# LQM: Linguistically Motivated Multidimensional Quality Metrics for Machine Translation

**Samar M. Magdy**[λ] **Fakhraddin Alwajih**[λ*] **Abdellah El Mekki**[λ*]
**Wesam El Sayed**[ξ] **Muhammad Abdul-Mageed**[λ,γ]

[λ]The University of British Columbia [ξ]Minia University [γ]Canada Research Chair in NLP and ML
{samar.ahmad, muhammad.mageed}@ubc.ca

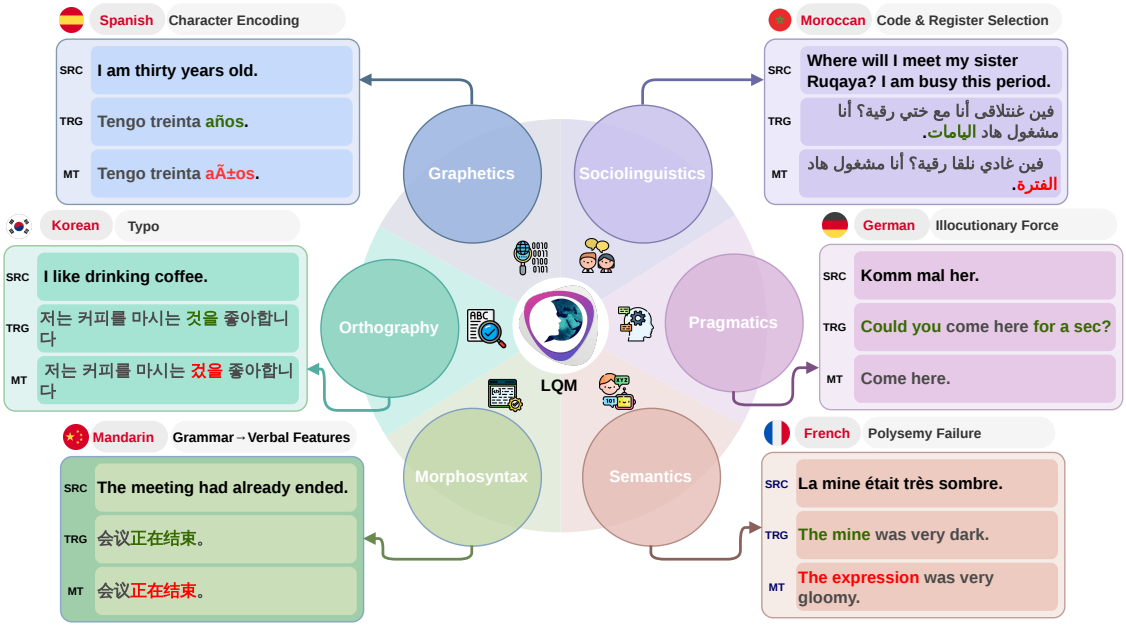


Figure 1: Cross-lingual examples illustrating the proposed LQM framework's linguistic levels, demonstrating its language-agnostic design and broad applicability beyond Arabic.

## Abstract

Existing MT evaluation frameworks, including automatic metrics and human evaluation schemes such as Multidimensional Quality Metrics (MQM), are largely language-agnostic. However, they often fail to capture dialect- and culture-specific errors in diglossic languages (e.g., Arabic), where translation failures stem from mismatches in language variety, content coverage, and pragmatic appropriateness rather than surface form alone. We introduce **LQM**: Linguistically Motivated Multidimensional Quality Metrics for MT. LQM is a hierarchical error taxonomy for diagnosing MT errors through six linguistically grounded levels: *sociolinguistics*, *pragmatics*, *semantics*, *morphosyntax*, *orthography*, and *graphetics* (Figure 1). We construct a bidirectional parallel corpus of 3,850 sentences (550 per variety) spanning seven Arabic dialects (*Egyptian*, *Emirati*, *Jordanian*, *Mauritanian*, *Moroccan*, *Palestinian*, and *Yemeni*), derived from conversational, culturally rich content. We evaluate six LLMs in a zero-shot setting and conduct expert span-level human annotation using LQM, producing 6,113 labeled error spans across 3,495 unique erroneous sentences, along with severity-weighted quality scores. We complement this analysis with an automatic metric (spBLEU). Though validated here on Arabic, LQM is a language-agnostic framework designed to be easily applied to or adapted for other languages. LQM annotated errors data, prompts, and annotation guidelines are publicly available at https://github.com/UBC-NLP/LQM_MT.

*Equal contribution



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 40470–40493

## 1 Introduction

Evaluating MT remains challenging, particularly when systems must preserve sociolinguistic and pragmatic constraints in addition to semantic content. While automatic metrics such as BLEU (Papineni et al., 2002), COMET (Rei et al., 2020), and chrF++ (Popović, 2017) provide rapid, scalable feedback, they primarily capture surface overlap or embedding similarity and can diverge from human judgments in cases where translation quality depends on variety choice, register, or cultural appropriateness (Yao et al., 2024; Chao, 2025). This has motivated increased use of human-in-the-loop evaluation to characterize failure modes (Brewster et al., 2025).

Among fine-grained human evaluation frameworks, MQM (Lommel et al., 2014) has become a widely adopted standard, offering a flexible error taxonomy for identifying translation issues. However, in complex diglossic languages such as Arabic (Ferguson, 1959; Bassiouney, 2020), we find that MQM-style categorizations can under-specify dialect- and culture-conditioned errors. In particular, when errors reflect mismatches in dialect, register, or pragmatic intent, surface-oriented error labels may not provide enough structure to consistently localize the linguistic level at which a model fails.

In this work, we analyze the limitations that arise when applying MQM to bidirectional MT involving Arabic dialects, and highlight three recurring challenges. *(i)* MQM provides an operational inventory of error types, but it does not explicitly index errors to linguistic levels, which can make it difficult to separate similar surface manifestations with different underlying explanations. *(ii)* The taxonomy offers limited built-in support for the systematic treatment of variation, including dialect choice, register, and culturally conditioned pragmatic constraints, which are central in dialectal Arabic. *(iii)* As a result, MQM annotations can be harder to translate into targeted corrective signals for improving dialectal Arabic MT, since distinct phenomena (e.g., pragmatic infelicity vs. semantic reference errors) may be grouped under broad categories.

To address these gaps, we introduce Linguistically Motivated Multidimensional Quality Metrics, dubbed **LQM**, a linguistically grounded taxonomy designed to diagnose MT errors in a way that is aligned with linguistic theory and practical annotation. We develop LQM through a two-pronged process aimed at balancing theoretical coverage with empirical adequacy: *Top-down:* We applied MQM to a pilot subset of our MT data and analyzed where existing categories and guidelines were insufficient to consistently capture dialectal and sociopragmatic phenomena. *Bottom-up:* We then performed iterative, data-driven refinement over observed errors (span-level), consolidating recurring patterns into categories organized by linguistic level. We synthesize these perspectives into a hierarchical framework spanning Sociolinguistics, Pragmatics, Semantics, Morphosyntax, Orthography, and Graphetics. Our contributions are: *(i) LQM framework:* We propose LQM, a linguistically grounded taxonomy that explicitly separates sociolinguistic and pragmatic phenomena from semantic and form-level errors, enabling more targeted diagnosis in diglossic and dialectal settings. The hierarchy supports two complementary annotation settings: a lightweight version for assigning broad error categories and a diagnostic version for labeling specific error types. *(ii) Parallel dataset:* We introduce a conversational, culturally rich parallel corpus covering seven Arabic varieties to stress-test models under dialect-sensitive conditions. *(iii) Multi-model evaluation:* We evaluate six LLMs with expert span-level human annotations (including severity-weighted scores) and complement this analysis with standard automatic metrics.

## 2 Related works

MQM has become a widely used framework for diagnostic MT evaluation and error taxonomy analysis (Lommel et al., 2014; Freitag et al., 2021). Recent work has pushed MQM-style evaluation toward finer granularity by localizing errors at the span level, including metric-based approaches such as xCOMET (Guerreiro et al., 2023) and prompted LLM evaluators such as GEMBA-MQM (Kocmi and Federmann, 2023; Fernandes et al., 2023). Related directions include agentic or multi-step evaluators and refinement pipelines (He et al., 2024; Wang et al., 2025a), as well as efforts to improve reliability by filtering or validating annotated spans via post-editing signals (Lu et al., 2025; Kocmi et al., 2024; Kreutzer et al., 2020).

In parallel, research in Automatic Post-Editing (APE) has evolved from classical formulations (Simard et al., 2007) to LLM-assisted pipelines (Bhattacharyya et al., 2023; Raunak et al., 2023).

Recent resources and protocols increasingly emphasize human-centered corrections and explainability, providing structured annotations or rationales that can support targeted diagnosis (Wasti et al., 2025; Jung et al., 2024; Alves et al., 2024). Large Reasoning Models (LRMs) further enable multi-step explanations for ambiguity resolution (Liu et al., 2025), motivating work that constrains or structures LLM-based MQM annotation, for example, via compressed taxonomies (ThinMQM) or tagged span annotation (Zhan et al., 2025; Yeom et al., 2025a; He et al., 2025; Wang et al., 2025b; Feng et al., 2025).

Despite these advances, applying general-purpose MQM-style frameworks to languages with diverse varieties such as Arabic remains challenging due to diglossia and dialectal variation (Ferguson, 1959; Bassiouney, 2020), as well as non-standardized orthographies across dialects (Habash et al., 2012). Dialectal MT research has leveraged resources such as Alexandria (Mekki et al., 2026) and MADAR (Bouamor et al., 2018), has increasingly studied LLM-based systems in specific dialectal settings (Yakhni and Chehab, 2025; Fernandes et al., 2023). Yet, morphological and syntactic variations continue to complicate error identification and interpretation (Zbib et al., 2012; Sajjad et al., 2020). Existing benchmarks such as Tarjamat (Kadaoui et al., 2023) and NADI 2024 (Abdul-Mageed et al., 2024) underscore persistent gaps for low-resource dialects, but they do not provide a linguistically leveled, dialect-sensitive diagnostic taxonomy for isolating how and where translations lose dialectal identity or pragmatic appropriateness.

These limitations motivate LQM, which organizes MQM-style error annotation by linguistic level (from sociolinguistics and pragmatics to form-level phenomena), enabling a more consistent diagnosis in dialectal Arabic settings.

## 3 Pitfalls of MQM

MQM (Lommel et al., 2014) is widely used for human diagnostic evaluation and is intentionally designed to be broadly applicable across languages. In our setting, bidirectional MT involving multiple Arabic dialects, we find that MQM's generic categories can under-specify error patterns that are primarily driven by variety choice and sociopragmatic constraints rather than surface form alone (Abdul-Nabi et al., 2024).

| MQM Category | MQM Subcategory | Count | Rate (%) |
|---|---|---|---|
| **Accuracy** | **Mistranslation** | **3,673** | **60.09** |
| Accuracy | Addition | 453 | 7.41 |
| Accuracy | Missing | 357 | 5.84 |
| Fluency | Grammar | 325 | 5.32 |
| Style | Unidiomatic style | 322 | 5.27 |
| Style | Awkward style | 292 | 4.78 |
| Terminology | Wrong term | 192 | 3.14 |
| Accuracy | Undertranslation | 130 | 2.13 |
| Fluency | Spelling | 125 | 2.04 |
| Accuracy | Overtranslation | 66 | 1.08 |
| Accuracy | Untranslated | 57 | 0.93 |
| Terminology | Inconsistent with term resource | 47 | 0.77 |
| Style | Language register | 31 | 0.51 |
| Style | Inconsistent style | 15 | 0.25 |
| Locale conv. | Currency format | 9 | 0.15 |
| Fluency/Ling. conv. | Punctuation | 8 | 0.13 |
| Fluency | Inconsistency/unintelligible | 5 | 0.08 |
| Fluency | Character encoding | 4 | 0.07 |
| Locale convention | Number format | 1 | < 0.02 |
| Terminology | Inconsistent use of terminology | 1 | < 0.02 |
| **Total** | | **6,113** | **100.00** |

Table 1: MQM error counts and rates aggregated over annotated spans. The Mistranslation row is highlighted to indicate its frequent use as a catch-all label in our setting.

