Stout in Humanities and Social Sciences (SHSS)
ISSN: 2979-2797
DOI: https://doi.org/10.61424/33995e23
Journal Homepage: https://stoutjournals.org/index.php/shss

Beyond Cultural Trivia: Calibrated Exemplar Retrieval, Mode-Aware Routing, and Selective Abstention for Cross-Cultural LLM Services

Daisy Zhu
Computer Science, Northeastern University, Boston, MA, USA
Corresponding Author: Daisy Zhu, **E-mail**: d.zhu.res@icloud.com

ARTICLE INFORMATION

Submitted: April 15, 2026
Accepted: July 13, 2026
Published: August 01, 2026
Volume: 2
Issue: 1

KEYWORDS

Large language models; cultural evaluation; CulturalBench; uncertainty calibration; selective prediction; abstention; retrieval; regional robustness.

ABSTRACT

Cross-cultural language-model services need more than high average accuracy: they need inference procedures that recognize ambiguous answer structures, communicate uncertainty, and avoid concentrating errors in underrepresented regions. This study evaluates such procedures on the 2024 public release of CulturalBench, comprising 1,227 multiple-choice questions and 4,908 option-level binary judgments from 45 countries or regions. A deterministic decision-layer test bed was used to isolate retrieval, score verification, calibration, and abstention from changes in proprietary model checkpoints. Word- and character-ngram scorers, global and country exemplar retrieval, score fusion, a score-stack verifier, an observable mode-aware router, temperature scaling, and confidence-based selective prediction were evaluated with country-stratified, question-grouped five-fold cross-validation. The text-only scorer achieved 40.51% accuracy on CulturalBench-Easy and 9.70% question-level exact match on CulturalBench-Hard. A position-only shortcut reached 53.63% and 51.75%, exposing substantial answer-position regularity. The mode-aware router increased these scores to 54.69% and 52.32%, respectively; temperature scaling preserved the predictions while reducing expected calibration error to 0.0234 on Easy and 0.0131 on Hard. At 80% coverage, the calibrated router achieved 56.62% Easy accuracy and 54.79% Hard exact match. Nevertheless, calibrated regional scores ranged from 44.09% to 70.37% on Easy and from 40.16% to 66.67% on Hard. Temperature scaling improved calibration, and abstention lowered risk at intermediate coverage, but neither high confidence nor aggregate performance established cross-cultural robustness. Position balance, multi-answer sensitivity, worst-region reporting, and group-aware coverage should therefore be treated as core requirements for evaluating cross-cultural language-model services.

1. Introduction

Large language models increasingly mediate public-facing communication, search, education, travel, and administrative assistance. A fluent response can still be socially inappropriate when it imposes a familiar cultural default. Cultural competence therefore includes recognizing local practices, allowing multiple valid behaviors, and deferring when confidence is insufficient. Studies document uneven cultural knowledge and alignment across languages and regions (AlKhamissi et al., 2024; Naous et al., 2024; Zhao et al., 2024).

CulturalBench tests cultural knowledge with human-authored and human-verified questions from 45 countries or regions. Easy poses one four-option choice; Hard converts the candidates into four True/False judgments and requires every judgment to be correct (Chiu et al., 2024). The latter exposes mode-seeking when several options are valid. In the 2024 release, 86 of 1,227 questions contain multiple gold-true options, creating a focused test of systems that collapse variable practices into one canonical answer.

Copyright: © 2026 The Author(s). This open access article is distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC- BY-NC-SA) License (https://creativecommons.org/licenses/by-nc-sa/4.0/), published by Stout Journals, London, United Kingdom

Three problems remain. Average accuracy can reward answer-position shortcuts. Raw confidence may not represent empirical correctness (Guo et al., 2017). Abstention can lower aggregate error while withholding service unevenly across groups. Selective prediction must therefore be assessed with regional coverage and worst-region outcomes (El-Yaniv & Wiener, 2010; Sagawa et al., 2020).

Retrieval and verification help only when their roles are defined precisely. Retrieval-augmented generation usually conditions a generator on retrieved evidence passages (Lewis et al., 2020). CulturalBench contains no passages or URLs, so this experiment tests labeled training-fold exemplar retrieval. Chain-of-verification methods generate intermediate checks; the present test bed instead measures agreement among independent scorers and an observable answer-mode signal (Dhuliawala et al., 2024).
The study asks how text scoring, exemplar retrieval, score fusion, and routing compare across the paired formats; how calibration and abstention change probability quality and selective risk; and whether weighted gains extend to country-macro, worst-country, and macro-region results. It reports out-of-fold predictions, proper scores, paired tests, risk–coverage curves, and audits of position and multi-answer effects. Cross-cultural evaluation should reward supported decisions, not frequent positions expressed with high confidence.

2. Literature Review

Cultural awareness is not a single attribute. Country labels can organize evaluation but cannot represent every identity, generation, subculture, or situation. Value benchmarks measure national differences, whereas norm-oriented evaluation asks whether judgments adapt to scenario context (Rao et al., 2025; Zhao et al., 2024). Group-stratified results are therefore useful without treating a country score as an essential cultural property.

Benchmarks operationalize complementary capabilities. BLEnD evaluates everyday knowledge across cultures and languages (Myung et al., 2024); NormAd tests contextual norm adaptation (Rao et al., 2025); CulturalBench pairs Easy and Hard knowledge formats (Chiu et al., 2024); and WorldValuesBench measures multicultural value awareness (Zhao et al., 2024). Their different targets show why no single score captures cultural capability.

Bias studies reinforce the need for broad coverage. Naous et al. (2024) documented differences between Arab and Western contexts, while AlKhamissi et al. (2024) found that prompting language and demographic conditioning affect measured alignment. CultureBank organizes community-contributed descriptions of practices and norms (Shi et al., 2024). Evaluation should therefore distinguish factual selection, contextual adaptation, confidence, and regional robustness.

Context-sensitive service design also requires more than static country labels. Dialogue memory can control what is retained and when it is forgotten, while federated topic-preference learning supports personalization without centralizing every user signal (Kuo et al., 2025; Sun et al., 2023). High-stakes access systems have paired retrieval and ranking with humanitarian and therapist-facing workflows (Y. Chen & Xu, 2026; Y. Zhang & H. Zhang, 2025a), while explicit strategy control and behavior retrieval make response policies more interpretable (Xin, 2026b; Y. Zhang & Z. Zhou, 2026). These systems motivate context retention and tailored access, but neither a stored preference nor a region label is evidence that a cultural claim is valid.

External knowledge offers one route to stronger responses. Genuine retrieval-augmented generation draws passages from a nonparametric memory (Lewis et al., 2020), and CultureBank illustrates a culturally relevant resource (Shi et al., 2024). Labeled training examples instead form exemplar memory. TF–IDF with cosine similarity provides a transparent retrieval scheme (Salton & Buckley, 1988), and retrieved demonstrations can support prediction when they are relevant (Rubin et al., 2022). CulturalBench supplies labels but no evidence passages.

Evidence-bearing retrieval has developed along several complementary dimensions. Private runbooks, bilingual incident records, and retrieval–summary integration show how a system can retrieve source material rather than merely similar labels (Y. Li, 2024; Liu et al., 2024; B. Zhang et al., 2024). Financial, legal, and humanitarian applications further constrain retrieval to domain-relevant evidence (Y. Chen & Xu, 2026; Y. Chen et al., 2025; J. Li & Zhou, 2026). Retrieval-policy research evaluates budgeted multi-hop search, hybrid reranking, and retrieval-quality prediction (Meng et al., 2026; Su et al., 2025; R. Zhang et al., 2025). Work on unanswerable questions and deterministic evidence chains then makes retrieval coverage, abstention, and provenance explicit (Jin, 2025b; Zhong, Chen, et al., 2025).

Output-side provenance is equally important. Word-level hallucination detection and accounting disclosure cards expose the sources behind individual claims (C. Li et al., 2024; K. Zhang et al., 2025), while numerical and accounting agents add domain-specific constraints (Z. Li et al., 2026; S. Zhou et al., 2026). Evidence ranking, retrieved normal prototypes, and deterministic failure narratives illustrate traceable verification in scientific and operational settings (Su, Chen, et al., 2026; Sun et al., 2026; Xin, 2025a). Review-grounded recommendation and command-safe DevOps copilots extend the same principle from explanation to

action (Chang et al., 2026; B. Zhang et al., 2026), while incident visualization cards keep evidence inspectable during human handoff (Nie et al., 2026).

