Stout in Languages, Linguistics and Literature (SLLL)
ISSN: 2979-2800
DOI: https://doi.org/10.61424/4jxy8974
Journal Homepage: https://stoutjournals.org/index.php/SLLL

Translation Equivalence Is Not Linguistic Equity: English–Arabic–Chinese Calibration and Cultural-Item Performance Gaps in a Compact Multilingual Language Model

Julian Li
Computer Science, Duke University, Durham, NC, USA
Corresponding Author: Julian Li, **E-mail**: julian.li_duke@gmail.com

ARTICLE INFORMATION

Submitted: April 10, 2026
Accepted: July 25, 2026
Published: August 01, 2026
Volume: 2
Issue: 1

KEYWORDS

Multilingual language models; calibration; cultural sensitivity; English; Arabic; Chinese; retrieval-augmented in-context learning; selective prediction

ABSTRACT

Parallel translation simplifies multilingual evaluation, but it does not guarantee equivalent model behavior. This study evaluated a quantized 135-million-parameter multilingual language model on the English, Arabic, and Chinese test slices of Global-MMLU-Lite. Six complete native and retrieval-augmented runs produced 2,400 item-level predictions. Cross-language inference used 390 genuinely aligned items, including equal numbers of culturally sensitive and culturally agnostic questions; a 387-item strict cohort excluded three questions with duplicated answer options. Native accuracy was 26.7% in English, 26.9% in Arabic, and 25.6% in Chinese, yet native-condition prediction agreement ranged from 3.3% to 52.8%, with Cohen's κ between −.058 and .015. Same-language retrieval changed accuracy by −1.0, −0.5, and +2.8 percentage points, respectively; none of the paired effects survived Holm correction. Culturally sensitive items were directionally harder in every native-language condition, although all bootstrap intervals included zero. Calibration differed more sharply than accuracy: native expected calibration error was .102 in English, .069 in Arabic, and .308 in Chinese. Cross-fitted temperature scaling reduced these values mainly by flattening probabilities toward the four-choice uniform distribution, while selective risk remained high. Native Arabic and Chinese prompts required 3.04 and 2.12 times as many tokens as English prompts. The results show that score parity can coexist with different decisions, confidence failures, and token burdens. Translation equivalence should therefore be evaluated jointly through accuracy, item-level agreement, calibration, selective risk, and token burden.

1. Introduction

Multiple-choice benchmarks are widely used to evaluate language-model knowledge. MMLU established an examination format spanning academic and professional subjects (Hendrycks et al., 2021), while multilingual extensions either translate common items or collect questions within particular educational settings. ArabicMMLU and CMMLU follow the latter approach (Koto et al., 2024; H. Li et al., 2024); Global MMLU combines translated content with cultural annotation (Singh et al., 2025). These designs answer different questions about language, knowledge, and culture.

Translation is an imperfect control. It can preserve a designated answer while changing lexical frequency, syntax, option plausibility, or access to learned knowledge. A culturally sensitive source item may also remain culturally specific in every language; its Arabic or Chinese rendering is not thereby a test of Arabic or Chinese cultural knowledge. Cultural sensitivity is consequently treated here as an item property rather than a property of the prompt language (Hershcovich et al., 2022; Naous et al., 2024).

Accuracy alone cannot reveal whether aligned versions elicit the same decisions or whether errors carry reliable confidence. Calibration separates predictive probability from correctness (Guo et al., 2017; Jiang et al., 2021), and multilingual models can show language-dependent reliability despite shared parameters (Ahuja et al., 2022; Yang et al., 2026). Expected calibration error, proper scoring rules, and selective risk therefore complement paired accuracy.

Seen this way, multilingual equity is a reliability problem rather than a score-comparison problem. Cross-domain work has paired prediction with calibrated uncertainty, distribution-free intervals, rejection policies, and risk-bounded operating rules in capacity planning, credit decisions, and medical vision–language classification (He et al., 2023; Y. Chen et al., 2023; Jin et al., 2024b; S. Chen et al., 2024; Lu & Zou, 2026). These application areas are not evidence about multilingual language models, but they establish a transferable measurement principle: acceptable top-label accuracy does not guarantee that confidence is usable for triage. Fair credit scoring likewise shows why subgroup interpretation and explanations must accompany a headline score (Xin, 2026d).

Language can also be treated as a structured domain shift: the intended content is shared, whereas script, lexical frequency, tokenization, and learned associations change. Studies of cross-cloud transfer, cross-dataset activity recognition, approximately shared features, calibrated fusion, and rank uncertainty show that common structure can coexist with environment-specific error or ordering (He et al., 2024; J. Zhang, 2025; Zhong, Pan, et al., 2025; Xin, 2025b; Zhong & Ling, 2024a, 2024b). A vision–language consistency study makes a related distinction between structural preservation and surface agreement (Y. Chen & Li, 2025). Here these works are methodological analogies, not direct evidence about translation.

This study evaluates SmolLM2-135M Multilingual Base under native zero-shot scoring, a reference-English pivot, and subject-restricted same-language retrieval-augmented in-context learning. It asks whether English, Arabic, and Chinese differ in forced-choice accuracy, agreement, calibration, cultural-item performance, and token burden; whether retrieval or the English representation changes those differences; and whether confidence can support selective answering. The compact checkpoint is a diagnostic case rather than a proxy for frontier systems. Item-level pairing ensures that translated versions remain dependent observations throughout calibration and resampling.

2. Literature Review

Global MMLU pairs translated questions with cultural-sensitivity annotation and shows that culturally sensitive and agnostic subsets can alter model rankings (Singh et al., 2025). Global-MMLU-Lite supplies a balanced compact subset (Cohere Labs, 2024), while Belebele demonstrates the broader value of parallel content for paired comparison (Bandarkar et al., 2024). Alignment controls source content but cannot eliminate translation-induced difficulty.

Native ArabicMMLU and CMMLU instead reflect their own educational settings (Koto et al., 2024; H. Li et al., 2024). BLEnD and Arabic–English cultural evaluations further show that region, framing, and prompt language can interact (Myung et al., 2024; Naous et al., 2024). OMoS-QA likewise documents the difficulty of cross-language correspondence between questions and evidence (Kleinle et al., 2024). The present design therefore pairs items by verified content and treats English as a reference representation, not as the output of a machine-translation system.

Cross-language evidence access becomes more consequential when a system remembers earlier sessions, personalizes retrieval, or supports sensitive dialogue. Confidence-gated memory controls whether content should be written or retrieved (Sun et al., 2023), federated topic-preference learning adds privacy constraints (Kuo et al., 2025), and counseling and conversational-recommendation systems couple retrieval with ranking or response generation (Y. Zhang & H. Zhang, 2025a; Xin, 2026b). A multilingual humanitarian RAG study is the closest deployment analogue because it joins evidence access, answerability, calibration, and interface-level trust across languages (Y. Chen & Xu, 2026). Strategy-controlled emotional-support dialogue and segment-level representation learning further illustrate that language behavior can vary by interaction context and subgroup (Y. Zhang & Zhou, 2026; Xin, 2026g).

Expected calibration error summarizes reliability but depends on binning; negative log-likelihood and the multiclass Brier score provide complementary proper scores (Brier, 1950; Guo et al., 2017). For forced choice, normalizing admissible label logits targets the evaluated decision, although arbitrary answer symbols can carry a prior. Contextual calibration estimates that prior from a content-free input (Zhao et al., 2021). Question-answering calibration can remain poor (Jiang et al., 2021), vary across languages (Ahuja et al., 2022), and fail to transfer from English (Yang et al., 2026).

Calibration also interacts with subgroup composition and decision policy. Uplift and privacy-robust incrementality studies foreground heterogeneous effects rather than a single population average (Mu et al., 2023; Bai, Wang, et al., 2026), while extreme-imbalance fraud and bankruptcy studies show how class structure can make aggregate accuracy misleading (Jin et al., 2024a; Y. Chen et al., 2024). Conservative off-policy selection offers a parallel for acting only when a policy is adequately supported (Ye et al., 2026). The present cultural-label contrasts do not use those methods, but they follow the same reporting discipline: disaggregate first, attach uncertainty, and avoid interpreting a small overall difference as behavioral equivalence.

Selective prediction ranks items by confidence and measures retained error as coverage changes (El-Yaniv & Wiener, 2010). Domain shift can disrupt that ranking (Kamath et al., 2020), while multilingual feedback produces language-dependent abstention behavior (Feng et al., 2024). Risk–coverage analysis therefore tests whether confidence has operational value beyond its absolute calibration.

Selective prediction must also be distinguished from safety refusal. Confidence-based withholding ranks uncertain cases; red-team guardrail updates, evidence-based self-verification, and multistage refusal instead target adversarial or harmful behavior (Zheng et al., 2023; Zheng et al., 2024; Zheng & Li, 2024). Dark-pattern explanation shows that apparently ordinary microcopy may encode a separate behavioral risk, while prompt-injection risk cards and attack-chain sequence models emphasize human oversight and traceability (Xu et al., 2025; Su, Rao, et al., 2026; Bai, Chen, et al., 2026). Numerical guardrails and generalist-agent trajectory diagnostics similarly argue for evaluating the control layer around an answer, not only the answer itself (Z. Li et al., 2026; Zhong et al., 2026). Accordingly, the risk–coverage curves below are evidence about confidence ranking, not claims of jailbreak robustness or safe conversational refusal.

Retrieval-augmented generation grounds responses in selected context (Lewis et al., 2020), including across languages (Asai et al., 2021), but multilingual RAG remains sensitive to prompt language, script, and evaluation design (Chirkova et al., 2024). Here retrieval selects solved same-language development examples rather than factual passages. Subject restriction limits irrelevant demonstrations, and language matching keeps retrieval quality distinct from translation quality.

Recent RAG work further separates retrieval success from answer faithfulness. Evidence-grounded DevOps QA and bilingual incident triage evaluate grounding in operational corpora (B. Zhang et al., 2024; Liu et al., 2024); evidence-chain systems add word-level attribution or deterministic provenance (C. Li et al., 2024; Jin, 2025a). Budgeted multi-hop retrieval makes context selection an explicit policy, while unanswerable-question RAG makes evidence sufficiency and abstention first-class outcomes (Su et al., 2025; Zhong, Chen, et al., 2025). Retrieval-quality prediction likewise treats the retriever as a component that must be evaluated rather than an assumed source of improvement (R. Zhang et al., 2025). These distinctions are central here because solved examples can change the model's answer distribution even when they do not supply the missing fact.

