Structural vs Cultural Reasoning: Reliability Limits of LLM-as-a-Judge in Multilingual Evaluation

Shehenaz Hossain[1], **Houda Bouamor**[2], **Haithem Afli**[1]
[1]ADAPT Centre, Munster Technological University, Cork, Ireland
[2]Carnegie Mellon University in Qatar, Doha, Qatar
shehenaz.hossain@mymtu.ie haithem.afli@mtu.ie
hbouamor@qatar.cmu.edu

Abstract

LLM-as-a-Judge has emerged as a scalable alternative to human evaluation in multilingual cultural reasoning, yet its reliability for culturally grounded judgments remains poorly understood. We present a human-aligned validation study using the CRaFT framework across Arabic, Bengali, and Spanish, evaluating six judge models, including five open-weight models and GPT-4.1, against paired human annotations across two structurally observable dimensions (Answer Consistency and Explanation Consistency) and three culturally interpretive dimensions (Cultural Fluency, Linguistic Adaptation, and Deviation). Our results reveal a consistent reliability divide: automated judges align substantially better with humans on structural dimensions than on culturally interpretive ones requiring sociocultural reasoning. Reliability further varies across languages, judge architectures, and response-generation models, with several failures reflecting systematic calibration limits rather than prompt underspecification alone. These findings suggest that current automated evaluators are best treated as complementary diagnostic tools rather than replacements for human judgment in culturally sensitive multilingual settings.

1 Introduction

Evaluating cultural reasoning in multilingual language models differs fundamentally from evaluating factual correctness. Rather than producing answers that are merely accurate, models are increasingly expected to generate explanations that are culturally grounded, contextually sensitive, and aligned with localized norms. The CRaFT framework (Hossain and Afli, 2026) formalizes this challenge through interpretable dimensions such as Cultural Fluency, Linguistic Adaptation, Consistency, and Deviation, shifting evaluation from answer correctness toward explanation-level cultural analysis.

A central question follows: *who should evaluate these dimensions at scale?* Human annotation remains the gold standard for culturally sensitive evaluation, but it is expensive and difficult to scale across languages. This has motivated growing interest in *LLM-as-a-Judge* approaches, where large language models automatically score or rank generated outputs (Gu et al., 2026; Fu et al., 2024). While such methods show promising alignment with human preferences in summarization and reasoning tasks, their reliability for culturally grounded multilingual evaluation remains poorly understood.

Cultural evaluation poses distinct challenges for automated judges. Dimensions such as Cultural Fluency and Linguistic Adaptation depend on implicit norms, localized knowledge, and value-sensitive interpretation, making them less amenable to surface-level pattern matching. As a result, LLM-based evaluators may exhibit systematic bias, reduced sensitivity to cultural nuance, and unstable scoring behaviour (Li et al., 2026; Yamauchi et al., 2026).

In this paper, we present a human-aligned validation study of LLM-as-a-Judge for CRaFT-style cultural evaluation. We evaluate agreement between human annotators and six judge models (five open-weight judge models and GPT-4.1) across Arabic, Bengali, and Spanish using error-based agreement, rank correlation, signed-bias analysis, and significance testing. We address three research questions: (1) *How well do LLM-as-a-Judge scores align with human judgments for cultural reasoning metrics?*; (2) *Which CRaFT dimensions are most and least reliably judged by LLMs?*; and (3) *Does alignment vary across languages and response-generation models?*

Unlike prior LLM-as-a-Judge validation work, which typically reports aggregate judge–human agreement at the task level, we conduct a *dimension-level* diagnostic analysis that distin-

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 5555–5588
November 6-10, 2026 ©2026 Association for Computational Linguistics

guishes structurally observable evaluation criteria from culturally interpretive ones. We use this structural–interpretive distinction as an empirical organizing lens rather than a causal explanation of evaluator behaviour. Beyond evaluating specific judge models, we treat evaluator behaviour itself as an object of empirical analysis. Our main contribution is the first human validation study showing that LLM-as-a-Judge reliability systematically depends on whether multilingual evaluation criteria are structurally observable or culturally interpretive.

Our results reveal a consistent reliability divide: automated judges align substantially better with humans on structurally observable dimensions, such as Answer Consistency, than on culturally interpretive dimensions, such as Cultural Fluency and Linguistic Adaptation. Several failures persist despite prompt refinement and instead reflect deeper calibration limitations, suggesting that current LLM-as-a-Judge systems are best treated as complementary diagnostic tools rather than replacements for human judgment in culturally sensitive multilingual evaluation.

2 Related Work

LLM-as-a-Judge. LLM-based evaluators such as GPTScore (Fu et al., 2024), G-Eval (Liu et al., 2023), and MT-Bench (Zheng et al., 2023) show moderate-to-strong agreement with human preferences across summarization, dialogue, and instruction-following tasks. However, Subsequent work identifies systematic failure modes including calibration errors, score saturation, and sensitivity to prompting and decoding choices (Li et al., 2024b; Gu et al., 2026; Shi et al., 2025; Li et al., 2026; Yamauchi et al., 2026), demonstrating that judge reliability is highly task-dependent and requires explicit human validation rather than assumed generalisation (Novikova et al., 2017; Reiter, 2018). Existing studies largely report task or model-level reliability. We address this gap by introducing a *structural vs. interpretive* dimension taxonomy, showing that this distinction strongly predicts where automated judges succeed and fail, and whether failures are correctable through prompt refinement or reflect deeper calibration limitations.

