

Is TunisianMMLU Actually Tunisian?

A Native-Speaker Audit of a Machine-Translated Dialect Benchmark

Fatma Ezzahra Jlali
Lycée Pilote Manouba
Manouba, Tunisia
fjlali91@gmail.com

Marek Šuppa
Comenius University Bratislava
Bratislava, Slovakia
marek@suppa.sk

Abstract

TunisianMMLU, the only MMLU-style benchmark for Tunisian Arabic (Derja), was machine-translated with no native-speaker validation. Scanning all 21,500 items with native-validated patterns, we flag 63.4% of the corpus; the largest signal, in 55.1% of items, is the Moroccan present-tense prefix, a form Tunisian does not have. The dominant defect is not translation noise but a different dialect. We then present the first native-speaker audit: 400 items stratified over all 44 subjects, labelled independently by two native speakers on six criteria. The non-author annotator marks 64.5% defective, both agree on 62.0%, and author adjudication yields 86.0%. Nearly all defects originate in translation, not in the source benchmarks. Six models corroborate the labels, scoring highest on items marked clean and lowest on unrepairable ones. Corrected text raises every above-chance model by 2.3 to 4.9 points (one gain survives multiple-test correction), leaves a chance-level Derja model unchanged, and narrows the gap between a dialect-adapted model and its base without changing rank order. We release TunisianMMLU-Verified: a corrected subset, corpus-wide flags, per-item labels, the annotation guideline, and all evaluation code.¹

1 Introduction

Evaluation claims about low-resource language varieties are only as reliable as the benchmarks that produce them (Singh et al., 2025; Gema et al., 2025). For Tunisian Arabic (Derja),² a variety with no standardized orthography and, to our knowledge, one MMLU-style benchmark, that

benchmark was produced by machine translation with no native-speaker validation stage, as its own documentation states (annotations_creators: machine-generated) (Ghezaiel and Lorré, 2025). To our knowledge, the extent to which it measures Tunisian language competence, rather than translation artifacts, has not previously been examined. It is the first evaluation resource of its kind for Derja, and the pipeline it documents is the one most low-resource benchmark efforts can afford. Our aim is to establish what it currently measures and to make it usable, not to discount the effort of building it.

Machine-translated benchmarks without native validation are increasingly common. Singh et al. (2025) showed that translated MMLU distorts multilingual evaluation across 42 languages and released human-verified translations, but state that dialects are “a topic we do not address here.” Altakrori et al. (2026) answered for five Arabic dialects—Syrian, Egyptian, Emirati, Saudi, and Moroccan—by manually translating MMLU-Redux. Tunisian is not among them. Meanwhile, Gema et al. (2025) estimate that 6.49% of MMLU itself contains errors at source: any translated descendant inherits those defects before translation adds its own.

We audit TunisianMMLU with the methodology these lines of work established, and one addition each of them lacked: an *error-origin* judgment that separates what translation broke from what was already broken, and a *repair-cost* reading that separates what a word swap fixes from what requires retranslation.

Our contributions are:

1. An *error-origin* label separating translation-introduced defects from those inherited from the source benchmarks, and a *repair-cost* verdict separating word-swap repairs from items needing retranslation—neither present

¹Data: <https://huggingface.co/datasets/fatmajlali/TunisianMMLU-Verified>; code and annotations: <https://github.com/jjlalli/tunisianmmlu-audit>.

²We write *Derja* for Tunisian Arabic (ISO 639-3 aeb) and *Darija* for Moroccan Arabic (ary), following community usage.

in prior benchmark audits.

2. A native-validated mechanical scan of the full corpus, giving a corpus-level view the 400-item audit cannot.
3. Evidence that the contamination is not neutral across models, including a dialect-adapted model whose advantage over its own base is visible only on corrected text.
4. TunisianMMLU-Verified: corrected items, per-item labels and verdicts, corpus-wide severity flags, the annotation guideline, and all evaluation code.

2 The TunisianMMLU Benchmark

TunisianMMLU (Ghezaiel and Lorré, 2025) comprises multiple-choice questions across 44 subject configurations, produced by translating subsets of MMLU (Hendrycks et al., 2021), ArabicMMLU (Koto et al., 2024), and DarijaMMLU (Shang et al., 2025) into Tunisian Derja using NLLB-200-3.3B (NLLB Team et al., 2022) followed by CODA orthographic normalization (Habash et al., 2012; Zribi et al., 2014). Its documentation declares the pipeline: `annotations_creators: machine-generated`, `language_creators: machine-translated`, with no native validation stage listed. The card reports 22,027 items, but the released data files total 21,500 (the discrepancy is undocumented). We scan that snapshot (accessed August 2026), and all corpus-wide figures below refer to it rather than to the card’s count. The released files mark every item as MMLU- or ArabicMMLU-derived, and no item is marked as DarijaMMLU-derived. The benchmark was built by the developers of the Labess Tunisian LLM to evaluate their own model in-dialect, and is released under CC BY-NC-SA 4.0.

Two structural properties matter for any analysis. First, **provenance and subject are fully confounded**: every subject configuration derives entirely from a single source benchmark (English-derived subjects such as `high_school_psychology` from MMLU; school-exam subjects such as `islamic_studies` from ArabicMMLU). The released files carry a `subject_darija` field with 41 distinct values, cutting across the 44 subject configurations; only two of its values appear under both sources (528 items). Consequently, all source-level comparisons in this paper are descriptive, not

causal. Second, the context field—supporting passages in the school-exam style—exists almost exclusively on ArabicMMLU-derived items (706 items, 602 of them in one subject), so item *shape* also differs by provenance.