Verification provides another control point. Chain-of-Verification asks explicit checking questions (Dhuliawala et al., 2024); RARR retrieves support before revising statements (Gao et al., 2023); and SelfCheckGPT detects inconsistency across sampled responses (Manakul et al., 2023). These generative procedures motivate independent checking but do not describe the tested algorithm. Here, a supervised score stack follows stacked generalization (Wolpert, 1992), and a mode-aware router uses the explicit multi-statement instruction.

Verification also includes adversarial and trajectory-level reliability. Continual red-teaming, evidence-based self-verification, and utility-preserving refusal test whether a guardrail remains useful under attack (Zheng & Li, 2024; Zheng et al., 2023; Zheng et al., 2024). Privacy and data-integrity cards and rubric-grounded scoring expose guardrail decisions for human review (Su, Rao, et al., 2026; Xin, 2026a). A broader audit analogy comes from operations and security: attack-chain stages, cost-aware routing, multi-source root-cause attribution, localized forensic narratives, and tool-use trajectory prediction make failures inspectable rather than reducing reliability to one confidence value (Bai et al., 2026; C. Li et al., 2026; Liu et al., 2023; Xin, 2026e; Zhong et al., 2026).
Calibration matters because abstention depends on confidence ranking. The Brier score evaluates the full predictive distribution (Brier, 1950; Gneiting & Raftery, 2007), ECE summarizes confidence–accuracy gaps, and held-out temperature scaling preserves predicted classes (Guo et al., 2017). Because fixed-bin ECE depends on binning, it should be read with reliability diagrams and proper scores (Vaicenavicius et al., 2019).

Cross-domain selective-decision studies provide methodological precedents, not evidence about cultural knowledge. Credit-risk and screening workflows combine probability calibration with rejection, distribution-free uncertainty, explanations, or fairness (Y. Chen et al., 2023; Jin, 2025a; Jin et al., 2024; Xin, 2026d). Infrastructure forecasting likewise uses calibrated uncertainty, conformal demand envelopes, and operational risk curves (S. Chen et al., 2024; He et al., 2023; Zhao et al., 2025). Perception, brain–computer-interface, and medical vision–language studies separate scores from decision confidence (Lu & Zou, 2026; Wu, Mi, et al., 2025; Xin, 2025b), while evidence-calibrated retrieval and conformal alert control connect uncertainty directly to abstention or alerting (Xin, 2026f; Zhong, Chen, et al., 2025).

Calibration alone does not remove distribution shift. Cross-cloud transfer and few-shot cold-start forecasting evaluate whether a decision rule survives a new workload or topology (He et al., 2024; He et al., 2025). Few-shot medical pretraining, edge-oriented activity transfer, quality-model distillation, and approximately shared-feature analysis offer related strategies for learning under limited target-domain supervision (Mi et al., 2025; J. Zhang, 2025; Zhong, Pan, et al., 2025; B. Zhou, Wang, et al., 2025). Operational and educational studies further link predictions to workload explanations or teacher-facing rationales (He et al., 2026; J. Zhang, 2026; Xin, 2026g). Theoretical work on rank aggregation and estimator uncertainty reinforces the need to quantify, rather than assume, stability (Zhong & Ling, 2024a; Zhong & Ling, 2024b).

Selective prediction trades coverage for lower error among accepted cases (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017). Unequal group confidence can make aggregate risk reduction geographically uneven. Worst-group analysis complements weighted averages by emphasizing the weakest predefined group (Sagawa et al., 2020). CulturalBench therefore warrants country-macro, worst-country, country-gap, minimum-coverage, and macro-region results.

When a system may abstain, the interface must make the reason visible. Medical explanation and safety cards combine uncertainty, evidence, and patient-facing risk communication (C. Li et al., 2025; Ye et al., 2025; Zhong, Wu, et al., 2025; B. Zhou, Li, et al., 2025). Scientific-search, product-recommendation, and counseling cards similarly expose ranking or recommendation evidence (Jin, 2025c; B. Zhang et al., 2025; Y. Zhang & H. Zhang, 2025b). Financial and infrastructure dashboards translate forecast risk into a visual hierarchy for review (Z. Li et al., 2025; Liu et al., 2025). These designs do not establish calibration by themselves; they make calibrated evidence and abstention legible to the person affected.

Related interface research clarifies how explanations become actionable. Text-grounded design rationales, visual briefs, and design-critic systems turn metadata or creative intent into editable decisions (Y. Li et al., 2025; Mu et al., 2026; Tu et al., 2025; Wu, Meng, et al., 2025). Dark-pattern detection shows why wording and presentation must also be audited for manipulative cues (Xu et al., 2025). For higher-consequence handoffs, incident and privacy-risk cards keep provenance, constraints, and escalation paths visible (Nie et al., 2026; Su, Rao, et al., 2026).

Evaluation must finally connect predictions to decision consequences. Uplift modeling and offline counterfactual evaluation distinguish predictive fit from the value of the policy selected from it (Mu et al., 2023; Mu et al., 2025). Privacy-safe budget optimization and profit-aware admission control add explicit resource and privacy constraints (Bai & Wu, 2026; Zhao et al., 2026).

In education, rubric-grounded scoring and early-warning rationales make the intervention pathway inspectable (Xin, 2026a; Xin, 2026c). Results from these domains do not transfer numerically to culture, but the evaluation principle does: report who receives a decision, who is deferred, and what cost or benefit follows.

3. Methodology

3.1 Research Design and Data

The experiment used the two 2024 CulturalBench CSV files and retained all records. Easy contains 1,227 four-option questions with letter labels and countries. Hard contains 4,908 binary option judgments for the same 1,227 identifiers. Thus, the files contain 6,135 rows, not 6,135 distinct questions, and match the 2024 release described by Chiu et al. (2024).

Hard identifies 1,141 single-mode and 86 multi-mode questions, with 1,326 positive and 3,582 negative judgments. A blank Spanish distractor was encoded as an empty token to preserve four-option structure. A Mexican prompt occurring under two identifiers was co-folded, and a repeated option in one Spanish Hard question was retained.

Countries were mapped to the benchmark's 11 macro-regions: North America, South America, Eastern Europe, Southern Europe, Western Europe, Africa, Middle East/West Asia, South Asia, Southeast Asia, East Asia, and Oceania. Because the CSVs contain neither topic labels nor evidence passages, the analysis reports neither topic-specific results nor external-evidence performance.

Table 1. Dataset profile for the paired 2024 CulturalBench files

Characteristic	Value	Unit
Easy questions	1227	question
Hard option judgments	4908	judgment
Unique countries/regions	45	country/region
Macro-regions	11	region
Single-mode questions	1141	question
Multi-mode questions	86	question
Hard positive labels	1326	judgment
Hard negative labels	3582	judgment
Easy blank option cells	1	cell
Hard blank option cells	1	cell
Repeated Easy question texts	2	row
Repeated Hard question-option pairs	2	row

Note. The Easy and Hard files share the same 1,227 question identifiers. Hard contains four option judgments per question.

3.2 Validation Protocol

Five outer folds were assigned at question level. Within each country, normalized prompts were ordered by a seeded hash and distributed round-robin, preserving country representation, co-folding duplicate prompts, and keeping each question's four Hard options together. In every iteration, three folds fitted base scorers, one fitted fusion, verification, and calibration, and one served as test data. The calibration fold was deterministically split for verifier and temperature fitting. Every question was tested once.

The seed was 20260729. One script regenerated vectorizers, classifiers, routing, calibration, metrics, and plots, while the run manifest recorded source checksums. The design prevents within-question leakage and test-label tuning.

