WHEN CONFIDENCE DOESN'T TRAVEL: CALIBRATION OF OPEN-WEIGHT LLMS ON AFRICAN-LANGUAGE MCQS UNDER PROMPT PERTURBATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Research question: Are open-weight LLMs well-calibrated on African-language multiple-choice questions, does their calibration survive prompt perturbations, and at what compute cost? **Answer:** Mostly no. The failure is model-specific in ways aggregate benchmarks hide, and the full audit costs under $40. We audit four open-weight instruct models (Llama-3.1-8B, Qwen2.5-7B, Gemma-2-9B, Mistral-7B) on 422,000 MCQ items spanning 24 Global-MMLU languages (9 African, 15 European) and 18 afrimmlu configurations, under 5 prompt perturbations, for $39.48 on serverless A100s. Accuracy drops 16–31 points from European to African language cells for every model. The calibration gap, measured as Expected Calibration Error (ECE), is real but model-dependent at baseline: large for Qwen (+0.140 ECE [0.099, 0.177]) and Gemma (+0.133 [0.109, 0.159]), absent for Llama, and reversed for Mistral. Perturbation effects flip sign across models; a format that calibrates one model can break another, and Gemma loses 20 accuracy points in African language cells simply when asked to "answer with a single letter", with a 26.1% rate of fallback cases (no valid answer letter emitted) across evaluated African cells that reaches 100% in some format_A afrimmlu cells. Miscalibration is largely repairable in-distribution: regularized logistic regression fit on $\sim$250 items per cell cuts held-out ECE to 0.04–0.10 depending on region. We provide code, parsed outputs, and an executable reproduction notebook.

1 INTRODUCTION

Multiple-choice benchmarks such as MMLU (Hendrycks et al., 2021) are the de facto yardstick for LLM capability, and their multilingual descendants, Global-MMLU (Singh et al., 2025), which covers 42 languages, and afrimmlu (IrokoBench; Adelani et al., 2024), have exposed large accuracy disparities between high-resource and African languages. But deployment decisions in African education and public services depend not only on *whether* a model is right, but on *whether it knows when it might be wrong*. A model that is 26% accurate on Amharic MCQs while reporting 44.5% mean confidence (as we observe for Llama-3.1-8B at baseline; 49.8% under a simple reformatting) is actively misleading.

Calibration, the alignment between a model's confidence and its empirical accuracy, is well-studied in English (Guo et al., 2017), and recent work has begun to extend it to multilingual settings: Yang et al. (2023) measure ECE of multilingual QA models across a dozen languages (including Swahili via TyDiQA), and N2C2-style cross-lingual in-context calibration (He et al., 2025) studies transfer of calibration signals. What remains unstudied is the configuration that matters most for African NLP practice: **open-weight models, MCQ benchmarks in African languages, a dataset-matched Africa-vs-Europe comparison, prompt-perturbation robustness, and per-language repair**, all under a budget realistic for under-resourced labs.

Contributions. (1) A calibration audit of open-weight LLMs built around a dataset-matched Africa-vs-Europe comparison: 4 models $\times$ 42 language cells $\times$ 5 perturbations $\times \leq$500 items (422k items), all on serverless GPUs for $39.48. (2) A language-paired statistical analysis showing the Africa–Europe calibration gap is estimable but *model-dependent*, including a reversed gap for Mistral-7B,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

so "multilingual calibration" cannot be assessed from a single model. (3) A perturbation-robustness study showing format effects transfer across neither models nor regions, including a gemma-2-9b failure mode (26% no-answer rate on African prompts) invisible to accuracy-only evaluation. (4) Evidence that miscalibration is cheaply repairable in-distribution (regularized logistic regression; held-out ECE 0.04–0.10) and a full reproducibility package (Kaggle dataset + executable notebook; pinned vLLM 0.6.3.post1/transformers 4.45.2).

2 Related Work

Calibration of neural classifiers is classically measured by Expected Calibration Error (ECE; Naeini et al., 2015) and repaired by regularized logistic regression, Platt-style sigmoid calibration (Platt, 1999), isotonic regression (Zadrozny and Elkan, 2002), or temperature scaling (Guo et al., 2017). We adopt regularized logistic regression, isotonic regression, and temperature scaling as repair baselines, evaluated held-out (§4, Repair).

Multilingual and African evaluation. Global-MMLU (Singh et al., 2025) documents accuracy and cultural-bias disparities across 42 languages (we use 24 of them); IrokoBench/afrimmlu (Adelani et al., 2024) provides human-translated MCQs for 16 African languages plus English/French reference configurations; InkubaLM (Tonja et al., 2024) studies small models for African languages. These works measure accuracy, not calibration. Closest to us, multilingual-QA calibration studies (Yang et al., 2023; He et al., 2025) measure ECE across languages but focus on extractive QA with API or fine-tuned models, do not isolate an Africa-vs-Europe gap, do not probe prompt-perturbation robustness, and do not evaluate per-language repair. To our knowledge, no prior work audits open-weight LLM calibration natively on African-language MCQ benchmarks across all four of these axes.

Delta vs our group's prior work (anonymized): an earlier study evaluated calibration of API-served models on a small English-only item set. This study differs in every axis: open-weight models only (no API dependency), 42 language cells, 422k items, a perturbation-robustness axis, and per-cell post-hoc repair.

3 Data and Models

Datasets. Global-MMLU (CohereLabs; Apache-2.0): 9 African (am, ha, ig, mg, ny, sn, so, sw, yo) and 15 European (cs, de, el, en, es, fr, it, nl, pl, pt, ro, ru, sr, sv, uk) test splits. afrimmlu (Masakhane, from IrokoBench): 18 configurations covering 16 African languages plus `eng` and `fra` reference configs (amh, ewe, fra, hau, ibo, kin, lin, lug, orm, sna, sot, swa, twi, wol, xho, yor, zul, eng). All labels are dataset-native; **no human annotation was performed**; MCQ text contains no PII (public benchmark items; see dataset cards). Following the datasets' provenance, all 27 non-European cells, including afrimmlu's `eng`/`fra` reference configurations, form the "Africa" aggregate (25 cells when these reference configurations are excluded; §5.1 reports both). We cap 500 items/language cell (1,000 for the initial Llama/Amharic pilot cells that scoped the study).

Models (open-weight, instruct variants): Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct, Gemma-2-9B-it, Mistral-7B-Instruct-v0.3. Larger open models (32B/27B) were excluded to respect the $50 budget (§6).

Perturbations (per item, same question): `baseline`; `format_A` ("Answer with a single letter"); `format_B` (terse "Q:" layout); `translate_prompt_en` (English instruction wrapper; question remains in-language, no MT); `fewshot_1` (one fixed exemplar from the dev/validation split, never the evaluated item, no answer leakage).

Inference. Modal serverless A100-80GB, vLLM 0.6.3.post1, greedy decoding (temperature 0), max 8 tokens, top-5 logprobs. Seed 42 fixes all resampling and split procedures (bootstrap, repair splits); decoding itself is deterministic. Total: 843 logged GPU runs (840 study cells plus 3 pilot runs), **$39.48** of a $50 hard cap enforced in code. Spend rows from the 2026-07-28 pilot are logged at a $1.50/hr estimate (∼$0.32 of the total); all later rows use the $3.73/hr list price that also underlies §5.5.