**Setup.** We applied MQM to translations from six Arabic-aware LLMs on a bidirectional translation dataset covering English and seven Arabic dialects. Six annotators—two senior linguists and four trained annotators—annotated a stratified sample of 50 sentences per model and translation direction. Each dialect was evaluated by annotators who are native speakers of that dialect. Full experimental details are provided in Section 5.

**Findings.** Table 1 reports aggregated MQM span annotations. A prominent pattern is the dominance of the *Mistranslation* label (60.09% of annotated errors). Based on qualitative inspection and annotator feedback, *Mistranslation* frequently functions as a catch-all label for heterogeneous phenomena that do not fit cleanly under other MQM tags, reducing diagnostic specificity regarding where the model fails.

Annotators reported that many such cases reflect three recurring sources of error: *(i) Pragmatic mismatches:* illocutionary force, discourse markers, vocatives, and honorifics. *(ii) Variety consistency:* defaulting to MSA or drifting into a different dialect. *(iii) Idiomatic usage:* mistranslation of proverbs and other fixed expressions.

## 4 LQM Framework

Motivated by the MQM analysis above, we introduce LQM, a linguistically grounded taxonomy

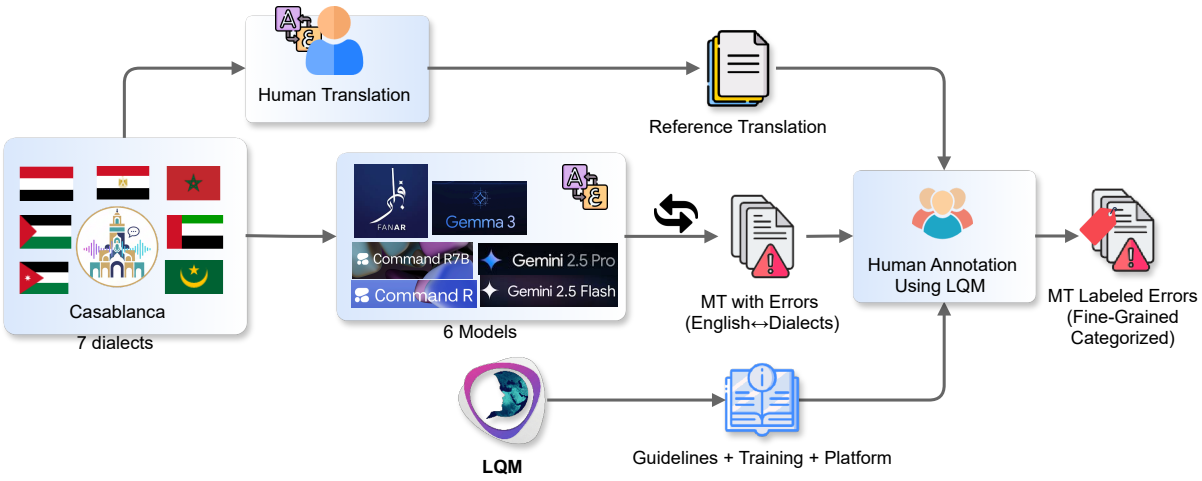


Figure 2: Data and annotation workflow. Casablanca dataset of seven Arabic dialects is translated by humans and evaluated by six LLM models; creating LQM guidelines and training support human annotation to produce fine-grained MT error labels.

for span-level human MT evaluation. LQM is designed to improve diagnostic precision in diglossic and dialectal settings by distinguishing sociolinguistic and pragmatic failures from semantic and form-level errors. Specifically, LQM organizes errors into six linguistic levels: *Sociolinguistics*, *Pragmatics*, *Semantics*, *Morphosyntax*, *Orthography*, and *Graphetics*. We summarize these levels below; full definitions and additional examples are provided in Appendix §A.

**(i) Sociolinguistics.** We place sociolinguistics at the top of the hierarchy to reflect communicative competence (Hymes et al., 1972): a translation may be grammatically well-formed yet inappropriate for the social context. This is especially consequential in Arabic, where diglossia makes variety choice functional rather than merely stylistic (Ferguson, 1959). LQM therefore introduces *code & register selection* errors with three subcategories: *(a) standardization interference (vertical mismatch)*, *(b) wrong dialect (horizontal mismatch)*, and *(c) register mismatch (tone/formality)*.

**(ii) Pragmatics.** While sociolinguistics targets broader social norms, the pragmatics level captures failures of communicative intent and implied meaning (Levinson, 1983), i.e., mismatches between sentence meaning and speaker meaning. We group these under *use, context, and cultural appropriateness* and include targeted subcategories: *(a) speech acts/illocutionary force*, *(b) code switching*, *(c) MWEs/proverbs*, *(d) discourse marker mismatch*, and *(e) vocatives/honorifics /titles* (Farwell and Helmreich, 1999).

**(iii) Semantics.** This level evaluates meaning transfer, and preservation of propositional content (Cruse, 1986). We distinguish: *(a) lexical semantics* (word meaning and lexical relations), *(b) propositional semantics* (truth-conditional content; avoiding unintended additions/omissions) (Soames, 1987), and *(c) discourse semantics* (cohesion and reference across sentences) (Halliday and Hasan, 2014; Kamp and Reyle, 2013). Example subcategories include *named entity*, *wrong term*, *polysemy failure*, and *cross-variety interference*; see Appendix§A for complete definitions.

**(iv) Morphosyntax.** This level captures violations of target-language structural constraints at the morphology-syntax interface (Radford, 2004). We separate: *(a) grammar* (e.g., agreement and inflectional features) (Corbett, 2006), including *verbal features* (tense, aspect, voice, mood, person) (Palmer, 2001) and *nominal features* (number, gender, case, definiteness, state); and *(b) constituent order* (Greenberg et al., 1963), including locale-sensitive reordering such as *address format* and *date format*.

**(v) Orthography/Writing Conventions.** This level evaluates written-form conventions and mechanical correctness (Derwing, 1992). Since dialectal Arabic lacks fully standardized orthography, we accept dialectal spellings recognized by native annotators as conventional in informal digital contexts (e.g., texting and social media).[1] We define five error types: *(a) spelling* (including *ty-*

[1]Given limited standardized dialect orthographies, we rely on annotator judgments for acceptability within emerging, socially shared informal conventions.

*pos/slips*), *(b) inconsistent spelling*, *(c) unconventional Spelling*, *(d) surface mechanics* (number, currency, time, telephone formats), and *(e) punctuation*.

**(vi) Graphetics.** The lowest level captures failures in the technical realization of the text code. We include *character encoding* errors, where the output is garbled due to encoding/decoding issues.

A comprehensive breakdown of LQM, including both the lightweight and diagnostic versions and their subcategories, is provided in Table A.1 (Appendix). We also report an external validation of LQM conducted by two linguists specializing in linguistics and translation studies in Appendix §B.

## 5 Experimental Setup

We conduct a case study to evaluate LQM in a stress-test setting for dialectal Arabic MT with Arabic-aware LLMs. Our experimental design mirrors the MQM analysis in Section 3, but replaces MQM with the LQM annotation protocol described in Section 4. Below, we describe the dataset, the evaluated models, and our human and automatic evaluation procedures. Figure 2 summarizes the data construction and LQM-based annotation workflow used in our experiments.

### 5.1 Dataset

As part of our contribution, we construct a new bidirectional parallel corpus of 3,850 sentences based on the manually transcribed conversational speech data introduced by Talafha et al. (2024). The corpus covers seven Arabic varieties: *Egyptian*, *Emirati (UAE)*, *Jordanian*, *Mauritanian*, *Moroccan*, *Palestinian*, and *Yemeni*. The dialectal transcripts were then professionally translated into MSA with support from native speakers of each dialect and subsequently translated into English.

The dataset is balanced with 550 sentences per variety. It is organized for bidirectional MT: Dialect→English (DA→EN) and English→Dialect (EN→DA). We report DA→EN results for all seven dialects. While for EN→DA, the Jordanian, and Yemeni dialects were excluded due to the lack of available native-speaker translators for target-side validation.

### 5.2 Evaluated LLMs

We select six Arabic-aware LLMs spanning both closed and open-weight systems. For closed-source models, we evaluate `Gemini-2.5-Pro` and `Gemini-2.5-Flash` (Team et al., 2023). For open-weight models, we evaluate `Fanar-9B` (Team et al., 2025), `Gemma-27B` (Team et al., 2024), `Command-A (111B)` (CohereLabs, 2024), and `Command-R7B (7B)`. All models are evaluated in a zero-shot setting using a fixed prompt (Appendix§C).

### 5.3 Human Labeling via LQM

**Annotation guidelines.** LQM relies on expert human annotation to yield both *(i)* fine-grained error diagnostics and *(ii)* severity-weighted quality scores. We developed annotation guidelines that define LQM categories and provide Arabic examples to support consistent span selection and labeling. Comprehensive annotation guidelines are available in the previously mentioned repository.

**Annotation scope.** Annotators applied LQM to a random sample of 50 translations per direction, drawn from the full set of outputs generated by the six models across dialects and directions. Across this annotated sample, annotators identified 3,495 unique translations containing at least one error and labeled 6,113 error spans.[2] Each labeled error includes: *(i)* span boundaries, *(ii)* LQM category/subcategory/sub-subcategory, *(iii)* severity (minor/major/critical), and *(iv)* an optional free-text explanation. Appendix §D describes the annotation process, annotator profiles, and quality assurance workflow. Figure 3 presents representative examples from our dataset across different dialects, highlighting the error spans for each level.

### 5.4 Evaluation Metrics

#### 5.4.1 LQM score derivation

Our LQM score equation is inspired by the MQM original scoring equation (Lommel et al., 2014), we follow the same severity weights $s_i$: *minor*=1, *major*=5, and *critical*=25. Then, we fix a weight of 1 across all the error types. For a translation with length $L$ (in words) and annotated errors $\{i\}$ with severity weights $s_i$, we compute:

$$\text{LQM\_Score} = \max\left(0,\ 100\left(1 - \frac{\sum_i s_i}{L}\right)\right)$$

This yields a normalized score in $[0, 100]$, where higher values indicate better translation quality.

[2] A sentence may receive multiple error labels, and the full set of 3,850 sentence pairs also includes outputs judged error-free.


Figure 3: LQM-annotated examples across the six linguistic levels. Severity and fine-grained error types are indicated in each example. Error spans are highlighted in red. Icons indicate model families: G Gemini, C Command, F Fanar, GM Gemma.

LQM scores are based on human judgments (span selection and severity assignment) rather than automatic matching metrics.

#### 5.4.2 Automatic surface metrics

We report automatic scores using *spBLEU* (Goyal et al., 2022), a sentence-level variant of BLEU (Papineni et al., 2002) to compute the correlation between LQM human annotated errors and automatic scores. We do not report model-based metrics such as COMET (Rei et al., 2020), since their performance and calibration may be less reliable for dialectal Arabic varieties that are underrepresented in common training and evaluation resources.

