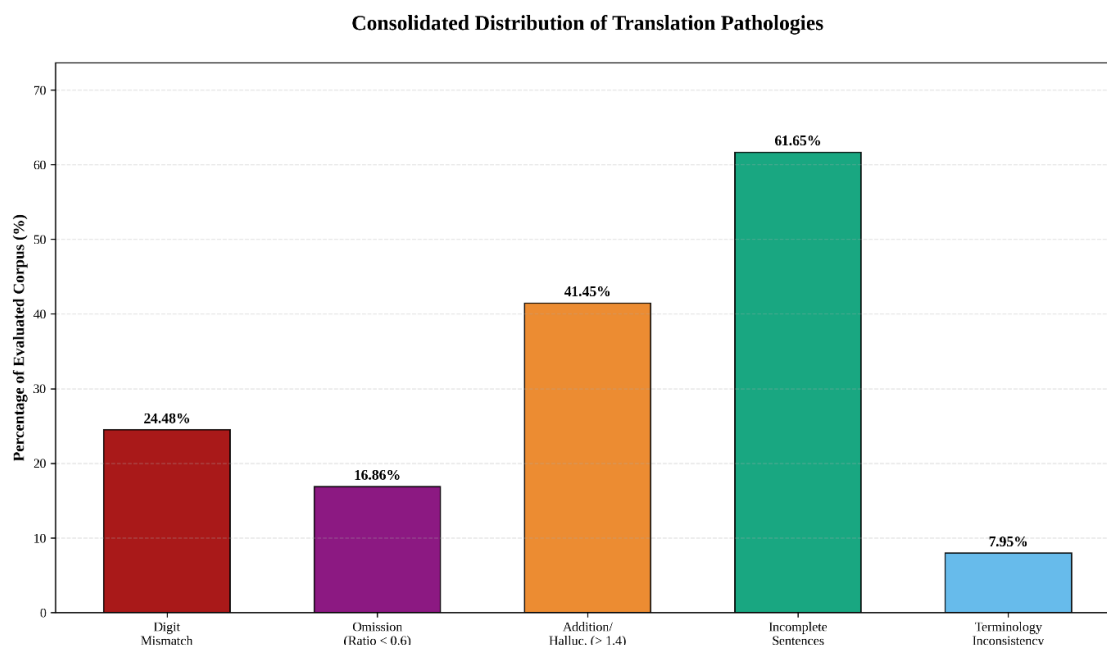


Figure 2. Consolidated Distribution of Translation Pathologies

Beyond aggregate quality metrics, diagnostic heuristics were applied to detect systematic translation pathologies across the evaluation corpus. Figure 2 presents the distribution of automatically detected translation errors, revealing substantial prevalence of multiple failure modes.

Incomplete sentence generation represented the most pervasive pathology, affecting 61.65% of the corpus. This indicates that the majority of TranslateGemma outputs failed to reproduce terminal punctuation present in English sources. Thus, this suggests systematic truncation or premature termination during decoding. This extraordinarily high rate represents a critical system-level failure that fundamentally undermines translation utility.

Addition/hallucination errors affected 41.45% of translations, and the criterion used to flag the error was sentence length ratios that exceeded 1.4 (Arabic words/English words). This substantial rate indicates frequent content expansion beyond source material, diverging markedly from the under-translation patterns documented in some large language model systems. The prevalence of additions suggests that TranslateGemma frequently generates verbose outputs or introduces material not present in the source text. This pattern aligns well with documented behavior of some large language models in other translation directions, which have been observed to produce verbose or explanatory outputs exceeding source content (Mohammed, 2025). Moreover, omission patterns were flagged when sentence length ratios fell below 0.6, and they affected 16.86% of translations. Although omission patterns are less prevalent than addition errors, this rate nonetheless indicates systematic content reduction in a substantial minority of translations.

Digit mismatch errors affected 24.48% of the corpus. Since UN documents extensively reference resolution numbers, article citations, dates, statistical data, and budgetary figures, this error rate demonstrates that approximately one quarter of all translations containing numerical information exhibit factual distortions. This represents an unacceptable failure rate for institutional translation contexts requiring absolute numerical accuracy.