Table 1: Region gaps at baseline (language-cell bootstrap, 95% CI). Pooled = all 5 perturbations. Bold = CI excludes zero; unbolded intervals include 0.

Model	Acc. Africa	Acc. Europe	Acc. gap [95% CI]	ECE Africa	ECE Europe	ECE gap (baseline) [95% CI]	ECE gap (pooled) [95% CI]
Llama-3.1-8B	0.243	0.478	−0.235 [−0.260, −0.206]	0.197	0.178	+0.018 [−0.025, 0.053] CI includes 0	+0.041 [0.018, 0.062]
Qwen2.5-7B	0.282	0.559	−0.278 [−0.309, −0.244]	0.453	0.314	**+0.140 [0.099, 0.177]**	+0.154 [0.133, 0.174]
Gemma-2-9B	0.307	0.620	−0.313 [−0.345, −0.279]	0.398	0.265	**+0.133 [0.109, 0.159]**	+0.107 [0.080, 0.139]
Mistral-7B	0.243	0.404	−0.161 [−0.187, −0.139]	0.207	0.293	**−0.087 [−0.113, −0.059]**	−0.017 [−0.037, 0.002] CI includes 0

4 METHOD

Confidence. For MCQ we follow the standard restricted-softmax protocol: confidence is the softmax over the A/B/C/D letter tokens at the first generated position; the prediction is the argmax letter. Because inference returns only the top-5 first-token logprobs, the softmax is computed over whichever of the four answer letters appear in that top-5 set. When no letter appears in the top-5 logprobs we record confidence 0.25 (chance) and flag the item (`conf_fallback`); fallback rates are reported per cell (§5.4). This top-5 restriction means confidence scales are condition-dependent: when fewer than 4 letters appear, the softmax is computed over a subset, which can inflate confidence relative to a full 4-way softmax. Within-model Africa-vs-Europe comparisons are relatively robust because the restriction affects both regions similarly, but cross-model confidence comparisons are biased by differential fallback rates (Gemma 26% Africa fallback vs Llama $<1\%$).

Metrics. Global ECE and MCE with 10 equal-width bins; class-conditional ECE for the correct-answer class (cwECE; not standard per-class averaged cwECE) computed on correct predictions only for the target class (class 1, answer B), measuring underconfidence on right answers, complementary to global ECE which captures wrong-but-confident failures. Full per-class cwECE averaged over all four answer classes requires per-class probability vectors not stored in the parquet files. Brier score. Point estimates are computed on the full sample; confidence intervals are bootstrap 95% CIs (1,000 resamples, seed 42; per-comparison coverage, we report many intervals and do not adjust for multiplicity, treating them as estimation rather than testing). Metric implementations are verified by unit tests in `test_metrics.py`.

Statistics. Region gaps are computed with language *cells* as the sampling unit: we resample cells within each region (1,000×) to CI the Africa−Europe difference. Note eight languages appear in both datasets (e.g., GMMLU `am` and afrimmlu `amh`) and thus contribute two cells; we treat cells, not unique languages, as units and report per-language tables for transparency. Perturbation effects are *paired by cell* (same languages under both prompts), avoiding cross-language confounds.

Repair. Regularized logistic regression (sklearn default C=1.0, L2), true unregularized Platt scaling, isotonic regression, and temperature scaling (NLL fit over logit(conf)/T) are fit on a random half of each cell (seed 42) and evaluated on the held-out half, no in-sample calibration claims. Logistic regression takes scalar confidence (0–1) as its sole input feature. We report means over cells from this single split; split-to-split variance is not modeled.

5 RESULTS

5.1 THE ACCURACY GAP IS MODEL-UNIVERSAL; THE CALIBRATION GAP IS MODEL-SPECIFIC

Every model loses 16–31 accuracy points from European to African cells (per-cell values in `africa_europe_gap.csv`). Calibration tells a subtler, scope-dependent story. At baseline, Qwen and Gemma are *additionally* miscalibrated in Africa, more overconfident exactly where they are least accurate. Llama's baseline gap is null but the pooled estimate excludes zero (+0.041), meaning its African calibration degrades under prompt variation even when baseline looks flat; it is overconfident everywhere (mean confidence 0.445 at 26.1% accuracy on Amharic; Figure 1). Mistral's baseline gap **reverses** (better calibrated in Africa) but attenuates to null when pooled. A single-model or single-prompt audit would have produced any of four different conclusions. Reliability diagrams for the remaining models appear in Appendix Figures 4–6.

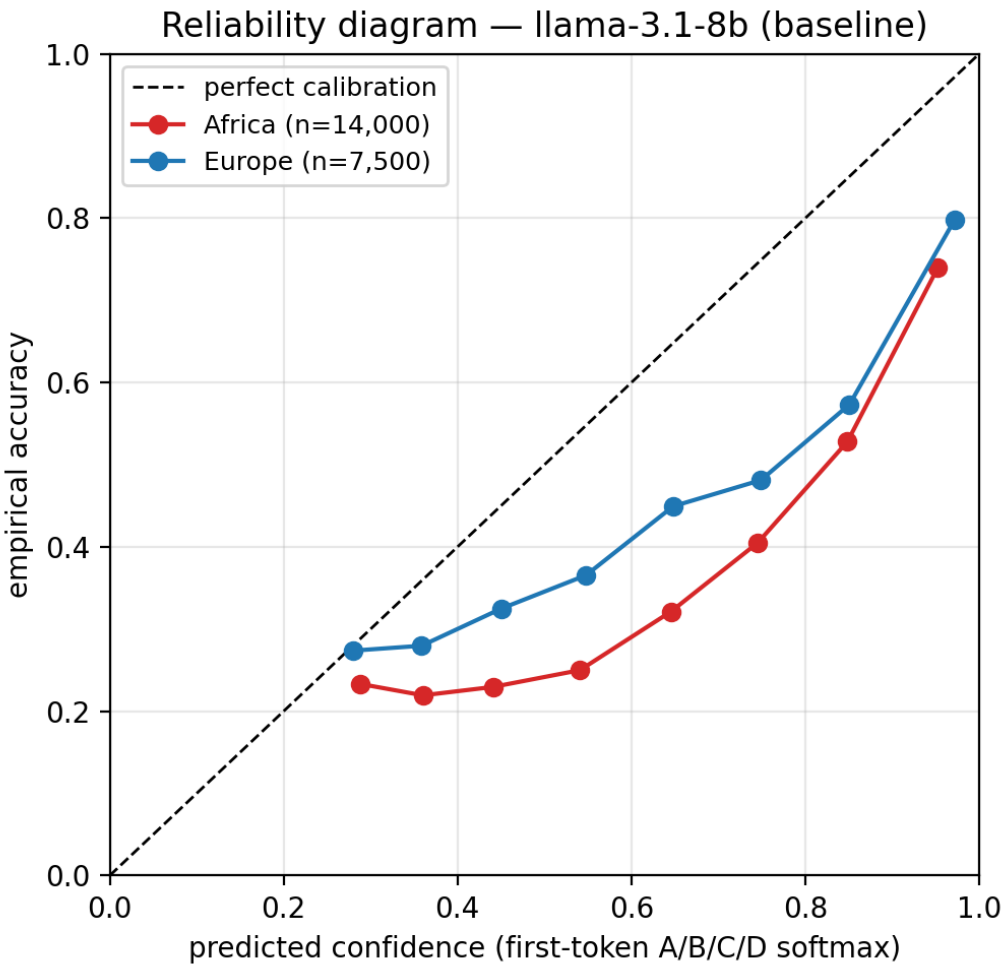

Figure 1: Calibration reliability diagram for Llama-3.1-8B at baseline. Both regions are shown; n refers to total items in each region.

