QUANTA: A Fine-Grained Dataset and Benchmark for Arabic Machine-Translation Error Analysis*

Afraa Alshammari[1], Ahmad Alahmari[3], Amal Alaboud[4], Jaber Ahmed Harthi[3], Karima Ali Shahwan Alzahrani[5], Ahnaf Adib[1], Patrick T. Brandt[2], Javier Osorio[6], Vito D'Orazio[7], Latifur Khan[1]

[1]Department of Computer Science, University of Texas at Dallas; [2]School of Economic, Political and Policy Sciences, University of Texas at Dallas; [3]Department of Foreign Languages, Jazan University; [4]Department of Foreign Languages, Taif University; [5]Independent Researcher; [6]School of Government and Public Policy, University of Arizona; [7]Department of Political Science, West Virginia University

Afraa.Alshammari@utdallas.edu, Aalahmari@jazanu.edu.sa, Asaboud@tu.edu.sa, Jharithi@jazanu.edu.sa, Kalzhrani153@gmail.com, Ahnaf.Adib@utdallas.edu, Pbrandt@utdallas.edu, Josorio1@arizona.edu, Vito.dorazio@mail.wvu.edu, Lkhan@utdallas.edu

Abstract

Automatic machine-translation (MT) metrics collapse heterogeneous errors into a single score. For Arabic, no fine-grained resource exists for diagnosing error types and severity. We introduce QUANTA, a 9,032-instance En to Ar MT-error dataset (738 real UNPC + 7,856 synthetic training instances, both structurally validated, plus a held-out 438-instance real Gold test set) annotated with MIZAN (ميزان, "scale"), a new 23-class MQM-compatible taxonomy organised in a four-level hierarchy: 23 leaf classes rolling up to 14 subtypes, 5 parent groups, and 4 TQA dimensions. To make the resource usable, we release the RAQEEB (رقيب, "observer") classifier, an AraBERT-v2 model for predicting MIZAN 23-class labels (5-fold Gold macro-F1 $= 0.614 \pm 0.017$). Data quality is supported by ETCA (Error Taxonomy Cross-vendor Audit), an LLM-as-judge check in which a different-vendor model audits QUANTA instances for anchor validity and label fidelity, curbing same-family self-preference bias, and by dual-annotator validation (Gwet's AC1 $=$ 0.97 synthetic; 0.83 real) on core taxonomy criteria. Ablations show that anchor-guided framing is decisive (without it, macro-F1 collapses to 0.114); that 738 structurally validated real samples are $\approx 12\times$ more sample-efficient than synthetic data; and that, restricted to the classes with real-tier training support, combining both tiers outperforms real (0.656) or synthetic (0.312) alone. Code, data, prompts, and checkpoints are publicly available.

1 Introduction

Machine Translation (MT) systems are now deeply embedded in information access, localization, and cross-lingual Natural Language Processing (NLP) pipelines, making the reliability of MT evaluation a central concern (Mukherjee and Shrivastava, 2025). MT evaluation is constrained by a trade-off between linguistic precision and scalability: human evaluation of MT output is generally considered accurate but is costly to scale, whereas automatic MT quality metrics (BLEU (Papineni et al., 2002), COMET (Rei et al., 2020), BERTScore (Zhang et al., 2020)) tend to collapse heterogeneous error phenomena from minor orthographic deviations to critical semantic distortions into a single opaque value. Compressing different MT error types into a single scalar both undercounts catastrophic semantic errors and overcounts cosmetic ones, a distinction a fine-grained, severity-aware taxonomy can make explicit. Properly assessing the quality of MT outputs is especially difficult for Arabic: its rich morphology (nunation, the definite article ال, strict gender agreement, agglutinative clitics) produces error phenomena absent from European-language MT evaluation (Al-Said, 2026; Habash, 2010), yet prior typologies (Munkova et al., 2025; Qiu et al., 2025) are language-agnostic at 3 to 7 error classes (Tafa et al., 2025). No fine-grained Arabic MT error benchmark with 20 or more classes exists.

QUANTA and MIZAN in one sentence. We release QUANTA, a fine-grained Arabic MT error dataset with full 23-class coverage, and MIZAN (ميزان, "scale"), its underlying multi-scale taxonomy of Arabic linguistic-shift errors, together with a released 23-class diagnostic classifier and a dual-annotator keyword-level validation protocol reaching almost-perfect inter-annotator agreement

*Code and data available at https://github.com/AALight/Quanta-Mizan

(Gwet's AC1 0.83–0.97) on the core taxonomy criteria. MIZAN is compatible with the Multi-dimensional Quality Metrics (MQM) framework: it adopts MQM's severity typology and rolls its 23 classes up to four Translation Quality Assessment (TQA) dimensions, but departs from MQM by adding four Arabic-specific leaf classes (Tanween Omission, Gender Disagreement, Definiteness Shift, Invalid Pattern) for which language-agnostic MQM provides no first-class error type. To keep the components distinct, the paper names four: **MIZAN** (the 23-class taxonomy), **QUANTA** (the dataset built and audited under it), **ETCA** (the cross-vendor LLM audit that validates it), and **RAQEEB** (the released classifier trained on QUANTA); Figure 1 is the overview, tracing raw MT output through anchor-constrained construction and validation to the QUANTA dataset.

Contributions. Together these form one integrated resource for fine-grained En to Ar MT error analysis, shown end-to-end in Figure 1: the MIZAN taxonomy, the QUANTA dataset built and audited under it, and the RAQEEB diagnostic classifier:

1. **MIZAN**: a 23-class taxonomy of Arabic linguistic-shift errors with a 4-level hierarchy (23 classes → 14 subtypes → 5 parent groups → 4 MQM-compatible TQA dimensions) and expert-calibrated severity weights.
2. **QUANTA**: a 9,032-instance fine-grained Arabic MT error dataset, applying the MIZAN error typology. QUANTA includes 738 expert-annotated *real* MT errors (tier T1) from the United Nations Parallel Corpus (UNPC) (Ziemski et al., 2016), 7,856 structurally validated *synthetic* instances (tier T2) covering all 23 classes, and a 438-instance Gold test set with full 23/23 coverage.
3. **ETCA** (Error Taxonomy Cross-vendor Audit): a quality audit in which a different-vendor LLM (Claude Sonnet 4, Anthropic, 2025) independently scores both the high-confidence generation seeds and the GPT-4o-generated T2 candidates (OpenAI, 2024); for the synthetic tier this means no model grades its own output, mitigating same-family self-preference bias (Kim, 2025; Wang et al., 2026). Verbatim prompts ship with the release (Appendix D); external validation against the human consensus (Gwet's AC1 = 0.62) is in §4.2.
4. **RAQEEB**: a fine-tuned AraBERT-v2 classifier that labels each MT error with its MIZAN class, released with the 5-encoder RAQEEB-Bench benchmark (5-fold mean Gold macro-F1 = 0.614 ± 0.017) as a diagnostic tool with documented per-class capability boundaries, not a sealed quality scorer.
5. **Human validation**: two expert linguists, not involved in taxonomy design, independently re-labelled 147 stratified instances, confirming that the MIZAN labels are reproducible (Gwet's AC1 = 0.97 on the core criteria for T2, 0.83 for T1).

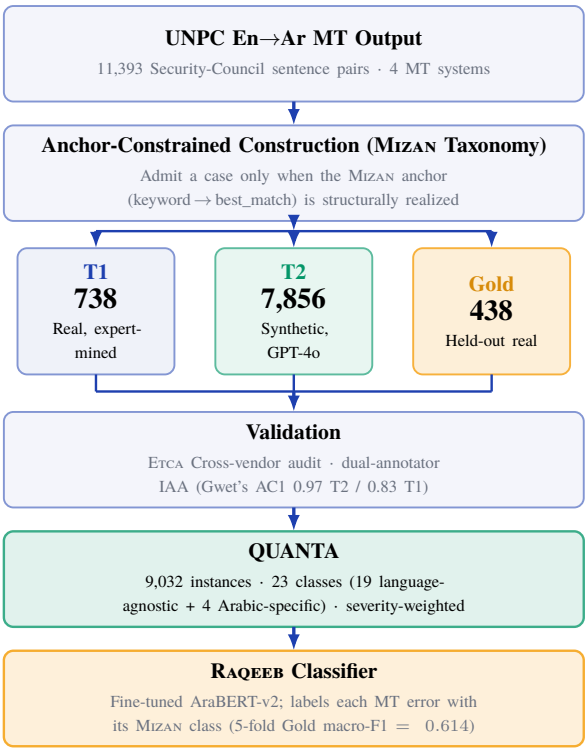

Figure 1: The QUANTA pipeline: from raw En→Ar MT output through MIZAN-guided anchor-constrained construction and validation (ETCA cross-vendor audit and dual-annotator IAA) to the QUANTA dataset and the released RAQEEB classifier.

2 Related Work

Arabic MT and scalar evaluation metrics. Arabic MT benchmarks (Sajjad et al., 2020; Nagoudi et al., 2022; Deutsch et al., 2025) and scalar metrics BLEU (Papineni et al., 2002), COMET (Rei et al., 2020, 2022), BERTScore (Zhang et al., 2020) are dominant in MT quality assessment. Despite their prevalence, they produce a single scalar per sentence and cannot decompose errors into linguistically meaningful categories. This is an important limitation when evaluating Arabic. Arabic morphological complexity (Habash, 2010) introduces phenomena (tanween,

definiteness, gender agreement, root-pattern violations) that scalar metrics fail to isolate. No prior work provides a fine-grained taxonomy with 20+ classes targeting Arabic-specific phenomena in MT.

Fine-grained MT error taxonomies. The MQM framework (Lommel et al., 2014, 2024) is the dominant reference for MT error annotation. Freitag et al. (2021) showed that MQM-based human evaluation correlates more strongly with professional translators' judgments than do scalar ratings. While earlier taxonomies introduced 5-class schemes (Vilar et al., 2006; Popović, 2011), more recent LLM-as-judge MQM variants include RUBRIC-MQM (Kim, 2025) and PPbMQM (Wang et al., 2026), both reporting LLM self-preference bias. For Arabic, the QALB shared task (Zaghouani et al., 2014; Mohit et al., 2014) targets learner grammar, not MT; and the critical-error taxonomy of Al Sharou and Specia (2022) is severity-oriented and cross-lingual rather than Arabic-specific. Closer to our setting, AMEANA (El Kholy and Habash, 2011) performs automatic error analysis of Arabic MT output, but as a morphological-feature diagnostic rather than an MQM-compatible, severity-weighted error taxonomy. For Arabic grammatical and learner-error annotation, ARETA (Belkebir and Habash, 2021) provides automatic error-type annotation, the Arabic Learner Corpus (Alfaifi, 2015) supplies expert learner-error annotation, and ArabiGEE (Elhady et al., 2026) offers a hierarchical error-explanation taxonomy; all three address native or learner text correction rather than translation errors. None of these is an MQM-compatible, severity-weighted MT-error taxonomy expressing tanween, definiteness, gender, and root-pattern phenomena at 20+ classes with token-level verifiable anchors, as Mizan provides. Finally, Lommel et al. (2024) argue that segment-level variance on fine-grained linguistic attributes is inherent in language; we draw on this in §5 to frame the lower-F1 classes our classifier reports as field-wide hard cases rather than classifier-specific failures.

Arabic-specific diagnostic evaluation. Osorio et al. (2024, 2025) benchmark domain-specific encoders on native vs. machine-translated UNPC text (Ziemski et al., 2016) and establish a statistically significant rarity loss in MT output that maps to our Terminology Substitution and Hypernym for Hyponym classes, plus dependency-distance shifts that map to the Syntactic Shifts parent group. Al-Olimat and Alshareef (2026) (ALPS) show that expert-curated diagnostic sets remain useful even when single-item inter-rater agreement is low, the same pattern we find in our own annotation (§4.3): near-perfect agreement on the core taxonomy criteria but lower agreement on more subjective judgments.

Synthetic data and validated error construction. LLM-generated training data (Schick and Schütze, 2021; Wang et al., 2023) and data-centric fine-tuning (Zha et al., 2023; Gao et al., 2021) can match or exceed larger noisy datasets, but the synthetic-to-real distributional gap is well documented (Chen et al., 2023). Validation frameworks specifically for synthetic MT-error data remain scarce: generic filtering does not guarantee that a claimed error actually occurs at a specific lexical anchor (Northcutt et al., 2021). The anchor-constrained validation pipeline that produced Quanta fills this gap: it admits a synthetic error only when the claimed error is structurally realised at its lexical anchor, with the cross-vendor Etca audit as an independent second check.