Similar separations appear across code intelligence, accounting review, scientific verification, legal governance, and personalized recommendation. Retrieval–summary integration keeps finding and explaining distinct (Y. Li, 2024); accounting-aware retrieval, prototype-based log explanation, and financial-search reranking expose different retrieval and ranking stages (Y. Chen et al., 2025; Xin, 2025a; Meng et al., 2026). Scientific claim verification, review-grounded recommendation, and multi-regulation legal RAG require a visible chain from source to conclusion (Su, Chen, et al., 2026; Chang et al., 2026; J. Li & Zhou, 2026). Retrieval-grounded HDFS narratives and cost-aware AIOps routing add deterministic explanation and routing constraints (Sun et al., 2026; C. Li et al., 2026). These are application examples rather than multilingual benchmark evidence, but collectively they motivate reporting the measured retrieval effect instead of presuming that more context is better.

Compact-model evidence also supports a narrow validity boundary. TinyLLM intrusion detection demonstrates that small checkpoints can be operationally useful while remaining capacity-limited (Lu & Zhou, 2024); multisource root-cause attribution and self-supervised LogBERT-style anomaly detection show how representation choices alter downstream diagnostics (Liu et al., 2023; Xin, 2026h). Conformal alert control and forensic sequence narratives place uncertainty and localization around those predictions (Xin, 2026e, 2026f). The comparison to security and operations is deliberately limited: it supports disciplined, task-specific evaluation of compact models, not transfer of their substantive performance to multilingual reasoning.

How reliability is communicated is a separate design problem. Medical explanation and safety cards organize confidence, uncertainty, and limitations for nonexpert readers (Zhong, Wu, et al., 2025; Zhou et al., 2025; T. Ye et al., 2025; C. Li et al., 2025). Recommendation and scientific-search cards similarly tie a model output to supporting evidence and calibrated trust (B. Zhang et al., 2025; Jin, 2025b). These proposals do not validate any interface in the present study; they explain why calibration gaps, selective risk, and cultural-item uncertainty should be visible rather than compressed into one accuracy number.

Other evidence-centered interfaces emphasize hierarchy, editable rationale, and human judgment. Accounting-disclosure cards, text-grounded design rationales, visual briefs, and risk dashboards make sources and uncertainty inspectable (K. Zhang et al., 2025; Wu et al., 2025; Tu et al., 2025; Z. Li et al., 2025). Capacity-dashboard cards translate forecast risk into an operational decision, whereas design-critic alignment warns that fluent model feedback can diverge from human judgment (Liu et al., 2025; Y. Li et al., 2025). Evidence-linked counseling cards, designer-editable briefs, and incident-visualization cards extend the same principle to sensitive or operational workflows (Y. Zhang & H. Zhang, 2025b; Mu et al., 2026; Nie et al., 2026).

3. Methodology

The experiment used the English, Arabic, and Chinese Global-MMLU-Lite test slices (Cohere Labs, 2024). Each has 400 items, 43 subjects, four answer options, a gold label, and cultural and topical metadata. Every slice contains 200 culturally sensitive and 200 culturally agnostic items; the supplied cultural label was retained as an item annotation.

Table 1 reports the identical broad-category composition of the three test slices.

Table 1. Broad-category composition of each 400-item language slice. CA = culturally agnostic; CS = culturally sensitive.

Category	CA	CS	Total
Business	29	29	58
Humanities	51	51	102
Medical	18	18	36
Other	28	28	56
STEM	23	23	46
Social Sciences	51	51	102

The full cohort supports language-specific summaries. Ten Chinese high_school_world_history rows differ in content from their English and Arabic identifier matches; removing them yields a 390-item aligned cohort with 195 culturally sensitive and 195 culturally agnostic cases. Cross-language and pivot analyses use this cohort.

A strict cohort also removes three questions with duplicated options: international_law/test/51 (A/B in all languages), business_ethics/test/44 (Arabic A/D), and miscellaneous/test/473 (Chinese C/D). Its 387 items comprise 193 culturally sensitive and 194 culturally agnostic cases. This exclusion avoids treating collapsed semantic choices as four distinct alternatives.

Figure 1 summarizes the audit and cohort construction.

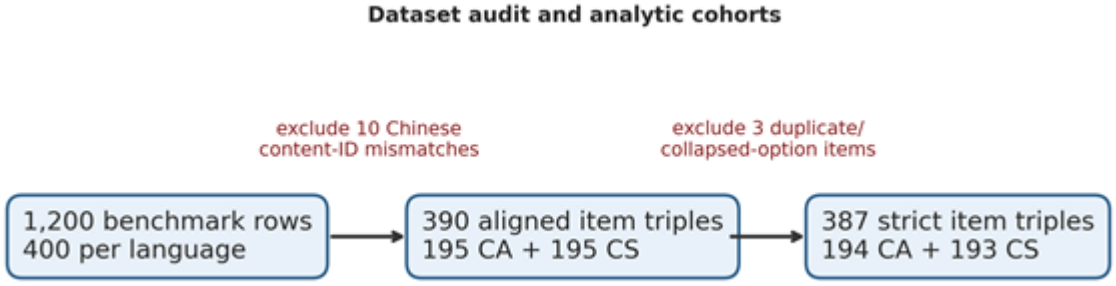

Figure 1. Dataset audit and construction of the full, aligned, and strict analytic cohorts.

The model was SmolLM2-135M Multilingual Base, a 50-language checkpoint (Agentlans, 2025; Hugging FaceTB, 2025), in GGUF Q4_K_M quantization (Radermacher, 2025). The file checksum was 52c76ba0c5572d6ba1a6ddc48ca18d58d3c8b995bdbf4260682e2ea783df5e90. CPU inference used llama-cpp-python 0.3.16, one thread, batch size 1,024, and seed 20260729. Native, English-retrieval, and Chinese-retrieval runs used 4,096-token contexts. Arabic retrieval used 8,192 after preflight identified a 5,621-token maximum. Every completed prompt fit its context without truncation.

Prompts kept a common information order: localized instruction, target marker, question, options A–D, and answer cue. Retrieval prompts placed two localized solved demonstrations before the target. Rather than generate text, the scorer extracted the final-position logits of the four Latin answer labels, verified as distinct single tokens 49–52. A same-language content-free prompt replaced the question and options with N/A; its logits were subtracted from native and retrieval logits, then normalized with a four-way softmax (Zhao et al., 2021). Retrieval used the same zero-shot baseline, without demonstrations. The top corrected probability supplied the prediction and confidence. Raw scores were retained for option-bias analysis, and input length included the beginning-of-sequence token.

The native zero-shot condition presented each item in its dataset language without demonstrations.

The reference-English pivot reused the English native prediction and probability vector for each aligned Arabic or Chinese identifier. It isolates the human-curated English representation without introducing a translation system or another model call.

The retrieval-augmented in-context learning condition selected from the five official same-language development items in each subject. Character-boundary TF–IDF used two- to five-character n-grams, minimum document frequency one, sublinear term frequency, L2 normalization, and cosine similarity. English was lowercased; Arabic and Chinese were unchanged. Stable source order broke ties, and the top two solved items became demonstrations. Test answers never entered the index, and the dataset's reference field remained annotation metadata.

Accuracy used the highest-probability label. Calibration used the complete four-class distribution: negative log-likelihood, the unscaled multiclass Brier score, and expected calibration error (ECE). ECE used 10 equal-count bins overall and five for cultural and category cells, with stable ordering of confidence ties. Reliability plots compare binwise mean confidence and accuracy.

Five folds were stratified by cultural label and category from English identifiers, then reused across conditions and languages. In each training portion, scalar temperature minimized NLL over $T \in [0.05, 20]$ and was applied to held-out corrected logits. Regimes were language-specific, pooled multilingual, and English transfer; pivot transfer used the English-native temperature (Ahuja et al., 2022; Yang et al., 2026). Native and retrieval fitting used 400 rows; pivot fitting used 390. Primary comparisons used the aligned cohort, cohort sensitivity used full and strict sets, and raw-option diagnostics used all 400 items.

Selective prediction was an offline withholding rule, not observed refusal. At coverage c, the $\lceil cn \rceil$ highest-confidence items were retained. Mean cumulative error defined area under the risk–coverage curve; fixed risks used 50%, 70%, 80%, and 90% coverage. Curves used pre-temperature contextual probabilities (El-Yaniv & Wiener, 2010).

Accuracy received 95% Wilson intervals. Aligned paired differences in accuracy, NLL, and Brier score were resampled 10,000 times within category-by-cultural strata (Efron & Tibshirani, 1994). Cultural contrasts used independent category-stratified bootstraps because the sensitive and agnostic sets were disjoint. ECE and selective-risk measures remained descriptive.

Contrasts were English minus Arabic, English minus Chinese, Arabic minus Chinese, retrieval minus native, pivot minus native, and culturally sensitive minus agnostic. Eleven paired accuracy tests used exact two-sided McNemar values (McNemar, 1947) with one Holm correction family (Holm, 1979). Paired prompt-token ratios received stratified bootstrap intervals; category analyses were secondary. Fold and resampling seed was 20260729.

The validity boundary is one compact quantized base model and one translated benchmark, with possible MMLU-like pretraining exposure. Quantization may alter logit spacing. Cultural annotations concern situated source knowledge, not whole communities, and the English pivot does not estimate machine-translation cost or error.

4. Results and Discussion

All six inference cells contained 400 predictions with 400 unique sample identifiers. None of the 2,400 scored prompts was truncated, including the longest Arabic retrieval prompt at 5,621 tokens. Table 2 links every completed run to the aligned and strict cohorts used in subsequent comparisons. This execution record matters because missing cases or language-specific truncation would otherwise change the composition of paired estimates.

Table 2. Prediction-file audit and analytic-cohort coverage.

Language	Condition	Rows	Unique IDs	Aligned n	Strict n	Truncated
English	Native	400	400	390	387	0
English	Retrieved	400	400	390	387	0
Arabic	Native	400	400	390	387	0
Arabic	Retrieved	400	400	390	387	0
Chinese	Native	400	400	390	387	0
Chinese	Retrieved	400	400	390	387	0

On the 390 aligned items, native-language accuracy remained close to the 25% chance level: 26.7% for English (95% CI [22.5%, 31.3%]), 26.9% for Arabic ([22.8%, 31.5%]), and 25.6% for Chinese ([21.6%, 30.2%]). Table 3 reports the corresponding confidence and proper-score metrics, and Figure 2 displays the Wilson intervals. None of the three native-language accuracy contrasts was significant after Holm correction. Aggregate scores alone would therefore suggest little linguistic disparity.