Heterogeneity within regions. Per-language baseline ECE ranges 0.207–0.505 for African languages and 0.120–0.395 for European languages (model-dependent). Variance is larger in Africa than Europe for Qwen (std 0.074 vs 0.057), Gemma (0.062 vs 0.016), and Mistral (0.057 vs 0.029); Llama shows the opposite pattern (0.041 vs 0.074) but with a near-zero mean gap (+0.018), confirming that the aggregate gap is not driven by a small subset of outlier languages.

Sensitivity to the eng/fra configs and dataset source. Excluding afrimmlu's `eng`/`fra` cells from the Africa aggregate (25 cells) leaves every conclusion intact: Qwen +0.141 [0.101, 0.184], Gemma +0.125 [0.105, 0.148], Mistral −0.097 [−0.118, −0.074], Llama +0.022 [−0.022, 0.058] CI includes 0. Restricting both regions to Global-MMLU only narrows the baseline accuracy gap from 0.272 to 0.198 and the cwECE gap from 0.283 to 0.182, but the model-dependent ordering persists (Qwen/Gemma positive, Llama near-zero, Mistral reversed), confirming that the aggregate gap is not driven solely by dataset source or the inclusion of high-resource reference configurations.

Complementary metrics (baseline, means over cells). cwECE[1] and Brier corroborate the regional pattern: Gemma cwECE 0.276 (Africa) vs 0.073 (Europe), Brier 0.402 vs 0.275; Qwen cwECE

[1] Class-conditional ECE computed on correct predictions only, measuring underconfidence on right answers. Full per-class cwECE averaged over all four answer classes requires per-class probability vectors not stored in the parquet files.

Table 2: Paired ΔECE vs baseline, African cells (27 pairs). Bold = 95% CI excludes zero; unbolded intervals include 0. †Gemma format_A has 100% fallback rate in afrimmlu cells; ECE is degenerate (all confidences forced to chance) and the negative delta is an artifact, not a calibration improvement.

Perturbation	Llama	Qwen	Gemma	Mistral
`format_A`	**+0.103**	**−0.054**	N/A†	**+0.207**
`format_B`	**−0.084**	+0.013 [−0.004, 0.032] CI includes 0	**−0.047**	**+0.020**
`translate_prompt_en`	**−0.041**	−0.028 [−0.069, 0.008] CI includes 0	**+0.060**	**+0.086**
`fewshot_1`	**−0.132**	**+0.105**	**−0.126**	**+0.029**

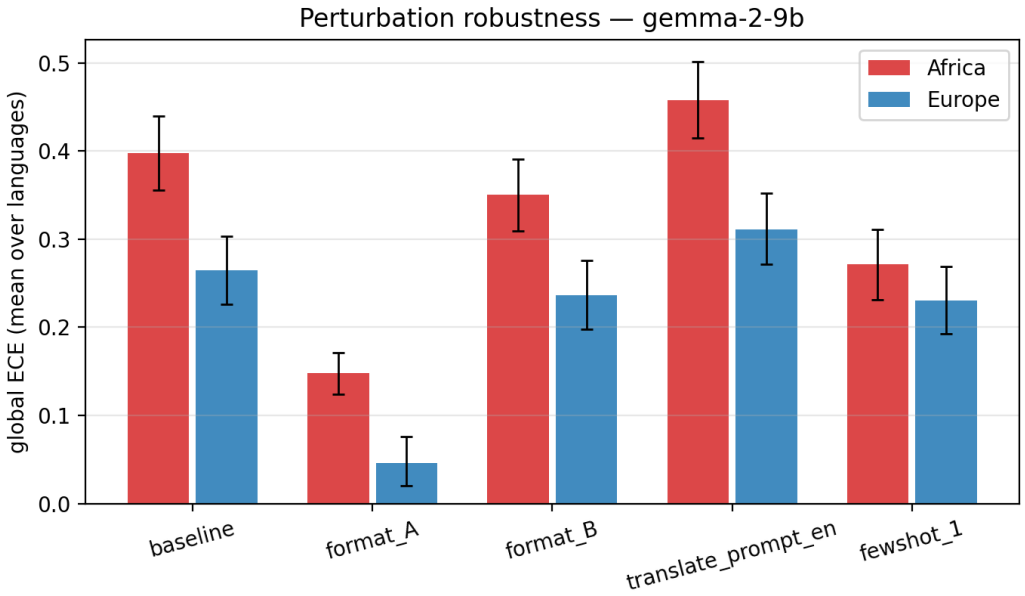

Figure 2: Perturbation robustness for Gemma-2-9B: global ECE (mean over languages) per prompt perturbation, shown separately for Africa and Europe. Error bars show bootstrap 95% CIs (1,000 resamples, seed 42).

0.247 vs 0.061, Brier 0.447 vs 0.316; Llama cwECE 0.547 vs 0.265; Mistral cwECE 0.552 vs 0.235. (MCE, cwECE, and Brier per cell with CIs: `metrics_table.csv`.)

5.2 Perturbation effects do not transfer

No perturbation helps or harms all models (per-model robustness curves: Figure 2 and Appendix Figures 7–9). `format_A` makes Gemma ECE undefined in African cells (100% fallback rate; all items assigned chance confidence because no A–D letter appears in top-5 logprobs) and simultaneously destroys accuracy (Δacc −0.199); it also worsens Llama calibration (+0.103 [0.078, 0.129]) and Mistral (+0.207). One-shot prompting calibrates Llama/Gemma but *worsens* Qwen (+0.105 [0.066, 0.142]): in-language exemplars can inflate misplaced confidence. Prompt-engineering advice that ignores model × language interaction is unreliable.

5.3 Miscalibration is repairable in-distribution, less so in Europe

Repair. Regularized logistic regression wins for all four models (median ECE reduction 58–86%). A 3-seed stability check on the same 210 cells per model confirms the winner is consistent: regularized logistic regression wins on at least two of three random splits for Llama (61.4%, 129/210), Qwen (80.5%, 169/210), Gemma (60.0%, 126/210), and Mistral (83.8%, 176/210), and the cross-

Table 3: Held-out ECE after post-hoc repair (fit on 250 items/cell; raw = mean over all cells/perturbations). "Reg. Logistic" = regularized logistic regression (sklearn default C=1.0, L2). Stability: regularized logistic wins on at least two of three random splits in 71.4% of cells (600/840; Table 7). †True unregularized Platt scaling (scipy.optimize.minimize, no penalty) evaluated on a 5-cell pilot (4 models, Amharic baseline and format_A): unregularized ECE 0.448-0.719 vs regularized 0.007-0.057. Full 840-cell evaluation not performed; data in outputs/platt_comparison.csv.