3 Audit Methodology

Items. Items carry two to five options (80.0% have four), so expected accuracy from uniform guessing is 27.9% over the corpus and 27.5% over the audited sample; we use “chance” in that sense throughout.

Sample. We drew 400 items stratified over all 44 subjects: a floor of four items per subject, with the remainder allocated proportionally to subject size, from a fixed seed (20260805). Both annotators worked the sample in an identical randomized order, so that any early stopping point still yields a valid random sample.

Annotators. Two native Tunisian speakers resident in Tunisia, with northern (Bizerte) and central-southern (Gafsa and Sidi Bouzid) dialect exposure. Annotator A is an author. Annotator B was recruited personally by A, has no NLP background, and was unaware of the study’s hypotheses, which makes hypothesis-consistent labeling less likely; she participated voluntarily and without payment, was offered co-authorship, and chose acknowledgment instead.

Instrument. Each item receives six binary labels: *semantic fidelity* (is the question coherent and answerable), *answer preservation* (is the marked key still uniquely correct), *Tunisian authenticity* (would a Tunisian produce this sentence), *MSA/Moroccan interference* (is the text free of intrusions from Modern Standard Arabic or Moroccan Darija), *orthographic naturalness* (is the spelling one a Tunisian would write), and *cultural validity* (does the premise survive in context). Each item then receives a verdict—OK, FIXABLE (repairable by a native speaker in under 90 seconds), or DISCARD (unrepairable without re-translation)—and, when defective, an origin label: TRANSLATION (the item would be sound in correct Derja), SOURCE (the item is defective in any language), or UNCLEAR. The origin label is necessary because a share of any translated benchmark’s defects predate translation and should not be attributed to the translation pipeline.

Protocol. Annotation was blind and independent, and agreement is computed with Cohen’s κ (Cohen, 1960) per label and on the binary broken/not-broken judgment, following Gema et al. (2025). Disagreements are adjudicated by annotator A, an author, while annotator B remains hypothesis-naïve. Corrections for FIXABLE items were written by annotator A. The 90-second criterion is a guideline threshold, not a timed measurement. Annotators judged items from the Derja text alone, without consulting source benchmarks. Suspected source-key errors were flagged but not overturned, and we report flags separately from discards. We release the guideline, whose worked examples were drawn exclusively from items outside the sample.

4 Audit Results

Error rate. After adjudication, 344 of 400 items (86.0%) are defective: 333 (83.2%) are FIXABLE and 11 (2.8%) are DISCARD, leaving 56 (14.0%) OK as written. Because adjudication was performed by annotator A, we also report the figure on which both annotators independently agree: 248 of 400 items (62.0%) were marked defective by both, and 45 (11.2%) acceptable by both. The defect rate of TunisianMMLU therefore lies between 62.0% and 86.0% depending on annotator strictness, and we treat 62.0% as the conservative bound. Annotator B alone marks 258 of 400 items (64.5%) defective, a figure that involves no judgment by an author. The two annotators disagree on the binary judgment for 107 items and on the three-way verdict for 119. Adjudication departed from A’s independent verdict on only 6 of those 119, so the adjudicated figure is close to a strict-annotator bound. The descriptive Wilson 95% interval on the sample is [82.3, 89.1] for the adjudicated rate and [57.2, 66.6] for the agreed floor, both sample-level intervals rather than design-based corpus estimates. The mechanical scan of Section 5, which involves no per-item human judgment, flags 67.5% of the sample, between the two (Figure 1). Defectiveness here is a claim about the text as Tunisian, not about answerability: answer preservation fails on 4.3% of items for annotator A and 1.5% for annotator B, so most defective items remain answerable, and the benchmark fails on its stated construct rather than as a set of questions.

Error origin. Of the 344 defective items, 330 (95.9%) are TRANSLATION-origin: the item

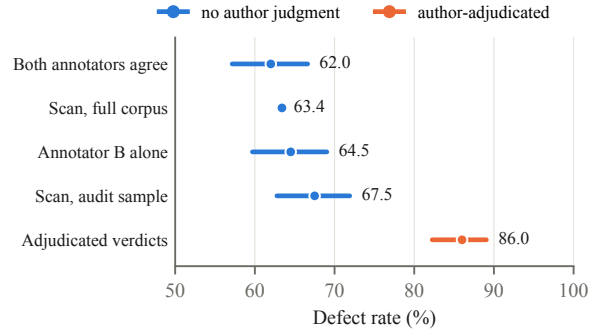


Figure 1: Defect-rate estimates on the audit sample and the corpus. Every estimate that involves no author judgment falls between 62.0% and 67.5%; only author adjudication reaches 86.0%. Bars are Wilson 95% intervals over the 400-item sample; the full-corpus scan is a census and has none.

would be sound in correct Derja. Seven (2.0%) are SOURCE-origin defects present before translation, and seven are UNCLEAR. Because annotators did not consult the source benchmarks, subtle source defects are undercounted—we detect fewer than the 6.49% Gema et al. (2025) report for MMLU—so 95.9% is an upper bound on the translation share. Defects in this benchmark are overwhelmingly attributable to the translation pipeline rather than inherited from the source benchmarks.

Agreement. On the binary defective/acceptable judgment, raw agreement is 73.2% with $\kappa = 0.32$. The disagreement is directional rather than random: annotator A marks 345 of 400 items defective and annotator B marks 258, and the largest disagreement cell is 92 items that A judged FIXABLE and B judged OK. The per-label pattern explains the gap: the most objective label, MSA/Moroccan interference, reaches $\kappa = 0.55$ (78.5% agreement), while the labels that ask whether a Tunisian would produce the text—Tunisian authenticity ($\kappa = 0.31$) and orthographic naturalness ($\kappa = 0.25$)—sit lowest. Tunisian Derja has no standard orthography, so these judgments have no external anchor and the two native annotators place the threshold of acceptability differently. Altakrori et al. (2026) identify the same variability as inherent to dialectal Arabic annotation. Two further labels, semantic fidelity and cultural validity, exceed 94% raw agreement but yield κ near zero because their distributions are nearly constant, which is why we report raw agreement alongside κ . Note that κ measures the independent passes, whereas released verdicts are post-adjudication.