Table 3. Accuracy and calibration metrics on the 390-item aligned cohort. ECE = expected calibration error; NLL = negative log-likelihood.

Language	Condition	n	Accuracy (%)	95% CI (%)	Confidence	ECE	NLL	Brier
English	Native	390	26.7	[22.5, 31.3]	0.367	0.102	1.448	0.777
English	Retrieved	390	25.6	[21.6, 30.2]	0.407	0.155	1.505	0.802
Arabic	Native	390	26.9	[22.8, 31.5]	0.322	0.069	1.417	0.765
Arabic	Retrieved	390	26.4	[22.3, 31.0]	0.468	0.234	1.543	0.838
Arabic	English pivot	390	26.7	[22.5, 31.3]	0.367	0.102	1.448	0.777
Chinese	Native	390	25.6	[21.6, 30.2]	0.564	0.308	1.714	0.910
Chinese	Retrieved	390	28.5	[24.2, 33.1]	0.551	0.267	1.822	0.898
Chinese	English pivot	390	26.7	[22.5, 31.3]	0.367	0.102	1.448	0.777

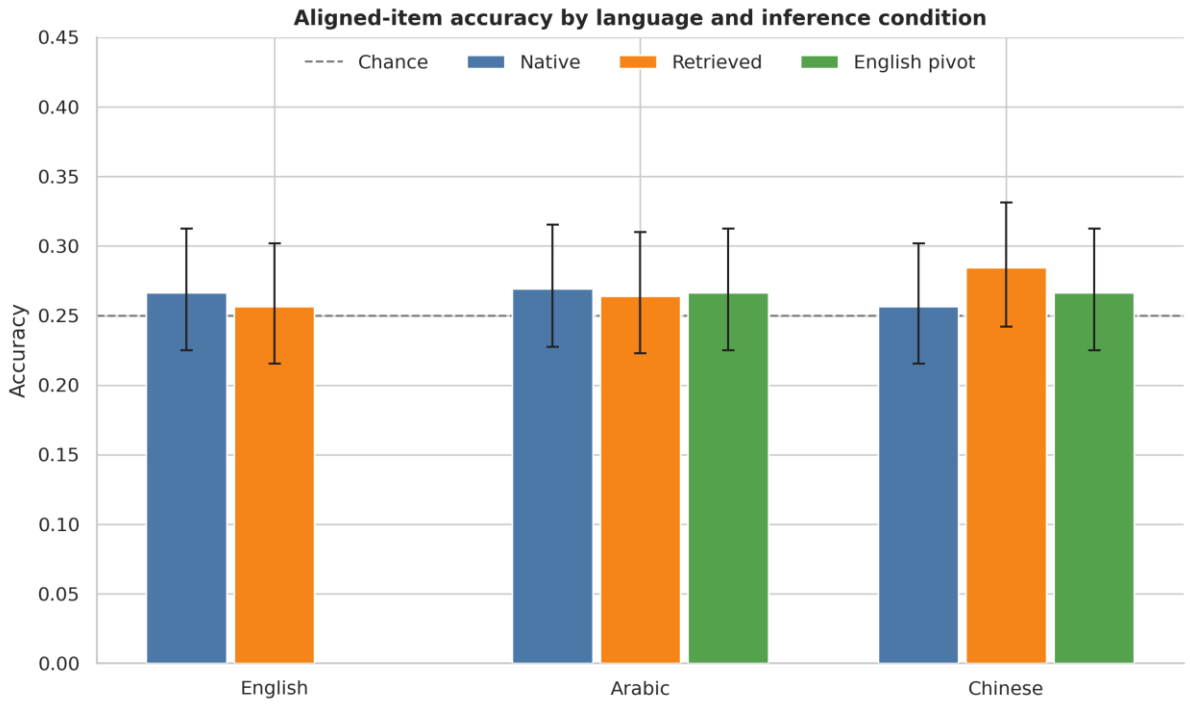

Figure 2. Aligned-cohort accuracy with 95% Wilson confidence intervals.

The item-level decisions told a different story. As Table 4 shows, native prediction agreement was 17.9% between English and Arabic, 52.8% between English and Chinese, and only 3.3% between Arabic and Chinese. Cohen's κ was −.058, .015, and −.001, respectively. Retrieved-condition agreement rose to 40.3%, 77.4%, and 48.7%, but κ remained modest at .026, .283, and .124. Similar accuracy thus arose from substantially different answer patterns. This distinction is consistent with the broader finding that multilingual benchmark scores can conceal language- and culture-conditioned behavior (Singh et al., 2025): translation held the intended answer constant, but it did not make the model's decision process invariant.

Table 4. Paired cross-language contrasts and answer agreement on the aligned cohort. Accuracy differences are in percentage points.

Comparison	Pairs	Accuracy Δ (pp)	95% CI (pp)	NLL Δ	Brier Δ	Agreement (%)	κ	Holm p
English native − Arabic native	390	-0.3	[-6.7, 6.2]	0.031	0.012	17.9	-0.058	1.000
English native − Chinese native	390	1.0	[-4.1, 5.9]	-0.266	-0.133	52.8	0.015	1.000
Arabic native − Chinese native	390	1.3	[-5.6, 8.2]	-0.297	-0.145	3.3	-0.001	1.000
English retrieved − Arabic retrieved	390	-0.8	[-6.4, 4.6]	-0.038	-0.036	40.3	0.026	1.000
English retrieved − Chinese retrieved	390	-2.8	[-6.2, 0.5]	-0.318	-0.097	77.4	0.283	1.000
Arabic retrieved − Chinese retrieved	390	-2.1	[-7.2, 3.1]	-0.279	-0.061	48.7	0.124	1.000

Retrieval did not yield a consistent gain. Relative to native inference, it changed English accuracy by −1.0 percentage point (95% CI [−5.6, 3.6], Holm-adjusted p = 1.000) and Arabic accuracy by −0.5 points ([−6.4, 5.4], p = 1.000). Chinese retrieval increased accuracy by 2.8 points ([0.8, 4.9]), but the exact paired test was not significant after correction (p = .211). Its NLL worsened by .108 while its Brier score improved by .011, giving a mixed probabilistic result. The English pivot changed Arabic accuracy by −0.3 points ([−6.4, 6.2], p = 1.000) and Chinese accuracy by +1.0 point ([−3.8, 5.9], p = 1.000). Both pivot cells inherited the 26.7% English accuracy by construction. Table 5 shows that neither intervention established a reliable improvement over direct native-language inference. This result accords with multilingual RAG research showing that added context can introduce language- and prompt-dependent effects rather than a uniform benefit (Chirkova et al., 2024).

The mixed retrieval effect is consistent with applied pipelines that separate evidence selection, diagnosis, explanation, and routing rather than treating them as one operation (Liu et al., 2023; Xin, 2025a; Nie et al., 2026; C. Li et al., 2026). The result therefore does not indict RAG generally; it shows that two subject-matched demonstrations were not a uniformly beneficial intervention for this compact model.

Table 5. Paired effects of same-language retrieval and the reference-English pivot. Accuracy differences are in percentage points.

Comparison	Pairs	Accuracy Δ (pp)	95% CI (pp)	NLL Δ	Brier Δ	Agreement (%)	κ	Holm p
English retrieved − English native	390	-1.0	[-5.6, 3.6]	0.057	0.025	61.3	0.298	1.000
Arabic retrieved − Arabic native	390	-0.5	[-6.4, 5.4]	0.126	0.073	25.9	-0.043	1.000
Chinese retrieved − Chinese native	390	2.8	[0.8, 4.9]	0.108	-0.011	88.7	0.032	.211
Arabic via English pivot − Arabic native	390	-0.3	[-6.4, 6.2]	0.031	0.012	17.9	-0.058	1.000
Chinese via English pivot − Chinese native	390	1.0	[-3.8, 5.9]	-0.266	-0.133	52.8	0.015	1.000

Calibration differences were larger than accuracy differences. Before temperature scaling, native English produced ECE = .102, NLL = 1.448, and Brier = .777; Arabic produced .069, 1.417, and .765; and Chinese produced .308, 1.714, and .910. Figure 3 displays the contextual reliability curves, while Table 6 compares those probabilities with all three cross-fitted temperature regimes. Temperature scaling left accuracy unchanged while reducing native ECE to .037–.051 for English and .051–.059 for both Arabic and Chinese; NLL converged to approximately 1.386–1.387 and Brier to .750. Most fold-specific temperatures reached or approached the upper optimization bound of 20, however, and mean confidence approached .25. The improved scores therefore came chiefly from flattening probabilities toward the four-choice uniform distribution, not from better ranking or additional knowledge. Calibration must be interpreted together with discrimination rather than alone (Guo et al., 2017; Jiang et al., 2021).

Table 6. Cross-fitted scalar temperature calibration for native inference on the aligned cohort.