3 The Mizan Taxonomy and Multi-Scale Framework

Motivation and design. General-purpose MT error taxonomies such as MQM (Lommel et al., 2014, 2024), Vilar (Vilar et al., 2006), and Hjerson (Popović, 2011) are language-agnostic 3–7-class schemes with no Arabic-specific morphological phenomena as first-class error types. **Mizan** (ميزان, "scale") fills this gap. Following the "two-pillar" view of Lommel et al. (2024), Mizan pairs a 23-class fine-grained error typology, organised as a 4-level hierarchy (23 classes → 14 subtypes → 5 parent groups → 4 MQM-compatible TQA dimensions), with expert-calibrated, MQM-compatible severity weights. Four design principles distinguish Mizan from these general-purpose taxonomies:

1. **Computational verifiability**: each class is defined by a specific keyword→best_match transformation verifiable at the token level (e.g., *Hypernym for Hyponym*: المحكمة الجنائية → المحكمة).
2. **Arabic-specific coverage**: four classes capture phenomena absent from MQM (Tanween Omission, Gender Disagreement, Definiteness Shift, Invalid Pattern).

Taxonomy	#Cls	Anchor	AR-spec.	Score
Vilar (Vilar et al., 2006)	5	×	×	×
Popović (Popović, 2011)	5	word	×	×
MQM Core (Lommel et al., 2014)	7	segment	×	✓
MQM Full (Lommel et al., 2024)	∼70	segment	×	✓
QALB (Zaghouani et al., 2014)	∼14	✓	grammar	×
GEMBA-MQM (Kocmi and Federmann, 2023)	22	span	×	MQM
PPbMQM (Wang et al., 2026)	top	span	×	MQM
Mizan (ours)	**23**	**anchor**	**4 cls**	**MQM**

Table 1: Mizan compared with major MT error taxonomies. #Cls: number of classes; Anchor: span-localisation granularity; AR-spec.: Arabic-specific classes; Score: severity/scoring scheme.

3. **MQM-compatible hierarchy**: a deterministic mapping from 23 classes to MQM-compatible TQA dimensions.
4. **Operational granularity**: the 23 leaf classes emerged bottom-up from expert-linguist analysis of 446 annotated UNPC MT-error instances rather than top-down.

Table 1 positions Mizan against prior taxonomies. The full per-class specification (definitions, severity, weights, parent group, TQA category) plus a per-class table of real Gold-test examples with Sentence_ID, MT-tool attribution, Arabic reference and MT-output excerpts, and expert justifications for every one of the 23 classes is in Appendix A.

Multi-scale framework. Mizan is an expert-calibrated, multi-scale taxonomy for diagnosing En→Ar MT errors. It integrates four operating dimensions: 23 anchor-observable classes, 14 subtypes, 5 linguistic-family parent groups, and 4 MQM-compatible TQA dimensions matching the MQM Core / MQM Full distinction of Lommel et al. (2014, 2024). Rendered as a tree in Figure 3 (Appendix A), this hierarchy lets a keyword-level error be traced upward through its parent group to the evaluation dimension it affects, the multi-scale property our classifier exploits through its L2 and L3 rollups (§5). The taxonomy was designed by two Arabic native-speaker researchers with doctoral qualifications (one in computational linguistics, one in translation). Instance-level dual-annotator validation is conducted independently by two further expert Arabic linguists, neither of whom participated in taxonomy design (§4.3).

Severity weights. Mizan's severity weights are assigned per class by the two expert linguists following MQM's severity typology (Minor to Critical) (Lommel et al., 2024). The per-class weights aggregate by parent group and by TQA category to 1.00 (Table 12); these totals are aggregations of the per-class weights, not separate scoring factors. These MQM-compatible weights supply the severity-penalty ingredient for segment-level quality scoring; computing such a score is out of scope here, where §5 instead reports the classifier's performance over the Mizan classes. The full per-class weight table and the rollup mapping are in Appendix A.

Why this matters: MT-quality stakes in formal-institutional text. The 23 Mizan classes capture MT failure modes that other metrics collapse into a single opaque value. Mizan makes each failure explicit by assigning it a class and a per-class severity weight s_i (examples per parent group in Table 2): النيجر بلد → النيجر منطقة is labelled Meaning Shift ($s_i = 5$), recategorising a state as a sub-state region; يهيب → يدعو is labelled Terminology Substitution ($s_i = 4$), weakening a binding Security Council directive ("calls upon" → "invites"); and إدانته → إدانتها is labelled Gender Disagreement ($s_i = 3$), an Arabic-specific agreement error that flips a possessive pronoun to feminine. Because each class carries its own severity weight, these failures stay distinct rather than averaged into one scalar, as high-stakes legal, medical, or diplomatic translation requires.

4 The Quanta Dataset

Quanta is the anchor-validated, Mizan-labelled dataset of En→Ar MT errors. It combines three complementary parts: a small set of real errors mined from UNPC translations (T1), a large synthetic set that fills every error class (T2), and a held-out expert-annotated Gold test set covering all 23 classes. Table 3 summarises the overall composition, and Table 4 the per-parent-group coverage. T1+T2 forms the training pool (8,594 instances); Gold (438 instances) is held out for evaluation with full 23/23 class coverage. Each released instance carries six fields, `Sentence_ID`, `Keyword`, `Best_Match` (or the sentinel `[OMITTED]`), `Sub-Subtype` (one of 23), `Parent group` (one of 5), and `TQA category` (one of 4); a worked example for each parent group is in Table 2.

Parent group	Class	Keyword → Best_Match	MT-quality impact (UN-domain gloss)	s_i
Semantic Div.	Total Omission	بعثة → [OMITTED]	UNOCI mandate: "mission" dropped, peacekeeping referent disappears.	5
Semantic Div.	Meaning Shift	النيجر بلد → النيجر منطقة	"Niger, a country" → "Niger, a region": the referent's category changes, a state rendered as a substate area.	5
Morphosyntactic	Gender Disagreement	إدانته → إدانتها	"its [the Council's] condemnation of violence against women and children": MT flips the possessive pronoun to feminine, breaking agreement.	3
Syntactic	Perfective to Progressive	حث → يحث	"urged" (past) → "urges" (present operative): legal force of resolution altered.	4
Pragmatic Div.	Terminology Substitution	يدعو → يهيب	"calls upon" (formal SC directive) → "invites" (weaker request).	4
Orthographic	Spelling Error	اربعة → أربعة	"four" written without its hamza (أربعة → اربعة): orthographic typo in a casualty count.	1

Table 2: Representative MIZAN examples from real En→Ar UNPC MT output, one per parent group plus a Total Omission; s_i is the per-class MIZAN severity weight (1 minor to 5 critical).

Split	Instances	Classes	Provenance
T1 (real)	738	12/23	UN expert
T2 (synthetic)	7,856	23/23	GPT-4o; structural audit
T1+T2	8,594	23/23	training
Gold test	438	23/23	held-out

Table 3: QUANTA Dataset Composition.

Per-parent-group class coverage. Table 4 breaks the 23 classes down by parent group, separating those that surface in real UNPC MT output from those filled for training by T2 synthesis (§4.1). Of the 23 classes, 15 appear in real MT and 8 are T2-only, while all 23 have Gold coverage. Orthographic Shifts has zero real-MT coverage, consistent with character-level errors being removed by post-editing in formal institutional translation, so this group is supplied entirely by synthesis.

Parent group	#Cls	Real-MT	T2-only	Gold
Semantic Divergence	7	4	3	✓
Morphosyntactic Shifts	4	3	1	✓
Syntactic Shifts	9	6	3	✓
Pragmatic Divergence	2	2	0	✓
Orthographic Shifts	1	0	1	✓
Total	23	15	8	23/23

Table 4: Per-parent-group class coverage in QUANTA.

4.1 Dataset Construction

Building a fine-grained error dataset poses a reliability problem (Lommel et al., 2014, 2024; Al-Olimat and Alshareef, 2026): real MT errors at any given MIZAN class are sparse, while synthetic errors are easy to generate but hard to trust, since the injected error may not land at the intended location (Kocmi and Federmann, 2023; Kocmi et al., 2024; Wang et al., 2026). QUANTA addresses this with an anchor-constrained validation pipeline whose core gate admits a candidate only when its MIZAN anchor is structurally realised: the keyword is present in the Arabic reference and its best_match substitution (or the sentinel [OMITTED] for omissions) appears in the MT output (Figure 1), paired with the cross-vendor ETCA audit (§4.2) as a second, independent control. The pipeline runs on the 11,393 UN Security Council sentence pairs of Osorio et al. (2025) across four MT systems.

The anchor keywords come from domain-contrastive keyness (Sketch Engine's simple-maths score against UNPC (Kilgarriff, 2009; Ziemski et al., 2016)), yielding 600 diagnostic keywords that two expert linguists extend with manually curated multi-word anchors before searching them across the four MT outputs.

From the 446-instance expert gold pool the pipeline extracts 177 unique keyword→best_match patterns; applying these across the corpus and deduplicating yields 2,770 candidate rows, of which 1,667 (60%) are rejected as structurally noisy and 794 pass at high confidence to form the T1 substrate, which bypasses ETCA and serves as the reliability anchor; removing the 56 that overlap the held-out Gold test set yields the final 738 T1 instances. Gold-overlap removal eliminates three classes entirely from the T1 training pool (Literal Translation, Perfective to Progressive, Tense Shift Under Negation), which are covered in training only through T2 that is generated using gold samples. The eight MIZAN classes that the pattern-based retrieval does not surface in the T1 pool (Meaning Shift, Partial Translation, Name Entity Error, Gender Disagreement, Wrong Structure, Active to Passive Voice, Noun to Adjective, Spelling Error) are filled for training by T2 synthesis under the same

anchor constraint, so 23/23 training coverage is a pipeline property rather than a retrieval property. These classes are not absent from real MT: all eight are attested in the held-out Gold test set, which is drawn entirely from real MT output (Meaning Shift alone contributes 70 instances), so every class is evaluated on real data even when its training signal is synthetic.

4.2 ETCA Cross-Vendor Audit

Same-family LLM-as-judge pipelines such as GEMBA-MQM (Kocmi and Federmann, 2023) suffer self-preference bias, where a judge systematically favours its own family's generations; routing generation and judging through different vendors is the general mitigation established by prior LLM-as-judge work (Kim, 2025; Wang et al., 2026). Our **Error Taxonomy Cross-vendor Audit (ETCA)** is a specific instantiation of this cross-vendor principle for taxonomy-grounded error auditing, so that no single model grades its own output: GPT-4o (OpenAI, 2024) generates synthetic candidates with 3-shot prompting seeded from a 79-row canonical T0 pool, and Claude Sonnet 4 (Anthropic, 2025) (a different vendor) scores each candidate on five dimensions: Anchor Validity, Phenomenon Clarity, Arabic Naturalness, Collateral Severity, and Seed Recommendation, aggregated into a **Data Quality Index (DQI)**, the normalised mean of the four substantive dimensions (excluding Seed Recommendation). ETCA runs on the 794 high-confidence generation seeds and on a stratified 414-instance T2 audit subset ($\approx 5\%$ of T2, spanning all 23 classes), where DQI ≥ 0.75 marks a high-confidence instance; the released 7,856-instance T2 split is selected by the structural re-classification audit (§4.4), not by a per-instance DQI threshold. The verbatim generation prompt and audit rubric are in Appendix D. ETCA's judgements align with the expert human consensus: on a 43-instance high-confidence subset adjudicated against the dual-annotator labels (§4.3), ETCA reaches Gwet's AC1 $= 0.62$ (substantial agreement on the Landis–Koch scale). This validates the auditor against expert human judgment rather than against another auditor LLM, whose agreement can be inflated by shared model biases; alignment with the human consensus is the stronger criterion for audit reliability. We therefore use ETCA strictly as a parallel dataset-quality audit rather than a stand-alone quality gate.