Terminology inconsistency, assessed through comparison against an expanded UN glossary of institutionally frozen terms, affected 7.95% of sentences containing recognized terminology. This represents the lowest error rate among diagnostic categories, but it nonetheless indicates that nearly 8% of translations containing standardized UN terms failed to preserve approved Arabic equivalents. Common patterns included substitution of standard terms with colloquial alternatives, use of literal translations rather than frozen institutional phrases, and inconsistent rendering of the same English term across different contexts. For example, "General Assembly" was inconsistently translated, sometimes using the standard "الجمعية العامة" and other times employing non-standard variants or paraphrases. Therefore, it may present a compromise of terminological stability which is essential for legal-administrative documents.

Table 2. Raw Error Samples from the Corpus

Source	Reference Translation	TranslateGemma Output	Error Type
9. Paragraph 444 of the report restates the Working Group's view that a court should not have compulsory jurisdiction.	تردد الفقرة ٤٤٤ من ٩ - التقرير رأي الفريق العامل والقاتل إنه ينبغي ألا تكون لأي محكمة ولاية قضائية إلزامية.	"Translate English to Arabic: 9. Paragraph 444 of the report restates the Working Group's view that a court should not have compulsory jurisdiction.	Incomplete sentence
		قانون العمل: التوظيف غير الرسمي	
As outlined in paragraph 444, a State becoming a party to the Statute would accept "certain administrative obligations".	وكما ورد في الفقرة ٤٤٤، ستقبل الدولة التي ستصبح طرفاً في النظام الأساسي "بعض الالتزامات الإدارية".	"Translate English to Arabic: As outlined in paragraph 444, a State becoming a party to the Statute would accept "certain administrative obligations".	Hallucination/ Addition
		أكدت البند 444 أن الدولة التي تصبح طرفاً في البند قد تقبل "بعض الالتزامات الإدارية".	
As noted by the Working Group, this act could be similar to acceptance of the Optional Clause of the ICJ Statute.	وكما لاحظ الفريق العامل، قد يكون هذا الاجراء القانوني المستقل مشابها لقبول الشرط الاختياري المنصوص عليه في النظام الأساسي لمحكمة العدل الدولية.	"Translate English to Arabic: As noted by the Working Group, this act could be similar to acceptance of the Optional Clause of the ICJ Statute.	Terminology inconsistency
		الترجمة: كما أشار إلى ذلك الفريق العامل، يمكن أن يكون هذا الفعل ممثلاً للاعتراف بالبنود الاختيارية من ميثاق المحكمة الدائمة للنظر في النزاعات الدولي.	

Source	Reference Translation	TranslateGemma Output	Error Type
32. As the Working Group notes in paragraph 493 of the report, article 15 (1) of the International Covenant on Civil and Political Rights (ICCPR) embodies the principle of nullo crimen sine lege.	٣٢ - وحسبما ذكر الفريق العامل في الفقرة ٤٩٣ من التقرير، تجسد المادة ١٥ (١) من العهد الدولي الخاص بالحقوق المدنية والسياسية مبدأ "لا جريمة إلا بناء على نص".	"Translate English to Arabic: 32. As the Working Group notes in paragraph 493 of the report, article 15 (1) of the International Covenant on Civil and Political Rights (ICCPR) embodies the principle of nullo crimen sine lege. The Working Group on Human Rights, in its report, notes that Article 15 (1) of the ICCPR, as translated into Arabic, in paragraph 493 of the report, embodies the principle of nullo crimen sine lege. The Working Group on Human Rights, in its report, notes that Article 15 (1) of the ICCPR, as translated into Arabic, is a key component in ensuring that all individuals have the right to a fair trial.	Digit mismatch/ incomplete sentences
		الترجمة: 32. كما يذكر الفريق العامل في الفقرة 493 من التقرير، فإن المادة 1	
The initiation of a case 61. In paragraph 512 of the report, the Working Group suggests that adopting the system of appointing an independent prosecutor would obviate the need for a preliminary hearing to test the evidence as a court could dismiss "frivolous or unsubstantiated charges".	٦١ - يشير الفريق العامل في الفقرة ٥١٢ من التقرير إلى أن الأخذ بنظام تعيين مدع عام مستقل سيؤدي إلى إلغاء الحاجة إلى جلسة تحقيق أولية لاختبار كفاية الأدلة نظراً لأنه يمكن للمحكمة أن ترفض "الاتهامات التافهة أو التي لا أساس لها".	"Translate English to Arabic: The initiation of a case 61. In paragraph 512 of the report, the Working Group suggests that adopting the system of appointing an independent prosecutor would obviate the need for a preliminary hearing to test the evidence as a court could dismiss "frivolous or unsubstantiated charges". Arabic: تأسيس قضية في الفقرة 512 من التقرير، يقترح الفريق العامل أن اعتماد نظام تعيين محام مستقل سيؤدي إلى إلغاء الحاجة إلى جلسة استماع أولية لاختبار الأدلة، حيث يمكن للمحكمة رفض "التهمة التي لا أساس لها من الصحة أو التي لا يوجد لها دليل".	Omission