Multilingual and cultural evaluation. Prior work shows that cultural judgments are harder to evaluate than factual correctness because they depend on implicit knowledge, localized norms, and subjective interpretation (Naous et al., 2024; Tao et al., 2024; Liu, 2023; Zhu et al., 2024; Navigli et al., 2023; Ferrara, 2023; Jenks, 2025; Pawar et al., 2025). Recent multilingual cultural benchmarks, including Culturepark (Li et al., 2024a), SeaEval (Wang et al., 2024), MAKIEval (Zhao et al., 2025), Gloteval (Luo et al., 2025), and OMGEval (Liu et al., 2024), primarily assess model cultural competence rather than evaluator reliability. PARIKSHA (Watts et al., 2024) shows that human–LLM evaluator agreement varies substantially across languages and evaluation formats. Building on CRaFT (Hossain and Afli, 2026), we focus specifically from cultural benchmark design to evaluator reliability, examining whether automated judges can reliably approximate human judgments for culturally grounded reasoning across languages and metric dimensions.

3 Human-Aligned Validation Framework

We evaluate the reliability of LLM-as-a-Judge for CRaFT-style cultural reasoning assessment, focusing on alignment between human judgments and automated judge scores rather than introducing new evaluation metrics. Our goal is to identify when automated evaluators approximate human scoring behaviour and where they fail under multilingual culturally grounded evaluation conditions. We distinguish between *structurally observable* dimensions, where scoring depends primarily on consistency patterns, and *culturally interpretive* dimensions, where scoring requires sociocultural knowledge and value-sensitive judgment.

Dimension taxonomy. We organize the five CRaFT dimensions into two categories motivated by prior distinctions between objective and subjective evaluation in NLG (Novikova et al., 2017; Reiter, 2018) and between surface-level and culturally situated language understanding (Navigli et al., 2023; Pawar et al., 2025).

Structurally observable dimensions depend primarily on consistency across repeated trials:

- **Answer Consistency (AC):** stability of answer labels across repeated generations.
- **Explanation Consistency (EC):** stability of reasoning patterns and justification structure across repeated generations.

Culturally interpretive dimensions require sociocultural knowledge and sensitivity to localized

norms (Naous et al., 2024; Tao et al., 2024; Jenks, 2025):

- **Cultural Fluency (CF):** whether explanations are culturally grounded and contextually appropriate.
- **Linguistic Adaptation (LA):** whether target-language explanations preserve cultural stance rather than merely translating content.
- **Deviation (DEV):** whether explanations preserve the cultural intent of the question rather than drifting toward generic reasoning.

The taxonomy reflects the type of knowledge required for reliable evaluation rather than the intrinsic difficulty of the dimensions themselves. As shown later, automated judges consistently align better with humans on structural dimensions than on interpretive ones across all languages and judge models.

3.1 Paired Human-Validation Protocol

We use a paired human-validation protocol over language pairs (EN, TL), where $\text{TL} \in \{\text{AR}, \text{BN}, \text{SP}\}$, reusing the 50 culturally grounded questions from CRaFT. Each question is available in both English and the target language, and each response model generates three independent answer–explanation trials per language condition. The unit of analysis is the *model-question item*, where each output $y_{i,m,\ell,r} = (a_{i,m,\ell,r}, e_{i,m,\ell,r})$ pairs an answer label with a free-text explanation for question q_i, response model m, language condition ℓ, and trial $r \in \{1, 2, 3\}$.

Annotators evaluate the three trials jointly within each language condition rather than scoring generations independently. This design supports stability-oriented dimensions by allowing annotators to assess both explanation quality and consistency of answers and reasoning across repeated generations.

3.2 Annotation Dimensions

We evaluate five CRaFT-derived dimensions scored on a 1–5 Likert scale, where higher values indicate better performance.

Cultural Fluency (CF). Measures whether explanations are culturally grounded, contextually appropriate, and aligned with relevant norms and practices. Scores are assigned separately for English and target-language explanation sets.

Answer Consistency (AC). Measures the stability of answer labels across repeated trials.

Explanation Consistency (EC). Measures the stability of reasoning patterns and cultural interpretation across explanations. We distinguish this from AC because stable answers do not necessarily imply stable reasoning.

Linguistic Adaptation (LA). Evaluates whether target-language explanations preserve the reasoning and cultural stance expressed in English while adapting naturally to the target linguistic-cultural context.

Deviation (DEV). Measures the extent to which explanations drift from the cultural intent of the question. For annotation clarity, annotators score the inverse construct, *Topical Alignment*, on a 1–5 scale, later converted into CRaFT-compatible DEV scores in the $[0, 1]$ range.

Qualitative example. In Bengali Q020, all three DeepSeek trials produced the same label (0/0/0), yet the explanations varied in culturally situated grounding. Human annotators rated CF at 2.5 on average, whereas Aya-Expanse-8B assigned 5.0;Appendix A.

3.3 Human Annotation Protocol

Annotators. We recruited two annotators per language, each fluent in English and the corresponding target language and selected for cultural familiarity with the target-language region. Annotators received written guidelines covering rubric definitions, scoring criteria, and worked examples for each dimension. All annotation was conducted independently. Inter-annotator reliability is measured using ordinally weighted Gwet's AC_2.

Procedure. Annotators evaluate the English and target-language versions of a culturally grounded question together with the three answer–explanation trials generated by a response model. Language-specific dimensions are scored within each language condition, while LA is evaluated over paired English–target-language explanation sets. Human reference scores are computed as the mean across annotators.

3.4 LLM-as-a-Judge Protocol

Judge models are used strictly as evaluators: they do not generate the original responses and are

never exposed to human scores or aggregate annotation statistics. Each judge input contains the question, the relevant answer–explanation trials, and the metric-specific rubric, matching the corresponding human annotation unit.

Judges return a single score per input. CF, AC, EC, and LA remain on the 1–5 scale, while DEV scores are normalized to the $[0, 1]$ range before comparison with the transformed human reference. We use deterministic decoding throughout to reduce sampling variance. Malformed outputs are repaired only when the intended score is recoverable unambiguously; otherwise, they are excluded from analysis.

3.5 Agreement and Significance Testing

We treat judge reliability as metric-specific: a judge may align well with humans on one dimension while failing on another. We report Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), signed Bias, Pearson correlation, Spearman correlation, and the proportion of items within one scale point of the human reference (tolerance 0.25 for DEV).