## 6 Results of Error Analysis

### 6.1 LQM Error Counts and Rates

Table 2 summarizes aggregated error counts and rates under LQM. Compared to MQM, which frequently assigned heterogeneous phenomena to the broad *Mistranslation* label, LQM reallocates these cases into linguistically interpretable categories. In particular, many instances previously collapsed under the MQM *Mistranslation* are separated into *(i)* sociolinguistic failures (e.g., code/register selection), *(ii)* pragmatic infelicities (e.g., discourse markers, vocatives, MWEs/proverbs), and *(iii)* semantic failures (e.g., named entities, polysemy, lexical coverage), enabling a more targeted diagnosis of where models break down. The primary limitations of current LLMs in dialectal Arabic–English translation appear to be linguistic rather than merely computational. We therefore provide additional linguistic analysis in Appendix§E.

### 6.2 Severity-weighted LQM Scores

Table 3 reports severity-weighted LQM scores (0–100) across models and directions. Overall, `Gemini-2.5-Pro` achieves the strongest performance, ranking first in 9 of the 12 evaluated directions. Performance is generally higher and more stable in DA→EN than in EN→DA, consistent with the additional constraint in EN→DA of maintaining the target dialect's sociolinguistic identity.

Across directions, EN→UAE yields the highest scores for most models, whereas EN→MOR is the most challenging setting, with most models scoring below 40 (with the `Gemini` variants as notable exceptions). At the model level, `Command-R7B` exhibits the weakest overall performance, with notably low scores on EN→UAE (21.94) and EN→MOR (17.00). `Fanar-9B` shows substantial variance across dialects, performing strongly on EN→UAE (78.32) but dropping sharply on EN→MOR (6.46). Finally, several other models lead in specific directions (e.g., *Command-A* on JOR→EN, `Gemini-2.5-Flash` on

| LQM Category | LQM Subcategory | LQM Subsubcategory | Count | Rate (%) |
|---|---|---|---|---|
| sociolinguistics | code & register selection | standardization interference (vertical mismatch) | 904 | 14.8 |
| semantics | lexical semantics | named entity | 572 | 9.4 |
| sociolinguistics | code & register selection | wrong dialect (horizontal mismatch) | 539 | 8.8 |
| semantics | lexical semantics | coverage: unknown term/dialect | 365 | 6.0 |
| semantics | propositional semantics | omission | 357 | 5.8 |
| semantics | lexical semantics | unnatural/ unidiomatic style | 322 | 5.3 |
| morphosyntax | grammar | verbal features | 295 | 4.8 |
| semantics | lexical semantics | awkward style | 292 | 4.8 |
| pragmatics | use, context, cultural appropriateness | mwes, proverbs & metaphors | 289 | 4.7 |
| semantics | discourse semantics | pronouns | 268 | 4.4 |
| semantics | propositional semantics | addition | 263 | 4.3 |
| semantics | lexical semantics | disambiguation: cross-variety interference | 246 | 4.0 |
| semantics | lexical semantics | wrong term | 192 | 3.1 |
| semantics | propositional semantics | hallucination | 190 | 3.1 |
| semantics | lexical semantics | undertranslation | 130 | 2.1 |
| orthography/ writing conventions | spelling | typo / slip | 125 | 2.0 |
| semantics | lexical semantics | disambiguation: polysemy failure | 120 | 2.0 |
| semantics | lexical semantics | transliteration | 115 | 1.9 |
| pragmatics | use, context, cultural appropriateness | speech acts mismatch | 104 | 1.7 |
| pragmatics | use, context, cultural appropriateness | forms of address (vocatives/honorifics/titles) | 97 | 1.6 |
| semantics | lexical semantics | overtranslated | 66 | 1.1 |
| semantics | lexical semantics | untranslated | 57 | 0.9 |
| semantics | discourse semantics | inconsistent with terminology resource | 47 | 0.8 |
| pragmatics | use, context, cultural appropriateness | discourse marker mismatch | 35 | 0.6 |
| sociolinguistics | code & register selection | register mismatch | 31 | 0.5 |
| morphosyntax | grammar | nominal features | 30 | 0.5 |
| pragmatics | use, context, cultural appropriateness | code switching | 19 | 0.3 |
| semantics | discourse semantics | inconsistent style | 15 | 0.2 |
| orthography/ writing conventions | surface mechanics | currency format | 9 | 0.1 |
| orthography/ writing conventions | punctuation | — | 8 | 0.1 |
| semantics | lexical semantics | unintelligible | 5 | <0.1 |
| graphetics | character encoding | — | 4 | <0.1 |
| orthography/ writing conventions | surface mechanics | number format | 1 | <0.1 |
| semantics | discourse semantics | inconsistent use of terminology | 1 | <0.1 |
| **Total** | | | **6,113** | **100.0** |

Table 2: LQM error counts and rates utilizing the LQM Diagnostic layer for fine-grained analysis.

MOR→EN, and `Fanar-9B` on EN→MAU), suggesting that relative strengths depend on both direction and dialect.

### 6.3 Correlations Between Automatic and Human Evaluation

Table 4 reports spBLEU scores per direction and model. We compute the correlation between spBLEU and human-derived LQM scores to quantify the agreement between surface-based metrics and expert judgments. We observe a weak positive association (Pearson $r = 0.289$, Spearman $\rho = 0.322$, $p < 0.001$), indicating that spBLEU captures only a limited portion of the variance in severity-weighted human assessments. This gap is expected in our setting, where major quality degradations often stem from sociolinguistic and pragmatic phenomena (e.g., dialect/register mismatches, vocatives/honorifics, discourse markers) that are not well-modeled by n-gram overlap. In Appendix F, we provide additional experiments examining the robustness of LQM to sentence length.

### 6.4 Inter-Annotator Agreement (IAA)

We report IAA using three complementary types of scores: span-detection scores, label-agreement scores on overlapping spans, and chance-corrected agreement measured with Cohen's kappa ($\kappa$). The analysis was conducted on 377 doubly annotated items from both translation directions. Following span-based MT evaluation work (Yeom et al., 2025b), we use overlap-based span agreement as the primary detection metric and exact span F1 as a stricter secondary metric. Agreement was strong under overlap-based matching (character-level F1 = 0.760; overlap span F1 = 0.821) but lower under exact span matching (F1 = 0.440), suggesting that annotators usually identified the same error regions while differing in precise span boundaries. On overlapping spans, agreement was highest for coarse error category (F1 = 0.662; $\kappa$ = 0.681), followed by severity (F1 = 0.630; $\kappa$ = 0.427)

| Direction | Fanar-9B | Command-A | Command-R7B | Gemma-27b | Gemini-2.5-flash | Gemini-2.5-pro |
|---|---|---|---|---|---|---|
| ENGLISH → DIALECT | | | | | | |
| ENG→EGY | 49.90 | 71.14 | 35.89 | 54.44 | 65.23 | **72.31** |
| ENG→MAU | **45.26** | 41.87 | 37.40 | 19.62 | 38.28 | 40.90 |
| ENG→MOR | 6.46 | 38.60 | 17.00 | 19.29 | 51.53 | **68.60** |
| ENG→PAL | 38.96 | 59.89 | 43.29 | 51.91 | 66.56 | **67.14** |
| ENG→UAE | 78.32 | 72.89 | 21.94 | 63.39 | 82.10 | **83.07** |
| DIALECT → ENGLISH | | | | | | |
| EGY→ENG | 53.57 | 60.88 | 45.56 | 68.69 | 74.45 | **75.89** |
| JOR→ENG | 65.27 | **73.26** | 58.22 | 69.76 | 72.27 | 66.19 |
| MAU→ENG | 40.76 | 56.38 | 43.90 | 61.45 | 59.26 | **63.88** |
| MOR→ENG | 40.98 | 64.45 | 51.50 | 62.34 | **72.15** | 70.32 |
| PAL→ENG | 64.78 | 73.18 | 62.62 | 72.89 | 73.42 | **79.47** |
| UAE→ENG | 43.85 | 66.61 | 45.08 | 54.06 | 62.65 | **67.09** |
| YEM→ENG | 61.28 | 62.90 | 58.07 | 67.00 | 70.76 | **73.41** |

Table 3: LQM scores (severity-weighted, 0–100) by model and direction. Higher is better.

| | Open-Weight Models | | | | Proprietary Models | |
|---|---|---|---|---|---|---|
| Direction | Fanar-9B | Command-A | Command-R7B | Gemma-27B | Gemini-2.5-Flash | Gemini-2.5-Pro |
| **English → Dialect** | | | | | | |
| ENG → EGY | 13.97 | 23.38 | 13.15 | 19.84 | 24.11 | **26.09** |
| ENG → MAU | 1.42 | 1.75 | 2.61 | 4.24 | 4.34 | **5.88** |
| ENG → MOR | 2.64 | 10.18 | 6.97 | 9.28 | 15.39 | **18.30** |
| ENG → PAL | 9.65 | 15.84 | 12.20 | 16.48 | 21.07 | **23.26** |
| ENG → UAE | 5.67 | 13.21 | 5.89 | 11.17 | 17.78 | **19.77** |
| **Dialect → English** | | | | | | |
| EGY → ENG | 28.89 | 31.62 | 26.70 | 27.54 | **32.18** | 31.47 |
| JOR → ENG | 29.19 | 31.78 | 26.56 | 28.99 | **32.22** | 31.88 |
| MAU → ENG | 10.13 | 12.82 | 8.96 | 11.19 | 16.03 | **16.59** |
| MOR → ENG | 16.99 | 23.19 | 17.64 | 19.27 | **24.00** | 23.34 |
| PAL → ENG | 25.67 | **31.17** | 23.47 | 25.87 | 30.56 | 27.12 |
| UAE → ENG | 20.91 | 27.18 | 19.83 | 23.36 | **27.89** | 26.90 |
| YEM → ENG | 22.54 | 23.98 | 20.17 | 22.09 | 25.97 | **26.26** |

Table 4: **spBLEU scores** across translation directions and model families. Higher is better. The best result in each row is highlighted in green and boldfaced.

and category+severity (F1 = 0.517; $\kappa$ = 0.509). Fine-grained error types were more variable (F1 = 0.484), and the strictest criterion—joint agreement on span, error type, and severity—yielded F1 = 0.388. Overall, the results indicate reliable error detection and coarse categorization, with greater variability in fine-grained labeling.

### 6.5 Analysis of Error Distributions and Model Attribution

In Table 5, we analyze the human-annotated corpus along two axes: *(i)* model-wise contribution to the total error mass, and *(ii)* the distribution of error types by translation direction. Figure A.1 (Appendix) summarizes model-wise error contributions and presents the overall distribution across LQM categories and fine-grained subcategories.