Model	Raw	Reg. Logistic	Isotonic	Temperature
Llama-3.1-8B	0.151	**0.064**	0.072	0.127
Qwen2.5-7B	0.405	**0.062**	0.073	0.234
Gemma-2-9B	0.287	**0.063**	0.063	0.205
Mistral-7B	0.281	**0.052**	0.065	0.225

seed standard deviation of mean repaired ECE is ≤ 0.007 for all models (Table 7). True unregularized Platt scaling[2] gives materially different results and is not recommended for small calibration sets. Split by region, repaired ECE is consistently *lower* in Africa (0.038–0.048) than in Europe (0.076–0.100): the same $\sim$250-item calibration set produces a larger reduction in ECE in the African cells. Two caveats: repair is validated on a same-distribution held-out split only, and a calibrated-but-25%-accurate model is not thereby deployment-fit (§6); calibration repair addresses trust, not capability. A single global temperature transformation may not capture calibration differences that vary across models, languages, or regions. Possible contributors include differences in tokenization, training data, and alignment, but these mechanisms are not established by this study.

5.4 Instruction-following failure is a hidden robustness axis

Gemma fails to produce any A–D letter in its top-5 first-token logprobs for **26.1% of African** vs 18.1% of European prompts (Qwen: 7.0%/4.1%; Llama and Mistral $\leq$0.31%), reaching 100% in some afrimmlu cells. Accuracy-only scoring silently converts these to wrong answers; for calibration auditing they are a distinct failure class, which we report per cell via `fallback_rate`. When fallback rate approaches 100%, all items are assigned chance confidence (0.25), which artificially deflates ECE, a model that emits no answer letters cannot meaningfully be assessed for calibration. We therefore exclude fallback items from ECE computations where fallback rate exceeds 50%, and report those cells as undefined.

5.5 Efficiency

Measured A100 wall-time cost per 1k items: Llama \$0.084, Qwen \$0.091, Mistral \$0.092, Gemma \$0.106 (Figure 3). Qwen delivers the best accuracy-per-dollar (0.393 mean acc at \$0.091/1k). The full audit, 422k items, 4 models, 5 perturbations, cost **\$39.48**, within reach of any research group.

6 Limitations

We lead with scope. (1) MCQ calibration uses a 4-way restricted softmax; it does not characterize open-ended generation. (2) `translate_prompt_en` wraps instructions in English; questions stay in-language, we do not evaluate machine translation. (3) Models are 7–9B; 32B/27B were dropped for budget, so no scale claims. (4) n=500/cell bounds per-language CI tightness. (5) Cost on US/EU-region GPUs proxies for efficiency; we do not meter energy on African infrastructure. (6) The perturbation set is a probe, not exhaustive. (7) Repair is validated in-distribution only, from a single random split per cell; a 3-seed stability check confirms the winner method (regularized logistic) is consistent across splits (Table 7), but repair claims apply to same-distribution held-out data only. (8) Region aggregates are coarse: "Africa" includes afrimmlu's eng/fra reference configs (sensitivity in §5.1), and eight languages contribute cells to both datasets. (9) We characterize; we

[2]True unregularized Platt scaling (scipy.optimize.minimize, no penalty) evaluated on a 5-cell pilot (4 models, Amharic baseline and format_A): unregularized ECE 0.448-0.719 vs regularized 0.007-0.057. Full 840-cell evaluation not performed; data in outputs/platt_comparison.csv.

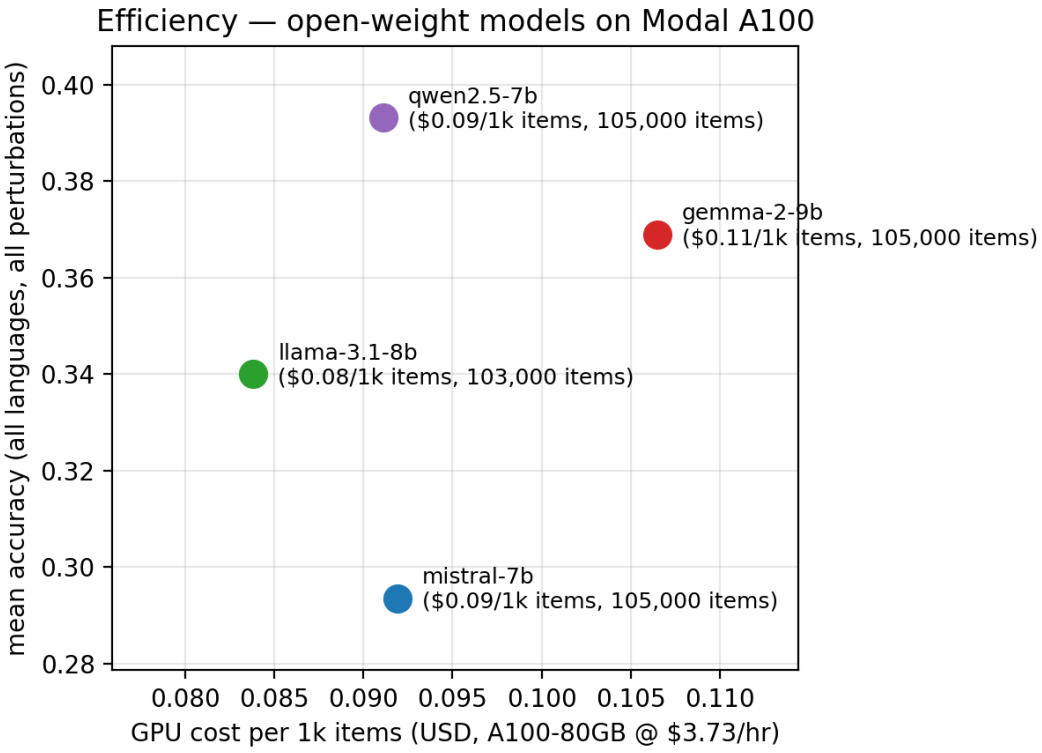

Figure 3: Efficiency cost curve: measured A100 wall-time cost per 1k items (`efficiency_cost_curve.png`).

do not prescribe deployment. (10) Confidence is computed over the subset of letters present in the top-5 first-token logprobs; this makes confidence scales condition-dependent and cross-model ECE comparisons are valid only within fallback-rate strata. (11) Subject/category metadata is not included in the parquet files; original Global-MMLU and afrimmlu datasets contain subject columns but we did not re-join them for this analysis.