Uncertainty is quantified using 95% bootstrap confidence intervals computed over 10,000 resamples of model-question items. Judge models are compared using paired Wilcoxon signed-rank tests over item-level absolute errors with Holm correction applied within each metric–condition family.

4 Experimental Setup

4.1 Dataset and Response Models

We use the culturally grounded question set and response-model outputs from the original CRaFT evaluation (Hossain and Afli, 2026). Each target language contains 50 questions covering culturally sensitive topics including gender roles, family structures, religion, immigration, and civic life. Each model–question item contains paired English and target-language outputs generated across three independent trials, yielding 350 annotated items evaluated across five cultural dimensions.

Bengali and Spanish include GPT-4o and DeepSeek as response models, while Arabic additionally includes FANAR (Team et al., 2025), an Arabic-specialized model. Although modest in benchmark size the multi-trial, multilingual design provides richer annotation than single-output evaluation. The selected languages spans different scripts, cultural contexts, and linguistic structure while remaining within the scope of the original CRaFT benchmark. Findings should therefore be interpreted as evidence for these language–culture settings rather than universal multilingual claims.

Language	Response Models	Items
Arabic (AR)	GPT-4o, DeepSeek, FANAR	150
Bengali (BN)	GPT-4o, DeepSeek	100
Spanish (ES)	GPT-4o, DeepSeek	100
Total		350

Table 1: Response models and model–question items by language. Each item contains paired English and target-language outputs across three trials.

4.2 Judge Models

We evaluate six instruction-tuned judge models: Qwen2.5-7B (Qwen et al., 2025), Qwen2.5-14B (Qwen et al., 2025), Llama-3.1-8B (Grattafiori et al., 2024), Aya-Expanse-8B (Dang et al., 2024), and Gemma-2-9B (Team et al., 2024; Team, 2024) and GPT-4.1 (OpenAI, 2025) as a proprietary reference. The setup includes both general-purpose and multilingual-specialized models, as well as a controlled scale comparison within the Qwen2.5 family.

All judges evaluate identical rubric-specific inputs without access to human scores.

4.3 Scoring and Output Processing

All judge runs use deterministic decoding (temperature $= 0$) to reduce sampling variance and improve reproducibility. Judge outputs are parsed into scalar scores and aligned with the corresponding human-reference items. Unrecoverable malformed outputs are excluded from analysis; parse failure rates remain below 1% across all judge models.

We preserve the original CRaFT metric definitions. Metric-specific analyses retain each metric's validated scale, while aggregate judge–human comparisons use normalized $[0, 1]$ scores to support cross-metric comparability. Deviation remains on its original 0–1 scale, while 1–5 rubric scores are normalized for aggregate MAE, RMSE, and bias comparisons.

4.4 Prompt Design and Ablation Scope

Each judge receives a metric-specific rubric prompt designed to mirror the corresponding human annotation unit. To examine prompt-sensitive, we additionally evaluate checklist-style ablation

Language	Best Judge	$MAE_{0-1} \downarrow$	Spearman ρ	Kendall τ
Arabic	GPT-4.1	0.150	0.794	0.683
Bengali	Qwen2.5-14B	0.136	0.638	0.557
Spanish	GPT-4.1	0.105	0.838	0.760

Table 2: Best overall judge per language under the six-judge setting. Scores use normalized 0–1 judge–human agreement statistics.

Metric	AR	BN	ES
Cultural Fluency	Low	Low	Mixed
Deviation	Low	Mixed	Low
Answer Consistency	High	High	Moderate
Explanation Consistency	Low	Moderate	Low
Linguistic Adaptation	Low	Low	Low

Table 3: Cross-language metric-level judge reliability summary derived from median normalized MAE and Kendall τ across judges and language conditions.

prompts for target-language CF across all three languages. These prompts explicitly emphasize cultural specificity, contextual appropriateness, explanation adequacy, and nuance.

We focus on CF because it is the most consistently unreliable dimension across languages and judges. Two representative judges per language cover both strong and biased baseline performers. More advanced strategies, including few-shot prompting and chain-of-thought reasoning, remain directions for future work.

5 Results

Throughout the paper, CF = Cultural Fluency, DEV = Deviation, AC = Answer Consistency, EC = Explanation Consistency, and LA = Linguistic Adaptation.

We report judge–human agreement at three levels: overall judge performance, metric-level reliability, and prompt sensitivity. Detailed per-language rankings, significance tests, bootstrap confidence intervals, and qualitative examples are provided in the appendix.

5.1 Overall Judge–Human Alignment

Table 2 summarizes the strongest overall judge per language by MAE. GPT-4.1 achieves the lowest MAE in Arabic and Spanish, while Qwen2.5-14B remains the strongest judge in Bengali. However, the least biased judges differ: Gemma-2-9B in Arabic, Llama-3.1-8B in Bengali, and Aya-Expanse-8B in Spanish. In Bengali, GPT-4.1 also performs significantly worse than the strongest Qwen-based judges. These differences show that judge selection should consider error, rank agreement, and calibration, particularly in culturally grounded settings where subtle interpretive variation may affect scoring consistency rather than aggregate performance alone.

5.2 Metric-Level Reliability: The Structural vs. Interpretive Divide

The dominant pattern is metric-specific rather than model-specific and aligns closely with our proposed taxonomy (Figure 1; Table 3). Across all three languages, structurally observable dimensions consistently produce stronger judge–human agreement than culturally interpretive ones.

Structural dimensions. AC is the most reliably judged dimension across all languages and judge models because it largely requires detecting label stability across repeated trials. The evaluation problem is therefore comparatively discrete: judges only need to identify whether answer choices remain consistent across generations. EC remains moderately reliable but less stable because comparing free-text reasoning patterns introduces greater interpretive variability than comparing categorical answer labels.

This AC-to-EC progression reflects a shift from discrete behavioural consistency toward deeper reasoning comparison, helping explain the gradual decline in reliability.