Language	Calibration	Accuracy (%)	ECE	NLL	Brier	Mean confidence
English	Contextual	26.7	0.102	1.448	0.777	0.367
English	Language-specific T	26.7	0.051	1.386	0.750	0.256
English	Pooled multilingual T	26.7	0.037	1.386	0.750	0.255
English	English-transfer T	26.7	0.051	1.386	0.750	0.256
Arabic	Contextual	26.9	0.069	1.417	0.765	0.322
Arabic	Language-specific T	26.9	0.051	1.387	0.750	0.253
Arabic	Pooled multilingual T	26.9	0.051	1.387	0.750	0.253
Arabic	English-transfer T	26.9	0.059	1.387	0.750	0.253
Chinese	Contextual	25.6	0.308	1.714	0.910	0.564
Chinese	Language-specific T	25.6	0.059	1.387	0.750	0.265
Chinese	Pooled multilingual T	25.6	0.051	1.387	0.750	0.264
Chinese	English-transfer T	25.6	0.059	1.387	0.750	0.265

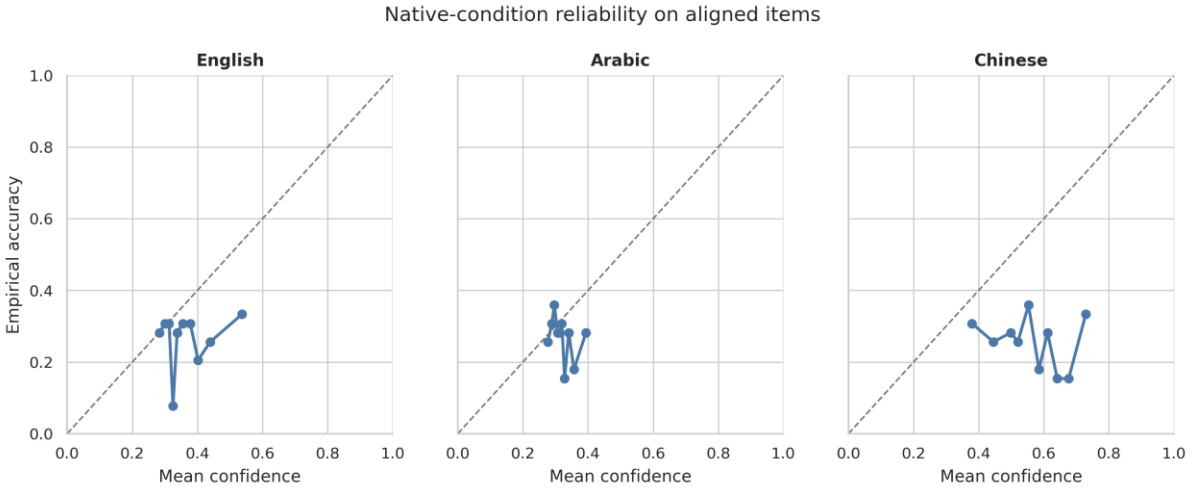

Figure 3. Contextual reliability diagrams for native predictions.

Confidence also supplied little basis for low-error selective answering. Native AURC was .709 for English, .742 for Arabic, and .758 for Chinese; retrieval yielded .759, .774, and .761. At 50% coverage, native risks were still 71.8%, 75.9%, and 77.9%, while retrieved risks were 76.9%, 77.4%, and 78.5%. At 90% coverage, native risk remained between 72.9% and 74.9%. Table 7 gives all fixed-coverage values, and Figure 4 traces the native-condition curves. Higher confidence did not identify a dependable low-risk subset. This offline result does not describe conversational refusal, but it reinforces prior evidence that confidence-based abstention can fail under multilingual or domain variation (Kamath et al., 2020; Feng et al., 2024).

Table 7. Selective-prediction risk across conditions on the aligned cohort. AURC = area under the risk–coverage curve.

Language	Condition	AURC	Risk @ 50%	Risk @ 70%	Risk @ 80%	Risk @ 90%
English	Native	0.709	71.8	74.7	74.0	73.5
English	Retrieved	0.759	76.9	75.8	75.0	73.8
Arabic	Native	0.742	75.9	74.7	73.4	72.9
Arabic	Retrieved	0.774	77.4	76.2	75.3	72.4
Arabic	English pivot	0.709	71.8	74.7	74.0	73.5
Chinese	Native	0.758	77.9	75.5	75.0	74.9
Chinese	Retrieved	0.761	78.5	74.4	73.4	72.6
Chinese	English pivot	0.709	71.8	74.7	74.0	73.5

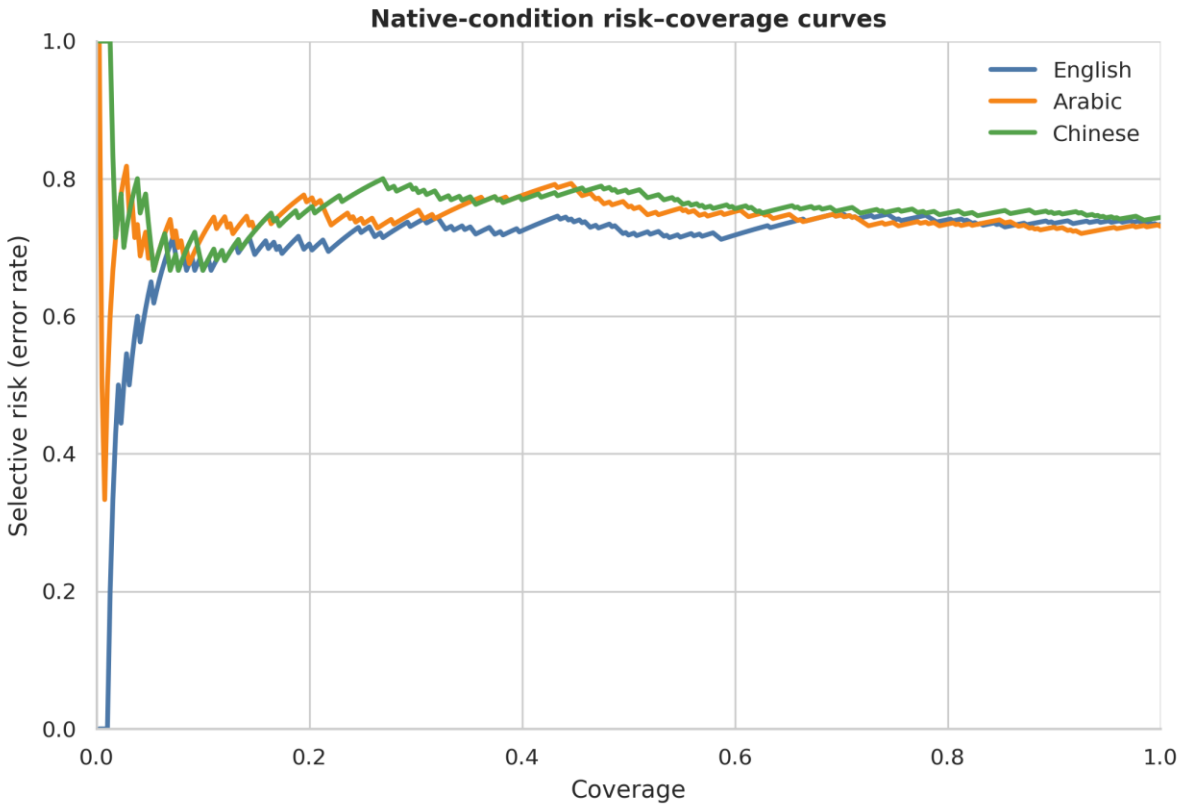

Figure 4. Native-condition risk–coverage curves.

Culturally sensitive items were directionally harder than culturally agnostic items in every native-language condition, although uncertainty was substantial. The sensitivity-minus-agnostic gap was −2.1 points for English (95% CI [−10.8, 6.7]), −7.7 points for Arabic ([−15.9, 1.0]), and −5.1 points for Chinese ([−13.3, 3.6]). Under retrieval, the gaps were −3.1, −0.5, and −5.6 points; under the English pivot, both Arabic and Chinese inherited gaps of −2.1 points. Every interval in Table 8 and Figure 5 crossed zero. The repeated negative direction is worth monitoring, but the experiment does not establish a culturally sensitive item penalty. Moreover, the annotation concerns the source knowledge required by each question; it should not be interpreted as a comprehensive measure of the cultures associated with the three prompt languages (Hershcovich et al., 2022; Naous et al., 2024).

Table 8. Accuracy differences between culturally sensitive and culturally agnostic aligned items.

Language	Condition	CA accuracy (%)	CS accuracy (%)	CS − CA (pp)	95% CI (pp)
English	Native	27.7	25.6	-2.1	[-10.8, 6.7]
English	Retrieved	27.2	24.1	-3.1	[-11.8, 5.6]
Arabic	Native	30.8	23.1	-7.7	[-15.9, 1.0]
Arabic	Retrieved	26.7	26.2	-0.5	[-9.2, 8.2]
Arabic	English pivot	27.7	25.6	-2.1	[-10.8, 6.7]
Chinese	Native	28.2	23.1	-5.1	[-13.3, 3.6]
Chinese	Retrieved	31.3	25.6	-5.6	[-14.4, 3.1]
Chinese	English pivot	27.7	25.6	-2.1	[-10.8, 7.2]

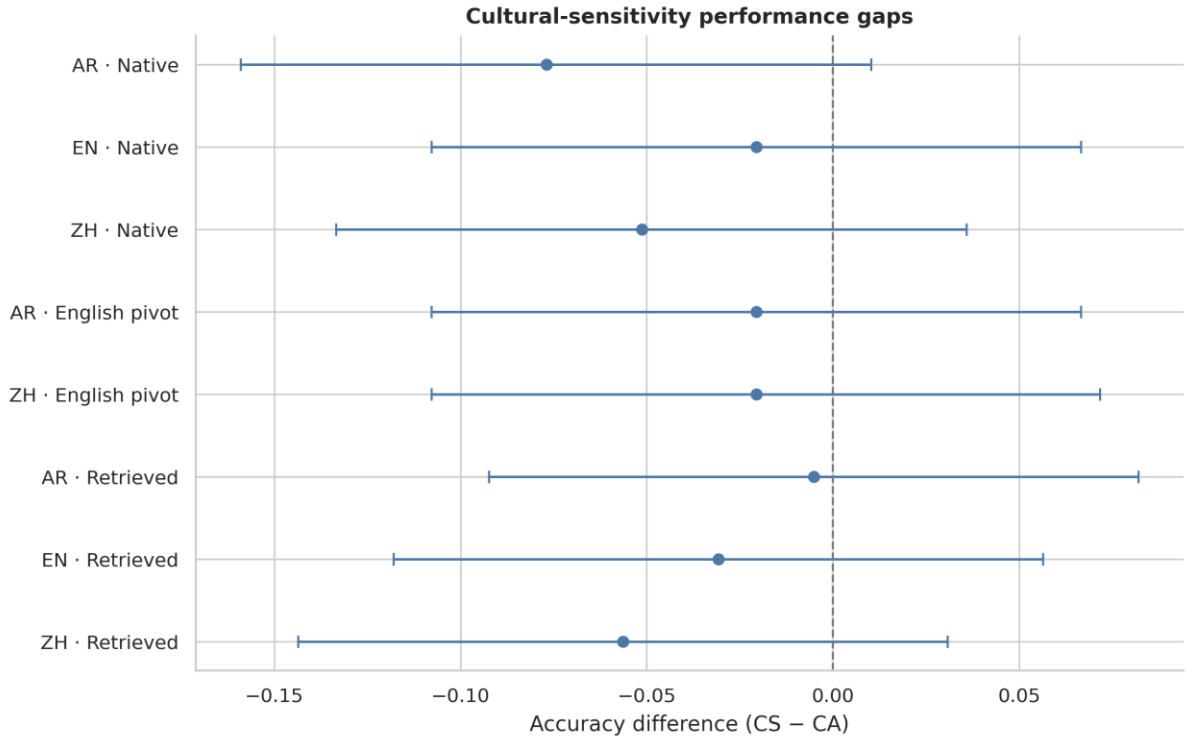

Figure 5. Culturally sensitive minus culturally agnostic accuracy gaps with 95% bootstrap intervals.