Original

إلا ما كانوش لوحات كيحدو السرعة، السرعة القصوى المسموح بها داخل المناطق المأهولة كتكون: Moroccan *ka-* present prefix (كتكون، كيحدو) and existential negation.

Corrected

إذا ما فاش علامات تحدد في السرعة، السرعة القصوى المسموح بها داخل المناطق العمرانية هي: Tunisian existential *and* bare imperfect تحدد.

“If there are no signs specifying the speed limit, the maximum speed allowed inside built-up areas is:”

Table 1: A FIXABLE item (driving test) before and after correction.

Pipeline defect families. Beyond dialect interference, annotation surfaced defect families that the benchmark’s documentation does not report. Placeholder corruption: the token الفراغ (“the blank”) in fill-in-the-blank exam items survives translation as the non-word فيراغ. Formatting artifacts: 456 items contain a literal \n in place of a line break, and 624 items contain the malformed token في for في. Duplicated options: items in which two answer options are identical, in some cases duplicating the keyed answer. Instruction leakage: translation instructions appearing inside item text. Religious scripture: Quranic text rendered into dialect rather than preserved in its canonical form; we restore these to the text of the Cairo (1924) edition in the Ḥaṣṣ عاصم عن reading, verified by both annotators, and treat them as a distinct defect category. Table 1 shows a representative FIXABLE item, and Appendix A shows further items for each family.

5 Corpus-Wide Scan

The audit covers 400 items, so to address whether it generalizes we scan all 21,500 items for mechanically detectable Moroccan markers and pipeline artifacts, flagging 13,640 of them (63.4%). The dominant signal is the Moroccan present-tense prefix *كت-* / *كي-*, present in 55.1% of items, where Tunisian Derja marks the present with the bare imperfect and has no such prefix. The scan also matches Moroccan function words (غادي، ديال)، and واخا forms, against Tunisian counterparts (باش، متاع)، the single-*lam* spelling الي of *إلي*³ and two pipeline artifacts (في، literal \n), with frequent Tunisian and MSA words sharing the

³A native speaker of Moroccan Arabic who reviewed the inventory rated this spelling rare in Moroccan usage; we count

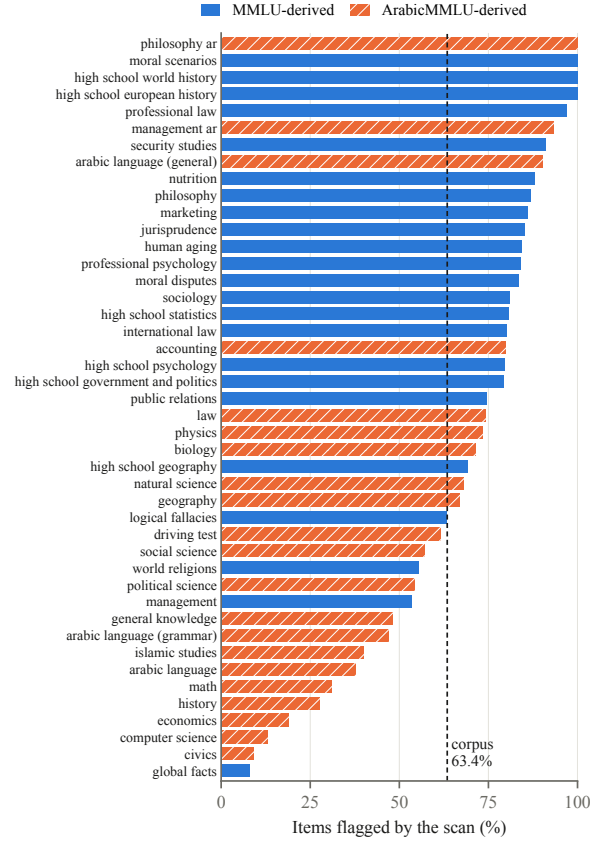


Figure 2: Share of items flagged by the marker scan in each of the 44 subject configurations, by source benchmark. The dashed line is the corpus rate. Every subject derives from a single source, so source and subject are confounded (Section 2).

same initial letters (كيميا، كآب، كيميا) excluded by list. The executable scanner, with its patterns, normalization rules and marker-wise counts, is part of the release. Several lexical markers are shared with Algerian varieties, and غادي is also a Tunisian locative adverb. The attribution to Moroccan therefore rests on the prefix, which Algerian largely lacks, while the defect claim requires only that the forms are non-Tunisian in the flagged functions. Annotator A reviewed 30 randomly sampled items flagged by the prefix alone and confirmed all 30. Against the adjudicated verdicts on the audit sample, the scan reaches 98.1% precision and 77.0% recall on the binary defective judgment (265 true positives, 5 false alarms, 79 misses, 51 true negatives, for 270 of 400 items flagged, 67.5%). Since the sample’s defect base rate is 86.0%, precision alone is uninformative. The false-alarm count is the diagnostic figure: the scan flags 5 of the 56 adjudicated-OK items and 2 of the 45 items both

it as a non-Tunisian spelling. It accounts for the five false alarms on the audited sample.

annotators passed. By source benchmark, 86.3% of MMLU-derived items are flagged against 51.6% of ArabicMMLU-derived items. Moroccan morphology on MMLU-derived items cannot be inherited from an English source, so unless items are mislabelled by provenance—which we cannot rule out, given that no released item is marked DarijaMMLU-derived—the pipeline introduced it (Figure 2). Eight subject configurations are at or above 90%, including `moral_scenarios` (895 of 895 items). The scan under-counts: absence of a string match does not indicate a clean item, and the annotators marked defective further items the scan does not reach. The scan-flag rate within the 400-item sample (67.5%) is consistent with the corpus-wide rate.