4.3 Dual-Annotator Validation

QUANTA is validated by two expert Arabic native-speaker linguists, who were not involved in designing the MIZAN taxonomy and who labelled each instance independently. On a stratified sample of 147 instances (55 T1, 92 T2), they reach Gwet's AC1 (Gwet, 2008) $=$ 0.97 on T2 and 0.83 on T1 across the three core taxonomy criteria (error existence, anchor identification, MIZAN class label), almost-perfect agreement on both tiers on the Landis–Koch scale (Landis and Koch, 1977). We use Gwet's AC1 as the primary statistic because the binary evaluation criteria have high prevalence by design, a regime in which Cohen's κ collapses artifactually (Feinstein and Cicchetti, 1990; Gwet, 2008).

IAA in the context of fine-grained linguistic annotation. Fine-grained linguistic annotation produces a characteristic IAA pattern: expert linguists agree near-perfectly on core taxonomy questions and diverge on more subjective attributes such as register fit and severity calibration. Lommel et al. (2024), in the 10th-anniversary MQM review, argue that segment-level variance on fine-grained linguistic attributes is inherent in language rather than a technical deficiency, because text admits multiple valid interpretations and error categorisation depends on the reader's background. Al-Olimat and Alshareef (2026) corroborate this for Arabic: their expert-curated ALPS diagnostic set reports pairwise Cohen's $\kappa = 0.23$ on individual items, while their adjudicated oracle labels reach 99.2% agreement and average human accuracy of 84.6%. Against that backdrop, our IAA is strong on the core taxonomy (T2 core-three mean $= 0.966$, T1 $= 0.827$) and patterned on the peripheral criteria as expected; the lower scores on subjective attributes reflect inherent reader-interpretation variance rather than a validation failure. Appendix B gives the full per-criterion table (severity, UN-register fit, aggregate means), the stratified-sampling design, and the inter-annotator-agreement protocol.

4.4 Post-Hoc Corrections

A post-hoc audit of the Partial Translation class surfaced a mis-labelling pattern: 188 candidates were excluded where the apparent "omission" was in fact a full translation of a different source word; 15 were relabelled from Partial Translation to Total Omission; 20 received corrected `Best_Match`

Encoder	Mean F1 ± std
AraBERT v2	**0.614 ± 0.017**
CAMeLBERT-Mix	0.597 ± 0.021
CAMeLBERT-MSA	0.587 ± 0.029
ConfliBERT	0.583 ± 0.021
XLM-R	0.565 ± 0.023

Table 5: 5-encoder benchmark on the 23-class MIZAN task (anchor-guided input, identical folds); mean ± std over 5 folds.

spans. The final counts in Table 3 reflect these corrections.

5 The RAQEEB Classifier

We release the RAQEEB classifier as a diagnostic evaluation tool with documented per-class capability boundaries, not as a one-number quality scorer. AraBERT v2 (Antoun et al., 2020) reaches 5-fold mean Gold macro-F1 $= 0.614 \pm 0.017$ on the 23-class MIZAN task; we report this transparently and partition the 23 classes into three operational tiers (Table 6) that scope how the classifier should be used.

Encoder comparison. Five encoders ingest anchor-guided input at a fixed learning rate of 5×10^{-5}: three Arabic-specific (AraBERT v2 (Antoun et al., 2020), CAMeLBERT-MSA, CAMeLBERT-Mix (Inoue et al., 2021)), one domain-adjacent (ConfliBERT (Hu et al., 2022)), one multilingual (XLM-R (Conneau et al., 2020)). All five cluster within a 0.049 macro-F1 spread (0.565–0.614; Table 5). We test every encoder pair two ways (Appendix G): a paired-bootstrap comparison with Benjamini–Hochberg false-discovery-rate (FDR) correction asks whether two encoders differ, and a Two One-Sided Tests (TOST) equivalence check ($\delta = 0.05$) asks whether they are practically the same. AraBERT v2 significantly outperforms ConfliBERT ($p_{adj} < 0.001$) and XLM-R ($p_{adj} = 0.007$), and CAMeLBERT-Mix significantly outperforms XLM-R ($p_{adj} = 0.015$); the three Arabic-specific encoders are mutually equivalent within ± 0.05 macro-F1. Arabic-specific pretraining thus provides a small but measurable advantage over the multilingual and domain-adjacent alternatives; ConfliBERT's political-conflict pretraining does not transfer to fine-grained error classification.

Tier reading. The Top tier (F1 $\geq$ 0.80) spans structural and morphological phenomena and, notably, the highest-severity semantic class, Total Omission (0.967), so the most damaging error type is among the most reliably detected, not the least. The Mid tier (0.50–0.80) covers classes that are reliable on aggregate macro-F1 but on which individual predictions still warrant confirmation. The Frontier tier (F1 $<$ 0.50) is dominated by semantically confusable class pairs that share a surface substitution pattern (Meaning Shift with Register Mismatch; Hypernym for Hyponym with Terminology Substitution); discriminating them needs richer lexical-semantic signal than the surface anchor pair carries (the relation between keyword and best_match), a natural extension of the anchor framing. These are field-wide hard cases, not classifier-specific failures: fine-grained schemes with many confusable classes carry a known per-class F1 penalty (Mekala et al., 2021; Suresh and Ong, 2021), and the underlying segment-level variance is intrinsic to language rather than a technical deficiency (Lommel et al., 2024; Al-Olimat and Alshareef, 2026). Production use should treat Frontier classes as candidates for human review.

Tier	Range	Classes (best fold F1)
Top	F1 $\geq$ 0.80	Total Omission (0.967), Wrong Word Order (0.960), Definiteness Shift (0.923)
Mid	0.50 $\leq$ F1 $<$ 0.80	Tanween Omission (0.788), Invalid Pattern (0.696), Register Mismatch (0.649), Literal Translation (0.526)
Frontier	F1 $<$ 0.50	Terminology Substitution (0.382), Meaning Shift (0.272), Hypernym for Hyponym (0.160)

Table 6: Operational tiers of the released RAQEEB classifier on the 10 MIZAN classes with $n \geq 10$ Gold support (full table: Appendix E).

Multi-level rollup. The MIZAN hierarchy is learnable at every granularity (Table 7), and not monotonically in the number of classes. The 14-subtype level (L3) is highest: we hypothesise an L3 learnability sweet spot, where merging confusable siblings that share a keyword→best_match transformation (by MIZAN's design) sheds the collisions that depress L4 without lumping the heterogeneous anchors that L2 does, consistent with evidence that the finest resolution is not automatically optimal and the best granularity is data-dependent (Pirovano et al., 2025; Chen et al., 2025). The wide L2 interval (± 0.062, overlapping L3) makes this ordering suggestive rather than significant.

Anchor framing and data efficiency. Two ablations isolate the headline signal (full details in Appendix H). Anchor framing: without the anchor pair the classifier collapses to macro-F1 $= 0.114$

Level	Granularity	Macro-F1
L4	23 leaf classes	0.614 ± 0.017
L3	**14 subtypes**	$\mathbf{0.676 \pm 0.021}$
L2	5 parent groups	0.624 ± 0.062
L1	4 TQA dimensions	0.656 ± 0.022

Table 7: MIZAN hierarchy learnability: AraBERT v2 macro-F1 (5-fold) at each rolled-up granularity.

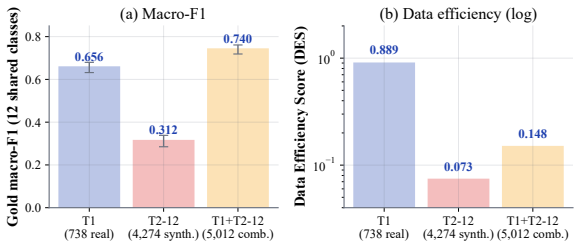

Figure 2: Data efficiency on the 12 T1-supported MIZAN classes (cf. Table 8): (a) 12-class Gold macro-F1 and (b) Data Efficiency Score (DES, log scale) for T1 (real), T2 (synthetic), and T1+T2. T1's DES of 0.889 versus T2's 0.073 is the $\approx 12\times$ sample-efficiency advantage of real data.

(paired-bootstrap $p < 0.0001$; controlled ablation, Appendix H), so anchor-guided framing is the essential signal for fine-grained Arabic MT error classification. Data efficiency: restricted to the 12 classes with T1 support, 738 structurally validated real samples reach macro-F1 = 0.656 versus 0.312 for the 4,274 synthetic instances in those classes, our Data Efficiency Score (DES, F1 per 1,000 samples; Eq. 2) of 0.889 vs. 0.073 ($\approx 12.2\times$); T1+T2 dominates either alone (= 0.740 on 5,012; non-overlapping 23-class confidence intervals), as visualized in Figure 2. Baselines confirm the task is not solvable off-the-shelf (Table 9): classical TF-IDF features cap at macro-F1 = 0.237 and the Arabic LLM ALLaM-7B (Bari et al., 2024) reaches only 0.050 zero-shot and 0.098 five-shot, against 0.614 for the fine-tuned anchor model.

Source	n	12-cls macro-F1	DES
T1 (real)	738	**0.656**	**0.889**
T2 (synthetic)	4,274	0.312	0.073
T1+T2	**5,012**	**0.740**	**0.148**

Table 8: Data ablation on the 12 T1-supported MIZAN classes: 12-class macro-F1 and DES (AraBERT v2, Format D, 5-fold CV).

Frontier-tier error analysis. The Frontier tier has structure: Meaning Shift reaches perfect precision (≈ 1.0) but recall 0.157 (11/70), its false negatives absorbed by semantically adjacent classes such as Register Mismatch, which acts as a catch-all for ambiguous semantic predictions (full distribution in Appendix E). Meaning Shift moreover has no real T1 training data, so its low recall partly reflects a synthetic-train / real-test gap that more real annotation would narrow (cf. the $\approx 12\times$ advantage), not a taxonomy defect.

Method	Macro-F1	Accuracy
Majority class	0.011	0.142
Stratified random	0.034 ± 0.008	0.054
Keyword rule	0.060	0.352
ALLaM-7B (zero-shot)	0.050	0.148
ALLaM-7B (5-shot)	0.098	0.228
Format A (BERT, blind)	0.114 ± 0.030	0.232
TF-IDF + LogReg	0.162 ± 0.007	0.192
TF-IDF + SVM	0.237 ± 0.013	0.275
AraBERT v2 + anchor (Format D)	$\mathbf{0.614 \pm 0.017}$	**0.593**

Table 9: Baseline comparison on the 438-instance Gold test set; all methods receive the same anchor-guided (Format D) input except Format A (the no-anchor control, main 5-fold run; the dedicated format-ablation run is Table 17). Classical methods trained on T1+T2; ALLaM-7B zero- and 5-shot; fine-tuned AraBERT v2 (bottom row) for reference.

Released checkpoint. The default deployable is fold-0 (single-fold macro-F1 = 0.635, matching the per-class results above); fold-3 (0.589) is released as the cross-domain checkpoint (full release in Appendix F).

6 Discussion

Framing and data quality outweigh model choice. Three results point the same way. First, anchor framing is the dominant lever: the keyword→best_match anchor lifts macro-F1 from a blind 0.114 to the released 0.614, dwarfing any gap between encoders. Second, provenance beats scale: 738 validated real instances are $\approx 12\times$ more sample-efficient than synthetic, so a small real set carries more signal than an order of magnitude more generated text. Third, encoder choice buys little: the three Arabic-specific encoders are statistically equivalent, and the only significant gaps are to the multilingual XLM-R and the domain-adjacent ConfliBERT. Together these say that, for fine-grained Arabic MT-error classification, how the input is framed and where the training data comes from matter far more than which encoder is used or how much synthetic data is added.

Implications for Arabic and low-resource MT evaluation. The practical lesson, for Arabic and for other morphologically rich, low-resourced languages, is to invest in anchor-style framing and targeted real annotation rather than in model size or synthetic volume (Zhou et al., 2023). Because the pipeline is language-agnostic outside the four

Arabic-specific classes (§3), this recipe should generalise beyond Arabic.

A diagnostic instrument, not a scalar score. The RAQEEB classifier decomposes errors across the MIZAN classes and flags the Frontier-tier semantic classes for human review rather than returning a single opaque number, and its multi-level structure lets low-confidence 23-class predictions back off to the more reliable 14-subtype level (§5). That decomposition is what scalar metrics cannot provide, and the reason this domain needs a fine-grained taxonomy rather than another scalar score.