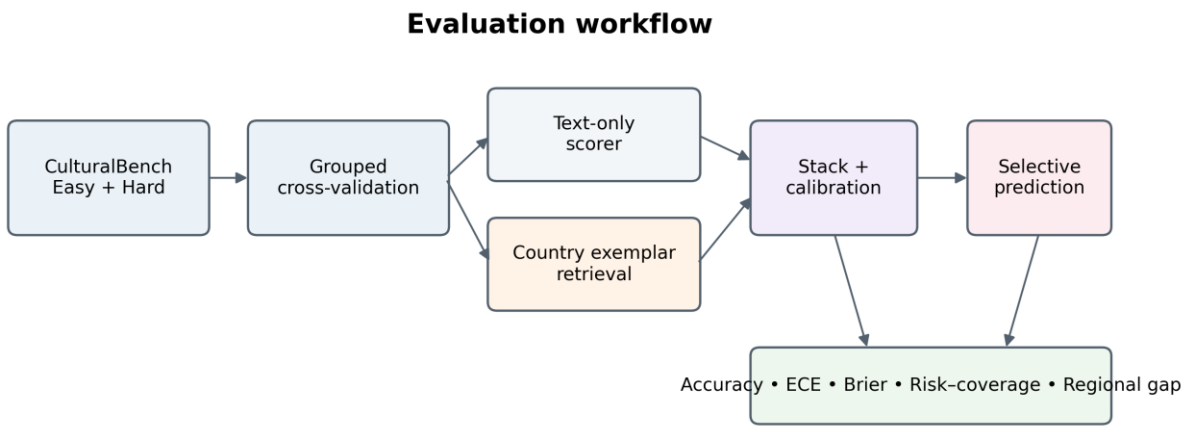

Figure 1. Question-grouped five-fold experimental workflow. Base fitting, verifier fitting, temperature fitting, and testing used disjoint question folds.

Table 2. Question-grouped five-fold evaluation protocol

Component	Specification
Evaluation unit	Question-grouped five-fold outer evaluation
Fold allocation	Country-stratified deterministic round-robin; duplicate prompts co-folded
Training/calibration/test	Three folds / one fold / one fold per outer iteration
Text-only scorer	Word TF-IDF (1–2 grams) + class-balanced logistic regression
Character scorer	Character TF-IDF (3–5 grams) + class-balanced logistic regression
Retrieval	Cosine TF-IDF, similarity-squared weighted vote, k=11
Score fusion	Calibration-fold grid search over text-scorer/retrieval weight
Score-stack verifier	Logistic stack over four scores, agreement deltas, and retrieval similarity
Mode-aware router	Observable Easy instruction routes multi-mode items to the character scorer and others to the position baseline
Calibration	Held-out temperature scaling for Easy and Hard probabilities
Selective prediction	Rank by maximum Easy confidence or minimum Hard option confidence
Random seed	20260729
Bootstrap resamples	2000

Note. All tuning and calibration operations used held-out folds; each question contributed to the test set once.

3.3 Decision-Layer Configurations

The position-only shortcut estimated the training-fold positive rate for each answer position. For Easy it yielded a four-class position distribution; for Hard it supplied a binary probability to each of the four positions. This deliberately weak baseline measured how much performance could be obtained without reading the cultural content.

The text-only scorer represented country, question, option letter, and candidate text with word TF–IDF unigrams and bigrams, followed by class-balanced logistic regression. The character scorer used boundary n-grams of length three to five with a separate logistic model. TF–IDF weighting followed Salton and Buckley (1988).

Global exemplar retrieval used cosine similarity over training question-option texts. The 11 nearest labeled examples contributed similarity-squared votes with prior smoothing. Country exemplar retrieval restricted this memory to the query country. Score

fusion combined text-only and country-exemplar probabilities, selecting its weight from 0.00 to 1.00 in 0.05 increments on calibration data. This follows retrieved-demonstration prediction (Rubin et al., 2022), but it is not open-corpus RAG.

The score-stack verifier fitted logistic regression to word, character, two exemplar, position, disagreement, and similarity signals (Wolpert, 1992). Every multi-mode Easy prompt, and no single-mode prompt, contains an instruction to select all appropriate statements. The mode-aware router therefore used the character scorer when this visible signal was present and the position scorer otherwise. The paired signal, never a Hard gold label, determined the Hard route.

Held-out temperature scaling divided Easy log probabilities or Hard logits by a positive temperature and preserved predicted labels. Selective prediction ranked Easy questions by maximum class probability and Hard questions by minimum confidence across four option judgments. Coverage from 100% to 50% was evaluated without changing predictions.

3.4 Metrics and Statistical Analysis

Easy evaluation reported accuracy, macro-F1, multiclass log loss, the unnormalized multiclass Brier loss, ten-bin ECE, country-macro accuracy, worst-country accuracy, and the best-minus-worst country gap. Hard option-level evaluation reported accuracy, balanced accuracy, macro-F1, area under the receiver operating characteristic curve, log loss, binary Brier loss, ECE, and the predicted positive rate.

Hard question-level exact match required all four option judgments to match the gold vector. This is stricter than option accuracy; a random balanced binary classifier has expected question exact match of 0.5 to the fourth power, or 6.25%. Jaccard similarity, set F1, and absolute error in the number of true options captured partial performance on multi-answer questions. Regional robustness used weighted, country-macro, worst-country, and gap measures. Selective risk equaled one minus accuracy among accepted questions, and minimum country coverage measured whether global abstention concentrated rejection in any country.

Calibration followed Guo et al. (2017), while Brier loss followed the proper-score convention of Brier (1950) and Gneiting and Raftery (2007). Fixed-bin ECE was interpreted jointly with reliability plots because its value is bin-dependent (Vaicenavicius et al., 2019). Following significance-testing guidance for natural-language processing experiments (Dror et al., 2018), paired accuracy differences were tested with exact McNemar tests, which use only discordant question outcomes (McNemar, 1947). Question-level percentile confidence intervals were estimated from 2,000 bootstrap resamples with replacement (Efron & Tibshirani, 1994). Statistical significance was assessed at two-sided $p < .05$.

4. Results and Discussion

4.1 Dataset Structure and Shortcut Exposure

The answer distribution was strongly position-dependent. Easy labels were A for 53.63% of questions, B for 20.70%, C for 15.40%, and D for 10.27%. In Hard, the positive rates by option position were 56.89%, 22.74%, 17.20%, and 11.25%; they sum to more than 100% because multi-mode questions can contain several positive options. Consequently, the position-only shortcut obtained 53.63% Easy accuracy and 51.75% Hard exact match. This performance did not measure cultural knowledge. It measured the regularity of answer placement and established a necessary adversarial baseline.

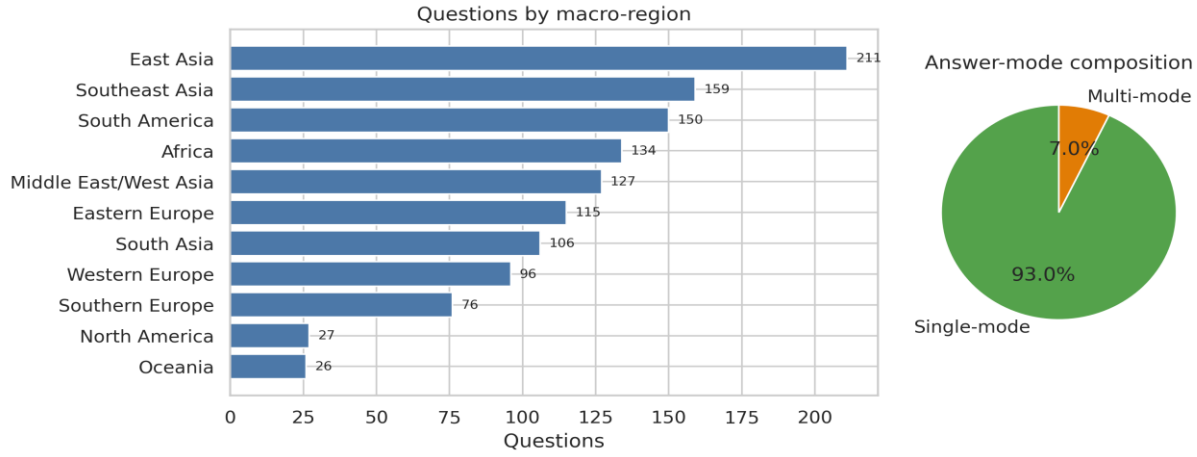

Figure 2. Dataset composition. Panel A shows question counts by macro-region; Panel B shows the single-mode and multi-mode question shares.

Figure 2 also shows the coverage imbalance that affects regional uncertainty. East Asia contributed 211 questions, while Oceania and North America contributed only 26 and 27. Country counts were still smaller in several cases. Weighted scores are therefore dominated by larger groups, whereas worst-country estimates have higher sampling variance. Reporting both is more informative than treating either as sufficient.