6 Model Evaluation

Setup. We evaluate six models on the 400 original items and on the corrected subset (389 items: corrected text for `FIXABLE` items, unchanged text for `OK` items, `DISCARD` items excluded). All comparisons between conditions are restricted to the same 389 items. Three open-weight models run locally with 4-bit quantization: Labess-7B (linagora/Labess-7b-chat-gguf, Q4_K_M), the Tunisian-adapted model whose developers released TunisianMMLU; ESPRIT-Derja-8B (ESPRIT-Group/ESPRIT-Derja-Qwen3-8B-v2), a Derja fine-tune of Qwen3-8B; and Qwen3-8B (Qwen/Qwen3-8B), its general-purpose base. Three commercial models run through their APIs at full precision: GPT-5.6 Sol (openai/gpt-5.6-sol), Claude Sonnet 5 (anthropic/claude-sonnet-5), and Gemini 3.1 Pro (google/gemini-3.1-pro-global); all API runs were made in August 2026. The prompt and greedy decoding are identical across models and conditions, answer extraction is format-constrained per backend, no model produced an unparsable answer, and all evaluation code is released.

Models corroborate the annotators. Before measuring what correction does, we check the audit labels against model behaviour on the original text (Table 3). Every one of the six models orders its accuracy `OK` > `FIXABLE` > `DISCARD`, and none exceeds 36.4% on the `DISCARD` items—frontier models included. With $n = 11$, the `DISCARD` column is indicative rather than precise (a Wilson interval on 4/11 spans roughly

Model	Orig.	Corr.	Δ	Flips
Gemini 3.1 Pro	91.0	93.3	+2.3	15
GPT-5.6	89.5	92.3	+2.8	29
Claude Sonnet 5	80.5	85.3	+4.9	39
Qwen3-8B	49.6	52.7	+3.1	60
ESPRIT-Derja	47.6	52.4	+4.9	55
Labess-7B	29.6	29.8	+0.3	93

Table 2: Accuracy (%) on the same 389 items under original and corrected text, and the number of items changing correctness between conditions. Δ is computed from item counts before rounding, so it can differ by 0.1 from the difference of the rounded columns.

[0.15, 0.65]), but six models of widely different capability agree with the two annotators about which items are broken beyond repair. This is an independent check on the labels that two annotators alone cannot provide.

Corrections help every model that is above chance. All five above-chance models improve on corrected text, by 2.3 to 4.9 points (Table 2); Labess, at chance, is unchanged. Rank order is identical in both conditions: zero of six models change rank. What changes is the spacing: Qwen3-8B’s lead over ESPRIT-Derja falls from 2.1 points to 0.3, and GPT-5.6’s deficit to Gemini narrows from 1.5 to 1.0. Under McNemar’s exact test on the discordant pairs, unadjusted p -values are .003 (Claude Sonnet 5), .015 (ESPRIT-Derja), .035 (Gemini), .061 (GPT-5.6) and .155 (Qwen3-8B); under Holm correction across the five tests only Claude Sonnet 5 remains significant at .05. We therefore treat the individual deltas as suggestive and the narrowing of between-model gaps as descriptive. The number of items changing correctness between conditions falls broadly with model strength, from 93 for Labess to 15 for Gemini, the one inversion being Qwen3-8B (60) and ESPRIT-Derja (55), which sit 2.1 points apart: the weaker the model, the more its answers vary with any change to the text.

The benchmark cannot measure its own target model. TunisianMMLU was released by Labess’s developers to evaluate that model in-dialect. Labess performs at chance under our harness (115 of 389 correct, 29.6%, against 27.5% expected from guessing; binomial $p = .36$)—including on the 56 items whose adjudicated verdict is `OK` (30.4%, versus 29.4% on `FIXABLE` items). Its own chat template was applied and a schema-constrained answer format removed pars-

Model	OK	FIX.	DISC.
	<i>n</i> =56	<i>n</i> =333	<i>n</i> =11
Gemini 3.1 Pro	92.9	90.7	27.3
GPT-5.6	92.9	88.9	36.4
Claude Sonnet 5	85.7	79.6	36.4
Qwen3-8B	62.5	47.4	36.4
ESPRIT-Derja	53.6	46.5	18.2
Labess-7B	30.4	29.4	9.1

Table 3: Accuracy (%) on original text by adjudicated verdict band.

ing failures, so the result is not a formatting artifact, though quantization effects cannot be excluded (Limitations). Its aggregate score is also unstable at the item level: 93 of 389 items (23.9%) change correctness between conditions while the aggregate moves 0.3 points, consistent with guessing. On the 11 DISCARD items Labess scores 9.1%, below chance, consistent with items that have no recoverable answer. Labess shows the same ordering as the other five models (Table 3), but from a chance-level baseline.

Contamination understates dialect-adapted models. ESPRIT-Derja’s gain concentrates on ArabicMMLU-derived items (+7.7, from 51.4% to 59.0%, $n = 222$) rather than MMLU-derived items (+1.2, $n = 167$), while Qwen3-8B gains uniformly (+3.0 and +3.2). The gain does not track marker density—the MMLU-derived items are the more heavily flagged (Section 5)—but the dialect-adjacent school-exam domain, and provenance, subject, and item shape are confounded (Section 2). We therefore read the result as consistent with suppression of the dialect-adapted model’s advantage, not as a causal estimate.