Broad-category results further showed that the language relationship depended on content. Table 9 and Figure 6 report native accuracy by category. Chinese reached 41.3% in STEM, compared with 32.6% for English and 13.0% for Arabic. Arabic reached 44.6% in the heterogeneous Other category, compared with 21.4% for English and 14.3% for Chinese. Chinese medical accuracy was 16.7%, below English at 30.6% and Arabic at 27.8%. Business and social-science scores were more closely grouped. Because several categories were small, these descriptive contrasts do not establish stable disciplinary strengths; they show that a single overall language score averages over different subject mixtures of success and error.

Table 9. Native accuracy by broad category on the aligned cohort.

Category	English (%)	Arabic (%)	Chinese (%)
Business	22.4	27.6	25.9
Humanities	30.4	22.8	28.3
Medical	30.6	27.8	16.7
Other	21.4	44.6	14.3
STEM	32.6	13.0	41.3
Social Sciences	24.5	26.5	25.5

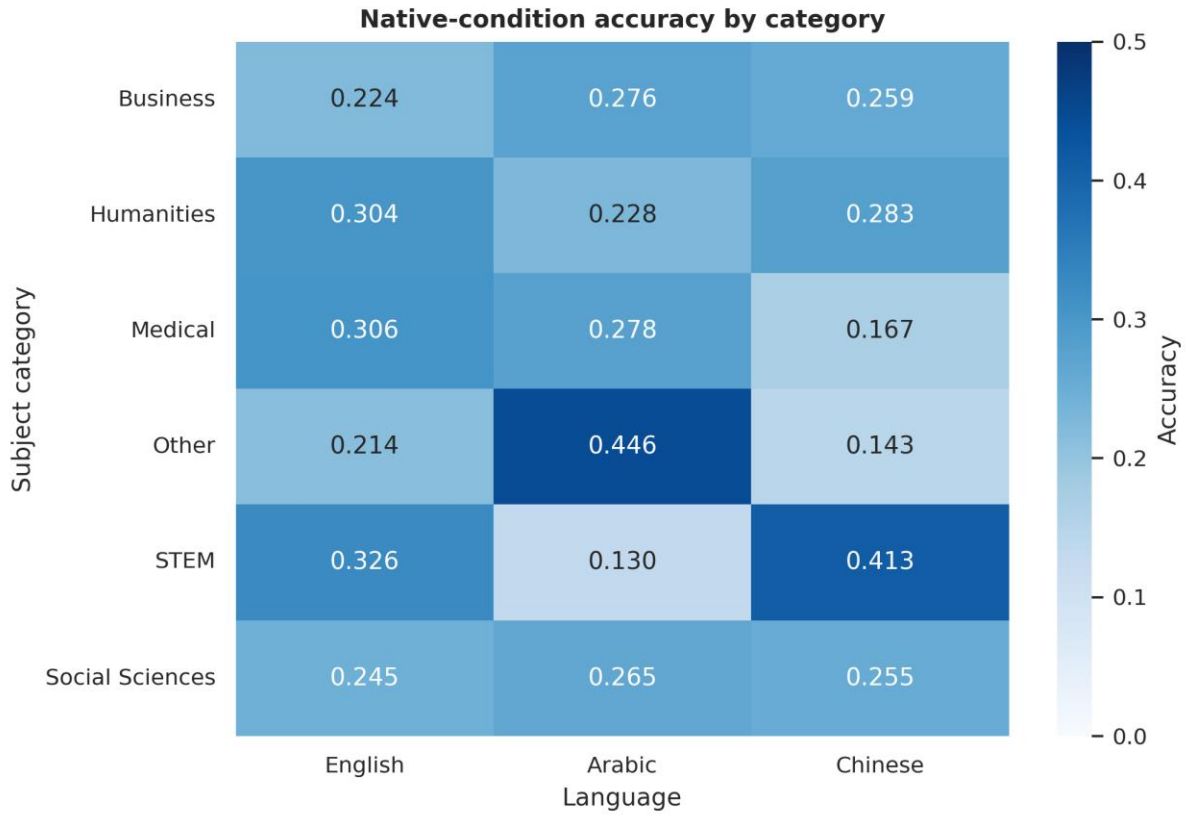

Figure 6. Heatmap of native accuracy by language and broad category.

Prompt length imposed an additional, operational disparity. Native prompts averaged 123.1 English tokens, 393.0 Arabic tokens, and 266.9 Chinese tokens. Arabic therefore required 3.04 times as many tokens as English (95% CI [3.00, 3.08]), while Chinese required 2.12 times as many ([2.09, 2.14]). Retrieval increased the means to 369.5, 1,205.4, and 767.7 tokens, with Arabic-to-English and Chinese-to-English ratios of 3.18 and 2.03. The medians and maxima in Table 10 confirm that the difference was not driven only by a few long questions. Figure 7 summarizes the paired distributions as boxplots, with outliers suppressed for legibility; the observed maxima remain in Table 10. These ratios combine script, tokenization, and surface-length effects, but under a fixed model they translate directly into unequal context-processing burden. This limitation is especially relevant because the model card cautions that the checkpoint retains an English tokenizer despite multilingual continued pretraining (Agentlans, 2025).

Operationally, these ratios are not cosmetic. Capacity research frames deployment under heterogeneous hardware, topology shift, calibrated risk, and human-readable capacity decisions (He et al., 2023, 2024; Liu et al., 2025). The present experiment did not measure latency, energy, or admission rates, but its two- to threefold token ratios identify a mechanism that can consume context and compute unequally under a fixed serving budget.

Table 10. Beginning-of-sequence-inclusive input-token burden on aligned items.

Language	Condition	Mean tokens	Median tokens	Maximum tokens	Token ratio to EN	Ratio 95% CI
English	Native	123.1	92	578	—	—
English	Retrieved	369.5	272	1643	—	—
Arabic	Native	393.0	282	2104	3.04	[3.00, 3.08]
Arabic	Retrieved	1205.4	874	5621	3.18	[3.14, 3.22]
Chinese	Native	266.9	201	1287	2.12	[2.09, 2.14]
Chinese	Retrieved	767.7	549	3776	2.03	[2.01, 2.06]

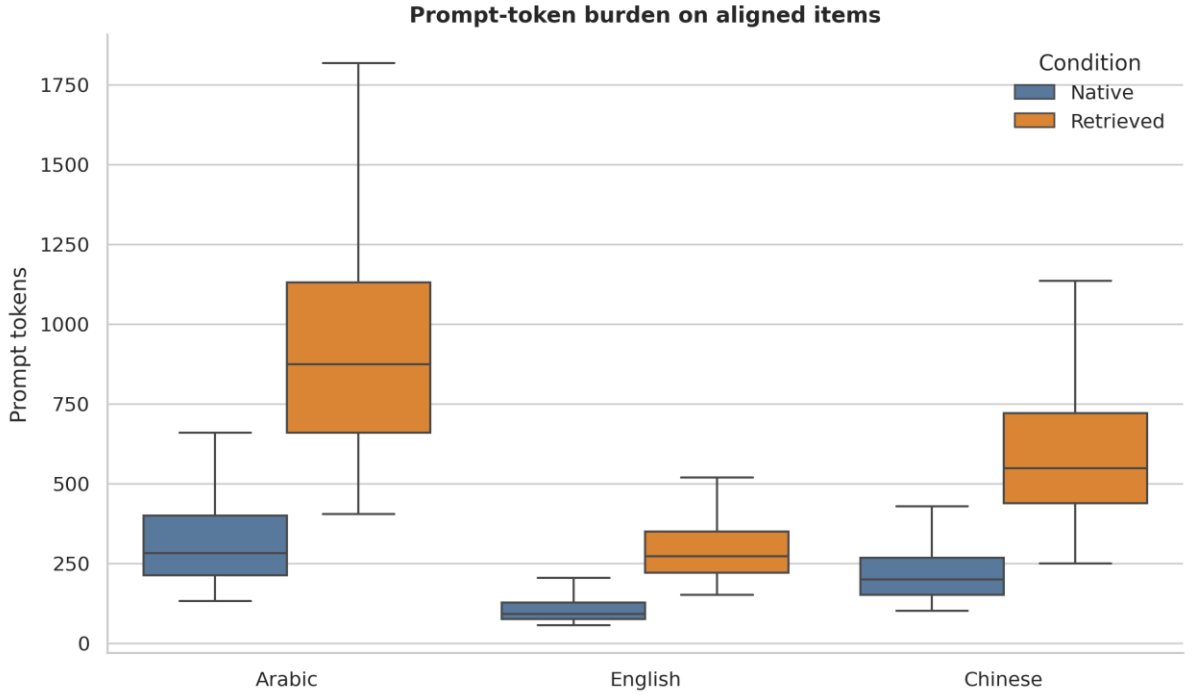

Figure 7. Prompt-token distributions by language and condition; outliers are suppressed in the display and observed maxima appear in Table 10.

The forced-choice logits also revealed strong answer-symbol priors. As Table 11 and Figure 8 show, unadjusted scoring predicted option A for 95.2% of English items, 100% of Arabic items, and 97.8% of Chinese items. Contextual correction improved accuracy and ECE, but residual concentration remained: English selected D for 53.5% of items, Arabic selected B for 67.5%, and Chinese selected D for 99.2%. These language-specific shifts help explain how almost equal accuracies could coexist with very low agreement. They support content-free label-prior correction in constrained multiple-choice prompting (Zhao et al., 2021), while showing that one subtraction did not restore a well-dispersed decision distribution.

Table 11. Raw and contextually corrected native answer-label distributions on all 400 items per language.