7 Conclusion and Future Work

We presented QUANTA, MIZAN, the ETCA cross-vendor audit, and the released RAQEEB classifier (5-fold mean macro-F1 $= 0.614$; fold-0 default at 0.635). Together they close the resource gap for fine-grained Arabic MT-error analysis: one download provides the taxonomy, dataset, audit protocol, and a baseline classifier for downstream probing.

Concrete downstream uses include calibrating MQM-compatible segment-level scores against human ratings and glossary-lookup remediation for Terminology Substitution. Code, data, prompts, the checkpoint, and all per-class artefacts are released under CC-BY-4.0.

No public Arabic MT resource annotates errors at MIZAN's 23-class resolution outside the UNPC domain; a $\sim$1,000-instance cross-register benchmark (medical, legal, literary) is the highest-leverage next data investment, drawing on existing English–Arabic parallel corpora (including domain-specific resources) as source material.

Several technical directions follow: augmenting the anchor with sentence-level semantic context for the Frontier-tier classes, the hierarchical-fallback deployment motivated by the L3 peak, a larger fine-tuned encoder, and broadening the ETCA audit across additional auditor vendors as a complementary robustness check on top of its existing expert-human validation. The framework is also designed to port beyond Arabic: 19 of the 23 MIZAN classes are language-agnostic (rooted in MQM) and only four are Arabic-specific (Tanween Omission, Gender Disagreement, Definiteness Shift, Invalid Pattern), so re-instantiating just those four language-specific leaves adapts the taxonomy to a new language while the language-agnostic backbone and the anchor-guided construction and audit pipeline carry over unchanged. On this basis MIZAN can be extended to dialectal Arabic and code-switching, and transferred to other morphologically rich languages such as Turkish and Kazakh (Yerkebulan et al., 2026) and to the reverse Ar→En direction.

Limitations

Domain, direction, and Arabic register. All experiments analyse En→Ar machine translation of UN Security Council Modern Standard Arabic exclusively. Other UN source languages, the reverse Ar→En direction, and dialectal Arabic varieties (Egyptian, Levantine, Gulf, Maghrebi) are out of scope and would require re-calibration of the four Arabic-specific classes.

Gold-test size and per-class support. The 438-instance Gold test set supports reliable per-class evaluation only for the 10 classes with $n \geq 10$; the remaining 13 are reported but not interpreted individually. All 23 are nonetheless evaluated at the L2 (parent-group) and L3 (subtype) rollup levels, where aggregation gives adequate support.

Human-annotation resources. Both the initial MIZAN annotation of QUANTA and the dual-annotator IAA validation ($n = 147$) were conducted by a small pool of expert annotators under the human-annotation, time, and financial constraints typical of an academic resource-paper effort. A larger annotation round with additional independent annotators is planned.

Single-LLM synthetic generator. The T2 synthetic split is generated by one LLM family (GPT-4o); the cross-vendor ETCA judge mitigates judge-side self-preference but does not eliminate generator-side stylistic artefacts. Because GPT-4o may have encountered UNPC text in pretraining, T2 sentences could echo memorised content. However, the headline results use the held-out real Gold set, so they are not affected by contamination. A multi-generator check (Claude, Gemini) is planned.

Acknowledgments

This research was supported in part by NSF award OAC-2311142 and NIST grant number 60NANB24D143. This work used Delta at NCSA / University of Illinois through allocation CIS220162 from the ACCESS program. We thank the expert Arabic annotators who contributed to the dual-annotator validation.

References

Hussein S. Al-Olimat and Ahmad Alshareef. 2026. ALPS: A diagnostic challenge set for Arabic linguistic & pragmatic reasoning. *arXiv preprint arXiv:2602.17054v1*.

Almoataz B Al-Said. 2026. Mizan: An innovative tool for building arabic morphological pattern dictionaries and enhancing lexical resource development. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 25(2):1–26.

Khetam Al Sharou and Lucia Specia. 2022. A taxonomy and study of critical errors in machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation (EAMT)*, Ghent, Belgium.

Abdullah Yahya G Alfaifi. 2015. *Building the Arabic learner corpus and a system for Arabic error annotation*. Ph.D. thesis, University of Leeds.

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513. Association for Computational Linguistics.

Anthropic. 2025. Introducing Claude 4. `https://www.anthropic.com/news/claude-4`.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT)*, pages 9–15.

M. Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya, Sultan Alrashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Majed Alrubaian, Ali Alammari, Zaki Alawami, AbdulMohsen O. Al-Thubaity, and 6 others. 2024. ALLaM: Large language models for Arabic and English. *arXiv preprint arXiv:2407.15390*.

Riadh Belkebir and Nizar Habash. 2021. Automatic error type annotation for Arabic. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 596–606, Online. Association for Computational Linguistics.

Jingzhou Chen, Junjie Huang, Peng Wang, Fengchao Xiong, Yuntao Qian, and Liang Xiao. 2025. Adaptive label granularity selection for remote sensing targets: Balancing accuracy and specificity. *IEEE Transactions on Geoscience and Remote Sensing*, 63.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2023. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Ahmed El Kholy and Nizar Habash. 2011. Automatic error analysis for morphologically rich languages. In *Proceedings of Machine Translation Summit XIII: Papers*.

Khaled Elhady, Omar Kallas, Nizar Habash, and Bashar Alhafni. 2026. Arabigee: A hierarchical taxonomy for arabic grammatical error explanation. *arXiv preprint arXiv:2606.10765*.

Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

Markus Freitag, George 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. In *Transactions of the Association for Computational Linguistics*, volume 9, pages 1460–1474.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3816–3830.

Kilem L. Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Nizar Habash. 2010. *Introduction to Arabic Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Yibo Hu, MohammadSaleh Hosseini, Erick Skorupa Parolin, Javier Osorio, Latifur Khan, Patrick T. Brandt, and Vito D'Orazio. 2022. ConfliBERT: A pre-trained language model for political conflict and violence. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 5469–5482.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: Diagnosing evaluation metrics for translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9540–9561. Association for Computational Linguistics.

Adam Kilgarriff. 2009. Simple maths for keywords. In *Proceedings of the Corpus Linguistics Conference*, volume 6, pages 41–55. University of Liverpool Liverpool.

Ahrii Kim. 2025. RUBRIC-MQM: Span-level LLM-as-judge in machine translation for high-end models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 147–165.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, pages 768–775.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Hannah Liu, Junghyun Min, En-Shiun Annie Lee, Ethan Yue Heng Cheung, Shou-Yi Hung, Elsie Chan, Shiyao Qian, Runtong Liang, Kimlan Huynh, Wing Yu Yip, York Hay Ng, Tsz Fung Yau, Ka Ieng Charlotte Lo, You-Wei Wu, and Richard Tzong-Han Tsai. 2025. SiniticMTError: A machine translation dataset with error annotations for sinitic languages. *Preprint*, arXiv:2509.20557.

Arle Lommel, Serge Gladkoff, Alan Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, Johani Innis, Lifeng Han, and Goran Nenadic. 2024. The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control. *arXiv preprint arXiv:2405.16969*.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. In *Tradumatica: Tecnologies de la Traducció*, 12, pages 455–463.

Dheeraj Mekala, Varun Gangal, and Jingbo Shang. 2021. Coarse2Fine: Fine-grained text classification on coarsely-grained annotated data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 583–594. Association for Computational Linguistics.

Behrang Mohit, Alla Rozovskaya, Nizar Habash, Wajdi Zaghouani, and Ossama Obeid. 2014. The first QALB shared task on automatic text correction for Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 39–47.

Ananya Mukherjee and Manish Shrivastava. 2025. Lost in translation? found in evaluation: A comprehensive survey on sentence-level translation evaluation. *ACM Computing Surveys*, 58(1):1–47.

Dasa Munkova, Lucia Benkova, Michal Munk, L'ubomír Benko, and Petr Hajek. 2025. From scores to insights: Predicting mt errors using reliable metrics and linguistic typology in slavic languages. *MethodsX*, page 103613.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. TURJUMAN: A public toolkit for neural Arabic machine translation. In *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) @ LREC2022*, pages 1–11, Marseille, France. European Language Resources Association.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

OpenAI. 2024. GPT-4o system card. `https://openai.com/index/gpt-4o-system-card/`.

Javier Osorio, Sultan Alsarra, Amber Converse, Afraa Alshammari, Dagmar Heintze, Latifur Khan, Naif Alatrush, Patrick T. Brandt, Vito D'Orazio, Niamat Zawad, and Mahrusa Billah. 2024. Keep it local: Comparing domain-specific LLMs in native and machine translated text using parallel corpora on political conflict. In *Proceedings of the 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 542–551.

Javier Osorio, Afraa Alshammari, Naif Alatrush, Dagmar Heintze, Amber Converse, Sultan Alsarra, Latifur Khan, Patrick T. Brandt, and Vito D'Orazio. 2025. The devil is in the details: Assessing the effects of machine-translation on LLM performance in domain-specific texts. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 315–332.

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 of the Association for Computational Linguistics (ACL)*, pages 311–318.

Davide Pirovano, Federico Milanesio, Michele Caselle, Piero Fariselli, and Matteo Osella. 2025. The advantage of fine-grained training. *arXiv preprint arXiv:2509.05130*.

Maja Popović. 2011. Hjerson: An open source tool for automatic error classification of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, 96:59–67.

Mengyang Qiu, Tran Minh Nguyen, Zihao Huang, Zelong Li, Yang Gu, Qingyu Gao, Siliang Liu, and Jungyeul Park. 2025. Multilingual grammatical error annotation: Combining language-agnostic framework with language-specific flexibility. *arXiv preprint arXiv:2506.07719*.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

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.

Ananya Sai B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar,

Mitesh M. Khapra, and Raj Dabre. 2023. IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228, Toronto, Canada. 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 (COLING)*, pages 5094–5107.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6943–6951.

Varsha Suresh and Desmond C. Ong. 2021. Not all negatives are equal: Label-Aware contrastive loss for fine-grained text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4381–4394. Association for Computational Linguistics.

Taofik O. Tafa, Siti Zaiton Mohd Hashim, Mohd Shahizan Othman, Hitham Alhussian, Maged Nasser, Said Jadid Abdulkadir, Sharin Hazlin Huspi, Sarafa O. Adeyemo, and Yunusa Adamu Bena. 2025. Machine translation performance for low-resource languages: A systematic literature review. *IEEE Access*, 13:72486–72505.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 697–702.

Sidi Wang, Sophie Arnoult, and Amir Kamran. 2026. Large language models as annotators for machine translation quality estimation. *arXiv preprint arXiv:2603.10775*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Gulnur Yerkebulan, Akerke Akanova, Zhantore Galymzhan, and Nazira Ospanova. 2026. A cefr-graded lexicon and morphology-aware benchmarks for kazakh lexical complexity prediction. *Technologies*, 14(6):346.

Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Ossama Obeid, Nadi Tomeh, Alla Rozovskaya, Noura Farra, Sarah Alkuhlani, and Kemal Oflazer. 2014. Large scale Arabic error annotation: Guidelines and framework. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 2076–2083.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2023. Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR)*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations Parallel Corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 3530–3534.

Appendices

Appendix roadmap. The appendices are organised in eleven sections: (A) Mizan taxonomy specification with the full per-class table, severity

weights, and the 4-level rollup; (B) dual-annotator validation protocol with full inter-annotator agreement detail; (C) construction and audit algorithms; (D) ETCA generation prompt and audit rubric verbatim; (E) full per-class classifier results; (F) reproducibility (split indices, hyperparameters, environment, code repository pointer); (G) encoder significance and equivalence (per-pair paired-bootstrap and TOST results); (H) anchor framing and data-efficiency ablations (Format A vs. Format D, classical and LLM baselines, with 23-class confidence intervals); (I) per-class T2 generation audit; (J) cross-language dataset comparison; (K) ethics and broader impact.

A MIZAN Taxonomy Specification

Table 10 lists all 23 MIZAN classes with definitions and representative examples. Table 12 gives the per-class severity weights and the full four-level rollup (subtype, parent group, TQA category). Figure 3 renders the taxonomy as a tree: all 23 leaf classes under their five parent groups, each leaf shaded by its default severity.