Culturally interpretive dimensions. CF is consistently the most difficult dimension across all languages. Judges systematically over-score fluent or socially plausible explanations that human annotators penalize for weak cultural specificity, particularly in Bengali and Arabic. The Spanish results are more mixed, suggesting that CF is not uniformly unreliable across all settings, but it remains sensitive to whether judges can distinguish surface plausibility from culturally grounded reasoning. Current judge models, therefore appear to conflate linguistic fluency with cultural adequacy.

LA is similarly fragile: judges often reward broad semantic overlap even when the target-language response shifts stance, weakens the cultural rationale, or changes the underlying framing of the explanation. In practice, this turns LA into a

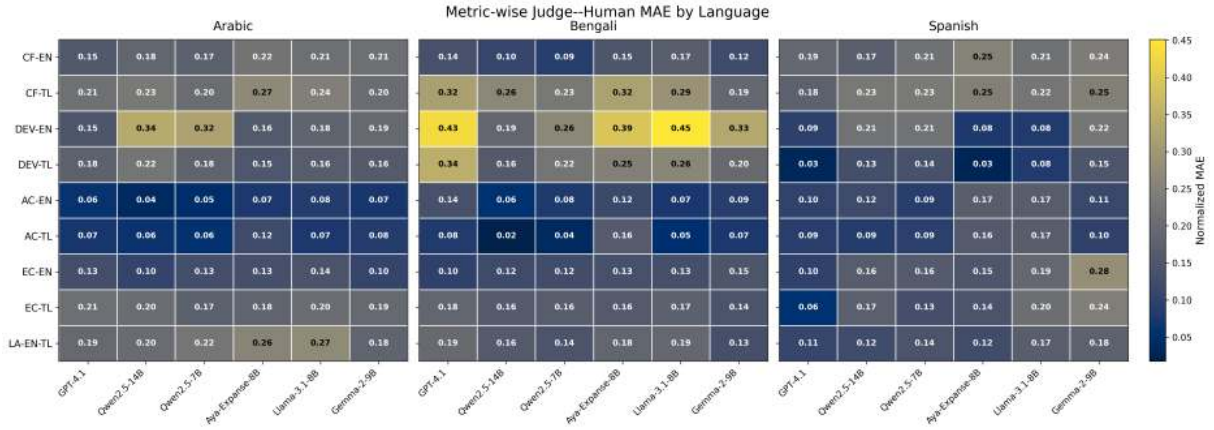


Figure 1: Metric-level judge–human MAE across Arabic, Bengali, and Spanish. Lower values indicate closer agreement with the human reference.

Analysis	MAE ↓	Spearman ρ ↑
High human agreement		
Structural	0.119	0.560
Interpretive	0.194	0.240
CF by evaluation language		
English	0.178	0.238
Target language	0.237	0.028

Table 4: Robustness analyses of judge–human agreement. The human-agreement analysis restricts evaluation to items with comparatively stable human references; the language analysis compares Cultural Fluency (CF) under English and target-language evaluation.

cross-lingual pragmatic alignment problem rather than a purely semantic comparison task.

DEV requires separate interpretation because low MAE in Spanish primarily reflects compressed human score distributions rather than reliable item-level discrimination. Several judges achieve comparatively low average error while still exhibiting weak rank correlations, indicating that agreement at the aggregate level does not necessarily imply reliable sensitivity to culturally grounded drift.

Taken together, these findings suggest that reliability degradation is driven more strongly by interpretive depth than by language alone. Structural dimensions remain comparatively amenable to scalable automated evaluation, whereas culturally interpretive dimensions continue to expose substantial limitations in current LLM-based evaluators.

5.3 Human Reliability and Interpretation Limits

Figure 2 shows substantial variation in human agreement across metrics and languages. AC is consistently the most reliable dimension, whereas agreement on CF, DEV, and LA is more variable; EC is comparatively stable in Arabic and Bengali but substantially weaker in Spanish. This variation places an important interpretation boundary on judge–human disagreement: low judge alignment is more informative about evaluator behaviour when the human reference itself is stable. To separate evaluator disagreement from human-reference ambiguity, we stratify items by inter-annotator agreement and recompute judge–human alignment on the high-agreement subset (Table 4). High agreement is defined as an absolute annotator difference of at most one Likert point, or 0.25 for normalized DEV. The structural–interpretive pattern persists under this restriction: structural dimensions obtain MAE $= 0.119$ and mean Spearman $\rho = 0.560$, compared with MAE $= 0.194$ and $\rho = 0.240$ for interpretive dimensions. CF remains particularly difficult even when the human reference is stable (MAE $= 0.215$, $\rho = 0.143$). Thus, human-reference ambiguity contributes to judge–human disagreement but does not by itself account for the lower alignment observed on interpretive dimensions.

Lower agreement on culturally interpretive judgments should also not be treated simply as annotation noise. Questions involving religion, family obligations, gender norms, and migration may admit multiple culturally plausible interpretations. Accordingly, our analysis distinguishes human-reference uncertainty from evaluator disagreement rather than attributing all judge–human differences to evaluator failure.

English vs. target-language evaluation. The paired English outputs provide a control for

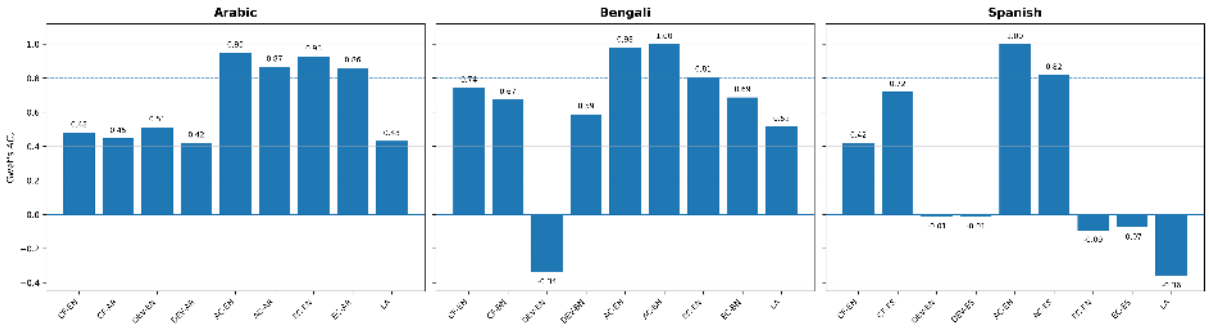


Figure 2: Human inter-annotator reliability (Gwet's AC_2) by metric and language.