**Model-wise error contribution.** In DA→EN, `Command-R` and `Fanar` contribute the largest share of errors, with `Command-R` peaking in Egyptian (26.7%) and `Fanar` peaking in Moroccan (23.3%) and Emirati (23.8%). In contrast, *Gemini-2.5-Pro* consistently contributes the smallest error mass across dialects (e.g., 8.5% in JOR→ENG). In EN→DA, `Fanar`, `Command-R`, and `Gemma` dominate the error mass. Peaks are observed for `Command-R` in EN→UAE (34.4%), `Fanar` in EN→MOR (33.0%), and `Gemma` in EN→MAU (22.7%). *Gemini-2.5-Pro* continues to contribute substantially less, reaching as low as 2.1% of the error mass in EN→MOR and 7.0% in

| Direction | PART I: Model Error Contribution (%) *(Who Failed?)* | | | | | | PART II: Error Type Distribution (%) *(Why?)* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Cmd-A | Cmd-R | Fanar | Gemma | Flash | Pro | Soc | Prag | Sem | Morph | Orth | Gra |
| | | | | | | ***Dialect → English*** | | | | | | |
| EGY → ENG | 8.0 | **26.7** | 26.6 | 14.8 | 12.2 | 11.8 | 2.4 | 24.3 | **67.7** | 5.4 | 0.2 | – |
| JOR → ENG | 12.3 | **26.2** | 8.9 | 25.5 | 18.6 | 8.5 | 0.4 | 14.8 | **82.6** | 1.6 | 0.4 | 0.2 |
| MAU → ENG | 18.1 | 18.1 | **20.0** | 19.5 | 14.0 | 10.3 | 0.7 | 3.4 | **92.3** | 3.4 | 0.2 | – |
| MOR → ENG | 15.5 | 21.6 | **23.3** | 19.2 | 8.4 | 12.0 | – | 10.4 | **85.1** | 3.9 | 0.6 | – |
| PAL → ENG | 15.0 | **23.3** | 18.4 | 18.8 | 15.0 | 9.5 | – | 10.9 | **87.0** | 1.4 | 0.6 | – |
| UAE → ENG | 13.4 | **23.8** | **23.8** | 14.9 | 13.0 | 11.2 | – | 10.8 | **84.5** | 4.1 | 0.4 | 0.2 |
| YEM → ENG | 16.3 | **18.8** | 18.6 | 13.5 | 17.4 | 15.4 | 0.7 | 11.9 | **84.4** | 2.8 | 0.2 | – |
| | | | | | | ***English → Dialect*** | | | | | | |
| ENG → EGY | 19.7 | **30.8** | 16.8 | 14.4 | 11.3 | 7.0 | **40.1** | 7.2 | 39.9 | 8.4 | 4.3 | 0.2 |
| ENG → MAU | 12.5 | 16.2 | 18.3 | **22.7** | 16.7 | 13.7 | **81.5** | 4.9 | 11.8 | 1.9 | – | – |
| ENG → MOR | 14.4 | 20.4 | **33.0** | 23.9 | 6.2 | 2.1 | **39.6** | 0.6 | 35.9 | 19.2 | 4.7 | – |
| ENG → PAL | 15.9 | 22.5 | **24.7** | 18.3 | 9.5 | 9.1 | **58.5** | 2.4 | 20.5 | 6.0 | 12.4 | 0.2 |
| ENG → UAE | 15.1 | **34.4** | 17.3 | 13.8 | 10.0 | 9.3 | **70.6** | 4.5 | 15.4 | 5.0 | 4.5 | – |

Table 5: **Diagnostic dashboard of dialectal failures.** Model error contribution (Part I) and error-type distribution (Part II) by translation direction. In *DA→EN* (top), **Semantic** errors dominate across dialects. In *EN→DA* (bottom), **Sociolinguistic** errors dominate overall, especially for **Mauritanian**, **Palestinian**, and **Emirati**, while **Egyptian** and **Moroccan** show a more mixed pattern, with **Semantic** errors remaining prominent and **Morphosyntax** also notable in Moroccan. *(Soc=Sociolinguistics, Prag=Pragmatics, Sem=Semantics, Morph=Morphosyntax, Orth=Orthography, Gra=Graphetics).*

EN→EGY.

**Error typology by direction.** We observe a direction-dependent shift in the dominant failure mode. In DA→EN, errors are primarily semantic: semantic categories account for 67.7% (Egyptian) up to 92.3% (Mauritanian) of the error mass, with particularly high concentrations in Maghrebi dialects (Mauritanian 92.3%, Moroccan 85.1%). Pragmatic errors form a notable secondary cluster in dialects such as Egyptian (24.3%) and Jordanian (14.8%), indicating difficulties with communicative intent even when the literal meaning is partially recovered. Morphosyntactic and orthographic errors are comparatively rare in DA→EN (often $\leq 5.4\%$), consistent with target-side normalization when generating English.

In EN→DA, the distribution shifts significantly toward sociolinguistic failures. This shift is most pronounced in EN→MAU, where sociolinguistic errors account for 81.5% of the error mass. Similar patterns are observed in EN→UAE (70.6%) and EN→PAL (58.5%). However, in Egyptian and Moroccan, semantic errors remain a high secondary failure mode (39.9% and 35.9%, respectively). We also observe increases in morphosyntactic and orthographic errors in generation (e.g., morphosyntax up to 19.2% for Moroccan; orthography 12.4% for Palestinian). Appendix G provides a more detailed analysis of the fine-grained error distributions per dialect.

## 7 Conclusion

We presented LQM, a linguistically grounded framework for MT evaluation that organizes errors by linguistic level, enabling diagnostic analysis in diglossic and dialect-rich settings. Crucially, while our empirical evaluation centers on Arabic, the underlying taxonomy is inherently language-agnostic and readily adaptable to other linguistic contexts. In a case study of seven Arabic dialects, we observed a direction-dependent shift in failure modes: DA→EN is largely driven by semantic breakdowns (67.7–92.3%), reflecting challenges in lexical coverage and meaning transfer from dialectal input, whereas EN→DA is dominated by sociolinguistic failures, with systems often defaulting to dialect-faithful varieties instead of maintaining the target dialect. Improving dialectal MT requires optimizing *both* semantic adequacy and sociolinguistic fidelity.

We also found that spBLEU aligns only weakly with severity-weighted LQM scores (Pearson $r = 0.289$, Spearman $\rho = 0.322$), consistent with n-gram overlap's insensitivity to pragmatic and dialect-identity errors emphasized by human annotation.

## Acknowledgments

We acknowledge support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 895-2020-1004), the Canadian Foundation for Innovation (CFI; 37771), the Digital Research Alliance of Canada,[3] and UBC ARC-Sockeye.[4]

We thank Aisha Alraeesi for annotating the Emirati Arabic data translated from English and Najla Hassan for annotating the Palestinian Arabic data translated from English. We are also grateful to Yayhay Mohamed Elhaj and Sidi Ebidi for annotating the English-to-Mauritanian direction, to Alcides Alcoba for support with the annotation platform, and to Abderahim Elmadany for his feedback. Finally, we thank the external linguists Ranada Hassan and Saudi Sadiq for independently validating the LQM framework.

## 8 Limitations

Our study has several limitations.

- **Dialect and direction coverage:** Although our study covers seven Arabic varieties overall, we report EN→DA evaluation for only five dialects—Egyptian, Emirati, Mauritanian, Moroccan, and Palestinian. We exclude Jordanian and Yemeni because we were unable to secure enough qualified native-speaker translators to validate outputs in those target dialects. Therefore, our conclusions about dialect preservation in the EN→DA setting apply only to the five evaluated varieties.
- **Domain specificity:** The corpus is derived from transcribed TV dialog, which is conversational and culturally grounded. Although this setting is well-suited for eliciting dialect- and pragmatics-related errors, results may differ in more formal or specialized domains (e.g., legal, medical, or technical translation) with distinct terminology and register constraints.
- **Non-standardized orthography:** Arabic dialect writing lacks fully standardized conventions. We accept spellings judged by native annotators to be common in informal digital contexts, but orthographic variation can complicate consistent span-level localization and make automatic evaluation less straightforward.
- **Prompting and inference settings:** All models are evaluated in a zero-shot setting with a fixed prompt. We do not systematically study the effects of alternative prompting strategies (e.g., few-shot exemplars, constrained output formats) or inference-time controls, which may change both overall quality and the distribution of LQM error types.

## Ethical Considerations

This work studies MT quality for Arabic dialects and introduces LQM, a linguistically grounded framework for span-level error annotation. Because dialectal data can reflect speakers' regional and social identities, we minimize the risk of sensitive attribute inference by *(i)* reporting results at the dialect/variety level rather than attempting to infer or annotate personal attributes such as gender, age, socioeconomic status, or education, and *(ii)* restricting annotations to translation errors and linguistic phenomena observable in text (e.g., code/register selection, pragmatic appropriateness, semantics, and form-level issues).

Our corpus is derived from previously released material and contains conversational content; we acknowledge that such data may include culturally specific expressions or potentially sensitive topics. Annotators were instructed to focus on translation quality rather than judge speakers or communities, and to provide brief explanations only when needed for clarity. We also recognize that resources for dialectal MT can be misused to generate targeted or stereotyped content; to mitigate this, we provide documentation emphasizing appropriate use and the limitations of automatic metrics for dialect identity and pragmatics, and we avoid presenting LQM as a tool for profiling individuals.

## References

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. NADI 2024: The fifth nuanced Arabic dialect identification shared task. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages

[3] https://alliancecan.ca

[4] https://arc.ubc.ca/ubc-arc-sockeye

709–728, Bangkok, Thailand. Association for Computational Linguistics.

Razan Abdul-Nabi, Rasha Obeidat, and Anas Bsoul. 2024. A survey on machine translation of low-resource arabic dialects. In *2024 15th International Conference on Information and Communication Systems (ICICS)*, pages 1–6.

Duarte M Alves, José P Pombal, Nuno M Guerreiro, and André FT Martins. 2024. xtower: Multilingual translation error explanation and correction with large language models. *arXiv preprint arXiv:2406.19482*.

John Langshaw Austin. 1975. *How to do things with words*. Harvard university press.

Reem Bassiouney. 2020. *Arabic sociolinguistics: Topics in diglossia, gender, identity, and politics*. Georgetown University Press.

Pushpak Bhattacharyya, Rajen Chatterjee, Markus Freitag, Diptesh Kanojia, Matteo Negri, and Marco Turchi. 2023. Findings of the wmt 2023 shared task on automatic post-editing. In *Conference on Machine Translation*.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and 1 others. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Ryan CL Brewster, Gabriel Tse, Angela L Fan, Marwa Elborki, Maiah Newell, Priscilla Gonzalez, Amitra Hoq, Crystal Chang, Maksud Chowdhury, Adiba Geeti, and 1 others. 2025. Evaluating human-in-the-loop strategies for artificial intelligence-enabled translation of patient discharge instructions: a multidisciplinary analysis. *NPJ digital medicine*, 8(1):629.

Dunstan Brown, Marina Chumakina, and Greville G Corbett. 2013. *Canonical morphology and syntax*. Oxford University Press.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Yang Chao. 2025. Natural language processing and deep learning in cross-cultural language acquisition: From machine translation to cultural context understanding. *Theoretical and Natural Science*, 92:13–18.

CohereLabs. 2024. Command R+: A 104b parameter open-weight model for rag and tool use.

Greville G Corbett. 2006. Introduction: Canonical agreement. In *Agreement*, pages 1–34. Cambridge University Press.

D Alan Cruse. 1986. *Lexical semantics*. Cambridge university press.

Bruce L Derwing. 1992. Orthographic aspects of linguistic competence. *The linguistics of literacy*, 21:193–211.

David Farwell and Stephen Helmreich. 1999. Pragmatics and translation. *Procesamiento del lenguaje natural, nº 24 (mayo 1999); pp. 18-39*.

Zhaopeng Feng, Shaosheng Cao, Jiahan Ren, Jiayuan Su, Ruizhe Chen, Yan Zhang, Zhe Xu, Yao Hu, Jian Wu, and Zuozhu Liu. 2025. Mt-r1-zero: Advancing llm-based machine translation via r1-zero-like reinforcement learning. *ArXiv*, abs/2504.10160.

Charles A Ferguson. 1959. Diglossia. *word*, 15(2):325–340.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André F. T. Martins, Graham Neubig, Ankush Garg, J. Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Conference on Machine Translation*.

Bruce Fraser. 1999. What are discourse markers? *Journal of pragmatics*, 31(7):931–952.

Bruce Fraser. 2009. An account of discourse markers. *International review of Pragmatics*, 1(2):293–320.

Markus Freitag, George F. Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 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.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Joseph H Greenberg and 1 others. 1963. Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*, 2:73–113.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luísa Coheur, Pierre Colombo, and André Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

John J Gumperz. 1982. *Discourse strategies*. 1. Cambridge University Press.

Nizar Habash, Mona T Diab, and Owen Rambow. 2012. Conventional orthography for dialectal arabic. In *LREC*, pages 711–718.

Michael AK Halliday. 1978. *Language as social semiotic*. London Arnold.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 2014. *Cohesion in english*. Routledge.