7 Responsible NLP

All data are public benchmarks with dataset-native labels (Global-MMLU: Apache-2.0; afrimmlu/IrokoBench: see dataset card); no new annotation, and MCQ items are public exam-style text without PII. The study's motivation is equitable reliability evidence for African-language NLP: who benefits is practitioners choosing open models for African languages, and the released per-language toolkit lets them verify claims rather than trust aggregates. Risks: (i) calibration scores could be over-read as deployment approval; we explicitly characterize only, and repaired models remain 24–31% accurate in African cells; (ii) aggregate Africa/Europe deficit framing can reinforce reductive narratives about African languages; we therefore release per-language tables, report the reversed Mistral gap and the eng/fra sensitivity, and frame gaps as model-specific behaviors, not language deficiencies.

8 Conclusion

Across the four evaluated open-weight LLMs, calibration on African-language MCQs is model-dependent rather than uniformly reliable, and the failure pattern is model-specific: accuracy gaps are universal, but calibration gaps range from absent at baseline (Llama) to severe (Qwen, Gemma) to reversed (Mistral). Prompt perturbations, including one-shot examples and answer-format in-

structions, shift calibration in directions that transfer across neither models nor regions, and can silence a model's answers entirely in low-resource languages. Miscalibration is nonetheless repairable in-distribution with $\sim$250 labeled items per language, and the repair winner (regularized logistic regression) is stable across random splits. Multilingual trustworthiness claims must therefore be made per model, per language, per prompt, and they are cheap to verify: our entire audit cost less than forty dollars. A single-model audit produces any of four qualitatively different conclusions, so multilingual calibration benchmarks must report per-model profiles rather than aggregate "multilingual" numbers. Possible contributors to the observed model-specificity include differences in tokenization, training data, and alignment, but these mechanisms are not established by this study. A single global temperature transformation may not capture calibration differences that vary across models, languages, or regions.

Reproducibility Statement

Reproducibility: Kaggle dataset (link anonymized for submission; parsed outputs, raw gzipped JSONs, code, final tables); executable notebook recomputes all tables and asserts equality with the shipped results. The Kaggle link will be revealed in the camera-ready version. Modal app + pinned deps in `src/`. Spend log: `outputs/modal_spend.json.gz`.

References

Adelani, D. I., et al. 2024. IrokoBench: A New Benchmark for African Languages in the Age of Large Language Models. arXiv:2406.03368.

Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. ICML.

Hendrycks, D., et al. 2021. Measuring Massive Multitask Understanding. ICLR.

Yang, Y.; Dan, S.; Roth, D.; Lee, I. 2023. On the Calibration of Multilingual Question Answering LLMs. arXiv:2311.08669.

Naeini, M. P.; Cooper, G. F.; Hauskrecht, M. 2015. Obtaining Well Calibrated Probabilities Using Bayesian Binning. AAAI.

He, J.; Yu, S.; Xiong, D.; Gutiérrez-Basulto, V.; Pan, J. Z. 2025. N2C2: Nearest Neighbor Enhanced Confidence Calibration for Cross-Lingual In-Context Learning. arXiv:2503.09218.

Platt, J. 1999. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *Advances in Large Margin Classifiers*.

Singh, S., et al. 2025. Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation. ACL 2025 (arXiv:2412.03304).

Tonja, A. L., et al. 2024. InkubaLM: A Small Language Model for Low-Resource African Languages. arXiv:2408.17024.

Zadrozny, B.; Elkan, C. 2002. Transforming Classifier Scores into Accurate Multiclass Probability Estimates. KDD.

AI use statement

In this work, we used generative AI tools for: (required) implement methods — specifically the experiment, parsing, and analysis pipelines; (required) design or provide feedback on research methodology or experiments — the multi-agent review and proofread pass identified scope mismatches, corrected table-to-CSV alignment, and fixed factual errors in the draft; (recommended) create or edit software code; (recommended) draft parts of a research paper and edit a research paper to improve readability — the abstract, introduction, and related work sections; (recommended) create or modify scientific figures or images — reliability diagrams and perturbation-robustness curves were

generated from analysis code; (recommended) creation of artifacts — the executable Kaggle reproduction notebook and dataset packaging; (recommended) summarize or analyse existing literature; (recommended) suggest experimental parameters; (recommended) sourcing/searching for information; (recommended) identify relevant literature; (recommended) format references; (recommended) suggest a structure for a research paper. We have not used generative AI tools for: generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, assist with translation, clean and reformat datasets, support qualitative and thematic data analysis, interpret results. The remaining required tasks are not applicable to this work. We have reviewed all AI-assisted work. LLM-generated code was verified with unit tests; AI-drafted text and figures were verified against source CSVs and experimental outputs by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

A Additional Tables

This appendix collects full perturbation metrics not shown in the main text.

Model	Perturbation	Global ECE↓	Acc	ΔAcc	Fallback	ΔFB	cwECE↓	ΔcwECE	Brier↓	ΔBrier	MCE↓	ΔMCE
gemma-2-9b	baseline	0.3977	0.307	0.000	0.046	0.000	0.276	0.000	0.402	0.000	0.693	0.000
	fewshot_1	0.2715	0.316	+0.008	0.000	-0.046	0.391	+0.115	0.294	-0.108	0.677	-0.016
	format_A	0.1481	0.109	-0.199	0.991	+0.945	0.720	+0.444	0.113	-0.290	0.156	-0.538
	format_B	0.3503	0.294	-0.013	0.008	-0.038	0.331	+0.056	0.346	-0.056	0.684	-0.010
	translate_prompt_en	0.4580	0.290	-0.018	0.259	+0.213	0.196	-0.079	0.475	+0.073	0.746	+0.052
llama-3.1-8b	baseline	0.1965	0.243	0.000	0.000	0.000	0.547	0.000	0.222	0.000	0.457	0.000
	fewshot_1	0.0648	0.285	+0.043	0.000	0.000	0.648	+0.101	0.202	-0.020	0.322	-0.135
	format_A	0.2998	0.265	+0.022	0.005	+0.005	0.430	-0.117	0.303	+0.081	0.705	+0.248
	format_B	0.1121	0.256	+0.013	0.000	0.000	0.625	+0.079	0.197	-0.024	0.356	-0.101
	translate_prompt_en	0.1551	0.255	+0.012	0.000	0.000	0.579	+0.032	0.210	-0.012	0.411	-0.047
mistral-7b	baseline	0.2066	0.243	0.000	0.000	0.000	0.553	0.000	0.246	0.000	0.707	0.000
	fewshot_1	0.2356	0.223	-0.021	0.000	0.000	0.535	-0.018	0.241	-0.005	0.702	-0.004
	format_A	0.4136	0.224	-0.019	0.014	+0.014	0.365	-0.189	0.393	+0.147	0.784	+0.077
	format_B	0.2264	0.259	+0.016	0.000	0.000	0.521	-0.032	0.265	+0.019	0.728	+0.022
	translate_prompt_en	0.2921	0.246	+0.003	0.001	+0.001	0.463	-0.091	0.300	+0.054	0.738	+0.031
qwen2.5-7b	baseline	0.4531	0.282	0.000	0.082	0.000	0.247	0.000	0.447	0.000	0.670	0.000
	fewshot_1	0.5577	0.295	+0.013	0.000	-0.082	0.131	-0.115	0.543	+0.096	0.658	-0.012
	format_A	0.3990	0.286	+0.005	0.000	-0.082	0.309	+0.062	0.387	-0.060	0.661	-0.009
	format_B	0.4665	0.299	+0.017	0.006	-0.075	0.223	-0.024	0.450	+0.004	0.645	-0.024
	translate_prompt_en	0.4248	0.274	-0.008	0.265	+0.183	0.273	+0.026	0.446	-0.001	0.690	+0.020

Table 4: Full perturbation metrics over African language cells (means and deltas vs baseline). †Gemma format_A: 100% fallback; cwECE is degenerate (all confidences forced to chance).