Weight invariants. Each class's severity level (Minor to Critical) and 1–5 score follow MQM's severity typology, assigned by the two expert taxonomy designers and mapped to the class's primary MQM dimension. The per-class weights in Table 12 are calibrated so that per-class partial sums recover both the 5-parent-group totals (Semantic Divergence 0.40, Morphosyntactic Shifts 0.10, Syntactic Shifts 0.10, Pragmatic Divergence 0.35, Orthographic Shifts 0.05; summing to 1.00) and the 4-TQA-category totals (Accuracy 0.42, Fluency 0.225, Terminology 0.28, Style & Register 0.075; summing to 1.00). The parent-group and TQA-dimension levels are therefore derived views, not separate multiplicative factors. These weights are MQM-compatible: they adopt MQM's severity typology and map onto its TQA dimensions, but computing a segment-level score from them is out of scope here.

Class-level severity distribution. Across the 23 MIZAN classes the severity levels are distributed as 2 Critical, 5 Severe, 7 Major, 8 Moderate, and 1 Minor (Table 12): roughly two-thirds of the taxonomy sits at Major or below, while the two Critical classes (Meaning Shift, Total Omission) carry the highest per-class weight. Table 11 gives the full per-class instance counts across the T1 (real), T2 (synthetic), and Gold tiers; the released artefacts include additional Gold examples beyond the one-per-class set in Table 10.

Worked example (weight asymmetry). The per-class weight×severity product spans two orders of magnitude: Meaning Shift ($s = 5$, $w_{\text{class}} = 0.12$) gives $5 \times 0.12 = 0.60$, while Active to Passive Voice ($s = 2$, $w_{\text{class}} = 0.0025$) gives $2 \times 0.0025 = 0.005$. This $120\times$ asymmetry encodes the domain-expert judgment that semantic errors damage MT quality far more than register-level voice shifts in formal UN text.

B Dual-Annotator Validation Protocol

B.1 Dual-Annotator Evaluation Design

We designed a keyword-guided dual-annotator evaluation protocol for the QUANTA error instances. Annotators see the claimed MIZAN class, keyword, and best-match span, and answer seven criteria:

1. Error presence;
2. Anchor accuracy;
3. Keyword-level severity;
4. Minimal-pair structure;
5. Correct MIZAN class label;
6. Naturalness;
7. Domain appropriateness (UN register).

Because the annotator is shown the claimed class and anchor, the protocol is verification-style rather than open annotation; both annotators judge every instance independently.

B.2 Stratified Sampling

The evaluation sample comprises 147 instances, drawn by stratified random sampling over MIZAN class and frozen before any model training: 55 T1 instances (real MT errors stratified across the available real-MT error classes) and 92 T2 instances (QUANTA-synthetic, 4 instances per class across all 23 classes). We stratify by class so that agreement is estimated across the full taxonomy rather than being dominated by the most frequent classes; the choice of Gwet's AC1 over Cohen's κ is motivated separately by the high prevalence of the binary criteria (§B.3).

B.3 Inter-Annotator Agreement

Table 13 reports per-criterion Gwet's AC1 values on the 55 T1 and 92 T2 instances. AC1 is the primary statistic because the binary evaluation criteria have high prevalence by design, a

Class	s_i	Definition	Sentence_ID (tool)	Keyword → Best_Match	AR reference excerpt	MT output excerpt
Semantic Divergence (7 classes, parent weight 0.40)						
Meaning Shift	5	Referential content or core semantic value changed	2005/s/ac_37/2005/_1455_/9/14 (Transformers)	النيجر منطقة → النيجر بلد	النيجر بلد ذو مساحة شاسعة تبلغ...	إن النيجر منطقة شاسعة تبلغ مساحتها...
Total Omission	5	Source element completely missing from MT output	2014/s/2014/430/50:1 (Google)	بعثة → [OMITTED]	... يقرر أن تشمل ولاية بعثة الأمم المتحدة في كو...	19 - يقرر أن تكون ولاية عملية الأمم المتحدة في كوت ديفوار كم
Partial Translation	4	Named entity or key term only partially rendered	2014/s/2014/740/38:1 (Transformers)	المكتب → مكتب الأمم المتحدة	29 - ووضع مكتب الأمم المتحدة المعني بالمخدرات وا...	29. قام المكتب، بالاشتراك مع جامعة...
Name Entity Error	4	Named entity incorrectly translated or transliterated	2006/s/2006/154/290:2 (Google)	صوت العالم → صوت الجهاد	...ة يوجه عام مثل موقع صوت الجهاد، أو تلك التي تقدم ا...	... بشكل عام، مثل موقع صوت العالم. الجهاد، أو تقديم ا...
Literal Translation	3	Word-for-word translation ignoring Arabic idiom	2006/s/2006/822/209:8 (Deep)	ساخن → محتدم	ودار جدال شفوي محتدم بين العقيد ريس والس...	...د مالك تبادلا شفهيا ساخنا.
Hypernym for Hyponym	2	Broader term used where source specifies narrower	2004/s/2004/226/118:2 (Deep)	منفصل → مستقل	...وبته وعنوانه في ملف مستقل عن الملفات الأخرى ل...	...لشخص وعنوانه في ملف منفصل.
Hyponym for Hypernym	2	Narrower term used where source specifies broader	2014/s/2014/740/81:1 (Google)	البحارة → الملاحين	...غير المشروط عن جميع الملاحين الأبرياء المحتجزين ...	...غير المشروط عن جميع البحارة الأبرياء المحتجزين ...
Morphosyntactic Shifts (4 classes, parent weight 0.10)						
Invalid Pattern	3	Morphologically impossible Arabic root-pattern form	2009/s/res/1904_2009_/23:1 (Google)	الأعتدة → معدات	...وما يتصل بها من معدات بجميع أنواعها، بما...	...للأسلحة والأعتدة ذات الصلة بجميع أنواعها...
Tanween Omission	2	Nunation omitted where grammatically required	2008/s/2008/324/97:2 (DeepL)	أولا → أولاً	أولاً، يرتبط التنفيذ الفعال لنظام الجزاءات...	أولا، حفز إنشاء المحكمتين المخصصتين...
Gender Disagreement	3	Masculine/feminine agreement violated	2010/s/prst/2010/25/819:1 (Transformers)	إدانتها → إدانته	...لأنه، رغم إدانته المتكررة للعنف الموجه ضد المرأة...	...على الرغم من إدانتها المتكررة للعنف ضد النساء...
Definiteness Shift	2	Incorrect ال addition or deletion	2005/s/2005/633/25:2 (Transformers)	السلام الإقليمي الدائم → سلام إقليمي دائم	...تحقيق سلام إقليمي دائم يتعين التعامل مع جميع الق...	...يتحقق السلام الإقليمي الدائم، لا بد من معالجة جميع الق...
Syntactic Shifts (9 classes, parent weight 0.10)						
Perfective to Progressive	4	Completed action rendered as ongoing	2009/s/2009/193/92:3 (Deep)	بحث → حث	...لذا فقد حث المجلس على التحول من اله...	...لك فإنه يحث المجلس على التحول من نهج...
Progressive to Perfective	4	Ongoing action rendered as completed	2005/s/2005/690/14:3 (DeepL)	كانت إريتريا تنظم → ظلت تنظم	فإريتريا ظلت تنظم وتدرب فئات شتى من ع...	كانت إريتريا تنظم وتدرب ثم ترسل إلى إ...
Tense Shift Under Negation	3	Tense changes incorrectly under negation	2011/s/2013/467/84:2 (Deep)	لن يكون → لا يكون	ولا يكون نظام الجزاءات فعالا...	إن نظام العقوبات لن يكون فعالاً إلا إذا تم ف...
Wrong Structure	3	Sentence structure violates Arabic syntactic patterns	2012/s/res/2083__2012_/178:1 (Google)	يرسل من اللجنة جميع → إحالة جميع المعلومات المعلومات	(ج) إحالة جميع المعلومات التي زودت بها اللجن...	(ج) يرسل من اللجنة جميع المعلومات المتعلقة بالقرار ال...
Wrong Word Order	3	Constituents in incorrect syntactic positions	2014/s/2014/943/35:3 (Deep)	القانون الإنساني → القانون الدولي الإنساني الدولي	...ماعات المسلحة بموجب القانون الدولي الإنساني العرفي.	...ماعات المسلحة بموجب القانون الإنساني الدولي العرفي.
Noun to Adjective	2	Source noun rendered as adjective	2011/s/2011/463/305:7 (Deep)	تحقيق → التحقيق	... سلطات خاصة في مجال التحقيق للسلطات المعنية بإن...	...لقانون فيها صلاحيات تحقيقية خاصة.
Adjective to Noun	2	Source adjective rendered as noun	2011/s/2011/245/51:2 (Google)	المستمرة → استمرار	...ي التحديات القضائية المستمرة التي تواجهها الدول ...	...تان، والثاني يتعلق باستمرار الطعون القضائية في ...
Active to Passive Voice	2	Active construction rendered in passive	2005/s/2005/60/563:4 (Transformers)	شن الهجوم → شنت الهجوم	...ات العسكرية هي التي شنت الهجوم أساسا، بينما قام ال...	شن الهجوم بشكل رئيسي من قبل ا...
Passive to Active Voice	2	Passive construction rendered in active	2002/s/2002/169/93:5 (Google)	نُقلت → تم نقل	...لموجودة في غوما، قد نُقلت إلى كيغالي وجيسيني.	وتم نقل طائرات البعثة الأرب...
Pragmatic Divergence (2 classes, parent weight 0.35)						
Register Mismatch	3	Formality level differs between source and target	2006/s/2006/822/209:8 (Deep)	تبادل كلامي → جدال شفوي	ودار جدال شفوي محتدم بين العقيد ري...	ودار تبادل كلامي ساخن بين العقيد ريس...
Terminology Substitution	4	Domain-specific term replaced with incorrect equivalent	2012/s/prst/2012/22/10:3 (Deep)	يدعو → يهيب	ويهيب بالأمين العام للأمم المتح...	ويدعو الأمين العام للأمم المتحد...
Orthographic Shifts (1 class, parent weight 0.05)						
Spelling Error	1	Orthographic error: typos, missing/extra letters	2005/s/2005/60/540:11 (Transformers)	اربعة → أربعة	... قد قتلا. وقتل أيضا أربعة أشخاص آخرين.	...اصغر قتلا ، كما قتل اربعة اشخاص آخرين.

Table 10: The full MIZAN 23-class taxonomy with one real En→Ar UN Gold-test example per class ([OMITTED] marks a Total Omission). The engine in parentheses after each Sentence_ID (Google, DeepL, Deep, or Transformers) is the MT system that produced that example.