7 Related Work

Audits of multilingual NLP data. Kreutzer et al. (2022) manually audited 205 language-specific corpora from five web-crawled multilingual datasets and found quality failures concentrated in low-resource languages, including at least 15 corpora with no usable text; their design pairs a human audit with automatic analyses, the structure we adopt at benchmark scale. A related response is to abandon translated evaluation entirely: Son et al. (2025) built KMMLU from original Korean exams because prior Korean benchmarks, translated from English, missed linguistic and cultural properties

of the language. For Tunisian Derja no original-exam alternative exists, which motivates auditing and repairing the translated benchmark instead.

Errors in MMLU and its descendants. Gema et al. (2025) re-annotated 5,700 MMLU items with an error taxonomy, estimating 6.49% of MMLU defective overall and up to 57% in single subjects, and showed that model rankings change when erroneous items are removed. Our audit extends this line to a machine-translated descendant, where the error taxonomy must additionally separate translation-introduced defects from inherited ones; our error-origin label finds the split to be 95.9% versus 2.0%.

Translated multilingual benchmarks. Singh et al. (2025) demonstrate at the scale of 42 languages that machine-translated MMLU distorts evaluation, that 28% of items require culturally sensitive knowledge, and that model rankings shift between full and culturally-sensitive subsets; they release professionally verified translations but explicitly exclude dialects. For African languages, IrokoBench (Adelani et al., 2025) and Uhura (Bayes et al., 2026) construct evaluation sets with human translation or verification rather than raw machine translation. Our work applies the verification standard these efforts establish to a dialect benchmark that lacked it.

Arabic and dialectal Arabic evaluation. ArabicMMLU (Koto et al., 2024) built native evaluation for Modern Standard Arabic from regional school exams, avoiding translation entirely. DarijaMMLU (Shang et al., 2025) produced a Moroccan benchmark by translation as part of the Atlas-Chat effort. AraDiCE (Mousi et al., 2025) created dialectal evaluation sets by machine translation followed by human post-editing. Its native evaluators judged raw dialectal MT output unacceptable for up to 60% of Egyptian items, the defect rate that post-editing was added to correct. TunisianMMLU’s pipeline has no such stage. DialectalArabicMMLU (Altakrori et al., 2026) manually translated MMLU-Redux into five Arabic dialects with per-dialect translator, reviewer, and adjudicator roles, deliberately avoiding machine translation, but Tunisian is not among its dialects. It reports its largest performance declines on Moroccan, the variety it also finds most lexically distant from MSA—and Moroccan is the dominant contaminant we find in TunisianMMLU. Robinson

et al. (2025) show that LLMs understand dialectal Arabic better than they produce it, which makes the quality of benchmark text itself a factor in dialect evaluation. TunisianMMLU is the only MMLU-style resource for Tunisian, so any in-dialect evaluation claim for Tunisian depends on its quality.

Dialect identification and dialectness. Fine-grained Arabic dialect identification is an established task with its own resources and shared tasks: MADAR (Bouamor et al., 2018) provides a 25-city parallel corpus, including Tunis, that documents exactly the lexical contrasts our scan relies on, and the NADI shared tasks (Abdul-Mageed et al., 2024) have moved to multi-label dialect identification because Arabic varieties overlap substantially in text. Keleg et al. (2023) argue further that dialectness is a continuous variable rather than a binary label. Two consequences for this work: membership judgments over closely related Maghrebi varieties are known to be hard, which is consistent with our lower agreement on the authenticity labels; and an automatic dialect-identification pass over the corpus would provide independent corroboration of our marker scan, which we leave to future work.

Tunisian orthography. CODA (Habash et al., 2012) and its Tunisian instantiation (Zribi et al., 2014) define conventional orthographies for dialectal Arabic, and TunisianMMLU’s pipeline cites CODA normalization. Our scan shows the normalization left the Moroccan present-tense prefix intact in over half the corpus, and our agreement analysis shows that in the absence of an adopted community standard, even native judgments of orthographic naturalness lack a shared anchor ($\kappa = 0.25$ despite 62.5% agreement).

8 TunisianMMLU-Verified

TunisianMMLU-Verified repairs the existing resource rather than replacing it: every artifact is keyed to the original item identifiers, so corrections and flags can be merged upstream instead of forking the benchmark. We release: the audited subset with per-item labels, verdicts, origins, and both annotators’ independent judgments; corrected Derja text for all 333 FIXABLE items (the question text for 329 of them and the options for 172, including three de-duplicated option lists); severity flags for all 21,500 items; the annotation guideline with worked examples;

the sampling seed and item identifiers; and all evaluation code. The artifacts are at <https://github.com/jjlalli/tunisianmmlu-audit> and <https://huggingface.co/datasets/fatmajlali/TunisianMMLU-Verified>; Appendix A shows worked examples by defect family. Licensing differs by artifact: the corrected Derja text is derived from TunisianMMLU and inherits CC BY-NC-SA 4.0 through its ShareAlike clause, so the non-commercial restriction propagates; the evaluation code is released under MIT, and the annotation guideline and per-item labels, which are newly authored, under CC BY 4.0.