Minggui He, Yilun Liu, Shimin Tao, Yuanchang Luo, Hongyong Zeng, Chang Su, Li Zhang, Hongxia Ma, Daimeng Wei, Weibin Meng, Hao Yang, Boxing Chen, and Osamu Yoshie. 2025. R1-t1: Fully incentivizing translation capability in llms via reasoning learning. *ArXiv*, abs/2502.19735.

Zhiwei He, Xing Wang, Wenxiang Jiao, Zhuosheng Zhang, Rui Wang, Shuming Shi, and Zhaopeng Tu. 2024. Improving machine translation with human feedback: An exploration of quality estimation as a reward model. *ArXiv*, abs/2401.12873.

Dell Hymes, JB Pride, and Janet Holmes. 1972. On communicative competence. sociolinguistics. *Eds. Pride, JB y J. Holmes*, pages 269–293.

Dahyun Jung, Sugyeong Eo, Chanjun Park, and Heuiseok Lim. 2024. Explainable ced: A dataset for explainable critical error detection in machine translation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 25–35.

Karima Kadaoui, Samar M. Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed Oumar El-Shangiti, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. TARJAMAT: Evaluation of bard and ChatGPT on machine translation of ten Arabic varieties. In *Proceedings of ArabicNLP 2023*, pages 52–75, Singapore (Hybrid). Association for Computational Linguistics.

Hans Kamp and Uwe Reyle. 2013. *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*, volume 42. Springer Science & Business Media.

Tom Kocmi and Christian Federmann. 2023. Gemba-mqm: Detecting translation quality error spans with gpt-4. *ArXiv*, abs/2310.13988.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popovi'c, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. *ArXiv*, abs/2406.11580.

Julia Kreutzer, Nathaniel Berger, and Stefan Riezler. 2020. Correct me if you can: Learning from error corrections and markings. In *European Association for Machine Translation Conferences/Workshops*.

Stephen C Levinson. 1983. *Pragmatics*. Cambridge university press.

Sinuo Liu, Chenyang Lyu, Minghao Wu, Longyue Wang, Weihua Luo, Kaifu Zhang, and Zifu Shang. 2025. New trends for modern machine translation with large reasoning models. *ArXiv*, abs/2503.10351.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Qingyu Lu, Liang Ding, Kanjian Zhang, Jinxia Zhang, and Dacheng Tao. 2025. Mqm-ape: toward high-quality error annotation predictors with automatic post-editing in llm translation evaluators. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5570–5587.

Elin McCready. 2019. *The semantics and pragmatics of honorification: Register and social meaning*, volume 11. Oxford University Press.

Abdellah El Mekki, Samar M Magdy, Houdaifa Atou, Ruwa AbuHweidi, Baraah Qawasmeh, Omer Nacar, Thikra Al-hibiri, Razan Saadie, Hamzah Alsayadi, Nadia Ghezaiel Hammouda, and 1 others. 2026. Alexandria: A multi-domain dialectal arabic machine translation dataset for culturally inclusive and linguistically diverse llms. *arXiv preprint arXiv:2601.13099*.

Geoffrey Nunberg, Ivan A Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Frank Robert Palmer. 2001. *Mood and modality*. Cambridge university press.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Andrew Radford. 2004. *English syntax: An introduction*. Cambridge University Press.

Vikas Raunak, Amr Sharaf, Hany Hassan Awadallah, and Arul Menezes. 2023. Leveraging gpt-4 for automatic translation post-editing. In *Conference on Empirical Methods in Natural Language Processing*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020. AraBench: Benchmarking dialectal Arabic-English machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5094–5107, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Deborah Schiffrin. 1987. *Discourse markers*. 5. Cambridge University Press.

John R Searle. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge university press.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007. Statistical phrase-based post-editing. In *North American Chapter of the Association for Computational Linguistics*.

Scott Soames. 1987. Direct reference, propositional attitudes, and semantic content. *Philosophical topics*, 15(1):47–87.

Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Mohamedou Cheikh Tourad, Rahaf Alhamouri, Rwaa Assi, Aisha Alraeesi, Hour Mohamed, Fakhraddin Alwajih, Abdelrahman Mohamed, Abdellah El Mekki, El Moatez Billah Nagoudi, Benelhadj Djelloul Mama Saadia, Hamzah A. Alsayadi, Walid Al-Dhabyani, and 8 others. 2024. Casablanca: Data and models for multidialectal Arabic speech recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21745–21758, Miami, Florida, USA. Association for Computational Linguistics.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, and 1 others. 2025. Fanar: An arabic-centric multimodal generative ai platform. *arXiv preprint arXiv:2501.13944*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Jiaan Wang, Fandong Meng, Yunlong Liang, and Jie Zhou. 2025a. Drt: Deep reasoning translation via long chain-of-thought. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6770–6782.

Jiaan Wang, Fandong Meng, and Jie Zhou. 2025b. Deep reasoning translation via reinforcement learning. *arXiv preprint arXiv:2504.10187*.

Adnan Wasti, Matthew Lee, Tausifa Alam, Sreyasi Ghosh, and Marine Carpuat. 2025. Translationcorrect: A human-centered post-editing framework for error-aware machine translation. In *Proceedings of ACL 2025 (to appear)*.

Malak Yakhni and Jeanine Chehab. 2025. Fine-tuning arabic llms for lebanese dialect translation and evaluation. *arXiv preprint arXiv:2405.12534*.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking machine translation with cultural awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13078–13096, Miami, Florida, USA. Association for Computational Linguistics.

Taemin Yeom, Yonghyun Ryu, Yoonjung Choi, and Jinyeong Bak. 2025a. Tagged span annotation for detecting translation errors in reasoning llms. *Proceedings of the Tenth Conference on Machine Translation*.

Taemin Yeom, Yonghyun Ryu, Yoonjung Choi, and JinYeong Bak. 2025b. Tagged span annotation for detecting translation errors in reasoning llms. In *Proceedings of the Tenth Conference on Machine Translation*, pages 878–886.

Rabih Zbib, Spyros Matsoukas, Richard Schwartz, John Makhoul, Enrique Jimenez, and Chad Malarkey. 2012. Machine translation of arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Runzhe Zhan, Zhihong Huang, Xinyi Yang, Lidia S. Chao, Min Yang, and Derek F. Wong. 2025. Are large reasoning models good translation evaluators? analysis and performance boost. *ArXiv*, abs/2510.20780.

# Appendices

We offer an additional structure as follows:

- LQM Framework §A
- LQM External Validation §B
- Prompts §C
- Data Annotation §D
- Linguistic Insights on LLMs Performance §E
- Robustness of LQM to Sentence Length §F
- LQM Fine-Grained Error Distribution §G

## A LQM Framework

Structured along the hierarchy of linguistic analysis, the LQM framework targets six distinct layers: **sociolinguistic**, **pragmatic**, **semantic**, **morphosyntactic**, **orthographic/writing conventions**, and **graphetic**.

**(i) Sociolinguistic**: LQM introduces the *code & register selection* type of error under the sociolinguistic level, explicitly penalizing three main subcategories: *(a) standardization interference (vertical mismatch)*, where the model reverts to use the standardized, "high," or prestige variety (e.g., Standard German, MSA) when a specific vernacular or "low" variety is requested. For example, the use of MSA appearing in Emirati data as قوم إبتعد عني ("Get up, stay away from me"), where إبتعد عني is an MSA expression rather than the Emirati one.

*(b) wrong dialect (horizontal mismatch)*: Where the model uses features specific to a different regional or social variety than the target (e.g., using Mexican Spanish slang in a translation for Spain, or Egyptian idioms in a Levantine text. For example, in the Jordanian sentence جاسر جا هنا أكتر من مرة، وهيدا ليه مليون سبب. ("Jasser came here more than once, and this has a million reasons."), the model mixes Jordanian with other dialects, using the Lebanese form هيدا and the Gulf form جا.

(Halliday, 1978)'s conceptualization of language as a "social semiotic," wherein register is defined as the specific language variety linked to the field, tenor, and mode of a situation. Accordingly, we include *(c) register mismatch (tone/formality)* as a distinct error type to ensure translations are evaluated not just on propositional content, but on their adherence to the tenor of discourse (e.g., formal vs. informal), a critical dimension often lost in NMT outputs that default to a neutral register (e.g., using "Tu" instead of "Vous" in French, or casual slang in a legal document).

**(ii) Pragmatics**: While sociolinguistics governs the broad social norms of interaction, the pragmatic level addresses failures in communicative intent and implied meaning, as well as the gap between sentence meaning and speaker meaning. We classify errors at this level under the error types of *use, context, and cultural appropriateness*, capturing instances where the translation is grammatically valid but functionally or culturally anomalous (Farwell and Helmreich, 1999). In high-context languages like Arabic, literal translation often yields semantically accurate but pragmatically hollow outputs. To capture these nuances, LQM introduces five targeted subcategories: *(a) speech acts/illocutionary force*: We evaluate whether the translation preserves the *illocutionary force* of the source based on Speech Act Theory (Austin, 1975; Searle, 1969). NMT models can flatten the pragmatic force of an utterance (e.g., converting a polite request into a direct command). It covers errors in interpreting speech acts (e.g., requests, threats, advice, jokes). For instance, the phrase ان شاء الله literally means "God willing", but its pragmatic function can vary significantly to mean "even if", "maybe", or "hopefully" depending on use.

*(b) code switching*: This category identifies failures where the model either ignores embedded foreign lexical items or forces them into Arabic morphology inappropriately, signaling a failure to recognize the shift in linguistic code. We categorize CS under pragmatics because language alternation serves specific communicative functions (Gumperz, 1982), often includes both situational and metaphorical switching. CS serves as a communicative strategy, linking it to social meaning and interactional goals. for example, the model translated 'I want you to bring me twelve bananas' into Arabic as أبيك تجيب لي ٢١ بنانة . This represents a code-switching error, where the English root is morphologically integrated into an Arabic verb form.

*(c) MWEs/proverbs*: Multi-word expressions (MWEs) often function as single non-compositional units. Consistent with theories of semantic non-compositionality (Nunberg et al.,

1994), this subcategory targets "literalism errors," where models translate the constituent parts of an idiom rather than its figurative meaning. For instance, the proverb و مش كل مره تسلم الجره was translated literally as 'and the jar doesn't stay safe every time.' This fails to convey the underlying warning (i.e., 'you won't get lucky every time'), resulting in a nonsensical output for the target audience.

*(d) discourse marker mismatch*: An inconsistency between the semantic/pragmatic meaning of a discourse marker and the context in which it is used (Schiffrin, 1987; Fraser, 2009). This can lead to a lack of coherence or an unintended meaning in a conversation or text. Cohesion in Arabic relies on discourse markers (e.g., *wa, fa*) that signal logical relations distinct from English (Fraser, 1999). We include this subcategory to capture errors where the model misinterprets these procedural cues, disrupting the text's argumentative structure or flow.For example, the model mistranslated the phrase والله in the sentence والله طلبك هو الي هيحدد as 'God willing, your request is what will determine.' This represents a pragmatic failure; the model interpreted والله as a literal religious oath (or confused it with إن شاء الله), failing to recognize its function here as a discourse marker used for emphasis/hedging.

*(e) vocatives/honorifics/titles*: This category evaluates the translation of forms of address, which serve crucial social functions in communication. We penalize failures to map honorifics or vocative particles (e.g., *Ya*) to their target equivalents, often resulting in a loss of the intended politeness level or social hierarchy (Brown and Levinson, 1987; McCready, 2019). This error frequently occurs when models treat titles as literal nouns rather than pragmatic markers of respect. In this example, the model failed to identify the vocative address in the sentence أنا أبويه، أحسن عشان النظافة, mistranslating it as 'I'm his father, it's better for hygiene.' The correct interpretation should be a direct address: 'Dad, I prefer it for the sake of hygiene.