4.2 Overall Predictive Performance

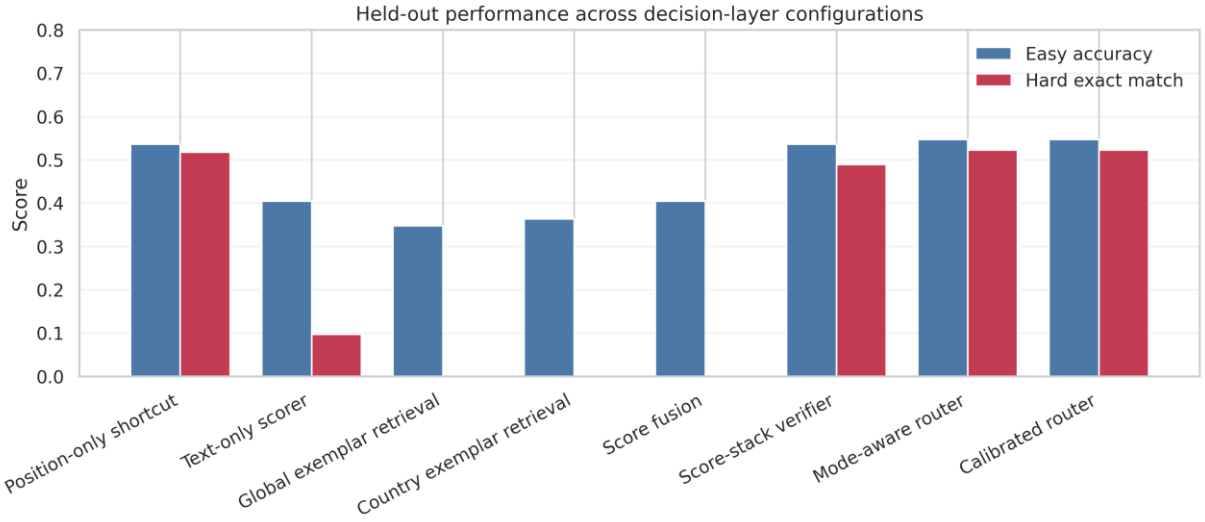

Figure 3. Main out-of-fold performance comparison. Easy uses accuracy; Hard uses question-level exact match.

Table 3. Out-of-fold performance on CulturalBench-Easy

Method	Accuracy	Macro-F1	Log loss	Brier	ECE
Position-only shortcut	0.5363	0.1745	1.1849	0.6363	0.0001
Text-only scorer	0.4051	0.3451	1.3458	0.7292	0.1230
Character scorer	0.4914	0.3378	1.3211	0.7165	0.2067
Global exemplar retrieval	0.3480	0.2922	1.3603	0.7363	0.0705
Country exemplar retrieval	0.3635	0.3085	1.3698	0.7413	0.0981
Score fusion	0.4051	0.3451	1.3458	0.7292	0.1230
Score-stack verifier	0.5363	0.1745	1.2182	0.6574	0.1436
Mode-aware router	0.5469	0.2220	1.1718	0.6286	0.0298
Calibrated router	0.5469	0.2220	1.1696	0.6275	0.0234

Note. Brier is the unnormalized multiclass Brier loss. ECE uses 10 equal-width confidence bins.

On Easy, the text-only word scorer reached 40.51% accuracy and 0.345 macro-F1. The character scorer performed better at 49.14% accuracy, while global and country exemplar retrieval reached 34.80% and 36.35%. Score fusion selected the word scorer exclusively in every Easy outer fold, so its result was identical to the text-only condition. The score-stack verifier reproduced the 53.63% position decision. The mode-aware router reached 54.69% accuracy and 0.222 macro-F1; calibration left these predictions unchanged while reducing log loss from 1.1718 to 1.1696, Brier loss from 0.6286 to 0.6275, and ECE from 0.0298 to 0.0234.

The low macro-F1 of the position-dominated methods is important. Their overall accuracy was high because they selected A frequently, but their performance across the four labels was uneven. The character scorer's lower accuracy but higher macro-F1 illustrates a different trade-off: it used more of the text signal and predicted a broader label set, yet it could not overcome the benchmark's strong positional prior. Thus, Easy accuracy alone would favor a method with limited answer diversity.

Table 4. Out-of-fold option-level performance on CulturalBench-Hard

Method	Accuracy	Balanced acc.	Macro-F1	AUC	Brier	ECE
Position-only shortcut	0.7643	0.6894	0.6938	0.7229	0.1660	0.0000
Text-only scorer	0.5952	0.5573	0.5428	0.5882	0.2364	0.2055
Character scorer	0.6353	0.6012	0.5842	0.6495	0.2280	0.2031
Global exemplar retrieval	0.7294	0.5002	0.4232	0.5519	0.1953	0.0075
Country exemplar retrieval	0.7298	0.5000	0.4219	0.5358	0.1962	0.0064
Score fusion	0.7296	0.5001	0.4226	0.5541	0.1958	0.0218
Score-stack verifier	0.7545	0.6905	0.6898	0.7439	0.2035	0.1975
Mode-aware router	0.7622	0.6958	0.6969	0.7462	0.1615	0.0227
Calibrated router	0.7622	0.6958	0.6969	0.7493	0.1611	0.0131

Note. AUC denotes area under the receiver operating characteristic curve. ECE uses 10 bins.

At the Hard option level, the calibrated router achieved 76.22% accuracy, 69.58% balanced accuracy, 0.697 macro-F1, and 0.749 area under the curve. Its predicted positive rate was 26.55%, close to the 27.02% gold rate. Temperature scaling reduced ECE from 0.0227 to 0.0131 and Brier loss from 0.1615 to 0.1611 while preserving all thresholded predictions. The position shortcut retained slightly higher raw option accuracy, 76.43%, but had lower balanced accuracy, macro-F1, and area under the curve. The comparison shows why raw option accuracy is insufficient under class and position imbalance.

Table 5. Out-of-fold question-level performance on CulturalBench-Hard

Method	Exact match	Jaccard	Set F1	Count MAE	Country macro
Position-only shortcut	0.5175	0.5416	0.5501	0.0807	0.5144
Text-only scorer	0.0970	0.2357	0.2975	1.2445	0.0941
Character scorer	0.1532	0.2991	0.3611	1.1312	0.1531
Global exemplar retrieval	0.0000	0.0006	0.0009	1.0807	0.0000
Country exemplar retrieval	0.0000	0.0000	0.0000	1.0807	0.0000
Score fusion	0.0008	0.0008	0.0008	1.0782	0.0004
Score-stack verifier	0.4890	0.5380	0.5552	0.1589	0.4879
Mode-aware router	0.5232	0.5437	0.5504	0.0856	0.5218
Calibrated router	0.5232	0.5437	0.5504	0.0856	0.5218

Note. Exact match requires all four option judgments for a question to be correct.

Question-level Hard scoring produced a sharper distinction. The text-only scorer achieved 9.70% exact match, and the character scorer achieved 15.32%. Pure exemplar retrieval predicted almost no positive options: global exemplar retrieval produced a 0.12% positive rate and zero exact-match questions, while country exemplar retrieval produced no positive predictions and zero exact match. Similar country labels and lexical overlap were not enough to recover the relevant cultural fact, and score fusion also produced near-all-negative decisions. In this setup, nearest in-domain examples did not supply the cultural facts needed for accurate Hard decisions.

The failure is consistent with the distinction between label-neighbor retrieval and evidence-bearing retrieval: source attribution and retrieval-coverage checks can support a claim only when the memory contains answer-relevant material (C. Li et al., 2024; Zhong, Chen, et al., 2025).

The score-stack verifier recovered 48.90% exact match. The mode-aware router reached 52.32%, slightly exceeding the 51.75% position shortcut, and calibration preserved that value. The routing gain was 0.57 percentage points with a bootstrap 95% interval of 0.16 to 1.06 points and an exact McNemar p value of .0156. Its mean Jaccard score was 0.5437 and mean set F1 was 0.5504. Although these are the best tested question-level results, most of the absolute score still came from the position structure. The appropriate interpretation is improved routing within this benchmark, not demonstrated cultural mastery.