Language	Score	Accuracy (%)	Mean confidence	ECE	Most predicted option	Option share (%)
English	Raw	23.0	0.511	0.281	A	95.2
English	Contextual	26.0	0.369	0.109	D	53.5
Arabic	Raw	23.5	0.785	0.550	A	100.0
Arabic	Contextual	26.8	0.322	0.072	B	67.5
Chinese	Raw	22.5	0.558	0.333	A	97.8
Chinese	Contextual	25.5	0.563	0.308	D	99.2

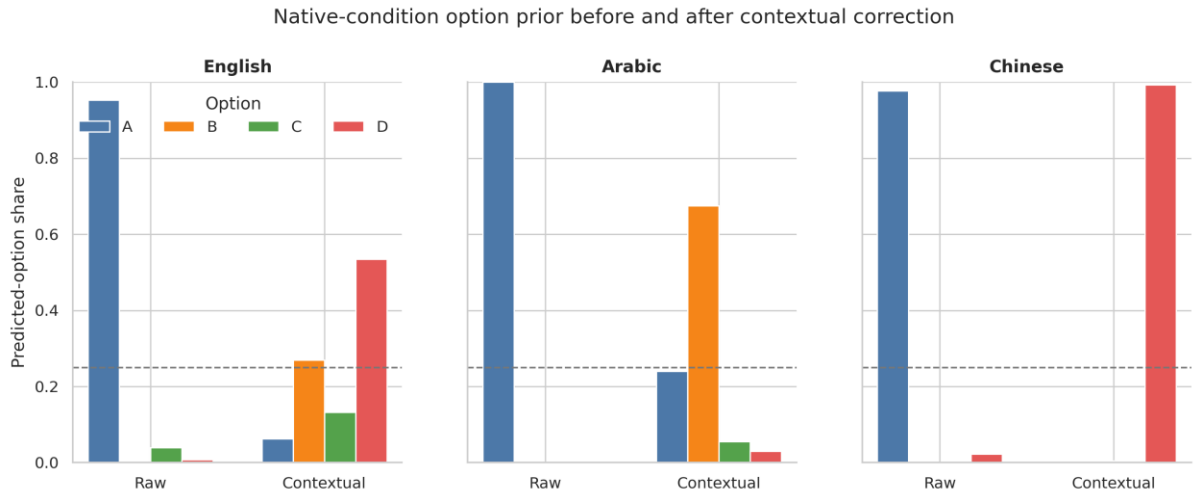

Figure 8. Answer-option shares before and after contextual correction.

Finally, the main findings were stable across cohort definitions. Table 12 shows native accuracies of 26.0%, 26.7%, and 26.6% for English in the full, aligned, and strict cohorts; the corresponding Arabic values were 26.8%, 26.9%, and 26.9%, and the Chinese values were 25.5%, 25.6%, and 25.8%. Retrieved and pivot estimates were similarly stable. The audit exclusions therefore protected the logic of paired comparison without creating the substantive pattern. Across cohorts, aggregate accuracy remained nearly equal while decision agreement, calibration, selective reliability, response-option preference, and token burden differed materially.

Human-facing auditability matters most when scores inform interventions. Rubric-grounded essay scoring, student-retention rationales, and teacher-facing course explanations all make the evidence trail and uncertain cases available for review (Xin, 2026a, 2026c; J. Zhang, 2026). For multilingual benchmarks, an analogous report should show language-specific accuracy, confidence calibration, risk–coverage, item-level disagreement, and token burden together. Such reporting does not turn a benchmark into a deployment interface; it prevents a single aggregate from standing in for linguistic equity.

Table 12. Accuracy sensitivity to the full, aligned, and strict cohort definitions.

Language	Condition	Full 400 (%)	Aligned 390 (%)	Strict 387 (%)
English	Native	26.0	26.7	26.6
English	Retrieved	25.0	25.6	25.8
Arabic	Native	26.8	26.9	26.9
Arabic	Retrieved	26.0	26.4	26.1
Arabic	English pivot	—	26.7	26.6
Chinese	Native	25.5	25.6	25.8
Chinese	Retrieved	28.2	28.5	28.7
Chinese	English pivot	—	26.7	26.6

5. Conclusions

Translation equivalence did not produce linguistic equity for the evaluated model. English, Arabic, and Chinese native accuracy differed by no more than 1.3 percentage points on aligned questions, yet native-condition answer agreement ranged from 3.3% to 52.8%, Chinese native ECE reached .308, and confidence-based withholding left error rates above 71% even at 50% coverage. Same-language retrieval and the English reference pivot did not yield a multiplicity-adjusted accuracy gain. Culturally sensitive questions were consistently but inconclusively harder, with every bootstrap interval spanning zero. Arabic and Chinese prompts also consumed substantially more context than English prompts.

These results make aggregate score parity an insufficient criterion for multilingual evaluation. At minimum, parallel benchmarks should report item-level agreement, probabilistic calibration, selective risk, answer-symbol diagnostics, and language-specific token burden alongside accuracy. Dataset audits must also verify semantic alignment rather than infer it from shared identifiers. Here, nearly identical aligned- and strict-cohort estimates show that the results were insensitive to the exclusions.

The scope remains model-conditional. SmolLM2-135M Multilingual Base is a compact, quantized base checkpoint, and its residual option collapse is not evidence about larger or instruction-tuned systems. The cultural annotation characterizes benchmark knowledge assumptions, not whole language communities. For this model and dataset, translated questions preserved their human answer key while eliciting different decisions, confidence structures, and token burdens. Multilingual evaluation should treat those dimensions as part of equity rather than secondary diagnostics.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers

Artificial Intelligence (AI) Use Disclosure: The authors declare that no artificial intelligence tools were used in the preparation of this manuscript.

References

[1] Agentlans. (2025). *SmolLM2-135M-multilingual-base* [Language model]. Hugging Face. https://huggingface.co/agentlans/SmolLM2-135M-multilingual-base

[2] Ahuja, K., Sitaram, S., Dandapat, S., & Choudhury, M. (2022). On the calibration of massively multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 4310–4323). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.290

[3] Asai, A., Yu, X., Kasai, J., & Hajishirzi, H. (2021). One question answering model for many languages with cross-lingual dense passage retrieval. *Advances in Neural Information Processing Systems, 34*, 7547–7560. https://proceedings.neurips.cc/paper/2021/hash/3df07fdae1ab273a967aaa1d355b8bb6-Abstract.html

[4] Bai, J., Chen, S., Zheng, D., & Kuo, M.-J. (2026). Interpretable attack-chain stage detection from AWS CloudTrail event sequences via linear models and HMM smoothing. *Information Electrical and Electronic Engineering, 6*(1), 28–43. https://doi.org/10.33474/infotron.v6i1.24923

[5] Bai, J., Wang, H., Wu, Q., & Zhang, B. (2026). Privacy-robust incrementality estimation in cookieless settings via uplift modeling: Reproducible evidence from the Hillstrom e-mail experiment. *Journal of Technology Informatics and Engineering, 5*(1), 17–38. https://doi.org/10.51903/jtie.v5i1.468

[6] Bandarkar, L., Liang, D., Muller, B., Artetxe, M., Shukla, S. N., Husa, D., Goyal, N., Krishnan, A., Zettlemoyer, L., & Khabsa, M. (2024). The Belebele benchmark: A parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 749–775). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.44

[7] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2

[8] Chang, X., Lu, Y., & Zhong, Z. S. (2026). Review-grounded explainable recommendation with faithfulness evaluation on Amazon reviews. *Journal of Electrical Engineering and Computer Science, 11*(1), 9–22. https://doi.org/10.54732/jeecs.v11i1.2

[9] Chen, S., He, S., & Sun, E. (2024). Risk-bounded GPU resource oversubscription via conformal demand envelopes in production AI clusters. *Journal of Advanced Computing Systems, 4*(5), 119–134. https://doi.org/10.69987/JACS.2024.40509

[10] Chen, Y., & Li, M. (2025). From hand-drawn sketches to interactive web prototypes: A reproducible vision-language approach with structural and visual consistency evaluation. *Journal of Technology Informatics and Engineering, 4*(2), 364–384. https://doi.org/10.51903/jtie.v4i2.490

[11] Chen, Y., & Xu, H. (2026). Trust-calibrated multilingual RAG for humanitarian information platforms: Empirical evaluation on OMoS-QA for migration information access. *International Journal of Graphic Design, 4*(1), 141–164. https://doi.org/10.51903/ijgd.v4i1.3552

[12] Chen, Y., Zhang, Y., & Sherman, M. (2024). Going concern and bankruptcy prediction under extreme class imbalance: Cost-sensitive learning, resampling, and focal loss with explainable financial-ratio portraits. *Journal of Advanced Computing Systems, 4*(4), 80–96. https://doi.org/10.69987/JACS.2024.40407

[13] Chen, Y., Zhang, Y., Chau, D., & Sherman, M. (2023). Credit card default risk tiering with probability calibration and uncertainty-driven rejection: A reproducible study on the UCI credit card clients dataset. *Journal of Advanced Computing Systems, 3*(4), 31–47. https://doi.org/10.69987/JACS.2023.30403

[14] Chen, Y., Zhou, S., & Lin, E. (2025). Accounting-aware evidence retrieval for institutional due diligence of tokenized trade receivable RWA. *Journal of Technology Informatics and Engineering, 4*(3), 649–663. https://doi.org/10.51903/jtie.v4i3.542

[15] Chirkova, N., Rau, D., Déjean, H., Formal, T., Clinchant, S., & Nikoulina, V. (2024). Retrieval-augmented generation in multilingual settings. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)* (pp. 177–188). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.knowllm-1.15

[16] Cohere Labs. (2024). *Global-MMLU-Lite* [Data set]. Hugging Face. https://huggingface.co/datasets/CohereLabs/Global-MMLU-Lite

[17] Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman & Hall/CRC. https://doi.org/10.1201/9780429246593

[18] El-Yaniv, R., & Wiener, Y. (2010). On the foundations of noise-free selective classification. *Journal of Machine Learning Research, 11*, 1605–1641. https://jmlr.org/papers/v11/el-yaniv10a.html

[19] Feng, S., Shi, W., Wang, Y., Ding, W., Ahia, O., Li, S. S., Balachandran, V., Sitaram, S., & Tsvetkov, Y. (2024). Teaching LLMs to abstain across languages via multilingual feedback. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 4125–4150). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.239

[20] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1321–1330). PMLR. https://proceedings.mlr.press/v70/guo17a.html

[21] He, S., Chang, X., & Sun, E. (2024). Cross-cloud transfer learning for AI training capacity forecasting under workload and topology distribution shift. *Journal of Advanced Computing Systems, 4*(1), 100–120. https://doi.org/10.69987/JACS.2024.40108

[22] He, S., Tu, H., & Liu, I. (2023). Safe PD capacity forecasting with time-series foundation models and calibrated uncertainty for heterogeneous GPU clusters. *Journal of Advanced Computing Systems, 3*(4), 48–66. https://doi.org/10.69987/JACS.2023.30404