Model	Aggregate	n cells	Acc	ECE gap	cwECE gap	Brier gap
gemma-2-9b	Africa (all 27)	27	0.307	+0.133 [0.109, 0.159]	+0.203	+0.127
	Africa (excl. eng/fra)	25	0.303	+0.125 [0.105, 0.148]	+0.198	+0.122
	Europe (all 15)	15	0.620	—	—	—
qwen2.5-7b	Africa (all 27)	27	0.282	+0.140 [0.099, 0.177]	+0.186	+0.131
	Africa (excl. eng/fra)	25	0.279	+0.141 [0.101, 0.184]	+0.189	+0.133
	Europe (all 15)	15	0.559	—	—	—
llama-3.1-8b	Africa (all 27)	27	0.243	+0.018 [−0.025, 0.053] CI includes 0	+0.282	+0.043
	Africa (excl. eng/fra)	25	0.241	+0.022 [−0.022, 0.058] CI includes 0	+0.287	+0.045
	Europe (all 15)	15	0.478	—	—	—
mistral-7b	Africa (all 27)	27	0.243	−0.087 [−0.113, −0.059]	+0.317	+0.011
	Africa (excl. eng/fra)	25	0.241	−0.097 [−0.118, −0.074]	+0.322	+0.013
	Europe (all 15)	15	0.404	—	—	—

Table 5: Sensitivity of region gaps to exclusion of afrimmlu `eng`/`fra` reference configs. ECE gaps are baseline-only; cwECE and Brier gaps are pooled means over all perturbations. CIs for the 25-cell aggregate are recomputed from the reduced cell set.

B Additional Figures

This appendix collects the per-model reliability diagrams and perturbation-robustness curves not shown in the main text.

Model	Original gap	After >50% fallback exclusion	Change	Cells excluded
gemma-2-9b	+0.1074	+0.1097	+0.0023	42 (format_A)
qwen2.5-7b	+0.1542	+0.1574	+0.0032	0
llama-3.1-8b	+0.0405	+0.0405	0.0000	0
mistral-7b	−0.0172	−0.0172	0.0000	0

Table 6: Fallback-stratum sensitivity: Africa−Europe ECE gap after excluding the 42 cells with fallback rate >50% (all Gemma format_A). The main qualitative conclusion (model-dependent gap, Qwen/Gemma positive, Llama near-zero, Mistral reversed) is unchanged.

Model	Method	Seed 42 mean ECE	Seed 123 mean ECE	Seed 456 mean ECE	Cross-seed std	Reg. Logistic win %
llama-3.1-8b	Reg. Logistic	0.0705	0.0761	0.0764	0.0033	61.4% (129/210)
	Isotonic	0.1066	0.1106	0.1131	0.0033	—
	Temperature	0.1183	0.1221	0.1163	0.0030	—
qwen2.5-7b	Reg. Logistic	0.0692	0.0572	0.0577	0.0068	80.5% (169/210)
	Isotonic	0.1164	0.1115	0.1107	0.0030	—
	Temperature	0.2316	0.2162	0.2160	0.0090	—
gemma-2-9b	Reg. Logistic	0.0846	0.0760	0.0792	0.0044	60.0% (126/210)
	Isotonic	0.1177	0.1182	0.1180	0.0003	—
	Temperature	0.2019	0.1891	0.1917	0.0066	—
mistral-7b	Reg. Logistic	0.0543	0.0633	0.0570	0.0046	83.8% (176/210)
	Isotonic	0.0866	0.0878	0.0826	0.0027	—
	Temperature	0.2234	0.2227	0.2153	0.0047	—

Table 7: Repair stability across 3 random held-out splits (seeds 42, 123, 456). Mean repaired ECE and cross-seed standard deviation reported per method. "Reg. Logistic win %" = fraction of cells where regularized logistic regression achieves the lowest mean ECE on at least two of three seeds.

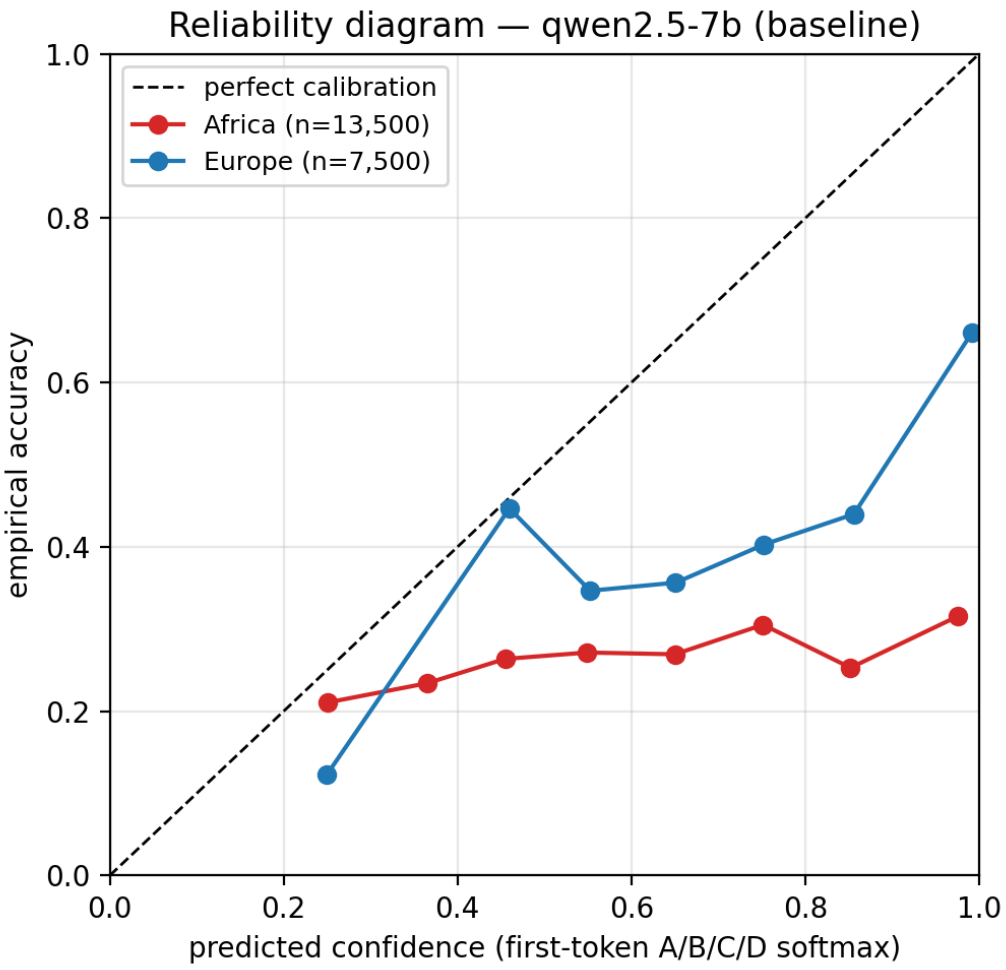

Figure 4: Calibration reliability diagram for Qwen2.5-7B at baseline. Both regions are shown; n refers to total items in each region.