**(iii) Semantics**: The Semantics level evaluates the fidelity of meaning transfer (Cruse, 1986), focusing on the preservation of propositional content. We classify these errors into three primary domains: *(i)* lexical semantics, *(ii)*propositional semantics, and *(iii)* discourse semantics. **(a) lexical semantics**: Under this level, we propose subcategories specifically tailored for MT in general, and for Arabic dialectal in particular. These include: *named entity*, failure of the model to translate proper nouns, geographical locations, or organizations correctly. For example, أم فخري where (Um Fakhri) is translated into (Um Fakhr). *wrong term*: This refers to the incorrect use of a term that violates domain-expert usage (e.g., legal or medical, etc.). For instance, the model translates the word comedians" in the phrase break with comedians is very different" into المهرّجين instead of ممثلين; *overtranslation*, where the model uses a specific word (hyponym) when the source used a general word (hypernym) (e.g., "Car" vs. "Ferrari"); *undertranslation*, where the model uses a general word when the source uses a specific word (e.g., "Ferrari" vs. "Vehicle"); *transliteration*: This category applies when the model provides a phonetic rendering of the source text rather than a semantic translation. For example, translating أيو الحمد لله الحمد لله المهم as "Ayou al-hamd lillah al-hamd lillah" instead of its English equivalent (e.g., Yes, thank God, anyway"); *unidiomatic/unnatural style*, where the output is grammatically correct but sounds unnatural; *awkward style*, where the style involves excessive wordiness or overly embedded clauses; *unintelligible*, where the output is garbled or incomprehensible; *measurement units*: use of an inappropriate measurement format for its locale; for example, converting units (lbs → kg) is a lexical change required to preserve semantic reality; *coverage: unknown term/dialect*: This category includes cases where the model fails to recognize a specific dialectal lemma. For example, in the sentence سيبوني أكمل يا عواطليه, the model mistranslated the Egyptian term عواطليه ('unemployed') as 'you crazy ones.' This indicates a coverage failure where the model lacked the necessary lexical representation for this dialect-specific term, resulting in incorrect English output; *disambiguation: polysemy failure*, where the model knows the word but selects the wrong meaning (your "Sign" vs "Knock down" example); and *disambiguation: cross-variety interference*, where the model knows the word but assigns it the Standard meaning instead of the Dialect meaning. For example, the model misinterpreted the homograph شرابين in the phrase شرابين حريمي أصفر و أحمر as 'drinks,' rendering it as 'Two girly yellow and red drinks.' The correct translation is 'two pairs of women's socks.' This represents a failure to distinguish the dialectal

term for socks (shurrab) from the standard term for drinks (sharab).

**(b) Propositional Semantics**: Truth-conditional semantics (Soames, 1987) defines the meaning of a sentence by specifying the conditions under which it is true or false. This category addresses the conservation of information, ensuring that the target text neither adds nor subtracts from the message of the source. We categorize these errors as follows: *addition*, where the translation contains information not found in the source; *omission*, where source content is missing from the target; and *untranslated*, where segments remain in the source language in the final output; and finally *hallucination*: This error occurs when the model generates content that is not supported by the source text, often producing output that is fluent but semantically unrelated. For instance, the model translated هلأ أختك متأخرها كبير into 'Now, your sister is very spoiled.' The concept of 'spoiled' is a hallucination; it has no basis in the source text, which actually refers to a 'large delay' or 'arrears' (متأخرها)

**(c) Discourse Semantics**: This domain includes four subcategories: *inconsistent use of terminology*, where the model uses multiple terms for the same concept in contexts where consistency is desirable; *inconsistent with terminology resource*, when the model uses a term that differs from term usage required by a specified termbase or other resource; *pronouns*, when the model fails to correctly render pronouns, causing changes in speaker or referent, or flipping the intended gender; and *inconsistent style*, when the style varies inconsistently throughout the text.

**(iv) Morphosyntax**: Evaluates adherence to the structural rules of the target language, focusing on the interface between morphology and syntax (Radford, 2004). We classify errors into two types: *grammar*: Addresses violations of morphological agreement (Brown et al., 2013; Corbett, 2006), capturing errors in *wrong number, gender, and verb tense*. We include two specific subcategories under this type: *verbal features*, where the translation violates target rules regarding tense, aspect, voice, mood, and person (Palmer, 2001).For example, the Moroccan phrase نتا دابا عا مخمخ was mistranslated as 'you've really cooked up something.' This indicates a mistranslation of syntactic form, where the model rendered the intended imperative mood as a declarative statement; and *nominal features*, which covers errors in number, gender, case, definiteness, and state (Idafa). For instance, the model failed to translate the subject in this Mauritanian sentence and changed it into something else يا خياتي راني لكم فاصل to be "Oh, my brothers, I'm taking a break". *constituent order*: Focuses on the linear arrangement of elements (Greenberg et al., 1963). We explicitly include *address format* and *date format* here, as these require specific reordering to match the target locale conventions rather than direct translation.

| **Category** (Lightweight LQM) | **Error Type** (Lightweight LQM) | **Subcategory** ***(Diagnostic LQM)*** | **Definition** |
|---|---|---|---|
| **Sociolinguistics** | Language Register | Standardization Interference | Use of MSA instead of the target variety. |
| | | Wrong Dialect | Output overlaps with or uses an incorrect dialect. |
| | | Register Mismatch | Formality level higher or lower than required. |
| **Pragmatics** | Use, Context, Cultural Appropriateness | MWEs/ Proverbs | Fails to deliver idiomatic translation; misuse of expression. |
| | | Code Switching | Failure to handle foreign words or recognize code shift. |
| | | Speech Acts/ Illocutionary Force | Intended illocutionary force or speaker intention not conveyed. |
| | | Discourse Marker Mismatch | Misuse of discourse markers affecting cohesion. |
| **Semantics** | Lexical Semantics | Named Entity | Failing to map the proper noun to the correct referent. |
| | | Wrong term | The term is invalid for the domain or creates a conceptual mismatch. |
| | | Overtranslation | Using a specific word ($Hyponym$) for a general source word ($Hypernym$). |
| | | Undertranslation | Using a general word ($Hypernym$) for a specific source word ($Hyponym$). |
| | | Transliteration | Incorrect phonetic rendering into the target language. |
| | | Unidiomatic Style | The style is grammatical but unnatural. |
| | | Awkward Style | Excessive wordiness or overly embedded clauses. |
| | | Unintelligible | The text is garbled or incomprehensible. |
| | | Measurement Units | The measurement format is inappropriate for the locale. |
| | | Coverage | The model fails to recognize a specific dialectal lemma. |
| | | Polysemy Failure | The model picks the wrong meaning for a polysemous word. |
| | | Cross-Variety Interference | Standard meaning assigned instead of the Dialect meaning. |
| | Propositional Semantics | Addition | The target includes information that is not present in the source. |
| | | Omission | Content present in the source is missing from the target. |
| | | Untranslated | The source segment is carried over without translation. |
| | | Hallucination | Adding new information that changes the facts. |
| | Discourse Semantics | Inconsistent use of terminology | Multiple terms used for the same concept. |
| | | Inconsistent with terminology resource | The usage of terms differs from the specified resource. |
| | | Pronouns | Incorrect pronoun causing a change in speaker/gender. |
| | | Inconsistent Style | The style varies inconsistently throughout the text. |
| **Morphosyntax** | Grammar (wrong number, gender, verb tense) | Verbal Features | Violates grammatical rules (Tense, Aspect, Mood, etc). |
| | | Nominal Features | Violates nominal rules (Number, Gender, Case, etc). |
| | constituent order | Address Format | Inappropriate address format for locale. |
| | | Date format | Inappropriate date format for locale. |
| **Orthography/ Writing conventions** | Spelling | Typo / Slip | Obvious mechanical error (e.g., typos). |
| | Inconsistent Spelling | - | Same word spelled differently within text. |
| | Unconventional Spelling | - | Spelling hard to read, even if phonetic. |
| | Surface Mechanics | Number Format | Inappropriate number format for locale. |
| | | Currency | Incorrect currency format for locale. |
| | | Time format | Incorrect time format for locale. |
| | | Telephone | Inappropriate telephone number format. |
| | Punctuation | - | Incorrect according to target conventions. |
| **Graphetics** | - | Character Encoding | Characters garbled due to incorrect encoding. |



Table A.1: Hierarchical classification of the LQM. Categories and error types represent the Lightweight LQM; subcategories represent Fine-grained analysis (Diagnostic LQM).

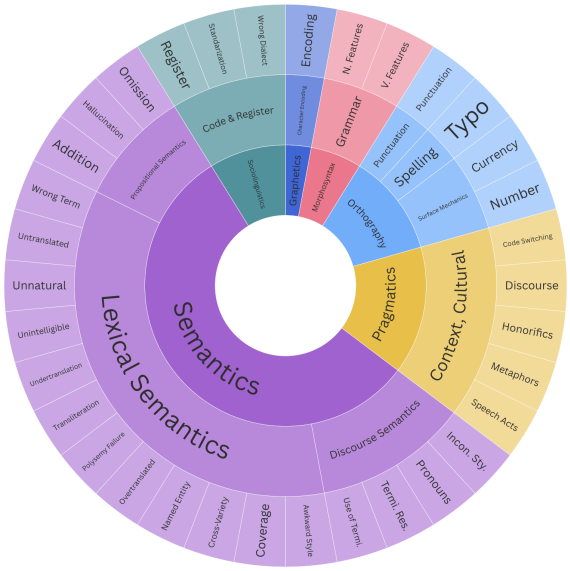


Figure A.1: Distribution of LQM error categories in our dataset, showing that semantic and lexical-semantic errors constitute the majority of labeled errors.

**(v) Orthography/Writing Conventions**: In this level, we evaluate the mechanical correctness of the written output, focusing on the visual representation of language (Derwing, 1992). We classify errors into five specific types: *spelling*: This category addresses deviations from standard orthographic norms. It includes the *typos/slips* subcategory for obvious mechanical errors (e.g., character insertions or deletions) that violate the fixed spelling rules of the target language. We also distinguish *between inconsistent spelling* and *unconventional spelling*. The category of *surface mechanics* governs non-lexical formatting conventions and comprises four subcategories: *number format*, *currency*, *time format*, and *telephone*. Finally, *punctuation* addresses cases where punctuation marks are missing, misused, or inconsistent with target language conventions.

**(vi) Graphetics**: The lowest level of the hierarchy addresses the technical realization of the text code. We identify one primary failure mode, which is *character encoding*, where the output text is garbled due to incorrect encoding or decoding processes. For example, in this model output أز يا±. Illustrative examples of LQM categories and subcategories are in Table G.2.

## B LQM External Review

To obtain external expert feedback on the proposed framework, we invited two linguists specializing in Arabic linguistics and translation to assess its conceptual soundness and practical applicability. Overall, both reviewers viewed the proposed LQM as a valid and promising linguistically motivated framework for machine translation evaluation. In particular, they highlighted its relevance for diglossic language settings, where variation across standard and non-standard varieties, register, and sociolinguistic meaning is central to translation quality. They identified this as a key strength of the framework, especially for Arabic and related contexts.

At the same time, the reviewers noted that some category boundaries, particularly those spanning semantic, pragmatic, and sociolinguistic dimensions, would benefit from further clarification. They also suggested streamlining some fine-grained subcategories to improve annotation consistency and usability. Taken together, their feedback supports the relevance of the framework for MT assessment in diglossic languages while also identifying areas for refinement. In response, we revised the category boundaries to reduce potential overlap and strengthened the definitions and illustrative examples associated with each category.