4.3 Multi-Answer Sensitivity

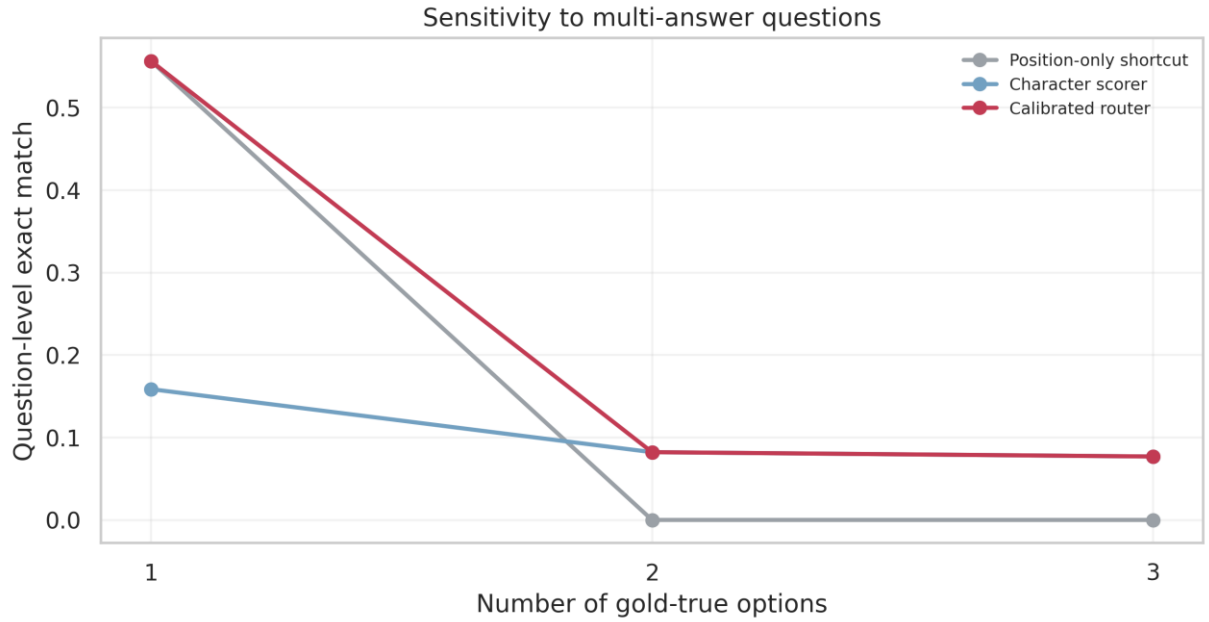

Figure 4. Sensitivity to answer mode. Multi-mode questions contain two or three gold-true options in the Hard representation.

Table 6. Performance by single-mode and multi-mode answer structure

Task	Method	Mode	N	Primary score	Secondary score
Easy	Position-only shortcut	Multi-mode	86	0.2674	—
Easy	Position-only shortcut	Single-mode	1,141	0.5565	—
Hard	Position-only shortcut	Multi-mode	86	0.0000	0.3430
Hard	Position-only shortcut	Single-mode	1,141	0.5565	0.5565
Easy	Character scorer	Multi-mode	86	0.4186	—
Easy	Character scorer	Single-mode	1,141	0.4969	—
Hard	Character scorer	Multi-mode	86	0.0814	0.3731
Hard	Character scorer	Single-mode	1,141	0.1586	0.2935
Easy	Calibrated router	Multi-mode	86	0.4186	—
Easy	Calibrated router	Single-mode	1,141	0.5565	—
Hard	Calibrated router	Multi-mode	86	0.0814	0.3731
Hard	Calibrated router	Single-mode	1,141	0.5565	0.5565

Note. Primary score is accuracy for Easy and exact match for Hard; the Hard secondary score is Jaccard similarity.

Multi-mode questions exposed the limitation of position-based success. The shortcut obtained 55.65% Hard exact match on 1,141 single-mode questions but 0% on the 86 multi-mode questions because it predicted only the dominant position. Its multi-mode Jaccard score was 0.343, indicating partial overlap without a fully correct answer set. The character scorer achieved 8.14%

multi-mode exact match and 0.373 Jaccard, while the calibrated mode-aware route achieved the same 8.14% exact match on questions with two or three true options. For Easy, the mode-aware router combined 55.65% single-mode accuracy with 41.86% multi-mode accuracy, producing the overall 54.69%.

Routing yielded statistically small but consistent gains over the position shortcut: 1.06 percentage points on Easy (95% bootstrap interval 0.24–1.88; p = .0241) and 0.57 points on Hard. Temperature scaling changed probabilities but no labels, so the paired accuracy difference between the mode-aware and calibrated routers was exactly zero. The result confirms that visible answer structure can guide a safer route, but the residual multi-mode exact match remains low. Systems intended for culturally variable questions need explicit multi-label reasoning rather than a single-mode prior with a corrective branch.

4.4 Calibration and Selective Abstention

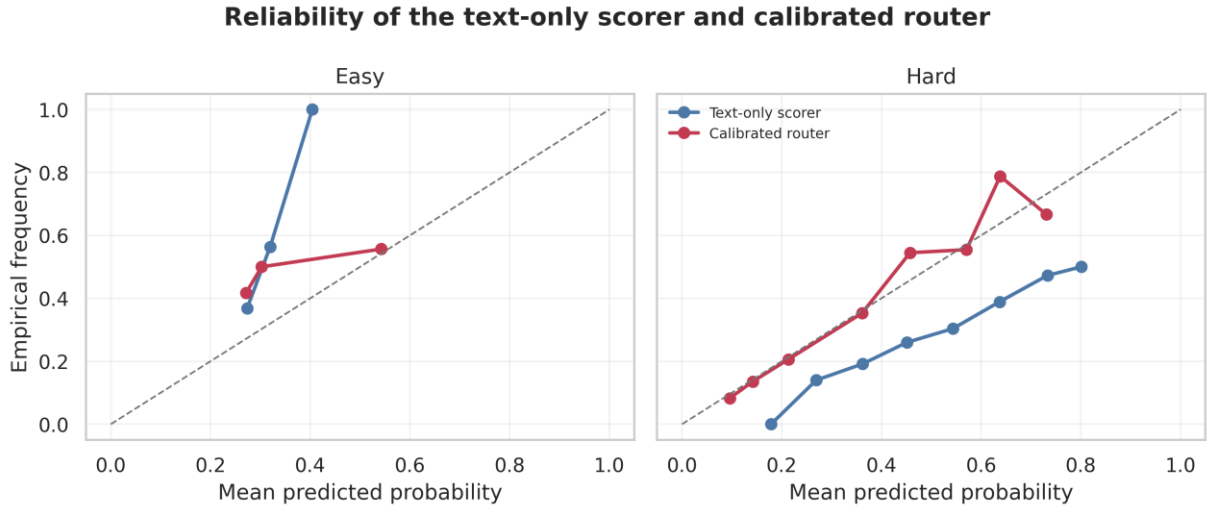

Figure 5. Reliability diagrams for the text-only scorer and calibrated router on Easy questions and Hard option judgments.

Reliability improved materially after calibration. On Easy, the text-only scorer had ECE 0.1230, whereas the calibrated router had ECE 0.0234. On Hard option judgments, the corresponding values were 0.2055 and 0.0131. Brier loss also fell from 0.7292 to 0.6275 on Easy and from 0.2364 to 0.1611 on Hard. Because the calibrated method also contains the mode-aware route and position feature, these full-pipeline differences are not attributable to temperature alone. The direct pre/post comparison is the mode-aware versus calibrated router: temperature preserved accuracy and reduced ECE by 0.0064 on Easy and 0.0096 on Hard. This pattern matches the cross-domain rationale for selective systems: probability quality must be evaluated separately from discrimination, and a rejection option is useful only when confidence tracks error (Y. Chen et al., 2023; Jin, 2025a; Jin et al., 2024).

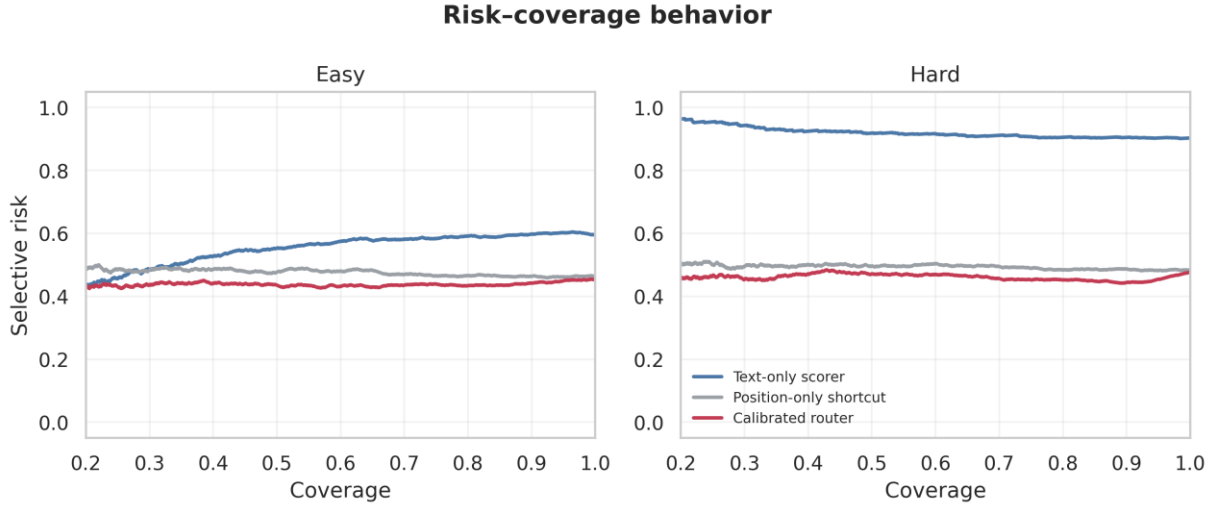

Figure 6. Risk–coverage curves for the text-only scorer and calibrated router. Lower selective risk is better.