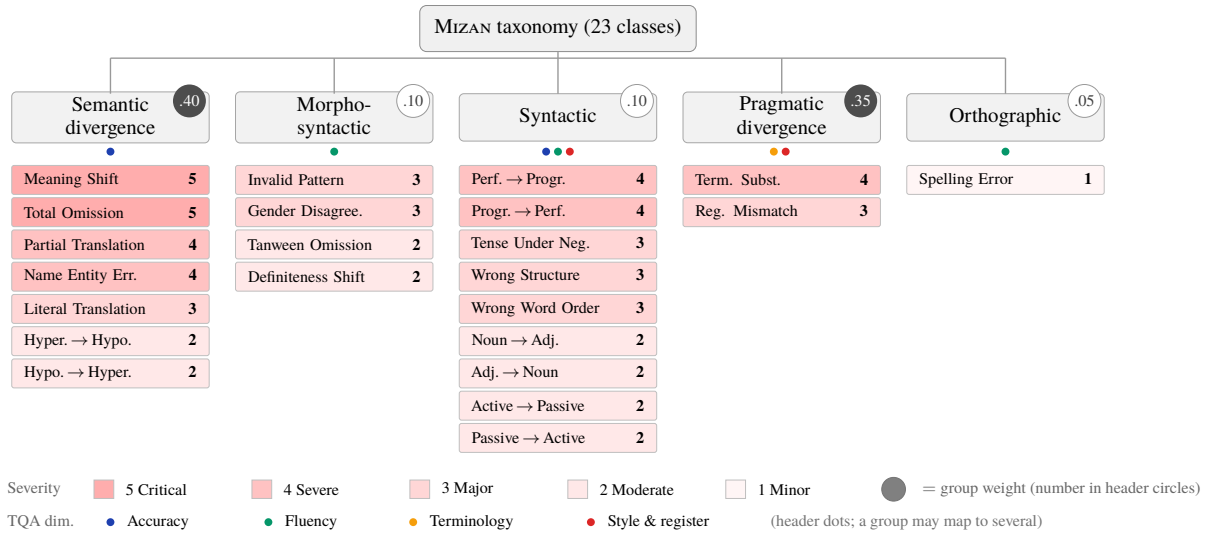

Figure 3: Mizan taxonomy: the 23 leaf classes grouped under their five parent groups; leaf shade and number encode severity, header dots the TQA dimension(s), the badge the parent-group weight. The full four-level rollup, including the 14 subtypes, is in Table 12.

regime in which Cohen's κ collapses artifactually (Feinstein and Cicchetti, 1990; Gwet, 2008). On both tiers, error correctness and anchor correctness reach almost-perfect agreement; label correctness reaches 0.930 on T2 and 0.680 on T1. T2 keyword-level severity reaches AC1 = 0.987 (supporting the severity-weight calibration); T2 UN-register fit reaches 0.608 (moderate, just below the substantial threshold); T1 keyword-level severity reaches 0.434 (moderate), reflecting expected divergence on fine-grained ordinal severity judgments against real-text contexts (Lommel et al., 2024). The mean AC1 over all reported criteria is 0.898 on T2 and 0.729 on T1; restricted to the three core taxonomy criteria (error, anchor, label), the mean rises to 0.966 and 0.827 respectively. All three core criteria meet or exceed the Landis–Koch substantial-agreement threshold on both tiers. Two of the seven workbook criteria, minimal-pair structure and naturalness, are recorded in the annotation protocol but omitted from Table 13 because both elicited systematic interpretation divergence between the two annotators (on minimal-pair, the annotator-mean positive rates were 0.90 versus 0.12); their low agreement reflects protocol ambiguity on those subjective judgments rather than unreliable taxonomy labels.

C Construction and Audit Algorithms

Algorithm 1 states the anchor-constrained construction that produces the T1 (real) and T2 (synthetic) tiers, and Algorithm 2 states the cross-vendor Etca audit. Both are described in prose in §4.1 and §4.2; the runnable implementation ships with the released repository.

D Etca Generation Prompt and Audit Rubric

This appendix reproduces verbatim the two prompts routed through the cross-vendor Etca audit (§4.2): the GPT-4o generation prompt and the Claude Sonnet 4 audit rubric.

Generation prompt (GPT-4o, 3-shot). The generator generates a UN-domain Arabic reference and its erroneous MT output for a target Mizan class. It is prompted with a system message (Figure 4), a three-shot exemplar block of (reference, class, MT output) triples drawn from the T0 canonical seed pool and stratified to cover the target Mizan family, and a per-sample instruction that re-states the anchor constraint and the requested output format (JSON only). Decoding parameters: temperature = 0.7, top-p = 0.95, max_tokens = 256; each candidate is generated independently with no rejection sampling at the generator stage.

Audit rubric (Claude Sonnet 4, five-dimension scoring). The auditor receives a (reference, MT, target class) triple and returns five integer scores in $\{1, \ldots, 5\}$ on the following dimensions:

1. **Anchor Validity**: does the keyword→best_match transformation realise the target class?
2. **Phenomenon Clarity**: is the error unambiguous, with no competing class label?

MIZAN class	T1	T2	Gold	Total
Semantic Divergence				
Meaning Shift	–	369	70	439
Total Omission	542	384	62	988
Partial Translation	–	166	7	173
Name Entity Error	–	369	2	371
Literal Translation	–	369	15	384
Hypernym for Hyponym	21	369	30	420
Hyponym for Hypernym	6	369	9	384
Morphosyntactic Shifts				
Invalid Pattern	3	263	10	276
Tanween Omission	15	369	14	398
Gender Disagreement	–	366	4	370
Definiteness Shift	11	320	12	343
Syntactic Shifts				
Perfective to Progressive	–	367	5	372
Progressive to Perfective	9	369	6	384
Tense Shift Under Negation	–	369	3	372
Wrong Structure	–	369	8	377
Wrong Word Order	2	369	13	384
Noun to Adjective	–	357	3	360
Adjective to Noun	2	359	3	364
Active to Passive Voice	–	368	4	372
Passive to Active Voice	2	369	8	379
Pragmatic Divergence				
Register Mismatch	64	365	96	525
Terminology Substitution	61	369	52	482
Orthographic Shifts				
Spelling Error	–	113	2	115
Total	**738**	**7,856**	**438**	**9,032**

Table 11: Per-class instance counts across the T1 (real), T2 (synthetic), and Gold tiers of QUANTA. "–" marks the 11 classes with no real T1 support (8 not surfaced by pattern-based retrieval plus 3 removed by Gold-overlap deduplication); these are covered in training only through T2 synthesis. Counts computed from the released dataset splits.

3. **Arabic Naturalness**: is the surface form fluent Arabic, modulo the injected error?
4. **Collateral Severity**: is the rest of the sentence held constant (lower is better)?
5. **Seed Recommendation**: would this candidate, if used as a T0 seed, mislead future generation?

The **Data Quality Index (DQI)** aggregates the four substantive dimensions (excluding Seed Recommendation, which is used only for downstream filtering) as a normalised mean:

$$\text{DQI} = \frac{1}{|D|} \sum_{d \in D} \frac{s_d - 1}{4} \in [0, 1], \qquad (1)$$

where D is the set of the four substantive dimensions, d indexes a dimension, and $s_d \in \{1, \ldots, 5\}$ is its integer score; each term $(s_d - 1)/4$ rescales a 1–5 score to $[0, 1]$, and DQI is their mean. Collateral Severity, where lower is better, enters on its raw 1–5 scale; audited instances sit near its floor (mean $\approx$ 1.0), so its effect on DQI is minor. DQI is computed on the audited subset (the 794 high-confidence seeds and a stratified 414-instance T2 audit subset, $\approx$ 5% of T2 covering all 23 classes) as a quality measure; acceptance into the released 7,856-instance T2 is determined by the structural re-classification audit (§4.4; of 8,044 generated candidates, 188 are excluded as Partial-Translation/Total-Omission confounds, yielding 7,856), not by the DQI threshold. The verbatim rubric is in Figure 5.

The audit prompt also requests a brief free-text justification per dimension; these justifications ship with the release data and are used in error analysis but not in the DQI score.

ETCA audit scores. Table 14 reports the per-dimension means. On the 414-instance T2 audit subset the substantive dimensions are high and Collateral Severity is near its floor; on the 794 generation seeds the scores are lower (the seeds are real-MT instances), and 664 of the 794 seeds are

Algorithm 1 Anchor-constrained construction

Require: aligned sentence pairs (r, m): Arabic reference sentence r, MT output sentence m; MIZAN classes $\mathcal{C}$; seed pool T_0. An anchor (k, b, c) instantiates class c's keyword→best_match transformation (k the reference-side form, b the MT-side form; token-level spans; b = `[OMITTED]` marks a Total Omission; per-class transformations in App. A). Mined anchors apply the frozen keyword→best_match patterns to the pairs (§4.1). $\text{OCCURS}(x, y)$ holds iff span x matches a span of y under a diacritic-/clitic-normalized comparison. Anchor gate $\text{REALIZED}(k, b, \textit{ref}, \textit{mt}) \equiv \text{OCCURS}(k, \textit{ref}) \wedge (\text{OCCURS}(b, \textit{mt}) \vee (b{=}\texttt{[OMITTED]} \wedge \neg\, \text{OCCURS}(k, \textit{mt})))$.

Ensure: validated instance set D (unique instances)

(ETCA audit, Alg. 2, runs in parallel; it does not gate admission.)

$D \leftarrow \emptyset$
/ T1: mine real MT errors (anchor gate only) */*
for each aligned pair (r, m) **do**
  for each mined anchor (k, b, c) **do**
    if $\text{REALIZED}(k, b, r, m)$ **then** ▷ anchor structurally realized
      $D \leftarrow D \cup \{(r, m, k, b, c, \text{T1})\}$
    end if
  end for
end for
/ T2: synthesize to full 23-class coverage; STRUCTURALAUDIT: structural-noise filter, §4.1 */*
for each class $c \in \mathcal{C}$ **do**
  for each of N_c generated candidates (rotating across UN sub-domains) **do**
    $(r, m', k, b) \leftarrow \text{GENERATE}(c, T_0)$ ▷ GPT-4o, 3-shot from T_0: generates r, m', anchor (k, b)
    if $\text{REALIZED}(k, b, r, m')$ **and** STRUCTURALAUDIT(r,m',k,b,c) **then**
      $D \leftarrow D \cup \{(r, m', k, b, c, \text{T2})\}$
    end if
  end for
end for
return D

MIZAN class	Severity	s_i	w_{class}	Subtype (L3)	Parent group	TQA category
Meaning Shift	Critical	5	0.1200	Semantic Shift	Semantic Divergence	Accuracy
Total Omission	Critical	5	0.1200	Translation Gap	Semantic Divergence	Accuracy
Partial Translation	Severe	4	0.0400	Translation Gap	Semantic Divergence	Accuracy
Name Entity Error	Severe	4	0.0400	Translation Gap	Semantic Divergence	Accuracy
Literal Translation	Major	3	0.0400	Idiomatic Shift	Semantic Divergence	Accuracy
Hypernym for Hyponym	Moderate	2	0.0200	Granularity Shift	Semantic Divergence	Accuracy
Hyponym for Hypernym	Moderate	2	0.0200	Granularity Shift	Semantic Divergence	Accuracy
Invalid Pattern	Major	3	0.0300	Morphological Error	Morphosyntactic Shifts	Fluency
Tanween Omission	Moderate	2	0.0200	Morphological Error	Morphosyntactic Shifts	Fluency
Gender Disagreement	Major	3	0.0300	Agreement Error	Morphosyntactic Shifts	Fluency
Definiteness Shift	Moderate	2	0.0200	Definiteness Shift	Morphosyntactic Shifts	Fluency
Perfective to Progressive	Severe	4	0.0150	Aspect Shift	Syntactic Shifts	Fluency
Progressive to Perfective	Severe	4	0.0150	Aspect Shift	Syntactic Shifts	Fluency
Tense Shift Under Negation	Major	3	0.0200	Tense Shift	Syntactic Shifts	Accuracy
Wrong Structure	Major	3	0.0200	Structural Error	Syntactic Shifts	Fluency
Wrong Word Order	Major	3	0.0200	Structural Error	Syntactic Shifts	Fluency
Noun to Adjective	Moderate	2	0.0025	Syntactic Category Shift	Syntactic Shifts	Fluency
Adjective to Noun	Moderate	2	0.0025	Syntactic Category Shift	Syntactic Shifts	Fluency
Active to Passive Voice	Moderate	2	0.0025	Voice Shift	Syntactic Shifts	Style & Register
Passive to Active Voice	Moderate	2	0.0025	Voice Shift	Syntactic Shifts	Style & Register
Register Mismatch	Major	3	0.0700	Register Mismatch	Pragmatic Divergence	Style & Register
Terminology Substitution	Severe	4	0.2800	Register Mismatch	Pragmatic Divergence	Terminology
Spelling Error	Minor	1	0.0500	Orthographic Issues	Orthographic Shifts	Fluency

Table 12: MIZAN taxonomy rollup and severity weights.

Criterion	Validates	T1	T2	L–K min
Error correct.	Error existence	**0.900**	**0.989**	≥ 0.81
Anchor correct.	Anchor identification	**0.900**	**0.978**	≥ 0.81
Label correct.	MIZAN class label	0.680	**0.930**	≥ 0.61
Severity	Severity calib.	0.434	**0.987**	≥ 0.61
UN register	T2 register fit	—	0.608	≥ 0.61
Core-three mean (err., anchor, label)		0.827	0.966	—
All-reported mean		0.729	0.898	—

Table 13: Inter-annotator agreement (Gwet's AC1) on 55 T1 and 92 T2 instances. L–K min is the Landis–Koch agreement band each criterion is expected to meet (almost-perfect ≥ 0.81 for error/anchor, substantial ≥ 0.61 otherwise); UN-register fit is T2 only.