## C Prompts

## D Data Annotation

Annotator selection went through multiple stages of quality assurance. First, we developed detailed annotation guidelines that included illustrative examples tailored to the seven Arabic dialects covered in our study. These examples were drawn from a wide range of regional varieties to ensure clarity and consistency for annotators from different Arab countries. Next, we designed an annotation interface that incorporated all LQM categories along with the subcategories introduced in our framework. We uploaded all model outputs to this interface to facilitate the annotation process. Before launching the full annotation round, we conducted an internal pilot in which we sampled a subset of the data. This allowed us to verify that the tool functioned properly and that all necessary features were available to support accurate and efficient annotation. During the pilot, we identified several model-generated outputs that were ambiguous or did not fit neatly into the existing error categories. In response, we added a comment box to the interface so that annotators could provide feedback on model behavior or flag unlisted error types, such as pragmatic errors. After refin-

**Prompt**

**Translation Prompt Generator (Both Directions ENG↔Dialect)**

```
def create_prompt(src, trg):
   if src == 'ENG':
      prompt = f'Translate the following English phrase into {langs_map[trg]}
Arabic dialect written in Arabic script.  Your response must only contain the
translated text, with no additional explanations or labels.'
   else:
      prompt = f'Translate the following {langs_map[src]} Arabic dialect phrase
into English.  Your response must only contain the translated text, with no
additional explanations or labels.'
   return prompt
```

Figure C.1: Prompt used to generate all dialectal variants in our dataset (EN→Dialect; Dialect→EN).

ing the tool and guidelines based on these insights, we onboarded all annotators by conducting live demonstrations that walked them through both the annotation procedures and the interface for labeling span-level errors. All annotators were invited to share feedback and comments, which we incorporated as part of our iterative quality assurance process.

**Annotator Profiles.** To ensure linguistic expertise and dialectal authenticity, we employed direction-specific annotation teams:

- **Dialect→English**: This direction was annotated by two senior linguists (Ph.D. holders) specializing in translation studies and linguistics, both of whom are native Egyptian Arabic speakers. To mitigate source-side ambiguity, annotators were provided with both the original dialectal text and its MSA equivalent as a reference. Approximately 40% of the dataset was finalized through collaborative live sessions to ensure full agreement, leveraging the MSA context to resolve nuanced dialectal expressions and improve overall comprehension.

- **English→Dialect**: This task involved four native speakers (Moroccan, Palestinian, Emirati, and Mauritanian), each holding an MA or Ph.D. in related fields. Each annotator labeled only their native dialect to ensure authentic judgments of naturalness. For the English-to-Egyptian direction, the annotation was performed by the same two linguists mentioned previously. For each of the two directions, 40% of the data were labeled by two annotators in live sessions with full agreement, once full agreement was reached, the rest of the data was annotated by a single annotator.

## E Linguistic insights on the LLMs performance

The primary limitations of current LLMs in dialectal Arabic–English translation are linguistic rather than merely computational. Our analysis shows that the most persistent errors stem from interpreting dialect-specific semantics, the dominant failure mode in the Dialect → English direction. Across dialects, Semantic errors account for 67.7%–92.3% of total error mass, indicating that the main bottleneck is weak lexical and conceptual mapping for idiomatic and culturally situated expressions. These often encode *illocutionary force* or social alignment that surface-level lexical correspondence cannot recover, producing translations that are formally plausible but pragmatically deficient. This is especially visible in high-resource dialects such as Egyptian, where *Pragmatic* errors reach 24.3%. Models also show systematic weakness in dialectal *morphosyntax*; even in comprehension, Morphosyntactic failures remain notable in dialects such as Mauritanian (3.4%), where non-canonical constructions diverge from MSA norms.

These challenges intensify and change structurally in the English → Dialect direction, where failure shifts from semantic decoding to deficient *sociolinguistic* authenticity. Models display a pronounced **"identity crisis,"** often defaulting to standardization (MSA-vertical mismatch) or producing hybrid forms (wrong dialect-horizontal mismatch). This is reflected in the rise of *Sociolinguistic* errors, peaking at 70.6% for UAE and 58.5% for Palestinian generation. The problem is especially pronounced in open-source architectures: while `Pro` contributes as little as 2.1% to the error mass in

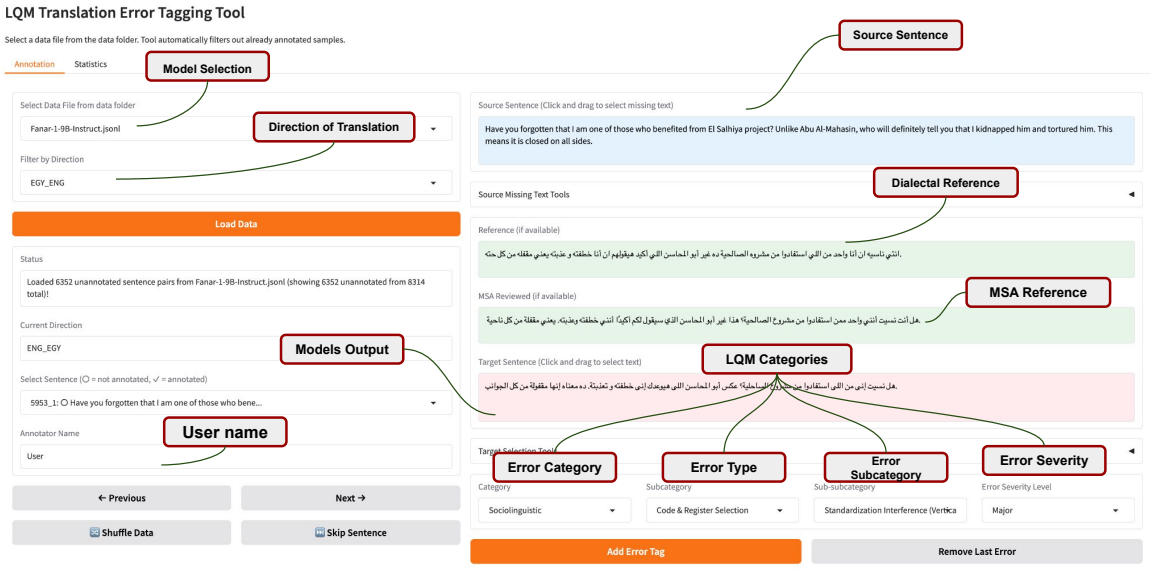


Figure D.1: Screenshot of our LQM annotation tool for error labeling.

Moroccan generation, open-source models such as `Fanar` and `Command-R` contribute up to 33.0% and 20.4%, respectively. This suggests a data-poverty effect in smaller or open-source models, yielding overgeneralized constructions and fewer culturally grounded pragmatic markers. Overall, these findings show that existing LLMs inadequately model the interaction of morphology, syntax, and sociolinguistic variation, motivating the more granular, linguistically informed evaluation provided by LQM.

## F Robustness of LQM to Sentence Length

Per-sentence normalization can inflate or dilute scores for very short or very long segments. To verify that our findings are not artifacts of such effects, we conduct three complementary analyses: *(i)* corpus-level micro-averaged LQM, *(ii)* a robustness check across length buckets, and *(iii)* a rank-stability test using Spearman $\rho$.

**Micro-averaged LQM.** Instead of averaging per-segment scores, we accumulate all error mass and all token counts across the corpus before dividing once:

$$\mathrm{LQM}_{\mu} = \max\Big(0,\ 100 - 100 \cdot \frac{\sum_s E_s}{\sum_s L_s}\Big) \qquad (1)$$

This "sum first, divide once" principle eliminates the sensitivity to segment length. Table F.1 reports micro-averaged scores for selected directions.

| Dir. | Fanar | Cmd-A | Cmd-R7B | Gemma | Flash | Pro |
|---|---|---|---|---|---|---|
| Egy→En | 49.4 | 62.5 | 39.7 | 73.0 | 77.4 | **77.5** |
| En→Egy | 53.6 | 70.8 | 9.9 | 56.6 | 68.1 | **71.9** |
| Pal→En | 67.1 | 78.0 | 64.3 | 74.7 | 76.1 | **81.0** |
| Uae→En | 24.0 | 66.2 | 36.8 | 53.2 | 55.3 | **66.5** |
| Yem→En | 64.0 | 57.8 | 52.9 | 69.3 | 72.1 | **75.0** |
| **Avg.** | 51.6 | 67.1 | 40.7 | 65.4 | 69.8 | **74.4** |

Table F.1: Micro-averaged LQM ($\mathrm{LQM}_{\mu}$) for selected directions. The model ranking is consistent with the per-sentence results in Table 3.

The model ranking under micro-averaged scoring is consistent with the per-sentence results reported in Table 3. `Gemini-2.5-Pro` ranks first in the majority of directions under both formulations, and `Command-R7B` remains the weakest model overall. The direction-level pattern—EN→MOR as the most challenging—is also preserved.

**Length-bucket analysis.** We split all segments into *short*, *medium*, and *long* using the 33rd and 66th percentiles of target-side token count (quantile bucketing ensures balanced counts per bucket). The cut-offs are: **short** $\leq$ 14 tokens ($n$=1,165), **medium** 15–22 tokens ($n$=1,233), and **long** $>$ 22 tokens ($n$=1,080). We recompute micro-averaged LQM within each bucket for every for each mfor each model–direction combination. Tables F.2 and F.3 show results for two representative direc-

| Model | Length bucket | | |
|---|---|---|---|
| | Short | Medium | Long |
| Gemini-2.5-Pro | **66.1** | **73.5** | 80.0 |
| Gemini-2.5-Flash | 66.0 | 71.1 | **81.3** |
| Gemma-27b | 58.9 | 65.0 | 78.3 |
| Command-A | 37.4 | 64.9 | 65.9 |
| Fanar-9B | 37.8 | 43.2 | 53.9 |
| Command-R7B | 0.0 | 45.3 | 48.3 |

Table F.2: Micro-averaged LQM by length bucket for Egy→Eng.

| Model | Short | Medium | Long |
|---|---|---|---|
| Gemini-2.5-Pro | **76.0** | **75.2** | 82.3 |
| Gemma-27b | 65.6 | 67.2 | 76.8 |
| Gemini-2.5-Flash | 58.7 | 66.8 | 79.5 |
| Command-A | 50.0 | 60.9 | **82.8** |
| Fanar-9B | 69.8 | 58.3 | 69.7 |
| Command-R7B | 48.1 | 44.9 | 68.7 |

Table F.3: Micro-averaged LQM by length bucket for Pal→Eng.

tions. The top-tier models (`Gemini-2.5-Pro`, `Gemini-2.5-Flash`) and the weakest model (`Command-R7B`) maintain their relative positions across all three length buckets in the vast majority of directions.

**Rank stability.** To quantify stability formally, we compute the Spearman rank correlation of model scores between pairs of length buckets for each direction (Table F.4). Averaged across all 12 directions, Spearman $\rho = 0.71$ (short vs. medium), 0.71 (medium vs. long), and 0.62 (short vs. long), confirming that model rankings are substantially preserved across length strata. The exceptions—Uae→Eng ($\rho \approx 0.06$–$0.43$) and Eng→Mau ($\rho \approx -0.15$–$0.70$)—reflect genuine variation in error profiles within those dialect pairs rather than scoring artifacts, consistent with the higher cross-model variability observed for these directions in Table 3.