Table 7. Selective accuracy, risk, and minimum country coverage

Task	Method	Target cov.	Selective acc.	Selective risk	Min. country cov.	Threshold
Easy	Text-only scorer	1.0000	0.4051	0.5949	1.0000	0.2500
Easy	Text-only scorer	0.8000	0.4084	0.5916	0.5789	0.2625
Easy	Calibrated router	1.0000	0.5469	0.4531	1.0000	0.2517
Easy	Calibrated router	0.9000	0.5584	0.4416	0.7143	0.5133
Easy	Calibrated router	0.8000	0.5662	0.4338	0.4286	0.5133
Easy	Calibrated router	0.7000	0.5646	0.4354	0.4286	0.5238
Easy	Calibrated router	0.6000	0.5685	0.4315	0.2857	0.5238
Easy	Calibrated router	0.5000	0.5651	0.4349	0.2727	0.5483
Hard	Text-only scorer	1.0000	0.0970	0.9030	1.0000	0.5000
Hard	Text-only scorer	0.8000	0.0947	0.9053	0.5714	0.5086
Hard	Calibrated router	1.0000	0.5232	0.4768	1.0000	0.5007
Hard	Calibrated router	0.9000	0.5557	0.4443	0.7143	0.5666
Hard	Calibrated router	0.8000	0.5479	0.4521	0.5714	0.5666
Hard	Calibrated router	0.7000	0.5437	0.4563	0.5455	0.5692
Hard	Calibrated router	0.6000	0.5305	0.4695	0.4545	0.5692
Hard	Calibrated router	0.5000	0.5293	0.4707	0.3636	0.5746

Note. Cases were ranked by held-out confidence. Minimum country coverage is the lowest retained fraction among 45 countries or regions.

Confidence ranking reduced selective risk for the calibrated pipeline, especially at moderate abstention. At full coverage, Easy accuracy was 54.69% and Hard exact match was 52.32%. At 80% coverage, Easy accuracy rose to 56.62% and Hard exact match to 54.79%; the corresponding risks were 43.38% and 45.21%. At 90% coverage, Hard exact match reached 55.57%. More aggressive abstention did not improve monotonically because confidence remained dominated by position and fold-specific structure. At 50% coverage, Easy accuracy was 56.51% and Hard exact match was 52.93%, below their best intermediate-coverage values.

The group coverage columns qualify the aggregate improvement. At 80% global coverage, the least-covered country retained only 42.86% of its Easy questions and 57.14% of its Hard questions. Thus, a single global threshold withheld substantially more cases from at least one country. A public-facing service should calibrate and audit thresholds by region or another relevant deployment group rather than assuming that global coverage is evenly distributed.

Evidence and risk-card research suggests a practical presentation layer for this audit: show the confidence basis, the acceptance or refusal decision, and the affected-group coverage together rather than displaying a decontextualized score (Jin, 2025c; Z. Li et al., 2025; B. Zhou, Li, et al., 2025).

4.5 Regional Robustness and Disparity

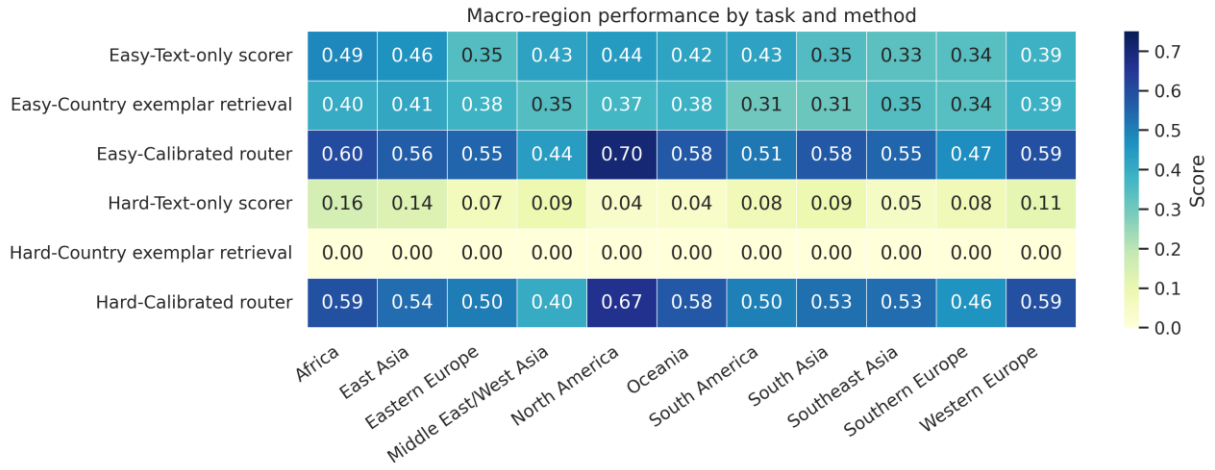

Figure 7. Macro-region performance heatmap for the evaluated decision-layer configurations.

Table 8. Calibrated-router accuracy by macro-region

Macro-region	N	Easy acc.	Hard exact
Africa	134	0.6045	0.5896
East Asia	211	0.5640	0.5355
Eastern Europe	115	0.5478	0.5043
Middle East/West Asia	127	0.4409	0.4016
North America	27	0.7037	0.6667
Oceania	26	0.5769	0.5769
South America	150	0.5133	0.5000
South Asia	106	0.5755	0.5283
Southeast Asia	159	0.5472	0.5346
Southern Europe	76	0.4737	0.4605
Western Europe	96	0.5938	0.5938

Note. Values are out-of-fold question-level accuracy for Easy and exact match for Hard.

The calibrated router's macro-region results were uneven. Easy accuracy ranged from 44.09% in Middle East/West Asia to 70.37% in North America. Hard exact match ranged from 40.16% to 66.67% across the same two regions. Africa reached 60.45% Easy accuracy and 58.96% Hard exact match; East Asia reached 56.40% and 53.55%. The ordering echoes prior CulturalBench findings that performance varies substantially by geography, but the present values describe this controlled decision layer rather than the frontier LLMs evaluated by Chiu et al. (2024).