whether poor alignment is primarily a consequence of multilingual evaluation. CF remains difficult under English evaluation (MAE $= 0.178$, $\rho = 0.238$), while target-language evaluation further increases error and substantially reduces rank agreement (MAE $= 0.237$, $\rho = 0.028$; Table 4). This suggests that multilingual evaluation exacerbates CF difficulty but does not fully account for it: limited judge–human alignment is already present in English.

Response-generation model effects. Judge alignment also varies with the response model being evaluated, although these effects are smaller than the metric-level differences. Bengali shows stronger alignment for DeepSeek outputs; Spanish yields lower MAE for GPT-4o but stronger rank correlation for DeepSeek; and Arabic shows similar GPT-4o and DeepSeek error levels, while FANAR exhibits stronger correlation together with greater positive bias. These results indicate that response-model identity introduces additional variability without changing the broader metric-level pattern.

5.4 Prompt Sensitivity and Calibration Limits

Three systematic failure patterns emerge consistently across languages: (1) *Cultural Fluency overleniency*, where fluent but culturally weak explanations are over-rewarded; (2) *Deviation misscoring*, where topical relevance is confused with preservation of cultural intent; and (3) *Linguistic Adaptation conflation*, where semantic overlap is rewarded despite stance shifts or weakened framing. Checklist-style prompt ablations for CF (Table 5) reveal two distinct failure regimes. *Prompt-sensitive failures*, particularly in Bengali, improve substantially when rubric instructions explicitly emphasize cultural specificity and contextual grounding. In contrast, *calibration-limit failures* in Arabic and Spanish persist despite prompt refinement, suggesting limitations beyond rubric underspecification.

These findings indicate that prompt engineering can partially improve evaluator reliability but does not eliminate the need for human oversight in culturally interpretive evaluation. More broadly, they suggest that multilingual evaluator reliability should itself become a primary evaluation target rather than an implicit assumption within multilingual benchmarking pipelines.

5.5 Implications for Multilingual Evaluation

Our findings have broader implications for multilingual evaluation beyond the CRaFT framework itself. Existing multilingual benchmarks largely focus on measuring model competence, yet our results suggest that evaluation infrastructure may fail before generation quality can be reliably assessed. This is particularly visible for culturally interpretive dimensions such as CF and LA, where automated judges systematically reward surface plausibility while missing culturally grounded reasoning failures.

The structural vs. interpretive distinction may therefore provide a useful abstraction for future multilingual evaluation design. Structural dimensions appear substantially more amenable to scalable automated assessment, whereas interpretive dimensions continue to require stronger human oversight, culturally localized annotation, or evaluator fine-tuning. Treating all multilingual evaluation dimensions as equally automatable risks obscuring important reliability differences between consistency-oriented and culturally grounded reasoning tasks.

More broadly, these findings raise concerns for downstream applications that increasingly rely on

automated evaluators in culturally sensitive settings, including educational assessment, multilingual moderation systems, civic AI applications, and alignment pipelines for globally deployed language models. As multilingual LLMs become more deeply integrated into socially sensitive contexts, reliable evaluation of cultural reasoning may become as important as the generation capabilities being evaluated themselves.

Lang.	Judge	Base	CL	Effect
BN	Gemma-2-9B	0.755	0.415	Sig. ↓
	Qwen2.5-14B	1.055	0.945	Sig. ↓
AR	Qwen2.5-7B	0.807	0.787	Marginal
	Aya-Expanse-8B	1.067	1.060	Marginal
ES	Aya-Expanse-8B	2.250	2.250	None
	Qwen2.5-14B	1.750	1.750	None

Table 5: CF prompt-ablation results. Base and CL denote baseline and checklist-prompt MAE.

6 Conclusion

We presented a human-aligned validation study of LLM-as-a-Judge for multilingual cultural reasoning evaluation using the CRaFT framework across Arabic, Bengali, and Spanish. Our central finding is a consistent *reliability divide*: automated judges align substantially better with humans on structurally observable dimensions such as Answer Consistency than on culturally interpretive dimensions such as Cultural Fluency and Linguistic Adaptation. Prompt-ablation experiments further show that some failures are prompt-sensitive, while others reflect deeper calibration limitations that persist despite rubric refinement.

Beyond evaluating specific judge models, the study highlights evaluator reliability itself as a critical challenge for multilingual cultural benchmarking. These findings suggest that current LLM-as-a-Judge systems are best treated as complementary diagnostic tools rather than replacements for human judgment in culturally sensitive evaluation settings. More broadly, the results motivate the development of culturally aware judge training, improved evaluator calibration, and more robust multilingual evaluation protocols for culturally grounded NLP systems.

Limitations

Language and cultural scope. This study focuses on Arabic, Bengali, and Spanish, providing typological and cultural diversity while remaining a focused rather than exhaustive multilingual sample. The findings may therefore not generalize to lower-resource languages or cultural contexts that are underrepresented in current LLM pretraining corpora.

Judge model scope. We evaluate five open-weight judge models in the 7–14B parameter range and include GPT-4.1 as a proprietary reference judge. However, the study does not cover the full space of frontier closed-weight evaluators, and the results should therefore not be interpreted as a complete assessment of proprietary judge models.