9 Discussion

Four practices follow from this audit for machine-translated benchmark construction. First, native validation before release: a two-annotator audit of a 400-item sample found between 62.0% and 86.0% of the sample defective; AraDiCE’s post-editing stage (Mousi et al., 2025) exists for the same reason. Second, marker scans as a release-time check: a validated string scan covered the full corpus at 98.1% precision and would have caught the dominant defect before release. Third, error-origin labeling, so that translation-introduced damage is not conflated with inherited source errors. Fourth, reporting model accuracy by verdict band: six models ordering OK > FIXABLE > DISCARD is a convergence check available to any audited benchmark. The broader reading is that unvalidated machine-translated benchmarks need not be neutrally noisy. On this one, the contamination is consistent with suppressing ESPRIT-Derja’s advantage over its own base model, and the benchmark scores Labess—the model it was released to evaluate—at chance. The problem compounds for work on language adaptation, the setting these benchmarks are most often built to serve. A dialect-adapted model evaluated on text written in the wrong dialect is measured least well on exactly the capability it was adapted for. None of this is specific to Tunisian or to this benchmark. The pipeline that produced it—machine translation into a low-resource variety with no native validation stage—is ordinary practice, and the four checks above are cheap enough to apply to any of its outputs.

Limitations

Our sample is 400 of 21,500 items, so subject-level conclusions are limited to coarse groups. Two annotators, both native Tunisian speakers, cannot represent regional variation within Tunisia, and we make no region-balance claim. Adjudication was performed by one annotator, who is also an author, so we report the two-annotator agreement floor (62.0%) and the non-author annotator’s rate (64.5%) alongside every adjudicated figure. Adjudication by an independent third native speaker, and a larger annotator pool, would remove this dependence and is the first extension we plan. Provenance and subject are fully confounded in the benchmark itself, so all source-level comparisons are descriptive. The severity classification is a regular-expression heuristic reviewed by one native speaker, not a linguistic analysis. Corrections cover the question field for 329 of the 333 FIXABLE items and the options for 172. The open-weight models are evaluated under 4-bit quantization, which may shift absolute scores, though the original-versus-corrected comparison within each model uses identical quantization and is unaffected. The dialect-adaptation contrast rests on a single base-model pair, Qwen3-8B and its Derja fine-tune, and replication across further pairs is future work. All model results come from a single run with one prompt template and a fixed option order. Multiple-choice accuracy is known to be sensitive to such choices, and the McNemar tests and identical-text OK items are only partial mitigations. Corrections repaired formatting artifacts and duplicated options as well as dialect, and corrected text may sit closer to Modern Standard Arabic, so part of each model’s gain may reflect general denoising rather than dialect repair. The absence of any gain for Labess argues against a pure-fluency account, but we do not decompose gains by defect type. Corrections were written by annotator A and not independently verified. A formatting error in the evaluated file, found after acceptance, left 173 of the 333 FIXABLE items in the corrected condition with the annotator’s option list (changed for 169 of them, unchanged for four) embedded in the question field and the original list repeated after it; gains on those items and on the 160 correctly formatted items differ by less than three points for every model, and the release contains both the file as evaluated and the repaired one. Audit percentages are sample-level quantities under the stated strati-

fied design rather than corpus projections, and the scan provides the corpus-level view. Finally, a corrected 400-item subset improves measurement but does not repair the full benchmark. Correcting the remaining items would require the kind of native post-editing effort applied by [Singh et al. \(2025\)](#) and [Altakrori et al. \(2026\)](#).

Ethics and AI-Use Statement

The audit protocol was designed with LLM assistance; all annotation, correction, and adjudication was performed entirely by the human annotators.

The study collected no data from human subjects. It did rely on human annotation, described in Section 3: annotator B gave informed consent, having been told the study’s purpose and how her labels would be used, and reviewed the description of her work reported here. Annotators judged published benchmark items only, so no personal or sensitive data was involved. No institutional ethics review was sought, as the work constitutes no human-subjects research under our local requirements.

The audited benchmark is publicly released under a license permitting this use. Religious texts encountered in the data were restored to their canonical form as described in Section 4.

Acknowledgments

We thank Mariem Masri, annotator B, for annotating the full sample, Yassine Chaouri (Mohammed V University) for reviewing the Moroccan marker inventory, and the anonymous reviewers for their comments. We would also like to acknowledge that the work has been partially funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I02-03-V01-00029.

References

- Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. [NADI 2024: The fifth nuanced Arabic dialect identification shared task](#). In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 709–728, Bangkok, Thailand.
- David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker,

- Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Ijeoma Chukwuneke, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukibi, Salomon Kabongo Kabenamualu, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, and 8 others. 2025. [IrokoBench: A new benchmark for African languages in the age of large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2732–2757.
- Malik H. Altakrori, Nizar Habash, Teresa Lynn, Younes Samih, Abed Alhakim Freihat, Kirill Chirkunov, Muhammed AbuOdeh, Radu Florian, Preslav Nakov, and Alham Fikri Aji. 2026. [DialectalArabicMMLU: Benchmarking dialectal capabilities in Arabic and multilingual language models](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 3199–3219, Palma de Mallorca, Spain. ELRA Language Resource Association.
- Edward Thomas Bayes, Israel Abebe Azime, Jesujoba Alabi, Jonas Kgomo, Tyna Eloundou, Elizabeth Proehl, Kai Chen, Imaan Khadir, Naome A. Etori, Shamsuddeen Hassan Muhammad, Choice Mpanza, Igneciah Pocia IP Thete, Dietrich Klakow, and David Ifeoluwa Adelani. 2026. [Uhura: A benchmark for evaluating scientific question answering and truthfulness in low-resource African languages](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 1485–1504, Palma de Mallorca, Spain. ELRA Language Resource Association.
- Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. [The MADAR Arabic dialect corpus and lexicon](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.
- Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. 2025. [Are we done with MMLU?](#) In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5069–5096, Albuquerque, New Mexico.
- Wajdi Ghezaiel and Jean-Pierre Lorré. 2025. [TunisianMMLU](#). Hugging Face dataset, <https://huggingface.co/datasets/linagora/TunisianMMLU>. LINAGORA. Accessed August 2026.
- Nizar Habash, Mona Diab, and Owen Rambow. 2012. [Conventional orthography for dialectal Arabic](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 711–718.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations (ICLR)*.
- Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. [ALDi: Quantifying the Arabic level of dialectness of text](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10597–10611, Singapore.
- Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. [ArabicMMLU: Assessing massive multitask language understanding in Arabic](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.
- Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wajah, Daan van Esch, Nasanbayer Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 2022. [Quality at a glance: An audit of web-crawled multilingual datasets](#). *Transactions of the Association for Computational Linguistics*, 10:50–72.
- Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arif Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. [AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4186–4218, Abu Dhabi, UAE. Association for Computational Linguistics.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, et al. 2022. [No language left behind: Scaling human-centered machine translation](#). *arXiv preprint arXiv:2207.04672*.
- Nathaniel Romney Robinson, Shahd Abdelmoneim, Kelly Marchisio, and Sebastian Ruder. 2025. [AL-QASIDA: Analyzing LLM quality and accuracy systematically in dialectal Arabic](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22048–22065, Vienna, Austria. Association for Computational Linguistics.