[23] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. In *International Conference on Learning Representations*. https://openreview.net/forum?id=d7KBjmI3GmQ

[24] Hershcovich, D., Frank, S., Lent, H., de Lhoneux, M., Abdou, M., Brandl, S., Bugliarello, E., Cabello Piqueras, L., Chalkidis, I., Cui, R., Fierro, C., Margatina, K., Rust, P., & Søgaard, A. (2022). Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 6997–7013). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.482

[25] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70. https://www.jstor.org/stable/4615733

[26] Hugging FaceTB. (2025). *SmolLM2-135M* [Large language model]. Hugging Face. https://huggingface.co/HuggingFaceTB/SmolLM2-135M

[27] Jiang, Z., Araki, J., Ding, H., & Neubig, G. (2021). How can we know when language models know? On the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics, 9*, 962–977. https://doi.org/10.1162/tacl_a_00407

[28] Jin, J. (2025a). Evidence-chain reliable RAG: Hallucination detection, source attribution, and deterministic provenance explanations. *Journal of Technology Informatics and Engineering, 4*(2), 520–533. https://doi.org/10.51903/jtie.v4i2.535

[29] Jin, J. (2025b). LLM-style evidence cards for scientific search interfaces: A UI/UX design framework for retrieval transparency, ranking trust, and visual evidence hierarchy. *International Journal of Graphic Design, 3*(2), 397–414. https://doi.org/10.51903/ijgd.v3i2.3698

[30] Jin, J., Huang, T., & Lu, S. (2024a). Cost-sensitive learning, simulated PU learning, and one-class autoencoding for extreme-imbalance credit card fraud detection. *Journal of Advanced Computing Systems, 4*(6), 64–73. https://doi.org/10.69987/JACS.2024.40605

[31] Jin, J., Huang, T., & Lu, S. (2024b). A model-risk-friendly probability of default workflow: Calibration, distribution-free uncertainty quantification, and SHAP explanations on the UCI credit card default dataset. *Journal of Advanced Computing Systems, 4*(6), 74–85. https://doi.org/10.69987/JACS.2024.40606

[32] Kamath, A., Jia, R., & Liang, P. (2020). Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5684–5696). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.503

[33] Kleinle, S., Prange, J., & Friedrich, A. (2024). OMoS-QA: A dataset for cross-lingual extractive question answering in a German migration context. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)* (pp. 231–248). Association for Computational Linguistics. https://aclanthology.org/2024.konvens-main.25/

[34] Koto, F., Li, H., Shatnawi, S., Doughman, J., Sadallah, A., Alraeesi, A., Almubarak, K., Alyafeai, Z., Sengupta, N., Shehata, S., Habash, N., Nakov, P., & Baldwin, T. (2024). ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 5622–5640). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.334

[35] Kuo, M.-J., Zheng, D., & Hires, J. (2025). Federated topic-preference learning for knowledge-grounded chat with differential privacy. *Journal of Technology Informatics and Engineering, 4*(2), 385–401. https://doi.org/10.51903/jtie.v4i2.502

[36] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[37] Li, C., Bai, J., & Wang, S. (2024). Evidence-chain reliable RAG: Word-level hallucination detection, source attribution, and provenance explanation for LLM applications. *Journal of Advanced Computing Systems, 4*(2), 76–92. https://doi.org/10.69987/JACS.2024.40207

[38] Li, C., Liu, G., & Zhao, Z. (2026). Cost-aware LLM-style routing for AIOps log analysis: Log parsing, anomaly detection, fault diagnosis, and incident summarization on LogEval task files. *Journal of Technology Informatics and Engineering, 5*(2), 91–103. https://doi.org/10.51903/jtie.v5i2.538

[39] Li, C., Zhou, B., & Gao, K. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX benchmark for explainable medical AI response interfaces. *International Journal of Graphic Design, 3*(2), 381–394. https://doi.org/10.51903/ijgd.v3i2.3709

[40] Li, H., Zhang, Y., Koto, F., Yang, Y., Zhao, H., Gong, Y., Duan, N., & Baldwin, T. (2024). CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 11260–11285). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.671

[41] Li, J., & Zhou, A. (2026). Multi-regulation RAG for AI product counsel: A legal governance framework for cross-border digital commerces. *Rule of Law Studies Journal, 2*(2), 105–123. https://doi.org/10.64780/rolsj.v2i2.225

[42] Li, Y. (2024). Findable then explainable: Retrieval–summary integration for code intelligence on a lightweight CodeSearchNet subset. *Journal of Advanced Computing Systems, 4*(7), 65–82. https://doi.org/10.69987/JACS.2024.40706

[43] Li, Y., Lu, S., & Zhao, L. (2025). LLM-as-design-critic: Aligning AI-generated UI feedback with human graphic design judgment. *International Journal of Graphic Design, 3*(1), 196–215. https://doi.org/10.51903/ijgd.v3i1.3661

[44] Li, Z., Zhang, K., & Wong, A. (2026). Numerical-reasoning guardrails for a quant research assistant: A compact reproducible benchmark using SEC and FRED data. *Journal of Technology Informatics and Engineering, 5*(2), 75–90. https://doi.org/10.51903/jtie.v5i2.541

[45] Li, Z., Zhou, S., & Zhou, Z. (2025). Financial risk dashboard design for institutional RWA investors: Visual hierarchy, chart comprehension, and explainability in FinChart-Bench. *International Journal of Graphic Design, 3*(1), 196–210. https://doi.org/10.51903/ijgd.v3i1.3715

[46] Liu, G., He, S., & Liu, I. (2023). LLM-augmented multi-source root cause attribution for CPU and network faults in microservices. *Journal of Advanced Computing Systems, 3*(6), 39–57. https://doi.org/10.69987/JACS.2023.30604

[47] Liu, G., He, S., & Wong, H. (2025). LLM-compatible visual brief cards for AI infrastructure capacity dashboards: A UI/UX framework for turning forecast risk into graphic design decisions. *International Journal of Graphic Design, 3*(1), 196–213. https://doi.org/10.51903/ijgd.v3i1.3723

[48] Liu, G., Li, C., & Zhang, E. (2024). OpsLLM for cloud incident triage: Bilingual RAG-based root cause analysis and alert summarization for AI infrastructure operations. *Journal of Advanced Computing Systems, 4*(4), 97–111. https://doi.org/10.69987/JACS.2024.40408

[49] Lu, S., & Zhou, D. (2024). TinyLLM-assisted intrusion detection for real-time IoT networks. *Journal of Advanced Computing Systems, 4*(8), 72–87. https://doi.org/10.69987/JACS.2024.40809

[50] Lu, S., & Zou, T. (2026). Uncertainty-aware medical vision–language classification on a lightweight MedMNIST-compatible biomedical patch benchmark. *Journal of Technology Informatics and Engineering, 5*(2), 1–19. https://doi.org/10.51903/jtie.v5i2.530

[51] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2), 153–157. https://doi.org/10.1007/BF02295996

[52] Meng, S., Chen, J., & Zheng, I. (2026). LLM-inspired offline reranking for financial search: Query rewriting, hybrid retrieval, and listwise relevance ranking on FiQA. *Journal of Technology Informatics and Engineering, 5*(1), 361–378. https://doi.org/10.51903/jtie.v5i1.537

[53] Mu, J., Lu, Y., & Hwang, E. (2026). Structured visual brief interfaces for advertising design: A UI/UX framework for turning creative intentions into designer-editable graphic design cards. *International Journal of Graphic Design, 4*(1), 192–208. https://doi.org/10.51903/ijgd.v4i1.3702

[54] Mu, J., Lu, Y., & Smith, M. (2023). LLM-assisted incrementality (uplift) modeling for email advertising: From feature interactions to interpretable audience–creative–channel policies. *Journal of Advanced Computing Systems, 3*(1), 31–48. https://doi.org/10.69987/JACS.2023.30103

[55] Myung, J., Lee, N., Zhou, Y., Jin, J., Putri, R. A., Antypas, D., Borkakoty, H., Kim, E., Perez-Almendros, C., Ayele, A. A., Gutiérrez-Basulto, V., Ibáñez-García, Y., Lee, H., Muhammad, S. H., Park, K., Rzayev, A. S., White, N., Yimam, S. M., Pilehvar, M. T., . . . Oh, A. (2024). BLEnD: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems, 37*, 78104–78146. https://doi.org/10.52202/079017-2483

[56] Naous, T., Ryan, M. J., Ritter, A., & Xu, W. (2024). Having beer after prayer? Measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 16366–16393). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.862

[57] Nie, J., Liu, G., Li, C., & Zou, T. (2026). Evidence-constrained incident visualization cards for distributed cloud logs: A UI/UX framework for turning Hadoop, OpenStack, and ZooKeeper logs into actionable SRE design interfaces. *International Journal of Graphic Design, 4*(1), 179–185. https://doi.org/10.51903/ijgd.v4i1.3703

[58] Radermacher, M. (2025). *SmolLM2-135M-multilingual-base-GGUF* [Quantized language model]. Hugging Face. https://huggingface.co/mradermacher/SmolLM2-135M-multilingual-base-GGUF

[59] Singh, S., Romanou, A., Fourrier, C., Adelani, D. I., Ngui, J. G., Vila-Suero, D., Limkonchotiwat, P., Marchisio, K., Leong, W. Q., Susanto, Y., Ng, R., Longpre, S., Ruder, S., Ko, W.-Y., Bosselut, A., Oh, A., Martins, A., Choshen, L., Ippolito, D., . . . Hooker, S. (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)* (pp. 18761–18799). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.919

[60] Su, W., Chen, S., & Qian, E. (2026). Narrative-aware scientific claim verification agent with evidence ranking for ClimateCheck. *Journal of Technology Informatics and Engineering, 5*(1), 327–340. https://doi.org/10.51903/jtie.v5i1.549

[61] Su, W., Chen, S., & Zhao, C. (2025). Budgeted multi-hop retrieval agent for compositional question answering: A retrieval-policy evaluation on the official MultiHop-RAG benchmark. *Journal of Technology Informatics and Engineering, 4*(3), 649–662. https://doi.org/10.51903/jtie.v4i3.543

[62] Su, W., Rao, H., & Ma, E. (2026). Privacy and data-integrity risk cards for LLM agents: A UI/UX design framework for secure human oversight under prompt-injection attacks. *International Journal of Graphic Design, 4*(1), 186–191. https://doi.org/10.51903/ijgd.v4i1.3699