Discussion

The evaluation results reveal a complex performance profile for TranslateGemma-4B on English–Arabic institutional translation. The model demonstrates capacity for semantic adequacy as measured by neural metrics, yet exhibits substantial deficiencies in surface-form accuracy, terminological precision, and structural completeness. This performance pattern reflects broader tensions in contemporary machine translation evaluation between fluency-oriented large language model architectures and the strict fidelity requirements of specialized translation domains.

Interpretation of Metric Divergence

The substantial divergence between surface-form metrics (BLEU: 6.95, chrF++: 33.28) and neural semantic metrics (BERTScore: 74.21, COMET: 71.60) merits careful interpretation in light of institutional translation requirements. High semantic similarity scores indicate that TranslateGemma captures the propositional content and general meaning of source sentences. Thus, it can produce Arabic text that conveys similar information but through alternative lexical and structural choices. This capacity aligns with documented strengths of large language models in semantic understanding and paraphrase generation (Khasawneh & Alsharif, 2025; Zouidine & Khalil, 2025).

However, the low BLEU score signals systematic divergence from institutionally approved translations in lexical selection, phrasal structure, and terminological conventions. For institutional documents, such divergence represents a critical deficiency rather than benign variation. UN translations operate within strictly controlled terminological frameworks where specific Arabic phrases correspond to legally defined English concepts (Ziems et al., 2016). Deviation from these frozen expressions, even when semantically plausible, undermines legal precision and consistency across UN documents. The metric divergence, thus, indicates that TranslateGemma produces semantically adequate paraphrases rather than institutionally compliant translations.

This interpretation finds support in comparative studies of human and machine translation for legal and institutional texts. Abdelaal and Al Sawi (2025) documented that neural machine translation systems produce technically correct Arabic while failing to adhere to legal discourse conventions and jurisdiction-specific terminology. Altakhaine et al. (2024) similarly observed that AI translation tools demonstrate competence in general semantic transfer while exhibiting deficiencies in specialized legal vocabulary and register maintenance. The present findings extend these observations to large language model architectures, and further suggests that semantic adequacy alone is insufficient for institutional translation contexts.

Error Pattern Analysis

The predominance of hallucination errors over omission errors represents a noteworthy convergence toward documented patterns in some large language model translation systems. Studies of ChatGPT and similar models have reported tendencies toward verbose outputs and explanatory additions (Mohammed, 2025; Sadiq, 2025). TranslateGemma's under-translation tendency may reflect several factors. First, the instruction-tuned architecture may prioritize conciseness as a learned preference from training data. Second, the fixed maximum token limit of 128 may impose constraints that encourage compression. Third, the model's attention mechanisms may inadequately weight longer-distance dependencies in complex institutional sentences.

The morphological complexity of Arabic contributes to these patterns in predictable ways. Arabic's rich inflectional system creates opportunities for systematic errors when source language constraints are underspecified (Abu-Ayyash, 2017; Al-Khalifa et al., 2024). The model's tendency toward omission may partially reflect avoidance of complex agreement requirements in longer, more syntactically intricate sentences. When faced with sentences containing multiple embedded clauses or coordination structures, the model may default to simplified outputs that omit challenging constructions.

Digit mismatch errors are particularly consequential for UN documents. Numerical information in institutional texts typically carries legal weight, whether identifying specific resolutions, delimiting time periods, specifying budgetary allocations, or citing article numbers. The observed error rate suggests systematic difficulty in preserving factual precision during translation. This finding aligns with broader concerns regarding large language model hallucination and factual accuracy (Alwazna, 2024). The digit mismatch pattern indicates that TranslateGemma cannot be reliably deployed for institutional translation without extensive human verification of numerical content.