Subset	Anchor	Clarity	Natural.	Collateral
T2 audit (414)	4.70	4.74	4.62	1.02
Seeds (794)	4.17	4.45	4.40	2.15

Table 14: ETCA per-dimension means (1–5) on the 414-instance T2 audit subset and the 794 generation seeds; for Collateral Severity, lower is better.

recommended. Per-instance ETCA scores ship with the release data.

External validation of ETCA. On a 43-instance high-confidence subset adjudicated against the dual-annotator consensus described in Appendix B, ETCA reaches Gwet's AC1 = 0.62 (substantial agreement on the Landis–Koch scale); the present paper uses ETCA strictly as a parallel dataset-quality audit.

Algorithm 2 ETCA cross-vendor audit

Require: candidate set X (each generated by GPT-4o); substantive dimensions $\mathcal{D}$ = {Anchor Validity, Phenomenon Clarity, Arabic Naturalness, Collateral Severity}, each scored on an integer 1–5 scale by an independent vendor (Claude Sonnet 4); Seed Recommendation is a separate flag, scored independently of $\mathcal{D}$ (scale defined in the audit rubric, App. D), not included in DQI

Ensure: Data Quality Index $\text{DQI}(x) \in [0, 1]$ for each $x \in X$

for each candidate $x \in X$ **do**
  $s \leftarrow \textsc{Score}(x,\ \mathcal{D} \cup \{\text{Seed Recommendation}\})$ ▷ Claude Sonnet 4, a different vendor
  $\text{DQI}(x) \leftarrow \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \frac{s_d - 1}{4}$ ▷ normalized mean of the 4 substantive dims
  mark x high-confidence **if** $\text{DQI}(x) \geq 0.75$
end for
return $\{\text{DQI}(x)\}_{x \in X}$

E Per-Class Classifier Results

E.1 Multi-Level Analysis

Aggregating the per-class F1 values by TQA dimension (the L4→L1 macro-F1 progression is in §5): Fluency is strongest ($\approx$ 0.74–0.84, dominated by the Arabic morphological classes Tanween Omission, Definiteness Shift, Invalid Pattern) and Terminology weakest ($\approx$ 0.38–0.51, single-class group). Anchor-guided framing peaks on morphosyntactic phenomena and hits its representational ceiling on semantic distinctions, consistent with the Frontier-tier confusion pattern reported in body §5.

E.2 Per-Class F1 ($n \geq 10$ Gold Support)

Three readings follow. First, **surface-cue classes saturate near ceiling** across all five encoders: Total Omission (0.855–0.967), Definiteness Shift (0.774–0.923), Wrong Word Order (0.880–0.960), Tanween Omission (0.743–0.963) show that anchor-guided framing is sufficient to learn Arabic-morphology phenomena that scalar MT

Class	n	AraBERT v2	CAMeL-MSA	CAMeL-Mix	ConfliBERT	XLM-R
Total Omission	62	**0.967**	0.959	0.959	0.855	0.913
Definiteness Shift	12	**0.923**	**0.923**	0.774	**0.923**	0.800
Wrong Word Order	13	**0.960**	0.917	0.917	0.917	0.880
Tanween Omission	14	0.788	0.897	0.743	0.743	**0.963**
Invalid Pattern	10	**0.696**	0.552	0.541	0.625	0.640
Register Mismatch	96	0.649	0.657	0.653	0.601	**0.667**
Literal Translation	15	0.526	0.577	**0.811**	0.588	0.423
Hypernym for Hyponym	30	0.160	**0.475**	0.471	0.423	0.400
Terminology Substitution	52	0.382	0.417	0.458	0.494	**0.506**
Meaning Shift	70	**0.272**	0.225	0.108	0.198	0.228

Table 15: Per-class F1 (best fold) on the 10 MIZAN classes with ≥ 10 Gold support; bold marks the highest F1 per row.

```
You are an expert Arabic linguist generating synthetic MT error
examples for a UN-domain benchmark.

YOUR ROLE:
Generate NEW examples of specific translation error types. You
will
see a few real examples of each error type, then create NEW
instances
with NEW Keyword/Best_Match pairs that follow the same pattern of
linguistic shift.

DOMAIN GUIDANCE:
Generate content that resembles authentic United Nations texts (e.
g.,
Security Council resolutions, Secretary-General reports, official
UN
statements). Focus on political, diplomatic, legal, humanitarian,
or
conflict-related topics with formal, neutral, institutional tone.

MINIMAL PAIR REQUIREMENT:
The 'ar' (clean) and 'mt_output' (error) sentences MUST be
identical
except for the specified anchor edit (Keyword->Best_Match, or
Keyword
omitted). Do not introduce additional changes to vocabulary,
structure, or meaning beyond the specified error.

=== CRITICAL: DIVERSITY REQUIREMENTS FOR MODEL TRAINING ===
1. LEXICAL DIVERSITY: Use a COMPLETELY DIFFERENT keyword than any
   shown in examples or listed as "already used"
2. SYNTACTIC VARIATION: Use a DIFFERENT sentence opening than
recent
   examples
3. DOMAIN ROTATION: Draw vocabulary from the specified UN
subdomain
4. NO REPETITION: Never reuse keywords, sentence structures,
templates

=== HARD CONSTRAINTS ===
- Generate NEW Keyword DIFFERENT from seed examples AND "already
used"
- Keyword MUST appear exactly ONCE in 'ar'
- Best_Match MUST appear in 'mt_output' (unless Best_Match is [
OMITTED])
- 'ar' and 'mt_output' must be IDENTICAL except the anchor edit
- Formal Arabic, institutional tone, 15-30 words
- Start sentence with a DIFFERENT opening than recent examples
- Output ONLY valid JSON with exactly 5 keys:
  {"Keyword":"...","Best_Match":"...","ar":"...","mt_output
":"...","en":"..."}
```

Figure 4: GPT-4o generation system prompt (verbatim).

metrics cannot isolate. Second, **no encoder dominates**: the highest per-class F1 is distributed across all five encoders. Third, **semantic-divergence classes remain hard**: Meaning Shift peaks at 0.272 and Hypernym for Hyponym peaks at 0.475 across all encoders, highlighting the Frontier-tier classes where future work on representation or retrieval-augmented classification would yield the largest gains; that this pattern holds across all five encoders frames these as field-wide representational limits rather than encoder-specific failures, and augmenting the anchor with sentence-level semantic context (§6) is the most targeted next step.

The remaining 13 MIZAN classes have $n < 10$ gold instances, making per-class confidence intervals too wide for individual conclusions.

F Reproducibility

All experiments in this paper are fully reproducible from the released artefacts at the public repository linked in the title footnote.

End-to-end regeneration. Beyond reloading the released splits, QUANTA is reproducible *by construction*: the deterministic construction pipeline (anchor-gate extraction, the verbatim generation and audit prompts, and the DQI $\geq$ 0.75 audit threshold; Appendix C) regenerates the dataset from the public UN Parallel Corpus (Ziemski et al., 2016), so every result can be reproduced end-to-end from public source data rather than only reloaded. No institution-restricted data is involved.

Classifier training. Every classifier result uses the same training procedure: stratified 5-fold cross-validation (`random_state` = 42) on T1+T2 with the held-out 438-instance Gold test, with classes of fewer than 5 instances retained entirely in training. We train with learning rate 5×10^{-5}, batch size 32, 5 epochs, AdamW optimiser, weighted cross-entropy loss, a linear warmup of 10% of total training steps, mixed-precision training via `torch.amp.autocast`, and a maximum input length of 512 tokens. The best validation-fold checkpoint is selected by macro-F1 and evaluated on the held-out Gold test.

Significance testing. Pairwise paired-bootstrap tests (10,000 resamples on fold-level differences) with Benjamini–Hochberg FDR correction test for differences, and Two One-Sided Tests (TOST, equivalence margin $\delta = 0.05$) test for practical equivalence; both back the encoder comparison in

```
Role
You are an expert Arabic linguistics auditor evaluating whether a
specific annotated linguistic phenomenon is correctly instantiated
 in
a machine-translated Arabic sentence pair.

You are NOT judging overall translation adequacy.
You are NOT reclassifying labels.
You are NOT fixing or rewriting text.
Your task is to audit anchor-level correctness and usability of
this
sample as a Tier2 generation seed.

You will receive
- error_label: annotated Sub-Subtype (fixed; do not change)
- keyword: lexical anchor expected in the clean sentence
- best_match: expected transformation in the MT output (may be a
  lexical form or the sentinel [OMITTED])
- clean_ar: original Arabic sentence
- error_ar: machine-translated Arabic sentence

Your task
Evaluate only whether the annotated phenomenon (error_label) is:
1) correctly instantiated via keyword / best_match,
2) clearly observable in the clean vs. error Arabic pair,
3) linguistically plausible as MT output,
4) suitable for use as a Tier2 generation seed.
Judge with respect to the given label only. Do NOT consider
alternative labels. Do NOT infer meaning from any other language.

Evaluation axes (return five values)
1) anchor_validity:       1 = contradicts label, 5 = perfect
alignment
2) phenomenon_clarity:    1 = barely observable, 5 = unambiguous
3) arabic_naturalness:    1 = broken, 5 = fluent plausible MT
Arabic
4) collateral_severity:   1 = no unrelated distortion, 5 = severe
5) tier2_seed_recommendation: 0 = do NOT use, 1 = safe to use

Decision guidance
- Collateral changes are acceptable if the target phenomenon is
intact.
- Rare classes are valid if correct.
- If anchors are ambiguous or weak, set tier2_seed_recommendation
= 0.
- Never rescue by imagining intent -- judge only what is present.

Output format (STRICT)
Return exactly five TAB-separated values in this order:
anchor_validity  phenomenon_clarity  arabic_naturalness
collateral_severity  tier2_seed_recommendation
Valid example:  5  4  4  2  1

Hard constraints: No explanations. No text. No JSON. No additional
numbers. No class reassignment. Any deviation invalidates the
output.
```

Figure 5: Claude Sonnet 4 audit rubric (verbatim).

Table 5, with full per-pair results in Appendix G. Bootstrap 95% confidence intervals on best-fold predictions use 1,000 resamples.

Released artefact pointer. The QUANTA dataset (T1, T2, and the held-out Gold test set), the frozen 5-fold splits, the released AraBERT v2 checkpoints (all five folds; fold-0 default), the ETCA verbatim prompts, the post-hoc audit manifests, and the full benchmark logs are released with the paper. SHA-256 checksums, pinned dependency versions, and per-experiment reproduction bundles (each with a script that re-scores the reported numbers) ship in the repository's `REPRODUCE.md`.

Environment. Python 3.10, CUDA 12.1; pinned dependency versions in the repository's `requirements.txt`.

G Encoder Significance and Equivalence

Beyond reporting mean macro-F1, we test all ten encoder pairs two ways (Table 16): a paired bootstrap (10,000 resamples on the five fold-level differences, fixed seed) with Benjamini–Hochberg FDR correction tests whether two encoders *differ*, and a Two One-Sided Tests procedure (TOST, equivalence margin $\delta = 0.05$ macro-F1) tests whether they are *practically equivalent*. We report both because a non-significant difference is not, on its own, evidence of equivalence.

Only three pairs are significant after correction, all involving the multilingual XLM-R or the domain-adjacent ConfliBERT, while the three Arabic-specific encoders (AraBERT v2, CAMeLBERT-MSA, CAMeLBERT-Mix) are mutually equivalent within ± 0.05 macro-F1. Together these validate AraBERT v2 as the released encoder: it is significantly stronger than the non-Arabic and domain-adjacent baselines (both $p_{adj} \leq 0.007$) and statistically no worse than any other Arabic-specific encoder, while attaining the highest mean. ConfliBERT, although it is AraBERT v2 further pre-trained on conflict-domain Arabic, does not gain from that adaptation on this UN-domain Modern Standard Arabic task, consistent with the known trade-offs of domain-adaptive pre-training.

The FDR correction is consequential: AraBERT v2 vs. CAMeLBERT-MSA is significant before correction ($p_{\text{raw}} = 0.031$) but not after correcting for the ten comparisons ($p_{adj} = 0.078$). Reading the two tests jointly (Figure 6), two pairs are genuinely different (significant and non-equivalent: the two significant XLM-R comparisons), four are genuinely equivalent (non-significant and within ± 0.05: the two within-family Arabic pairs involving CAMeLBERT-Mix and the two CAMeLBERT–ConfliBERT pairs), AraBERT v2 vs. ConfliBERT is significant but small (its $+0.032$ gap sits near the margin, $p_{\text{TOST}} = 0.077$), and the remaining three are inconclusive at five folds.