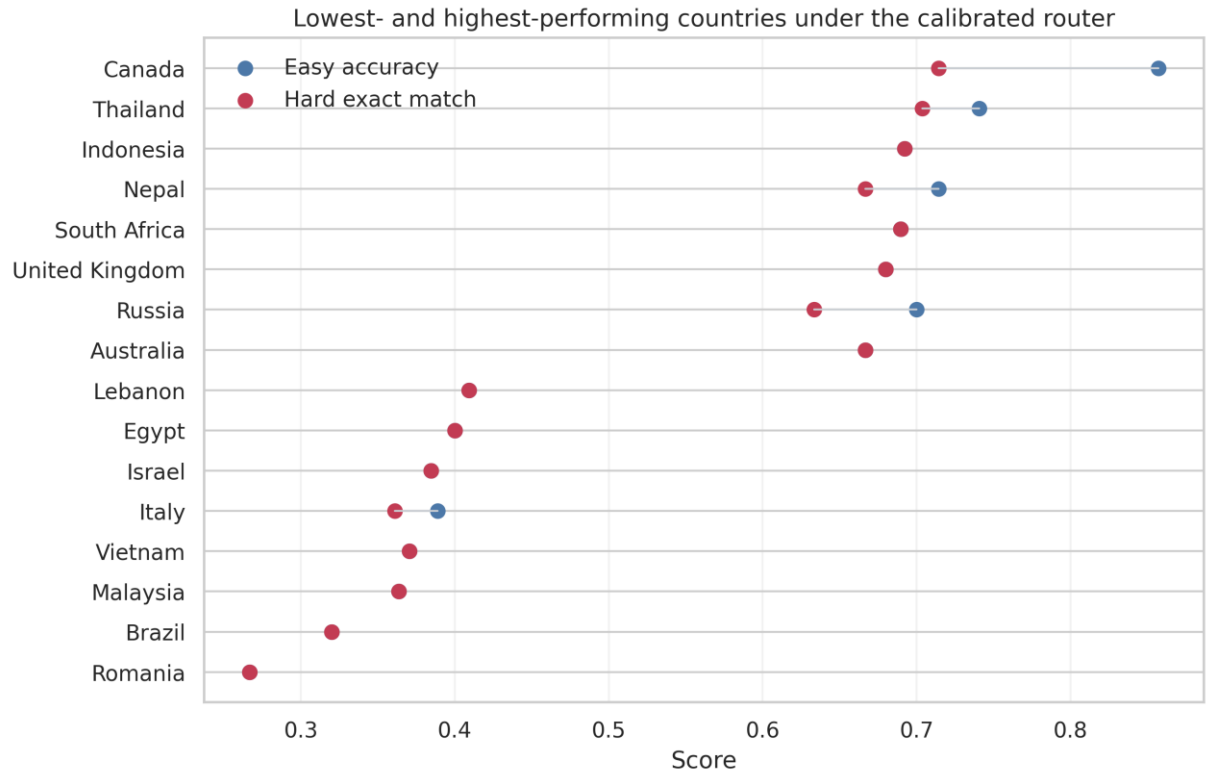

Figure 8. Country-level calibrated-router performance, ordered separately for Easy accuracy and Hard exact match.

Table 9. Country-level robustness and disparity summary

Task	Method	Weighted	Country macro	Worst country	Country gap
Easy	Position-only shortcut	0.5363	0.5348	0.2667	0.5905
Easy	Text-only scorer	0.4051	0.4127	0.2222	0.4921
Easy	Country exemplar retrieval	0.3635	0.3675	0.1538	0.5734
Easy	Mode-aware router	0.5469	0.5452	0.2667	0.5905
Easy	Calibrated router	0.5469	0.5452	0.2667	0.5905
Hard	Position-only shortcut	0.5175	0.5144	0.2667	0.4370
Hard	Text-only scorer	0.0970	0.0941	0.0000	0.2222
Hard	Country exemplar retrieval	0.0000	0.0000	0.0000	0.0000
Hard	Mode-aware router	0.5232	0.5218	0.2667	0.4476
Hard	Calibrated router	0.5232	0.5218	0.2667	0.4476

Note. Gap is the best-minus-worst country score. Weighted values reflect the observed question distribution.

Country-level measures tell a similar story. For the calibrated router, weighted Easy accuracy was 54.69%, country-macro accuracy was 54.52%, and the worst-country accuracy was 26.67%; the best-minus-worst gap was 59.05 points. On Hard, weighted exact match was 52.32%, country-macro exact match was 52.18%, the worst-country score was 26.67%, and the gap was 44.76 points. Figure 8 shows that Canada and Thailand were among the higher-scoring countries, whereas Romania, Brazil, Malaysia, Vietnam, Italy, Israel, Egypt, and Lebanon were in the lower tail for at least one format. Because some country samples are small, the rank order should be interpreted as an audit signal, not a stable cultural hierarchy.

The retrieval conditions did not close these gaps. Country exemplar retrieval had 36.75% country-macro Easy accuracy and zero Hard country-macro exact match. Restricting neighbors to the same country reduced the candidate pool but did not add factual evidence. This supports a practical design principle: a culturally scoped retrieval filter is useful only when the underlying memory contains independently relevant and answer-bearing material.

Answer-bearing documents, explicit retrieval budgets, and source-level provenance are therefore prerequisites for calling a future extension culturally grounded RAG (J. Li & Zhou, 2026; Su et al., 2025; B. Zhang et al., 2024).

4.6 Ablation and Statistical Interpretation

Table 10. Ablation of routing, verification, and calibration components

Task	Method	Primary score	ECE	Brier
Easy	Text-only scorer	0.4051	0.1230	0.7292
Easy	Score fusion	0.4051	0.1230	0.7292
Easy	Score-stack verifier	0.5363	0.1436	0.6574
Easy	Mode-aware router	0.5469	0.0298	0.6286
Easy	Calibrated router	0.5469	0.0234	0.6275
Hard	Text-only scorer	0.0970	0.2055	0.2364
Hard	Score fusion	0.0008	0.0218	0.1958
Hard	Score-stack verifier	0.4890	0.1975	0.2035
Hard	Mode-aware router	0.5232	0.0227	0.1615
Hard	Calibrated router	0.5232	0.0131	0.1611

Note. Primary score is Easy accuracy or Hard question-level exact match. Calibration preserves predicted labels.

Table 11. Paired bootstrap intervals and exact McNemar tests

Task	Baseline	Comparison	Difference	CI low	CI high	Exact p
Easy	Text-only scorer	Calibrated router	0.1418	0.1059	0.1785	< .001
Easy	Position-only shortcut	Mode-aware router	0.0106	0.0024	0.0188	.0241
Easy	Mode-aware router	Calibrated router	0.0000	0.0000	0.0000	1.0000
Hard	Text-only scorer	Calibrated router	0.4262	0.3953	0.4580	< .001
Hard	Position-only shortcut	Mode-aware router	0.0057	0.0016	0.0106	.0156
Hard	Mode-aware router	Calibrated router	0.0000	0.0000	0.0000	1.0000

Note. Confidence intervals use 2,000 question-level bootstrap resamples. P values are two-sided exact McNemar tests.

The ablation sequence clarifies which components mattered. Score fusion did not improve Easy over the text-only scorer and achieved 0.08% Hard exact match. The score-stack verifier raised Easy accuracy to 53.63% and Hard exact match to 48.90% by learning the strong positional signal. The mode-aware router added a further 1.06 and 0.57 points. Temperature scaling changed no predictions but produced the lowest ECE and Brier loss. Relative to the text-only scorer, the final pipeline improved Easy accuracy by 14.18 points (95% bootstrap interval 10.59–17.85) and Hard exact match by 42.62 points (39.53–45.80); both exact McNemar tests had $p < .001$.

These large improvements require a restrained interpretation. They demonstrate that an inference layer can exploit answer structure, verification features, and calibration more effectively than a sparse text-only scorer. They do not establish that the system acquired corresponding cultural knowledge. The strongest baseline ignored question meaning, and the mode-aware route used an explicit formatting signal. This is precisely why the shortcut baseline, multi-mode subset, macro-F1, balanced accuracy, and worst-region metrics are necessary. A result that disappears after option rebalancing or on multi-answer items should not be presented as a broad advance in cultural reasoning.

The same separation between measured utility and safety appears in adversarial guardrail and agent-trajectory evaluations, where refusal behavior, failure traces, and downstream usefulness are audited jointly (Zheng & Li, 2024; Zheng et al., 2024; Zhong et al., 2026).

The study has four bounded limitations. First, the estimates characterize the deterministic decision layer rather than any particular proprietary large-model checkpoint. Second, exemplar retrieval is not external evidence retrieval, because the CSV files provide no evidence corpus. Third, country and macro-region labels are coarse evaluation groups rather than complete representations of culture. Fourth, the 2024 snapshot contains answer-position imbalance and small country samples. These properties are part of the measured dataset and directly motivate the proposed reporting standard.

5. Conclusions

This study conducted a complete, leakage-controlled evaluation of retrieval, verification, calibration, and abstention on the paired 2024 CulturalBench files. The analysis covered all 1,227 Easy questions and all 4,908 Hard option judgments, retained source anomalies without changing record counts, and generated question-level out-of-fold predictions for every method.

The principal empirical result is not that a complex model solved cross-cultural reasoning. It is that answer structure materially changes what a score means. A position-only shortcut reached 53.63% Easy accuracy and 51.75% Hard exact match, yet failed every Hard multi-mode question. The mode-aware router raised overall scores to 54.69% and 52.32% and recovered limited multi-answer performance. Temperature scaling then reduced ECE to 0.0234 and 0.0131 without altering predictions. Confidence gating lowered risk at intermediate coverage, but country coverage became uneven and macro-region gaps remained large.