[63] Sun, X., Lu, Y., & Chen, J. (2023). Controllable long-term user memory for multi-session dialogue: Confidence-gated writing, time-aware retrieval-augmented generation, and update/forgetting. *Journal of Advanced Computing Systems, 3*(8), 9–24. https://doi.org/10.69987/JACS.2023.30802

[64] Sun, X., Zhong, Z. S., & Wu, Q. (2026). Retrieval-grounded HDFS log anomaly detection and deterministic failure narrative generation. *Journal of Computer Systems and Applications, 3*(1), 15–30. https://doi.org/10.64229/j6d7fr94

[65] Tu, H., Zhao, S., & Zhou, A. (2025). Visual brief cards for advertising design: A structured UI/UX framework for turning creative intentions into graphic design decisions. *International Journal of Graphic Design, 3*(1), 210–226. https://doi.org/10.51903/ijgd.v3i1.3714

[66] Wu, Q., Meng, S., & Zhao, J. (2025). Text-grounded LLM-assisted design rationale interfaces: Turning advertising layout metadata into explainable UI/UX decision cards. *International Journal of Graphic Design, 3*(1), 216–240. https://doi.org/10.51903/ijgd.v3i1.3713

[67] Xin, Q. (2025a). Explaining OpenStack failure-injection log anomalies with retrieved normal prototypes. *Emerging Information Science and Technology, 6*(2). https://doi.org/10.18196/eist.v6i2.31232

[68] Xin, Q. (2025b). Uncertainty-aware late fusion for 3D perception (confidence calibration + fusion rule learning). *Journal of Technology Informatics and Engineering, 4*(1), 215–238. https://doi.org/10.51903/jtie.v4i1.485

[69] Xin, Q. (2026a). Auditable automated essay scoring and formative feedback: A rubric-grounded pipeline for secondary and higher education. *Journal of Applied Artificial Intelligence in Education, 2*(1), 1–19. https://doi.org/10.66053/jaaie.v2i1.348

[70] Xin, Q. (2026b). Behavior retrieval plus response generation for interpretable conversational personalized recommendation. *International Journal of Electrical, Energy and Power System Engineering, 9*(2), 120–136. https://doi.org/10.31258/ijeepse.9.2.120-136

[71] Xin, Q. (2026c). Early-warning analytics with LLM intervention rationales for student retention decisions: Classroom interaction modeling with xAPI-Edu-Data and dropout/success prediction. *Interdisciplinary Journal of Pedagogy and Research in Media Technology, 2*(1). https://doi.org/10.64268/inspire.v2i1.117

[72] Xin, Q. (2026d). Explainable and fair credit risk scoring with counterfactual explanations: A reproducible evaluation on the German Credit dataset (HELOC-motivated). *Journal of Information and Technology, 14*(2). https://doi.org/10.32664/j-intech.v14i02.2228

[73] Xin, Q. (2026e). Host-based intrusion detection with system call sequences: Window localization and forensic narratives. *AVITEC, 8*(2), 325–334. https://doi.org/10.28989/avitec.v8i2.3973

[74] Xin, Q. (2026f). Log anomaly detection with conformal alert control and evidence-grounded incident ticket generation. *AVITEC, 8*(2), 247–264. https://doi.org/10.28989/avitec.v8i2.3974

[75] Xin, Q. (2026g). Self-supervised customer representation learning for segmentation and next-purchase prediction on UCI Online Retail. *Journal of Information and Technology, 14*(1), 20–37. https://doi.org/10.32664/j-intech.v14i01.2229

[76] Xin, Q. (2026h). Self-supervised log anomaly detection with LogBERT-style transformers: Full empirical evaluation on a reproducible SynHDFS benchmark. *Journal of Electrical Engineering and Computer Science, 11*(1), 23–35. https://doi.org/10.54732/jeecs.v11i1.3

[77] Xu, H., Chen, Y., & Med, A. (2025). Automatic detection and explanation of dark patterns from interface microcopy: Empirical comparison of BERT-style encoders, RoBERTa-style encoders, and LLM-style decoders on the ec-darkpattern dataset. *Journal of Technology Informatics and Engineering, 4*(3), 590–612. https://doi.org/10.51903/jtie.v4i3.491

[78] Yang, Y., Dan, S., Roth, D., & Lee, I. (2026). On calibration of multilingual question answering LLMs. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=4klghu2PTj

[79] Ye, T., Chang, X., & Zhong, E. (2025). Uncertainty-aware breast ultrasound explanation cards: A visual communication framework for image-based AI diagnostic support using BreastMNIST_224. *International Journal of Graphic Design, 3*(2), 365–380. https://doi.org/10.51903/ijgd.v3i2.3701

[80] Ye, T., Mu, J., & Hunter, J. (2026). Off-policy evaluation and conservative policy selection for slot-level dynamic bidding and ranking on the Open Bandit Dataset (Small). *Journal of Technology Informatics and Engineering, 5*(1), 178–199. https://doi.org/10.51903/jtie.v5i1.503

[81] Zhang, B., Rao, H., & Zhao, D. (2024). Evidence-grounded RAG for cloud-native DevOps: Hallucination-resistant AIOps question answering over private operations documents. *Journal of Advanced Computing Systems, 4*(3), 109–125. https://doi.org/10.69987/JACS.2024.40308

[82] Zhang, B., Ren, Y., & Zou, J. (2025). LLM-style explainable e-commerce recommendation cards: A UI/UX design framework for trust-calibrated product recommendation. *International Journal of Graphic Design, 3*(2), 381–396. https://doi.org/10.51903/ijgd.v3i2.3697

[83] Zhang, J. (2025). From general human activity recognition to volleyball-oriented wearable transfer learning: Cross-dataset evidence from UCI HAR and WISDM for domain adaptation and edge deployment. *Journal of Technology Informatics and Engineering, 4*(1), 263–283. https://doi.org/10.51903/jtie.v4i1.524

[84] Zhang, J. (2026). Early warning, grade prediction, and teacher-facing LLM-ready explanations toward an open volleyball course: Reproducible evidence from four public education datasets. *Journal of Technology Informatics and Engineering, 5*(2), 20–44. https://doi.org/10.51903/jtie.v5i2.525

[85] Zhang, K., Chen, Y., & Qian, A. (2025). Evidence-grounded accounting disclosure review cards: A visual communication framework for LLM-style explanations over SEC financial statements and notes. *International Journal of Graphic Design, 3*(2), 395. https://doi.org/10.51903/ijgd.v3i2.3710

[86] Zhang, R., Wen, Z., Wang, C., Tang, C., Xu, P., & Jiang, Y. (2025). Quality analysis and evaluation prediction of RAG retrieval based on machine learning algorithms [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2511.19481

[87] Zhang, Y., & Zhang, H. (2025a). A therapist-facing session copilot for live counseling support: Reasoning-guided retrieval and ranking from multi-turn counseling dialogues. *Journal of Technology Informatics and Engineering, 4*(2), 464–486. https://doi.org/10.51903/jtie.v4i2.547

[88] Zhang, Y., & Zhang, H. (2025b). Visualizing the right counseling support: Evidence-linked recommendation cards for explainable mental health intake interfaces. *International Journal of Graphic Design, 3*(1), 214–229. https://doi.org/10.51903/ijgd.v3i1.3722

[89] Zhang, Y., & Zhou, Z. (2026). Strategy-aware therapist imitation for emotional support dialogues: A reproducible ESConv study for LLM response control. *Advances in Educational Technology and Psychology, 10*(2), 92–97. https://doi.org/10.23977/aetp.2026.100213

[90] Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 12697–12706). PMLR. https://proceedings.mlr.press/v139/zhao21c.html

[91] Zheng, D., & Li, C. (2024). Behavior-level jailbreak resistance via multi-stage refusal + utility preservation. *Journal of Advanced Computing Systems, 4*(1), 83–99. https://doi.org/10.69987/JACS.2024.40107

[92] Zheng, D., Li, C., & Davidson, H. (2023). Continual red-teaming for in-the-wild jailbreaks via online guardrail updates and guardrail distillation. *Journal of Advanced Computing Systems, 3*(2), 35–49. https://doi.org/10.69987/JACS.2023.30203

[93] Zheng, D., Zhang, B., & Geibel, J. (2024). VerifySafe: Toxicity-safe agent responses under adversarial prompts with evidence-based self-verification. *Journal of Advanced Computing Systems, 4*(1), 67–82. https://doi.org/10.69987/JACS.2024.40106

[94] Zhong, Z. S., & Ling, S. (2024a). Improved theoretical guarantee for rank aggregation via spectral method. *Information and Inference: A Journal of the IMA, 13*(3), Article iaae020. https://doi.org/10.1093/imaiai/iaae020

[95] Zhong, Z. S., & Ling, S. (2024b). Uncertainty quantification of spectral estimator and MLE for orthogonal group synchronization [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2408.05944

[96] Zhong, Z. S., Chen, J., Zhong, E., & Sun, X. (2025). Evidence-calibrated RAG for unanswerable question answering: Retrieval coverage, abstention calibration, and hallucination-proxy analysis on SQuAD 2.0. *Journal of Technology Informatics and Engineering, 4*(2), 502–520. https://doi.org/10.51903/jtie.v4i2.536

[97] Zhong, Z. S., Li, C., & Rao, H. (2026). Trajectory reliability prediction for generalist AI agents: Tool-use failure analysis and success forecasting on ZClawBench. *Journal of Technology Informatics and Engineering, 5*(1), 341–360. https://doi.org/10.51903/jtie.v5i1.539

[98] Zhong, Z. S., Pan, X., & Lei, Q. (2025). Bridging domains with approximately shared features. In Y. Li, S. Mandt, S. Agrawal, & E. Khan (Eds.), *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics* (Vol. 258, pp. 559–567). PMLR. https://proceedings.mlr.press/v258/zhong25a.html

[99] Zhong, Z. S., Wu, Q., & Mi, G. (2025). Uncertainty-aware medical image explanation cards: LLM-generated visual explanations for AI-assisted radiology interfaces. *International Journal of Graphic Design, 3*(2), 415–436. https://doi.org/10.51903/ijgd.v3i2.3616

[100] Zhou, B., Li, C., & Liu, L. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX design framework for rubric-based medical risk communication. *International Journal of Graphic Design, 3*(2), 365–380. https://doi.org/10.51903/ijgd.v3i2.3696