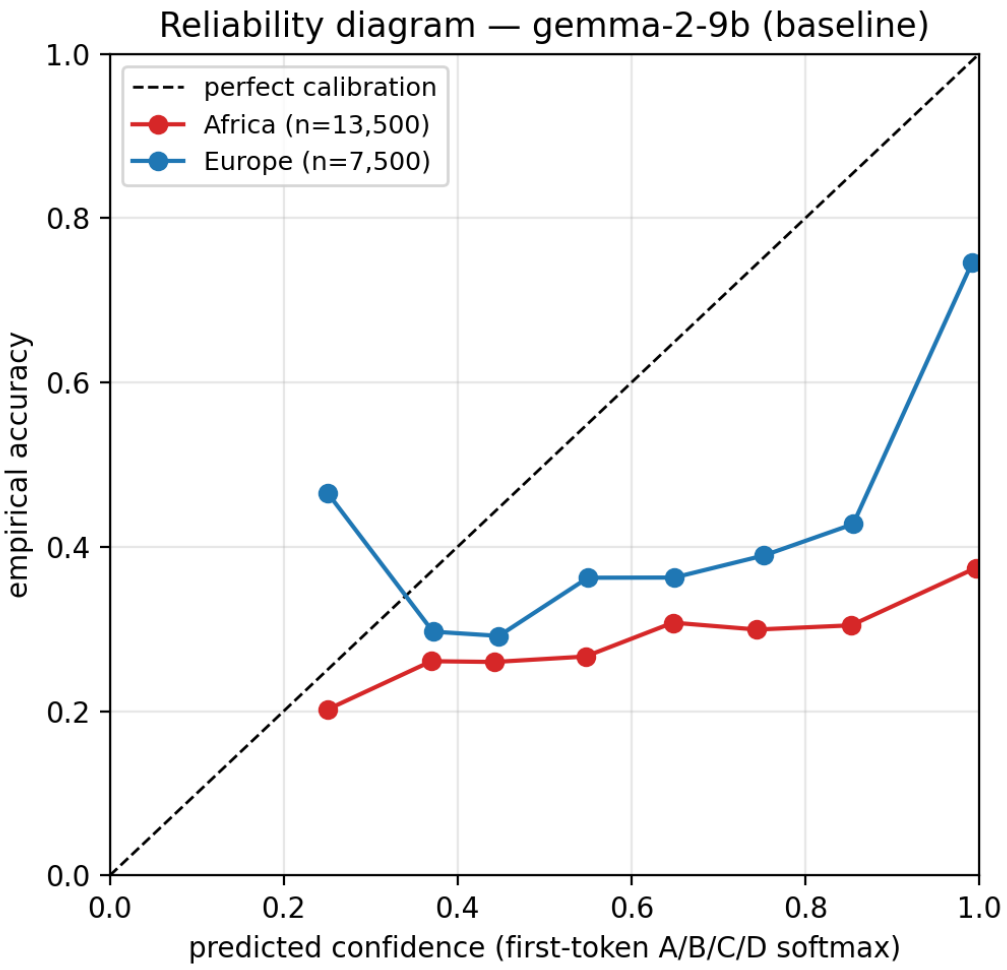

Figure 5: Calibration reliability diagram for Gemma-2-9B at baseline. Both regions are shown; n refers to total items in each region.

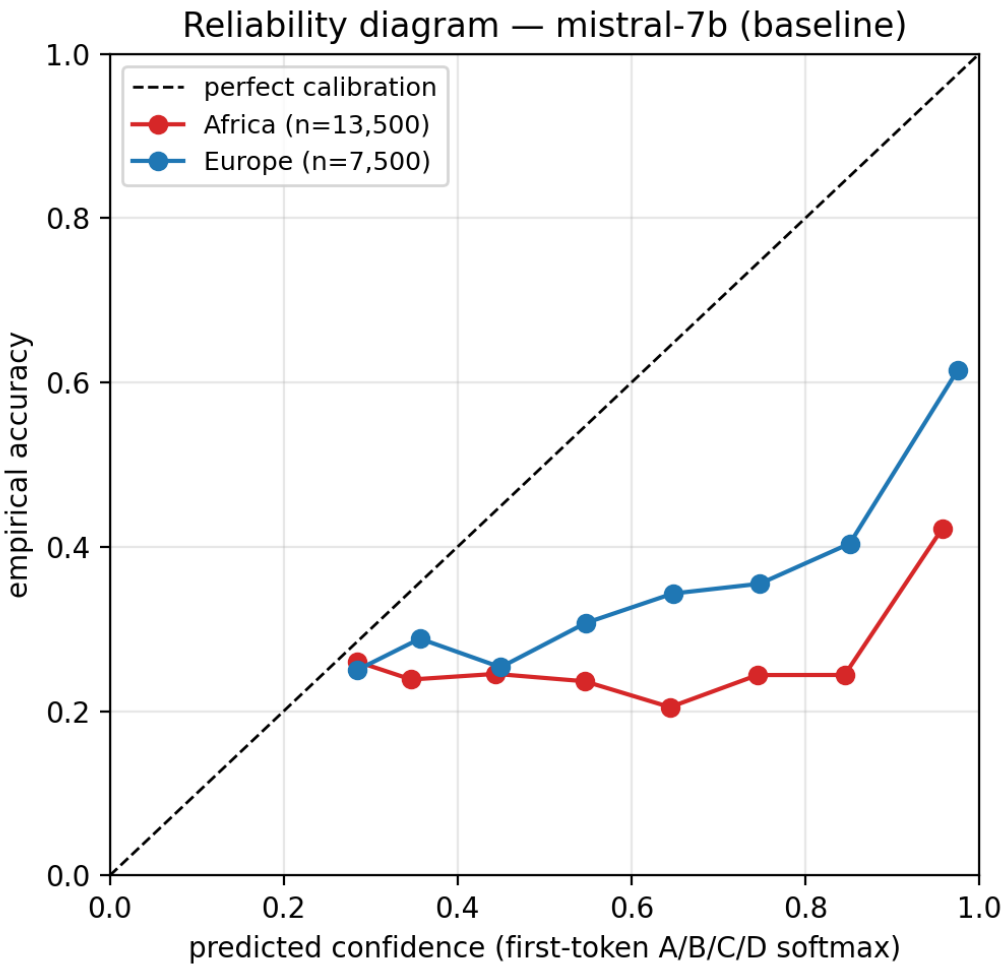

Figure 6: Calibration reliability diagram for Mistral-7B at baseline. Both regions are shown; n refers to total items in each region.