H Anchor Framing and Data-Efficiency Ablations

This appendix gives the full numerical support for the two body-level claims in §5: that anchor-guided framing is the essential input signal, and that structurally validated real data is roughly an order of magnitude more sample-efficient than size-matched synthetic data on the 12 classes for which T1 has training support.

Pair	ΔF1	95% CI	p_{adj}	Diff?	p_{TOST}	Equiv?
AraBERT v2 vs. ConfliBERT	$+0.032$	$[+0.015, +0.052]$	< 0.001	yes	0.077	marginal
AraBERT v2 vs. XLM-R	$+0.050$	$[+0.018, +0.069]$	0.007	yes	0.490	no
CAMeLBERT-Mix vs. XLM-R	$+0.032$	$[+0.009, +0.057]$	0.015	yes	0.136	no
AraBERT v2 vs. CAMeLBERT-MSA	$+0.027$	$[+0.003, +0.050]$	0.078	no	0.087	marginal
AraBERT v2 vs. CAMeLBERT-Mix	$+0.017$	$[-0.004, +0.041]$	0.303	no	0.034	yes
CAMeLBERT-Mix vs. ConfliBERT	$+0.014$	$[-0.012, +0.035]$	0.355	no	0.026	yes
ConfliBERT vs. XLM-R	$+0.018$	$[-0.012, +0.043]$	0.355	no	0.056	marginal
CAMeLBERT-MSA vs. XLM-R	$+0.022$	$[-0.015, +0.059]$	0.355	no	0.139	no
CAMeLBERT-MSA vs. CAMeLBERT-Mix	-0.010	$[-0.037, +0.017]$	0.574	no	0.030	yes
CAMeLBERT-MSA vs. ConfliBERT	$+0.004$	$[-0.021, +0.033]$	0.773	no	0.023	yes

Table 16: Pairwise encoder comparison: the paired-bootstrap difference test (Benjamini–Hochberg FDR-corrected p_{adj}; significant when $p_{adj} < 0.05$) and the TOST equivalence test (margin $\delta = 0.05$ macro-F1; equivalent when $p_{\text{TOST}} < 0.05$). ΔF1 is the mean fold-level difference (first encoder minus second).

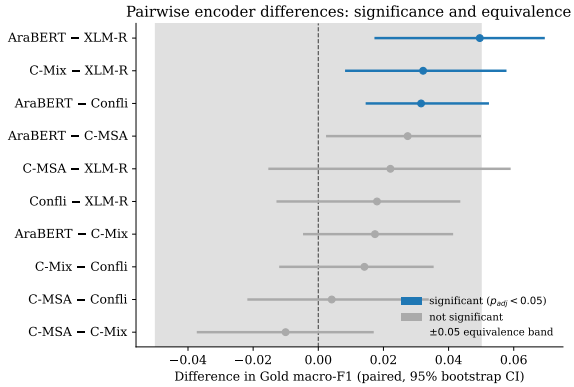

Figure 6: Pairwise encoder differences in Gold macro-F1 with 95% bootstrap CIs. A CI clearing the dashed zero line indicates a significant difference; the shaded band marks the ± 0.05 TOST equivalence margin, so both tests are visible in one view.

H.1 Data Ablation

The data ablation in Table 8 (§5) trains and evaluates each configuration (T1, T2, T1+T2) on the same 12 T1-supported Mizan classes, so the sample-efficiency comparison is not confounded by class coverage. The Data Efficiency Score is

$$\text{DES} = \frac{\text{F1}_{12}}{n/1000}, \tag{2}$$

where F1_{12} is the 12-class macro-F1 and n the number of training instances, i.e., macro-F1 per 1,000 samples; it is a transparent size-normalised contrast, not a scaling law. Bootstrap 95% confidence intervals on the 23-class macro-F1 (1,000 resamples on best-fold Gold predictions) confirm the combination: T1+T2 $[0.568, 0.671]$ does not overlap T2 $[0.401, 0.485]$. Real data contributes per-instance precision, synthetic data contributes volume and class-label coverage, and the combination yields the strictly best operating point.

H.2 Format Ablation (Anchor vs. Blind)

Table 17 compares AraBERT v2 under two input formats on identical folds and hyperparameters. Format A (blind): `[CLS] ar mt [SEP]`, the classifier sees only the Arabic reference and the MT output. Format D (anchor-guided): `[CLS] keyword best_match [SEP] ar mt [SEP]`, the classifier additionally receives the Mizan anchor pair that localises the error.

Format	Mean F1 ± std	Best fold
A (blind: `ar mt`)	0.1135 ± 0.030	0.1609
D (anchor: `kw bm + ar mt`)	$\mathbf{0.6185 \pm 0.023}$	**0.6397**
Δ	**+0.505** (**+445**%)	+0.479

Table 17: Input-format ablation on AraBERT v2, a controlled blind-vs-anchor run with its own 5-fold split; anchor-guided framing improves macro-F1 by $+0.505$ ($+445\%$, paired-bootstrap $p < 0.0001$). The Format-D mean here (0.6185) is from this ablation run, distinct from the main-benchmark mean (0.614).

The anchor is not a minor enhancement: it is the essential signal that enables fine-grained Arabic MT error classification.

I T2 Per-Class Generation Audit

Table 18 reports the T2 per-class count against the 369-sample-per-class generation target. Fourteen classes hit the target exactly; nine fell short, of which three are materially short ($< 90\%$ of target). The materially-short classes are hard to synthesise under the anchor constraint for class-specific mechanistic reasons (Spelling Error requires a minor orthographic perturbation that leaves context intact; Invalid Pattern requires a morphological mismatch the generator often over-repairs; single-token Definiteness Shift is blocked by the clitic-matcher's current scope limit). Weighted cross-entropy mitigates the residual class imbalance at training time.

J Cross-Language Dataset Comparison

Table 1 compares Mizan against prior error *taxonomies*; Table 19 instead positions the Quanta

Mizan class	Count	% of 369
Active-to-Passive Voice	368	99.7%
Perfective-to-Progressive	367	99.5%
Gender Disagreement	366	99.2%
Register Mismatch	365	98.9%
Adjective-to-Noun	359	97.3%
Noun-to-Adjective	357	96.7%
Definiteness Shift	**320**	**86.7%**
Invalid Pattern	**263**	**71.3%**
Spelling Error	**113**	**30.6%**
14 classes at target	369	100.0%

Table 18: T2 per-class shortfall against the 369-sample generator target: 14 classes hit it, 9 fell short (3 materially, $< 90\%$, bold). Generation-stage counts (8,044), before the 188-confound exclusion (§4.4).

dataset against fine-grained MT-error and diagnostic datasets in other languages. ACES (Amrhein et al., 2022) and DEMETR (Karpinska et al., 2022) are perturbation-based sets for meta-evaluating MT *metrics*; the WMT MQM corpus (Freitag et al., 2021), IndicMT-Eval (Sai B et al., 2023), and SiniticMTError (Liu et al., 2025) provide human error annotation for European, Indian, and Sinitic languages respectively; and WMT24++ (Deutsch et al., 2025) supplies references and post-edits for 55 languages including Arabic, but without error-type labels. Fine-grained MT-error annotation is thus an established direction across several languages, yet none of these resources targets Arabic with an anchor-verified, error-typed dataset paired with a diagnostic classifier. To our knowledge, Quanta is the first fine-grained ($\geq$ 20-class) Arabic MT-error dataset.

K Ethics and Broader Impact

Data source and licensing. Quanta is derived entirely from the United Nations Parallel Corpus (Ziemski et al., 2016), the public corpus of institutional UN documents also used by Osorio et al. (2025). The Quanta dataset (the Mizan annotations and taxonomy), the released Raqeeb classifier, and the verbatim Etca prompts are publicly available under the CC-BY-4.0 license at the repository linked in the title footnote (`https://github.com/AALight/Quanta-Mizan`); the underlying UN source text remains subject to the UNPC terms.

Annotation ethics and team. This work does not involve human research subjects; all annotators are authors of the paper. The Mizan taxonomy was designed by two authors with doctoral-level expertise in translation and in linguistics; the dual-annotator validation (Appendix B) was performed independently by two further authors, both expert Arabic linguists (one doctoral-, one master's-level), who were not involved in taxonomy design, each instance labelled independently by both in per-instance CSV workbooks. No personally identifying information about contributors is reported beyond academic degree and research specialisation. The UN content is formal institutional text with no graphic, harmful, or distressing material.

Scope and cultural sensitivity. Mizan describes general Arabic morphological phenomena, but the released Raqeeb classifier is calibrated and validated on formal Modern Standard Arabic (the UN Security Council domain); dialectal use would require additional validation.

LLM use disclosure. T2 synthetic instances are generated by GPT-4o (OpenAI, 2024); the Etca audit uses Claude Sonnet 4 (Anthropic, 2025). Cross-vendor generator-judge separation (§4.2) is deliberate and mitigates same-family LLM self-preference bias (Kim, 2025; Wang et al., 2026). Model outputs are not treated as ground truth: they are structurally validated and gold-evaluated against expert-annotated instances.

Intended use and broader impact. The released Raqeeb classifier is designed as a *diagnostic tool* for fine-grained Arabic MT error-type analysis in high-quality formal-text translation domains (institutional, legal, medical). It is not a scalar quality score and is not intended to replace human translators, to rank MT systems without expert review, or to serve as an automatic MT-quality gate in production without human review. The 23-class classifier has documented Frontier-tier limitations (§5); deployment outside the formal-text scope it was calibrated on requires recalibration and additional validation. Arabic is one of the lowest-resourced languages for fine-grained MT evaluation despite being a top-10 global language by speaker count. Existing taxonomies operate at 3–7 top-level categories and do not capture Arabic-specific morphological phenomena. Quanta and Mizan address this gap with a 23-class taxonomy, a publicly released 9,032-instance dataset, and a documented cross-vendor LLM-as-judge audit protocol that the community can replicate, build on, and re-calibrate to other formal-text domains, for which the released taxonomy, data, and classifier provide the starting point. Positive-impact applica-

Dataset	Lang. pair(s)	Size	Error classes	Validation / notes
ACES (Amrhein et al., 2022)	146 pairs (en-x, x-en, x-y)	36,476 examples	68 phenomena (MQM ontology)	Synthetic contrastive; MT-metric challenge set
DEMETR (Karpinska et al., 2022)	10 pairs (X→En)	∼31,000 examples	35 perturbation types	Synthetic minimal pairs; MT-metric diagnostic
WMT MQM (Freitag et al., 2021)	2 pairs (En→De, Zh→En)	3,418 segments (1,418 + 2,000)	22 MQM categories	Human span annotation (MQM)
IndicMT-Eval (Sai B et al., 2023)	5 pairs (En→Indic)	7,000 annotations (1,400/lang.)	13 MQM categories	Human MQM (span, severity)
SiniticMTError (Liu et al., 2025)	4 pairs: En→Mandarin, Cantonese, Wu; Mandarin→Hokkien	3,633 sents (2,009 + 1,402 + 68 + 154, pilots)	9 error types	Human span, type, severity
QUANTA (Ours)	1 pair (En→Ar)	9,032: 738 real + 7,856 synth + 438 gold	23 MIZAN classes	Dual-annotator AC1 0.83–0.97; ETCA audit; anchor-constrained

Table 19: QUANTA positioned against fine-grained MT-error and diagnostic datasets in other languages. ACES and DEMETR are synthetic sets for meta-evaluating MT *metrics*; WMT MQM, IndicMT-Eval, and SiniticMTError are human error-annotated sets for European, Indian, and Sinitic languages. for WMT MQM the size is the WMT20 En–De and Zh–En source segments annotated with MQM by professional translators; SiniticMTError's Wu (68) and Mandarin–Hokkien (154) splits are small pilots. QUANTA is the only Arabic, anchor-verified, error-typed dataset at 20+ classes, and is released with a diagnostic classifier.

tions include reducing annotator workload in large-scale Arabic MT evaluation by pre-filtering structurally valid candidates and supporting expert post-editors with per-error-type diagnostic signals at the token level.