**Summary.** Both the micro-averaged formulation and the length-bucket analysis confirm that our main conclusions—`Gemini-2.5-Pro` leading overall, EN→MOR as the most challenging direction, and direction-dependent shifts in error type—are **robust to sentence length effects** and are not driven by score distortion in short segments.

| Direction | S vs M $\rho$ | M vs L $\rho$ | S vs L $\rho$ |
|---|---|---|---|
| Egy→Eng | $0.829^{*}$ | $0.886^{*}$ | $0.886^{*}$ |
| Eng→Egy | $0.943^{**}$ | 0.771 | $0.829^{*}$ |
| Eng→Mau | −0.696 | −0.029 | −0.145 |
| Eng→Mor | $1.000^{**}$ | 0.791 | 0.791 |
| Eng→Pal | $0.943^{**}$ | $0.886^{*}$ | 0.771 |
| Eng→Uae | 0.600 | $0.886^{*}$ | $0.829^{*}$ |
| Jor→Eng | 0.657 | 0.714 | $0.829^{*}$ |
| Mau→Eng | $0.829^{*}$ | $0.886^{*}$ | $0.886^{*}$ |
| Mor→Eng | 0.657 | 0.771 | $0.829^{*}$ |
| Pal→Eng | 0.657 | 0.600 | 0.200 |
| Uae→Eng | 0.145 | 0.429 | 0.058 |
| Yem→Eng | 0.600 | $0.886^{*}$ | 0.714 |
| **Mean** | **0.71** | **0.71** | **0.62** |

Table F.4: Spearman $\rho$ of model rankings across length buckets. S = Short, M = Medium, L = Long. $^{*}\,p < 0.05$; $^{**}\,p < 0.01$.

## G LQM Fine-Grained Error Distribution Across Models and Per Direction

Table G.1 shows the error distribution across the 6,113 samples. It reveals a sharp divide between translation directions, with Dialect-to-English accounting for 58.6% of failures compared to 41.4% in **English-to-Dialect**. The most critical generational hurdle in English-to-Dialect is related to *standardization interference*, which represents 35.62% of errors in this direction. This indicates a pervasive MSA bias, where models default to formal or prestige varieties rather than maintaining the requested dialectal authenticity. This sociolinguistic failure is further complicated by horizontal dialectal bleed, where features from different regional varieties overlap, leading to a 21.29% error rate in dialect target selection (wrong dialect).

In the **Dialect-to-English direction**, failures are predominantly *semantic* and *pragmatic*. *Named Entity* recognition is the primary bottleneck at 13.49%, likely driven by the lack of standardized dialectal orthography, which makes proper noun recovery inconsistent. Furthermore, the model frequently fails to capture the speaker's social intent, with *speech acts* (2.54%) and *vocatives* (2.49%) together accounting for over 5% of errors in Dialect-to-English. While these stylistic and lexical issues vary by direction, morphosyntactic struggles—specifically *verbal features*—remain a persistent technical barrier across the board, appearing at significant rates in both decoding (3.04%) and generation (7.35%) tasks.

| | Error Sub-category | Dialect → Eng (%) | Eng → Dialect (%) |
|---|---|---|---|
| Socioling. | Standardization Interference | 0.00 | 35.62 |
| | Wrong Dialect (Horiz. Mismatch) | 0.00 | 21.29 |
| | Register Mismatch | 0.59 | 0.39 |
| Pragmatic | MWEs/Proverbs | 6.28 | 2.53 |
| | Speech Acts & Illocutionary Force | 2.54 | 0.51 |
| | Vocatives/Honorifics/Titles | 2.49 | 0.32 |
| | Discourse Marker Mismatch | 0.89 | 0.12 |
| | Code Switching | 0.11 | 0.59 |
| Semantic | NE (Named Entities) | 13.49 | 3.52 |
| | Coverage: Unknown Term/Dialect | 9.33 | 1.22 |
| | Omission | 8.94 | 1.46 |
| | Unnatural/Unidiomatic Style | 6.65 | 3.32 |
| | Awkward Style | 4.61 | 5.02 |
| | Addition | 5.86 | 2.09 |
| | Pronouns | 6.84 | 0.91 |
| | Disambiguation: Cross-Variety | 6.67 | 0.28 |
| | Wrong Term | 3.18 | 3.08 |
| | Hallucination | 4.47 | 1.18 |
| | Undertranslation | 2.90 | 1.03 |
| | Disambiguation: Polysemy Failure | 3.02 | 0.47 |
| | Transliteration | 2.96 | 0.36 |
| | Overtranslation | 1.65 | 0.28 |
| | Untranslated | 1.42 | 0.24 |
| | Inconsistent Term. Resource | 1.23 | 0.12 |
| | Inconsistent Style | 0.03 | 0.55 |
| | Unintelligible | 0.03 | 0.16 |
| | Inconsistent Use of Terminology | 0.03 | 0.00 |
| Morph. | Verbal Features | 3.04 | 7.35 |
| | Nominal Features | 0.28 | 0.79 |
| Ortho. | Typo/Slip | 0.14 | 4.74 |
| | Currency | 0.14 | 0.16 |
| | Punctuation | 0.08 | 0.20 |
| | Number Format | 0.00 | 0.04 |
| Graph. | Character Encoding | 0.06 | 0.08 |

Table G.1: Fine-grained LQM error distribution. Bar lengths are proportional to the percentage of total errors per translation direction.

| Category | Error Type | Subcategory | Dir | Source | Target | Model |
|---|---|---|---|---|---|---|
| **Sociolinguistics** | Code & Register Selection | Standardization Interference | EN → MO | My phone was silent so I did not see your calls. | باش ما شفتش صامت تلفوني كان ديالك ديال المكالمات. | ♦ Gemma |
| | | Wrong Dialect | UA → EN | May Allah protect you. You are good and blessed. | ومبارك. زين الله يحفظك. إنت | ★ Gemini |
| | | Register Mismatch | EN → MO | What are you doing, man? | أ بنادم آش كادير | ★ Gemini |
| **Pragmatics** | Use, Context, Cultural Approp | MWEs/ Proverbs | EG → EN | بقولك ايه ما تيجي ناكل لقمه سوا عشان يبقى عيش و ملح؟ | Tell you what, why don't we grab a bite together so there's bread and salt between us? | ★ Gemini |
| | | Code Switching | EN → EG | It's not called a casino, it's called a nightclub, ma'am. | مش اسمه كازينو، اسمه نايت كلوب يا مدام. | ★ Gemini |
| | | Speech Acts/ Illoc Force | MO → EN | ياك؟ دابا نتا مشيتي و جمعتي و سمعتي هاد التخربيق. | Right? Now you went, gathered, and listened to this nonsense. | ♦ Gemma |
| | | Discourse Marker Mismatch | MO → EN | الصراحة بابا انا شدني لو تخاك و تبلوكيت اوغوزمون مهدي عتق الموقف | To be honest, dad, I'm really embarrassed about what happened with that girl and the photos, and honestly, Mehdi saved the situation. | ♦ Gemma |
| | | Vocatives /Honorifics/Titles | YE → EN | يعني ايش يا اباه ، خلاص عاتجسمو ، ونورث ، وتوجع امورنا سابره. | So what, man, just toughen up, inherit [this], and make our affairs difficult. | ♦ Gemma |
| **Semantics** | Lexical Semantics | Named Entity | EN → PA | والله ذمتك وسيعة يا ابو عطا يا دكر آ | By God, your heart is wide, Abu Atta, you rooster! | ▲ Fanar |
| | | Wrong term | EN → PA | والله لو بحجوا مرة و مرتين و ثلاث و بضلوا يسعوا و يطوفوا و يسعوا و يطوفوا ما بنغفرلهم ذنب بعد اللي عملوه في ولاد البلد | by God, even if they go for Hajj once, twice, three times, and keep circumambulating and walking, we won't forgive them for their sin after what they did to the children of the town. | ▲ Fanar |
| | | Overtranslation | MA → EN | ذا اشوي عايشين منو وكارين منو دار وذا كامل طامع انت فينا ظرك نعطوك عشرين آلف | You see how we live, we are barely surviving, and you want everything, you think we will give you twenty thousand? | ▲ Fanar |
| | | Undertranslation | YE → EN | خليها على الله بس، امانه ان قلبي مغدغد غدغده عليه | Leave it to God, but honestly, my heart is restless with worry about it. | ▲ Fanar |
| | | Transliteration | EG → EN | سعد عارف ناصف مدكور زي ما حنا عارفينه بالزبط | Saad Arafa Nasif Madkor, as we all know. | ▶ Command-R |
| | | Unnatural Style | MO → EN | المرة الجاية غا ندير فحالك غاندير راحة بحالا سخفت. | Next time I'll do like you, I'll take it easy like I'm weak or tired. | ▶ Command-R |
| | | Awkward Style | PA → EN | يا سلام يا سلام كل البلد ورا و سلامة قدام! بس انا بقيت مشغول منت عارف | Peace, peace, the whole country is behind and safety is in front! But I've become busy, I don't know. | ▶ Command-R |
| | | Unknown Term/ Dialect | UA → EN | شو كان ييب حق المتوه ؟، و عيد شو كان ييب حق المتوه ؟ | What did he bring for the guest? What did he bring for the guest? | ▶ Command-R |
| | | Polysemy Failure | JO → EN | هات جاي بوقع لك و بوقع لعشرة مثلك كمان شو يعني فكرك أنا خويف يعني؟ | Give it to me, I will sign it for you and ten people like you. Do you think I am afraid? | ♦ Gemma |
| | | Cross-Variety Interference | JO → EN | ديري بالك على الصندوق يا آنسه ليلى بدي كل شي مظبوط و دقيق | Take care of the box, Miss Leila. I want everything to be accurate and precise. | ▶ Command-A |
| | Propositional Semantics | Addition | MA → EN | آن سياني عندي جيتو من امو ! جيتو من خوه جيتو من ظهرو | I have a problem with my mother, my brother, and my father. | ▲ Fanar |
| | | Omission | UA → EN | وين حرمنا المصون بويه زين زقروها ليي مشتاق ليها، سيري زقريها عمو | but where is our private space? I miss it, go, show it to me, uncle. | ▲ Fanar |
| | | Untranslated | JO → EN | طيب يا يوسف يعني خليها لبكرة الصبح. | ، لتاُس دِت تمررو يوسف صكَي، مرتِنِ | ▲ Fanar |
| | | Hallucination | EG → EN | أنا برضو انتي مش كنتي انتي اللي قاعدة . | I also thought you weren't the one sittir you stupid woman. | ▶ Command-A |
| | Discourse Semantics | Pronouns | MA → EN | شوف خليك لي ذلي قلت لك ول ذلي ليهي تقول لي جوابي فيه. | Look, stay with me, I told you, and stay with her | ▶ Command-A |
| **Morphosyntax** | Grammar (number, gender, tense) | Verbal Features | MO → EN | شوف ا موحا سير الله يرضي عليك عند دوك الناس لي كا يصاوبولنا هاد الحويجات | may God be pleased with you. Look, there are people who are helping us with these needs | ▲ Fanar |
| | | Nominal Features | EN → EG | I came to check on you, my love, because we heard women's feet going up the stairs. I was afraid about you, Kuka. | جيت أطمن عليكي يا حبيبتي، حريم طالعة رجلي عشان سمعنا على السلم. كنت خايفة عليكي يا كوكا. | ▶ Command-A |

Continued on next page...

**Table G.2 – continued from previous page**

| Category | Error Type | Subcategory | Dir | Source | Target | Model |
|---|---|---|---|---|---|---|
| **Orthography/ Writing conventions** | Unconventional Spelling | | EN → PA | What do I know, uncle? Is Ma'rouf the imposter better than us!? | شو بتعرف أنا، يا عمّي؟ معقول معرووف أحسن منا؟!الدّعيّ | ♦ Gemma |
| | Surface Mechanics | Currency | EN → UA | The fellow will give me ten rubies, father. I told you I went and talked to him, and now she has agreed. | ، يبه.ياقوت الرجال بيعطيني عشر قلت لك رحت وكلمته، و الحين وافقت. | ♦ Gemma |

Table G.2: LQM error types across all linguistic levels examples from covered dialects.