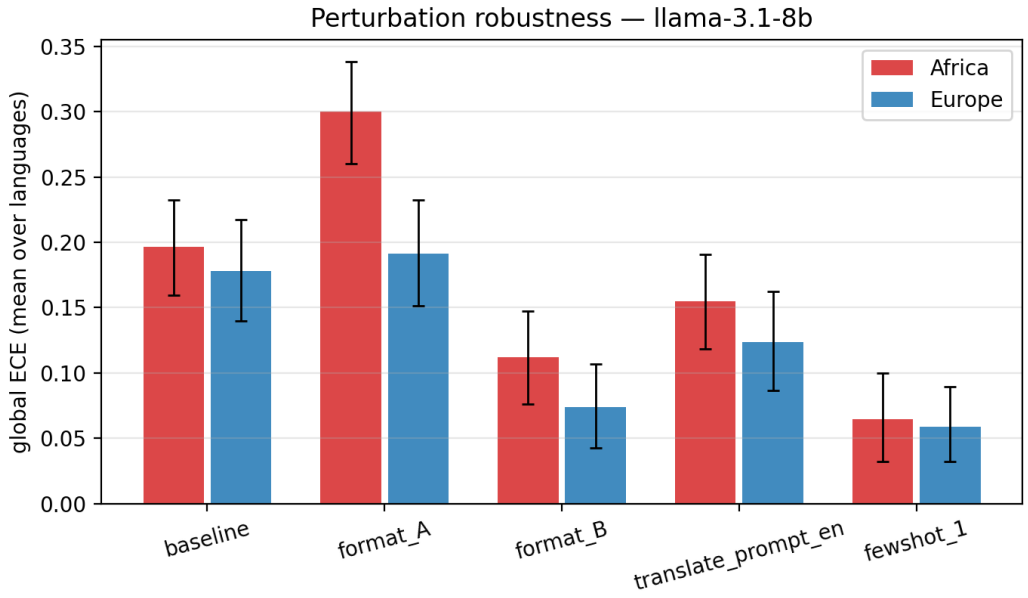

Figure 7: Perturbation robustness for Llama-3.1-8B: global ECE (mean over languages) per prompt perturbation, shown separately for Africa and Europe. Error bars show bootstrap 95% CIs (1,000 resamples, seed 42).

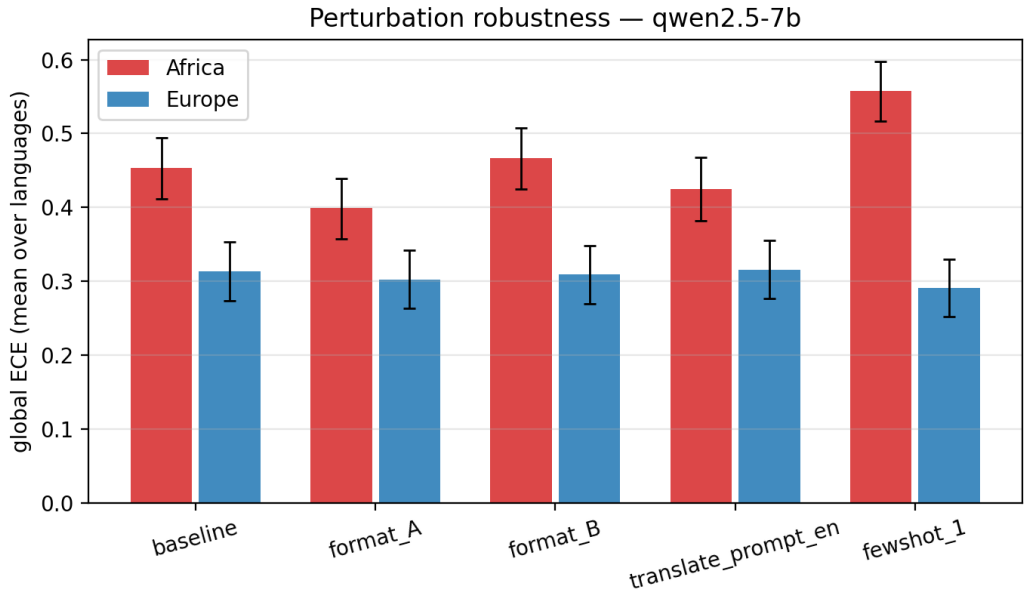

Figure 8: Perturbation robustness for Qwen2.5-7B: global ECE (mean over languages) per prompt perturbation, shown separately for Africa and Europe. Error bars show bootstrap 95% CIs (1,000 resamples, seed 42).

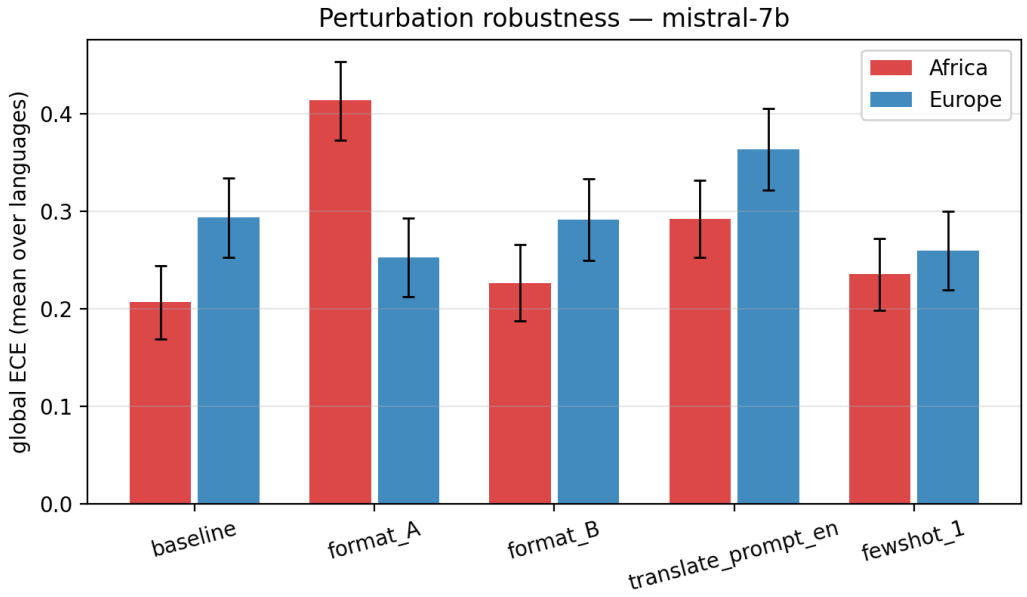

Figure 9: Perturbation robustness for Mistral-7B: global ECE (mean over languages) per prompt perturbation, shown separately for Africa and Europe. Error bars show bootstrap 95% CIs (1,000 resamples, seed 42).