Terminology inconsistency represents perhaps the most fundamental barrier to institutional deployment. The United Nations maintains extensive multilingual terminology databases precisely to ensure inter-document consistency and legal precision across language versions. Terms such as "General Assembly," "Security Council," and "Member States" possess fixed Arabic equivalents that must be used invariably across all official documents. TranslateGemma's substitution of these standard terms with alternatives violates core institutional translation principles. This inconsistency would necessitate comprehensive post-editing to achieve publication standards.

Implications for Arabic Machine Translation

The findings contribute to ongoing discussions regarding the suitability of large language models for specialized translation domains. While TranslateGemma demonstrates clear competence in semantic understanding and general fluency production, its deficiencies in terminological precision and structural accuracy align with persistent challenges documented across neural machine translation systems for Arabic (Ameur et al., 2020; Zakraoui et al., 2021). The results suggest that current open-source large language model architectures, even when explicitly designed for translation tasks, require domain-specific fine-tuning or constrained decoding mechanisms to achieve institutional translation standards.

Several technical approaches could potentially address observed deficiencies. First, fine-tuning on domain-specific institutional corpora with explicit terminology constraints could improve lexical selection consistency (Alghamdi et al., 2024). Second, integration of terminology databases through constrained generation or retrieval-augmented approaches could enforce standardized term usage. Third, calibration of length controls and attention mechanisms could mitigate systematic under-translation patterns. Fourth, incorporation of morphological analyzers and agreement checkers could reduce grammatical errors specific to Arabic inflection.

From a workflow perspective, the results suggest that TranslateGemma in its current form may serve effectively as a first-draft generation tool within human-in-the-loop translation pipelines rather than as a standalone automated translation system. The model's capacity for semantic adequacy could accelerate professional translators' work by providing initial renditions that capture general meaning, which translators could then refine for terminological precision, structural accuracy, and stylistic conformity. This hybrid approach aligns with recommendations from recent comparative studies of human and AI translation (Moneus & Sahari, 2024; Badah et al., 2025).

Metric Reliability Considerations

The evaluation results highlight ongoing methodological challenges in automatic machine translation assessment for Arabic. The divergence between BLEU and neural metrics demonstrates that different evaluation paradigms prioritize distinct dimensions of translation quality. BLEU and chrF++ capture terminological precision and surface-form fidelity, which are paramount for institutional translation. BERTScore and COMET capture semantic adequacy and fluency, which are prioritized in general translation evaluation. For specialized domains such as legal and institutional translation, surface-form metrics may provide more reliable quality indicators than neural semantic metrics, despite the latter's stronger correlation with human adequacy judgments in general domains (Bouamor et al., 2014; Guzmán et al., 2016). This domain-specific metric validity consideration underscores the importance of multi-dimensional evaluation frameworks that incorporate both surface-form and semantic measures.

Conclusion

This study may provide one of the first independent academic evaluations of TranslateGemma-4B for English–Arabic translation using professionally validated institutional reference translations. The evaluation employed a multi-metric framework incorporating both surface-form measures and neural semantic metrics, supplemented by diagnostic error analysis targeting omissions, hallucinations, and terminology inconsistencies. The findings reveal that TranslateGemma achieves moderate semantic adequacy as measured by BERTScore and COMET. These metrics indicated capacity to preserve general meaning across translation. However, the model exhibits substantial deficiencies in surface-form accuracy (BLEU: 6.95), terminological precision (7.95% inconsistency rate), and structural completeness (16.86% potential omission rate).

The performance profile indicates that TranslateGemma functions effectively as a semantic paraphrase generator but may not amount to the level of an institutional translation system meeting United Nations publication standards. The systematic divergence from frozen terminology, frequent omission of content elements, and preservation failures for numerical information collectively demonstrate that current large language model approaches, even when explicitly optimized for translation, require substantial refinement for high-stakes institutional deployment. The model may serve productively within translation workflows as a first-draft generation tool, but cannot substitute for professional institutional translation without comprehensive post-editing.