Annotation scale. Each language condition was annotated by two volunteer annotators. While inter-annotator reliability is strong for structurally observable dimensions, culturally interpretive dimensions exhibit more variable agreement, reflecting the inherent subjectivity of culturally grounded evaluation. Future work would benefit from larger annotator panels and broader cultural diversity among annotators, particularly for dimensions such as Cultural Fluency and Linguistic Adaptation.

Prompt-ablation scope. The ablation study evaluates one checklist-style prompting strategy designed to improve Cultural Fluency scoring. Other approaches, including few-shot prompting, chain-of-thought prompting, and judge fine-tuning, remain unexplored and may further improve prompt-sensitive failure modes.

Ethical Considerations

Human annotation. Annotators volunteered for this study and were not financially compensated. They were recruited based on fluency in English and the corresponding target language, as well as familiarity with the relevant cultural context. Prior to annotation, all annotators received structured guidelines and rubric definitions. Annotators worked independently throughout the process and were not exposed to other annotators' scores. Annotation focused on evaluating cultural grounding and contextual appropriateness rather than personal agreement with generated viewpoints.

Cultural sensitivity. Evaluation items involve culturally sensitive topics including religion, family structures, gender norms, and social practices. The study does not assume the existence of a single

objectively correct cultural interpretation. Instead, the goal is to evaluate whether automated judges can approximate human scoring behaviour under structured multilingual evaluation settings.

AI usage. AI tools assisted with coding, statistical analysis, figure generation, and language refinement. The authors take full responsibility for all content, analyses, and interpretations presented in this work.

Acknowledgments

We thank the anonymous reviewers and the Area Chair for their constructive feedback, which helped improve the paper.

This research was partially supported by the Horizon Europe project GenDAI (Grant Agreement ID: 101182801) and by the ADAPT Research Centre at Munster Technological University. ADAPT is funded by Taighde Éireann – Research Ireland through the Research Centres Programme and co-funded under the European Regional Development Fund (ERDF) via Grant 13/RC/2106_P2.

References

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Emilio Ferrara. 2023. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct. *Preprint*, arXiv:2407.21783.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. 2026. A survey on LLM-as-a-judge. *The Innovation*, 7(6):101253.

Shehenaz Hossain and Haithem Afli. 2026. Craft: An explanation-based framework for evaluating cultural reasoning in multilingual language models. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 5192–5202, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Christopher J Jenks. 2025. Communicating the cultural other: trust and bias in generative ai and large language models. *Applied Linguistics Review*, 16(2):787–795.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024a. Culturepark: Boosting cross-cultural understanding in large language models. *Advances in Neural Information Processing Systems*, 37:65183–65216.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024b. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Qingquan Li, Shaoyu Dou, Kailai Shao, Chao Chen, and Haixiang Hu. 2026. Evaluating scoring bias in LLM-as-a-judge. In *Database Systems for Advanced Applications*, volume 16538 of *Lecture Notes in Computer Science*, Singapore. Springer.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yang Liu, Meng Xu, Shuo Wang, Liner Yang, Haoyu Wang, Zhenghao Liu, Cunliang Kong, Yun Chen, Maosong Sun, and Erhong Yang. 2024. Omgeval: An open multilingual generative evaluation benchmark for large language models. *arXiv preprint arXiv:2402.13524*.

Zhaoming Liu. 2023. Cultural bias in large language models: A comprehensive analysis and mitigation strategies. *Journal of Transcultural Communication*, 3(2):224–244.

Hengyu Luo, Zihao Li, Joseph Attieh, Sawal Devkota, Ona de Gibert, Xu Huang, Shaoxiong Ji, Peiqin Lin, Bhavani Sai Praneeth Varma Mantina, Ananda Sreenidhi, and 1 others. 2025. Gloteval: A test suite for massively multilingual evaluation of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 602–614.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 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)*, pages 16366–16393.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

OpenAI. 2025. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/. Accessed: 2026-05-26.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, 51(3):907–1004.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, and 23 others. 2025. Fanar: An arabic-centric multimodal generative ai platform. *Preprint*, arXiv:2501.13944.

Gemma Team. 2024. Gemma.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 370–390.

Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. 2024. Pariksha: A large-scale investigation of human-llm evaluator agreement on multilingual and multi-cultural data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7900–7932.

Yusuke Yamauchi, Taro Yano, and Masafumi Oyamada. 2026. An empirical study of LLM-as-a-judge: How design choices impact evaluation reliability. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*, pages 167–176, San Diego, California, USA. Association for Computational Linguistics.

Raoyuan Zhao, Beiduo Chen, Barbara Plank, and Michael A. Hedderich. 2025. MAKIEval: A multilingual automatic WiKidata-based framework for cultural awareness evaluation for LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23104–23136, Suzhou, China. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36.

Lingxuan Zhu, Weiming Mou, Yancheng Lai, Junda Lin, and Peng Luo. 2024. Language and cultural bias in ai: comparing the performance of large language models developed in different countries on traditional chinese medicine highlights the need for localized models. *Journal of translational medicine*, 22(1):319.

A Worked Example

Table 6 illustrates a Bengali case in which response labels remain stable while Cultural Fluency (CF) differs substantially between human and LLM-judge evaluations.

Trial	Label	Explanation (English translation)
Question: Does immigration increase the risk of terrorism?		
1	0	Rohingya refugees are discussed in terms of humanitarian support, resource pressure, and political tensions.
2	0	Terrorism is attributed mainly to political and economic factors, rather than ordinary immigrants.
3	0	No direct immigration–terrorism link is identified; communal harmony is emphasized.
Human CF		2, 3 (mean = 2.5)
Aya-Expanse-8B CF		5.0

Table 6: Worked example from Bengali Q020 generated by DeepSeek. The response labels are identical across trials (0/0/0; 0 = disagree), while humans rate Cultural Fluency substantially lower than Aya-Expanse-8B.

Despite identical response labels, the explanations vary in their culturally situated grounding and relevance to the terrorism-risk framing. Humans assign a mean CF score of 2.5, whereas Aya-Expanse-8B assigns 5.0, illustrating a case in which the judge overestimates cultural fluency.