For cross-cultural LLM services, four evaluation practices follow. Answer-position and format-only baselines should accompany semantic models. Hard multi-label exact match and partial set metrics should accompany Easy accuracy. Calibration should be assessed with reliability diagrams, ECE, and proper scores rather than confidence alone. Finally, abstention should report country or region coverage and worst-group outcomes, because global risk reduction can conceal unequal access. These practices turn CulturalBench from a cultural-trivia scorecard into a sharper audit of whether a language-model service is accurate, calibrated, and appropriately cautious across cultural contexts.

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] AlKhamissi, B., ElNokrashy, M., Alkhamissi, M., & Diab, M. (2024). Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 12404–12422). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.671

[2] 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 Electronics Engineering, 6*(1), 28–43. https://doi.org/10.33474/infotron.v6i1.24923

[3] Bai, J., & Wu, Q. (2026). Privacy-safe marketing mix modeling and budget optimization under identifier loss: A controlled simulation study. *International Journal of Electronics and Communications Systems, 6*(1), 83–95. https://doi.org/10.24042/ijecs.v6i1.30533

[4] 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

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

[6] 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

[7] 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

[8] 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

[9] 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

[10] Chiu, Y. Y., Jiang, L., Lin, B. Y., Park, C. Y., Li, S. S., Ravi, S., Bhatia, M., Antoniak, M., Tsvetkov, Y., Shwartz, V., & Choi, Y. (2024). CulturalBench: A robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of LLMs [Preprint]. *arXiv*. https://arxiv.org/abs/2410.02677v1

[11] Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2024). Chain-of-Verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 3563–3578). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.212

[12] Dror, R., Baumer, G., Shlomov, S., & Reichart, R. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1383–1392). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1128

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

[14] 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

[15] Gao, L., Dai, Z., Pasupat, P., Chen, A., Chaganty, A. T., Fan, Y., Zhao, V., Lao, N., Lee, H., Juan, D.-C., & Guu, K. (2023). RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 16477–16508). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.910

[16] Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4878–4887). Curran Associates. https://proceedings.neurips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html

[17] Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association, 102*(477), 359–378. https://doi.org/10.1198/016214506000001437

[18] 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* (Vol. 70, pp. 1321–1330). PMLR. https://proceedings.mlr.press/v70/guo17a.html

[19] 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

[20] He, S., Li, C., & Rao, H. (2025). Few-shot cold-start workload forecasting for new AI inference tenants with time-series foundation models. *Journal of Technology Informatics and Engineering, 4*(1), 306–324. https://doi.org/10.51903/jtie.v4i1.546

[21] He, S., Nie, J., & Li, C. (2026). Power-aware inventory planning for AI infrastructure using job-level forecasting and LLM workload explanations. *Journal of Technology Informatics and Engineering, 5*(1), 341–359. https://doi.org/10.51903/jtie.v5i1.548

[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] Jin, J. (2025a). Calibrated resume–job matching for trustworthy LLM-assisted recruiter screening: Pairwise matching, probability calibration, and selective refusal on two public recruitment datasets. *Journal of Technology Informatics and Engineering, 4*(3), 625–648. https://doi.org/10.51903/jtie.v4i3.529

[24] Jin, J. (2025b). 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

[25] Jin, J. (2025c). 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

[26] Jin, J., Huang, T., & Lu, S. (2024). 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

[27] 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

[28] 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. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459–9474). Curran Associates. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[29] 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

[30] 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

[31] 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

[32] 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

[33] 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

[34] 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

[35] 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

[36] 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

[37] 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

[38] 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

[39] 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

[40] 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

[41] Manakul, P., Liusie, A., & Gales, M. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 9004–9017). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.557

[42] 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

[43] 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

[44] Mi, G., Ye, T., & Wood, D. (2025). A lightweight medical foundation model for cross-modal multi-task pretraining and parameter-efficient few-shot transfer on MedMNIST. *Journal of Technology Informatics and Engineering, 4*(3), 572–589. https://doi.org/10.51903/jtie.v4i3.492

[45] 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

[46] 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

[47] Mu, J., Ye, T., & Patel, P. (2025). Offline counterfactual evaluation for advertising and recommendation slot policies: A reproducible study on the Open Bandit Dataset (small). *Journal of Technology Informatics and Engineering, 4*(3), 521–543. https://doi.org/10.51903/jtie.v4i3.500

[48] 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

[49] 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

[50] 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

[51] Rao, A., Yerukola, A., Shah, V., Reinecke, K., & Sap, M. (2025). NormAd: A framework for measuring the cultural adaptability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 2373–2403). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.120

[52] Rubin, O., Herzig, J., & Berant, J. (2022). Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2655–2671). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.191

[53] Sagawa, S., Koh, P. W., Hashimoto, T. B., & Liang, P. (2020). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *International Conference on Learning Representations*. https://openreview.net/forum?id=ryxGuJrFvS

[54] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

[55] Shi, W., Li, R., Zhang, Y., Ziems, C., Yu, S., Horesh, R., de Paula, R. A., & Yang, D. (2024). CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 4996–5025). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.288

[56] 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

[57] 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

[58] 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

[59] 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

[60] 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

[61] 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

[62] Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., & Schön, T. (2019). Evaluating model calibration in classification. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* (Vol. 89, pp. 3459–3467). PMLR. https://proceedings.mlr.press/v89/vaicenavicius19a.html

[63] Wolpert, D. H. (1992). Stacked generalization. *Neural Networks, 5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

[64] 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

[65] Wu, Q., Mi, G., & Wood, D. (2025). Calibration-light subject-independent motor imagery BCI via self-supervised pretraining and Conformer. *Journal of Technology Informatics and Engineering, 4*(1), 239–262. https://doi.org/10.51903/jtie.v5i1.493

[66] Xin, Q. (2025a). Explaining OpenStack failure-injection log anomalies with retrieved normal prototypes. *Emerging Information Science and Technology, 6*(2), 125–146. https://doi.org/10.18196/eist.v6i2.31232

[67] 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

[68] 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

[69] 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

[70] 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), 9–27. https://doi.org/10.64268/inspire.v2i1.117

[71] Xin, Q. (2026d). Explainable and fair credit risk scoring with counterfactual explanations: A reproducible evaluation on the German Credit dataset (HELOC-motivated). *J-INTECH (Journal of Information and Technology), 14*(2), 215–231. https://doi.org/10.32664/j-intech.v14i02.2228

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

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

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

[75] 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

[76] 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

[77] 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

[78] 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

[79] Zhang, B., Sun, X., Liu, G., & Zhou, B. (2026). LLM-style DevOps copilot for cloud-native troubleshooting: Retrieval-augmented runbook generation and command-safety evaluation. *Journal of Technology Informatics and Engineering, 5*(2), 104–118. https://doi.org/10.51903/jtie.v5i2.534

[80] 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

[81] 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

[82] 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

[83] 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

[84] 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

[85] 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

[86] 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

[87] Zhao, S., Bai, J., & Roberson, D. (2025). Multi-horizon GPU demand forecasting with workload semantics and operational risk curves: An empirical study on Alibaba Clusterdata GPU trace. *Journal of Technology Informatics and Engineering, 4*(3), 544–571. https://doi.org/10.51903/jtie.v4i3.498

[88] Zhao, S., Ren, Y., & Chang, X. (2026). Profit-aware spot GPU admission control with cost-sensitive loss and evidence-grounded policy memos for AI workload supply-demand matching. *Journal of Technology Informatics and Engineering, 5*(2), 45–59. https://doi.org/10.51903/jtie.v5i2.545

[89] Zhao, W., Mondal, D., Tandon, N., Dillion, D., Gray, K., & Gu, Y. (2024). WorldValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 17696–17706). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.1539/

[90] 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

[91] 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

[92] 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

[93] 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

[94] 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

[95] 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), iaae020. https://doi.org/10.1093/imaiai/iaae020

[96] 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

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

[98] 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

[99] 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

[100] Zhou, B., Wang, H., & Chang, X. (2025). Distilling VMAF into an edge-deployable quality predictor: A pilot shot-level proxy with LLM-ready quality tokens. *Journal of Technology Informatics and Engineering, 4*(2), 447–463. https://doi.org/10.51903/jtie.v4i2.522

[101] Zhou, S., Chen, Y., & Lee, K. (2026). Accounting-aware evidence-constrained agents for disclosure, settlement, and secondary-market risk monitoring in tokenized RWA infrastructure. *Journal of Technology Informatics and Engineering, 5*(2), 60–74. https://doi.org/10.51903/jtie.v5i2.544