Guokan Shang, Hadi Abdine, Yousef Khoubrane, Amr Mohamed, Yassine Abbahaddou, Sofiane Ennadir, Imane Momayiz, Xuguang Ren, Eric Moulines, Preslav Nakov, Michalis Vazirgiannis, and Eric Xing. 2025. [Atlas-chat: Adapting large language models for low-resource Moroccan Arabic dialect](#). In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 9–30, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. [Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2025. [KMMLU: Measuring massive multitask language understanding in Korean](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4076–4104, Albuquerque, New Mexico. Association for Computational Linguistics.

Inès Zribi, Rahma Boujelbane, Abir Masmoudi, Mariem Ellouze, Lamia Hadrich Belguith, and Nizar Habash. 2014. [A conventional orthography for Tunisian Arabic](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*.

A Worked examples by defect family

Tables 4–6 show items from the audited sample as released and, for FIXABLE items, as corrected in TunisianMMLU-Verified. Item identifiers are `subject::row` in the released files. Glosses are ours.

Defect	Item	As released	Corrected	Gloss
Moroccan present prefix كـ-	biology::1356	كنتنج حبوب اللقاح بواسطة الانقسام داخل أكياس حبوب اللقاح فالمترك	حبوب اللقاح تتكوّن بالانقسام داخل أكياس حبوب اللقاح في المتك.	Pollen grains are produced by division inside the pollen sacs in the anther.
Moroccan present prefix, negated	biology::485	وحدة من هاذي ما كنتستعملش فالتفاعلات الضوئية الحلقية هي :	وحدة من هاذي ما تتستعملش في التفاعلات الضوئية الحلقية:	One of these is not used in the cyclic light reactions:
future; غادي complementiser; بي for immoral حرام	moral_disputes::262	شنوة الهدف الرئيسي متاع بينس في المقال متاعو 'شنوة الاستنساخ غادي يضر الناس؟' bash يستعمل الحجج التائجة باش بين الي الاستنساخ حرام باش يعترض على الحجج التائجة ضد الاستنساخ باش يستعمل صيغة الإنسانية متاع كانط باش بين الي الاستنساخ حرام باش يستعمل صيغة الإنسانية متاع كانط باش بين الي الاستنساخ مقبول أخلاقيا	شنوة الهدف الرئيسي متاع بينس في مقالو "بشنوة" باش يضر الاستنساخ الناس؟" bash يستعمل الحجج التائجة باش بين الي الاستنساخ غير أخلاقي باش يعترض على الحجج التائجة ضد الاستنساخ باش يستعمل صيغة الإنسانية متاع كانط باش بين الي الاستنساخ غير أخلاقي باش يستعمل صيغة الإنسانية متاع كانط باش بين الي الاستنساخ مقبول أخلاقيا	What is Pence's main aim in his article 'How will cloning harm people?'
possessive; دبال calqued what percentage	global_facts::50	حتى لسنة 2019، شكون الي من المية في المية دبال الناس الي يقولوا باللازم تصدق ربي باش يكون عندك أخلاق ومبادئ مليحة؟	حتى لعام 2019، قدّاش من المية من الناس يقولوا الي لازم تؤمن بربي باش يكون عندك أخلاق ومبادئ باهية؟	As of 2019, what percentage of people say you must believe in God to have good morals and values?
existential; كين reads as third in Tunisian	biology::210	كاينين تلت أنواع مختلفين متاع العضلات هما	فما ثلاثة أنواع مختلفين متاع العضلات هو ما:	There are three different types of muscle; they are:
in a name; هاد in the options -; الشبي	high_school_psychology::41	السيد غوردون تضرر ليه الجزء الخلفي من القصر الجبهي الأيمن. نتيجة لهاد الشبي، ما قدرش options: يفهم المعلومات الي كيسمع يفهم المعلومات الي كيقرا يتكلم بشكل مفهوم يحرك يدو اليسرى	السيد غوردون تضرر عندو الجزء الخلفي من القصر الجبهي الأيمن. نتيجة لهذا، ما نتجش: options: يفهم المعلومات الي يسمعها يفهم المعلومات الي يقراها يتكلم بشكل مفهوم يحرك يدو اليسرى	Mr Gordon suffered damage to the back of his right frontal lobe. As a result he could not:
Moroccan indefinite المسار	natural_science::573	عملية انتقال الطاقة من كائن حي لكائن آخر في واحد المسار محدد هي:	عملية انتقال الطاقة من كائن حي لكائن حي آخر في مسار محدد هي:	The transfer of energy from one living organism to another along a specific path is:

Table 4: FIXABLE items, dialect interference: Moroccan morphology and function words, with the native correction.