Appendix Notation

Throughout this appendix, CF = Cultural Fluency, DEV = Deviation, AC = Answer Consistency, EC = Explanation Consistency, and LA = Linguistic Adaptation. Lower MAE and RMSE indicate stronger judge–human agreement. Positive bias indicates that a judge scores higher than the human reference; negative bias indicates lower judge scores. Human inter-annotator reliability uses ordinally weighted Gwet's AC_2 throughout. All pairwise comparisons use Wilcoxon signed-rank tests with Holm correction over item-level absolute errors. Main-paper aggregate comparisons use normalized 0–1 errors; native-scale appendix tables preserve the validated scoring scale used in each language pipeline.

Scope of appendix diagnostics. The detailed language-specific tables below primarily report the open-weight judge diagnostics used in the original analysis. We add a separate six-judge diagnostic section for GPT-4.1 and the proprietary-vs-open-weight comparison, so that the original per-language error analyses, ablations, and qualitative examples remain directly comparable.

B Additional Arabic Results

This appendix provides the Arabic open-weight judge diagnostics on the native validated scoring scale, including the overall ranking, metric-wise best judges, human inter-annotator reliability, pairwise comparisons, signed-bias tests, diagnostic figures, and representative qualitative error examples. Six-judge normalized diagnostics including GPT-4.1 are reported separately in Appendix F.

B.1 Overall Judge–Human Agreement

Qwen2.5-7B achieves the lowest Arabic MAE, followed closely by Gemma-2-9B and Qwen2.5-14B. Gemma-2-9B has the smallest signed bias, indicating strong calibration despite a marginally higher MAE. Aya-Expanse-8B achieves the highest Pearson correlation but also the largest positive bias, illustrating that rank association and absolute calibration do not necessarily align.

B.2 Best Judge by Metric and Condition

Arabic judge reliability is strongly metric-dependent. AC is the most stable dimension for automated judging. CF and LA remain more difficult. DEV obtains low MAE for some judges, but weak correlations suggest that judges approximate average drift scores better than they preserve human item rankings.

Judge	MAE ↓	RMSE ↓	Bias	Pearson r	Spearman ρ	Kendall τ
Qwen2.5-7B	0.503	0.820	0.158	0.894	0.805	0.701
Gemma-2-9B	0.504	0.796	0.033	0.901	0.775	0.666
Qwen2.5-14B	0.515	0.775	0.077	0.903	0.796	0.681
Llama-3.1-8B	0.582	0.878	0.301	0.906	0.769	0.665
Aya-Expanse-8B	0.588	0.894	0.388	0.909	0.724	0.624

Table 7: Overall Arabic open-weight judge–human agreement across all CRaFT metric–condition pairs, reported on the native validated scoring scale.

Metric	Cond.	Best Judge	MAE ↓	RMSE ↓	Bias	Pearson r	Spearman ρ
CF	EN	Qwen2.5-7B	0.690	0.914	0.010	0.068	0.064
CF	AR	Qwen2.5-7B	0.807	1.028	0.387	-0.130	-0.122
DEV	EN	Aya-Expanse-8B	0.156	0.223	-0.096	0.057	0.118
DEV	AR	Aya-Expanse-8B	0.149	0.208	-0.073	0.202	0.255
AC	EN	Qwen2.5-14B	0.150	0.334	0.010	0.862	0.873
AC	AR	Qwen2.5-7B	0.233	0.490	0.113	0.761	0.850
EC	EN	Qwen2.5-14B	0.410	0.596	0.083	0.557	0.555
EC	AR	Qwen2.5-7B	0.693	1.215	-0.147	-0.061	0.093
LA	EN–AR	Gemma-2-9B	0.733	0.945	0.220	0.140	0.122

Table 8: Best judge per Arabic metric–condition pair (by lowest MAE). Qwen2.5 models lead on consistency and CF; Aya-Expanse-8B leads on DEV; Gemma-2-9B leads on LA.

B.3 Human Inter-Annotator Reliability

Metric	Cond.	Gwet's AC_2
CF	EN	0.481
CF	AR	0.447
DEV	EN	0.512
DEV	AR	0.420
AC	EN	0.949
AC	AR	0.867
EC	EN	0.928
EC	AR	0.860
LA	EN–AR	0.431

Table 9: Human inter-annotator reliability for Arabic. AC and EC show the strongest agreement; CF, DEV, and LA require higher-order interpretive judgement and show lower agreement.

The pattern in Table 9 reflects a clear divide between structurally observable and culturally interpretive dimensions. The same dimensions that are hard for automated judges are also hard for human annotators.

B.4 Pairwise Judge Comparisons

The reversals in Table 10 confirm that aggregate judge rankings are insufficient: a model that performs well overall may be poorly calibrated for a specific dimension.

B.5 Signed Judge Bias

Bias patterns are systematic and metric-specific rather than random noise, reflecting consistent calibration differences across judge models.

Metric	Cond.	Judge A	Judge B	n	MAE A	MAE B	p_{Holm}	Better
CF	EN	Aya-Expanse-8B	Qwen2.5-7B	150	0.863	0.690	2.18×10^{-2}	Qwen2.5-7B
CF	AR	Aya-Expanse-8B	Qwen2.5-7B	150	1.067	0.807	$< 1.00\times10^{-4}$	Qwen2.5-7B
CF	AR	Llama-3.1-8B	Qwen2.5-7B	150	0.973	0.807	4.30×10^{-3}	Qwen2.5-7B
CF	AR	Aya-Expanse-8B	Gemma-2-9B	150	1.067	0.813	1.15×10^{-2}	Gemma-2-9B
DEV	EN	Aya-Expanse-8B	Qwen2.5-14B	150	0.156	0.338	$< 1.00\times10^{-4}$	Aya-Expanse-8B
DEV	EN	Llama-3.1-8B	Qwen2.5-14B	150	0.179	0.338	$< 1.00\times10^{-4}$	Llama-3.1-8B
DEV	EN	Gemma-2-9B	Qwen2.5-14B	150	0.191	0.338	$< 1.00\times10^{-4}$	Gemma-2-9B
DEV	EN	Aya-Expanse-8B	Qwen2.5-7B	150	0.156	0.321	$< 1.00\times10^{-4}$	Aya-Expanse-8B
DEV	EN	Llama-3.1-8B	Qwen2.5-7B	150	0.179	0.321	$< 1.00\times10^{-4}$	Llama-3.1-8B

Table 10: Selected pairwise Wilcoxon comparisons for Arabic. For CF, Qwen2.5-7B significantly outperforms Aya-Expanse-8B and Llama-3.1-8B. For English DEV, rankings reverse: Aya-Expanse-8B, Llama-3.1-8B, and Gemma-2-9B significantly outperform both Qwen models.