Several limitations constrain the generalizability of these findings. First, the evaluation was restricted to a single domain (United Nations institutional documents), which may not represent TranslateGemma's performance on general-domain or news texts. Second, the study relied exclusively on automatic metrics and diagnostic heuristics without direct human evaluation of translation quality. Although automatic metrics provide scalable assessment, human judgment remains the gold standard for translation quality evaluation. Third, the evaluation examined only one model variant (4B instruction-tuned) from the TranslateGemma family; larger variants (12B, 27B) may demonstrate different performance characteristics. Fourth, the zero-shot evaluation paradigm, despite being methodologically appropriate for assessing intrinsic capabilities, may not reflect optimal performance achievable through careful prompt engineering or few-shot learning. In other words, the zero-shot configuration may be a limitation in itself since the model received no examples or demonstrations within the prompt (which was only “translate from English to Arabic”). Thus, prompt engineering and few-shot learning approaches may be best tackle this limitation since such approaches allow for the in-prompt inclusion of examples or

demonstrations which may reflect specific UN (or other domain-related) terminology, and translation techniques and demonstrations.

Future research should address these limitations through several complementary directions. First, human evaluation by professional translators and domain experts would provide direct assessment of translation adequacy, fluency, and terminological appropriateness. Second, comparative evaluation against other contemporary machine translation systems, including commercial systems and alternative large language models, would contextualize TranslateGemma's relative performance. Third, fine-tuning experiments on domain-specific corpora could assess the model's capacity for adaptation to institutional translation requirements. Fourth, extension to additional Arabic text domains, including news, literature, and scientific texts, would clarify whether observed patterns generalize beyond institutional discourse. Fifth, investigation of alternative prompting strategies and constrained decoding approaches could identify optimization paths for improving terminological consistency and structural accuracy. These research directions would collectively advance understanding of large language model capabilities and limitations for Arabic machine translation across diverse application contexts.

Acknowledgments

We would like to acknowledge Google for providing access to Google Colaboratory virtual platform, which enabled GPU acceleration for model inference and evaluation. The computational resources made available through this platform were essential for conducting the experiment reported in this article.

References

- Abdelaal, N., & Al Sawi, I. (2025). A comparative evaluation of machine translation vs. human translation for legal texts: A case study of translations between English to Arabic. *Comparative Legilinguistics*, 63, 186–223. <https://doi.org/10.14746/cl.2025.63.1>
- Abu-Ayyash, E. A. (2017). Errors and non-errors in English-Arabic machine translation of gender-bound constructs in technical texts. *Procedia Computer Science*, 117, 73–80. <https://doi.org/10.1016/j.procs.2017.10.095>
- Ahmed, O., Tawfik, K., & Tohamy, M. (2025). Figurative language as a semantic barrier in the Arabic/English translation of United Nations General Assembly speeches. *British Journal of Translation Linguistics and Literature*, 5(2), 55–85. <https://doi.org/10.54848/mvgwe536>
- 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>
- 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>
- Alkhatib, M., & Shaalan, K. (2018). The key challenges for Arabic machine translation. In K. Shaalan, A. E. Hassanien, & F. Tolba (Eds.), *Studies in computational intelligence* (pp. 139–156). Springer. https://doi.org/10.1007/978-3-319-67056-0_8
- Alqudsi, A., Omar, N., & Shaker, K. (2012). Arabic machine translation: A survey. *Artificial Intelligence Review*, 42(4), 549–572. <https://doi.org/10.1007/s10462-012-9351-1>