Defect	Item	As released	Corrected	Gloss
في → ف corruption: [فراغ] for [فيراغ], في فيبلاصة بلاصة	arabic_language_ (grammar)::227	اختار اسم الإشارة المناسب باش تحطو فيبلاصة [فيراغ] في الجملة هادي: شنوة [فيراغ] (ملعقة)	اختار اسم الإشارة المناسب باش تحطو في بلاصة [فراغ] في الجملة هادي: شنوة [فراغ] (ملعقة) ؟	Choose the demonstrative that fits the [blank] in this sentence: What is [blank]? (a spoon)
artifact في	natural_science:: 395	أي المراحل تعد الأطول في دورة حياة الإنسان؟	اناهي مرحلة تعتبر الأطول في دورة حياة الإنسان؟	Which stage is the longest in the human life cycle?
Literal \n; ما ، كيتاخدو ; كاينش كيشاركو	management_ar::24	اختار الجواب الصحيح بنعم ولا لا من ما كاينش شي \n\nهاذا الشي الي جاي: احتياطات كيتاخدو بعين الاعتبار منين كيشاركو الأفراد	اختار الجواب الصحيح بنعم ولا لا من الي جاي: ما فاش احتياطات تتاخذ بعين الاعتبار وقتلي يشاركوا الأفراد.	Choose yes or no: no precautions are taken into account when individuals take part.
Instruction leakage	civics::137	عملية التغيير الاجتماعي ترتبط بـ ترجم للتونسية	عملية التغيير الاجتماعي ترتبط بـ:	The process of social change is linked to [Now translate to Tunisian]
Instruction leakage	computer_science:: 430	العنصر الرئيسي لقواعد البيانات الآن ترجم إلى التونسية	العنصر الرئيسي في قواعد البيانات هو:	The main element of databases [now translate into Tunisian]
Duplicated option (هم twice); A: DISCARD, B: OK, adjudicated FIXABLE after de-duplication	arabic_language_ (grammar)::295	أَي من التكمالات التالية هي الصحيحة بدلا من [فيراغ] جاو التلاميذ، وقابل [فيراغ] الأستاذ. options: هم هما هو و	شنوة من التكمالات هادي هي الصحيحة في بلاصة [فراغ]؟ جاو التلاميذ، وقابل [فراغ] الأستاذ. options: هم هما هو و	Which completion is correct in place of [blank]? The pupils came, and the teacher met [blank].
Quranic text paraphrased into dialect; restored to the canonical text	arabic_language:: 588	جملة قرآنية وحدة بس فيها ضمير نصب منفصل options: ((هو الله اللي ما فإ إله غير هو)) ((وبهاد الشي تأمرت وأنا أول المسلمين)) ((وحنا قراب ليه أكثر من جبل الوريد)) نعبدو))	جملة قرآنية وحدة برك فيها ضمير نصب منفصل: options: ﴿ هو الله الذي لا إله إلا هو ﴾ ﴿ وبذلك أمرت وأنا أول المسلمين ﴾ ﴿ ونحن أقرب إليه من جبل الوريد ﴾ ﴿ إياك نعبد ﴾	Only one Quranic sentence contains a detached accusative pronoun:

Table 5: FIXABLE items, pipeline defects: the في → ف corruption, formatting artifacts, instruction leakage, duplicated options, and religious text.

Verdict	Item	As released	Gloss
SOURCE-origin, DISCARD: the key is wrong in any language (plastics are petrochemical; key says textile)	social_science::168	صناعة البلاستيك في الوطن العربي كتنتمي للصناعات :- options: الهندسية البترولية المعدنية النسيجية key: النسيجية	The plastics industry in the Arab world belongs to the ... industries:
SOURCE-origin, DISCARD: Moroccan legal substance (الحراسة النظرية, 96-hour regimes of the Moroccan CPP), not only Moroccan language	law::80	أجل الحراسة النظرية: options: 48 ساعة في الجرائم العادية كتنبدا من ساعة التوقيف قابلة للتمديد مرة وحدة لمدة 24 ساعة بإذن مكتوب من النيابة العامة 96 ساعة في الجرائم التي كتتمس بأمن الدولة الداخلي ولا الخارجي قابلة للتمديد مرة وحدة بإذن مكتوب من النيابة العامة 96 ساعة إلا كان الأمر متعلق بجرمة إرهابية قابلة للتمديد زوز مرات لمدة 96 ساعة بناء على إذن مكتوب من النيابة العامة كل الأجوبة صحيحة key: كل الأجوبة صحيحة	The period of police custody:
TRANSLATION-origin, DISCARD: right-to-left rendering scrambled the coordinates and minus signs; no reading yields the key	physics::152	إيلا كان موضع متاع ثلاثة د لجسام متشابهين هوما (2،2) ، - 4 - 4 ، - 4 ، - 5 ، 7 ، (،) موضع مركز الكتلة غادي يكون options: 3 - ، 1 - () 3 1 ، - () 3 ، 1 () 3 - ، 1 () key: 3 - ، 1 - (	If the positions of three similar bodies are (2,2), (-4,-4), (-5,7), the centre of mass will be
OK (both annotators)	civics::23	شنة موقف الشريف الحسين من الدولة العثمانية قبل ما يعلن على الثورة؟ options: العداء للاتحاديين العداء للدولة العداء للسلطان كل شي الي تقال صحيح key: العداء للاتحاديين	What was Sharif Hussein's stance toward the Ottoman state before he declared the revolt?

Table 6: DISCARD items by error origin, and an item both annotators passed.