- Altakhaineh, A. R. M., Alghathian, G. A., & Jarrah, M. M. (2024). A comparative study of accuracy in human vs. AI translation of legal documents into Arabic. *International Journal of Language and Law (JLL)*. <https://doi.org/10.14762/jll.2025.063>
- Alwazna, R. Y. (2024). The use of automation in the rendition of certain articles of the Saudi Commercial Law into English: A post-editing-based comparison of five machine translation systems. *Frontiers in Artificial Intelligence*, 6. <https://doi.org/10.3389/frai.2023.1282020>
- 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>
- Badah, A. M., Khalaf, C. N., Dwaikat, F. J., & AlQbailat, N. (2025). The usage of artificial intelligence in legal translation: Bridging the gap between law and language. *Ampersand*, 16, 100248. <https://doi.org/10.1016/j.amper.2025.100248>
- Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Association for Computational Linguistics. <https://aclanthology.org/W05-0909/>
- Bechara, H., Manohara, K., & Jankin, S. (2024). Creating and evaluating a multilingual corpus of UN General Assembly debates. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*. European Association for Machine Translation. <https://aclanthology.org/2024.eamt-1.52/>
- Bouamor, H., Alshikhabobakr, H., Mohit, B., & Oflazer, K. (2014). A human judgement corpus and a metric for Arabic MT evaluation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 207–213). Association for Computational Linguistics. <https://doi.org/10.3115/v1/d14-1026>
- Diab, N. (2021). Out of the BLEU: An error analysis of statistical and neural machine translation of WikiHow articles from English into Arabic. *CDELTA Occasional Papers in the Development of English Education*, 75(1), 181–211. <https://doi.org/10.21608/opde.2021.208437>
- Guzmán, F., Bouamor, H., Baly, R., & Habash, N. (2016). Machine translation evaluation for Arabic using morphologically-enriched embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Association for Computational Linguistics. <https://aclanthology.org/C16-1132/>
- Khasawneh, R. R., & Alsharif, B. B. (2025). A comparative study of AI-powered tools for Arabic-English and English-Arabic translation. *Journal of Language Teaching and Research*, 16(6), 2025–2035. <https://doi.org/10.17507/jltr.1606.24>
- Mohammed, T. A. S. (2025). Evaluating translation quality: A qualitative and quantitative assessment of machine and LLM-driven Arabic–English translations. *Information*, 16(6), 440. <https://doi.org/10.3390/info16060440>
- Moneus, A. M., & Sahari, Y. (2024). Artificial intelligence and human translation: A contrastive study based on legal texts. *Heliyon*, 10(6), e28106. <https://doi.org/10.1016/j.heliyon.2024.e28106>
- Nassar, H. (2025). Challenges of post-editing in English to Arabic machine translation of technical texts: A study of technological and linguistic barriers. *International Journal of Linguistics Literature & Translation*, 8(4), 01–15. <https://doi.org/10.32996/ijllt.2025.8.4.1>
- Popović, M. (2015). chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation* (pp. 392–395). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W15-3049>
- Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural*

- Language Processing (EMNLP) (pp. 2685–2702). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.213>
- Hadla, L. S., Hailat, T. M., Al-Kabi, M. N. (2015). Comparative study between METEOR and BLEU methods of MT: Arabic into English translation as a case study. *International Journal of Advanced Computer Science and Applications*, 6(11). <https://doi.org/10.14569/ijacsa.2015.061128>
- Sadiq, S. (2025). Evaluating English-Arabic translation: Human translators vs. Google Translate and ChatGPT. *Journal of Languages and Translation*, 12(1), 67–95. <https://doi.org/10.21608/jltmin.2025.423147>
- Shquier, M. M. A., & Sembok, T. M. T. (2008). Word agreement and ordering in English-Arabic machine translation. In 2008 International Symposium on Information Technology (pp. 1–10). IEEE. <https://doi.org/10.1109/itsim.2008.4631625>
- Vilar, D., & Black, K. (2026, January 16). TranslateGemma: A new suite of open translation models. Google. <https://blog.google/innovation-and-ai/technology/developers-tools/translategemma/>
- Zakraoui, J., Saleh, M., Al-Maadeed, S., & Alja'am, J. M. (2020). Evaluation of Arabic to English machine translation systems. In 2020 11th International Conference on Information and Communication Systems (ICICS) (pp. 185–190). IEEE. <https://doi.org/10.1109/icics49469.2020.239518>
- 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, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In International Conference on Learning Representations. <https://openreview.net/forum?id=SkeHuCVFDr>
- Ziemski, M., Junczys-Dowmunt, M., & Poulighen, B. (2016). The United Nations Parallel Corpus v1.0. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). European Language Resources Association. <https://aclanthology.org/L16-1561/>
- Zouidine, M., & Khalil, M. (2025). Large language models for Arabic sentiment analysis and machine translation. *Engineering Technology & Applied Science Research*, 15(2), 20737–20742. <https://doi.org/10.48084/etasr.9584>

Author's Biography

Yassine El Rhaffouli (yassine.elrhaffouli@etu.uae.ac.ma) is a PhD candidate and researcher in linguistics and translation studies with a strong interest in computational approaches to language analysis. His research focuses on machine translation evaluation, natural language processing, corpus-based translation studies, and translation pedagogy.

Hicham Boughaba (h.boughaba@uae.ac.ma) is a professor of linguistics and translation studies whose research spans applied linguistics, translation theory, psycholinguistics, and language pedagogy. His scholarly interests include discourse analysis, translation quality assessment, corpus-based studies, and the integration of digital and AI-driven tools in language and translation education.