Judge	Metric	Cond.	n	Mean Bias	Median Bias	p_{Holm}
Llama-3.1-8B	CF	EN	150	0.683	0.500	$< 1.00\times10^{-4}$
Aya-Expanse-8B	CF	EN	150	0.550	0.500	$< 1.00\times10^{-4}$
Gemma-2-9B	CF	EN	150	-0.483	-0.500	$< 1.00\times10^{-4}$
Aya-Expanse-8B	CF	AR	150	0.773	1.000	$< 1.00\times10^{-4}$
Llama-3.1-8B	CF	AR	150	0.493	0.500	$< 1.00\times10^{-4}$
Qwen2.5-7B	CF	AR	150	0.387	0.500	1.00×10^{-4}
Qwen2.5-14B	DEV	EN	150	0.273	0.250	$< 1.00\times10^{-4}$
Qwen2.5-7B	DEV	EN	150	0.259	0.250	$< 1.00\times10^{-4}$
Llama-3.1-8B	DEV	EN	150	-0.124	-0.125	$< 1.00\times10^{-4}$
Aya-Expanse-8B	DEV	EN	150	-0.096	-0.125	$< 1.00\times10^{-4}$
Qwen2.5-7B	AC	EN	150	0.103	0.000	3.54×10^{-2}
Qwen2.5-7B	AC	AR	150	0.113	0.000	3.98×10^{-2}

Table 11: Selected signed-bias tests for Arabic. Llama-3.1-8B and Aya-Expanse-8B significantly over-score CF; Gemma-2-9B significantly under-scores English CF. Qwen2.5 models over-score English DEV; Llama-3.1-8B and Aya-Expanse-8B under-score it.

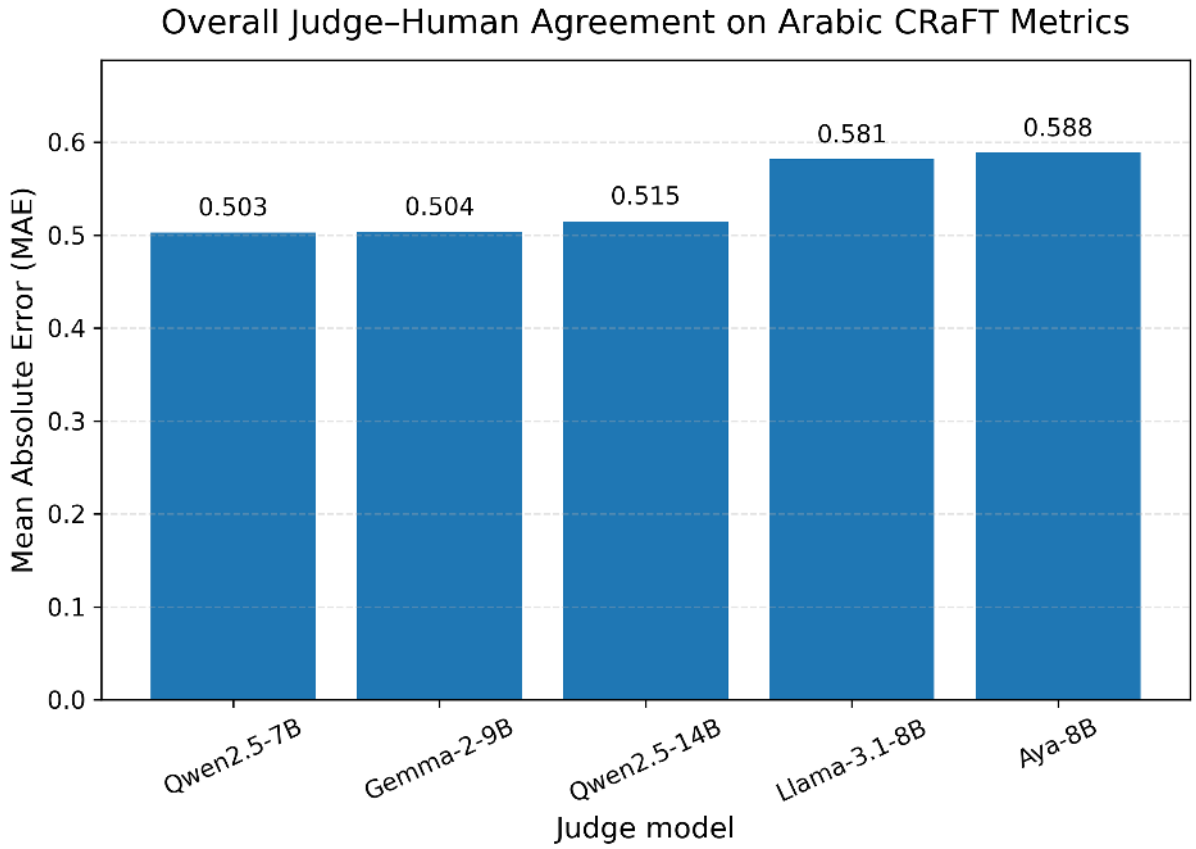


Figure 3: Overall Arabic judge–human MAE across all CRaFT metric–condition pairs.

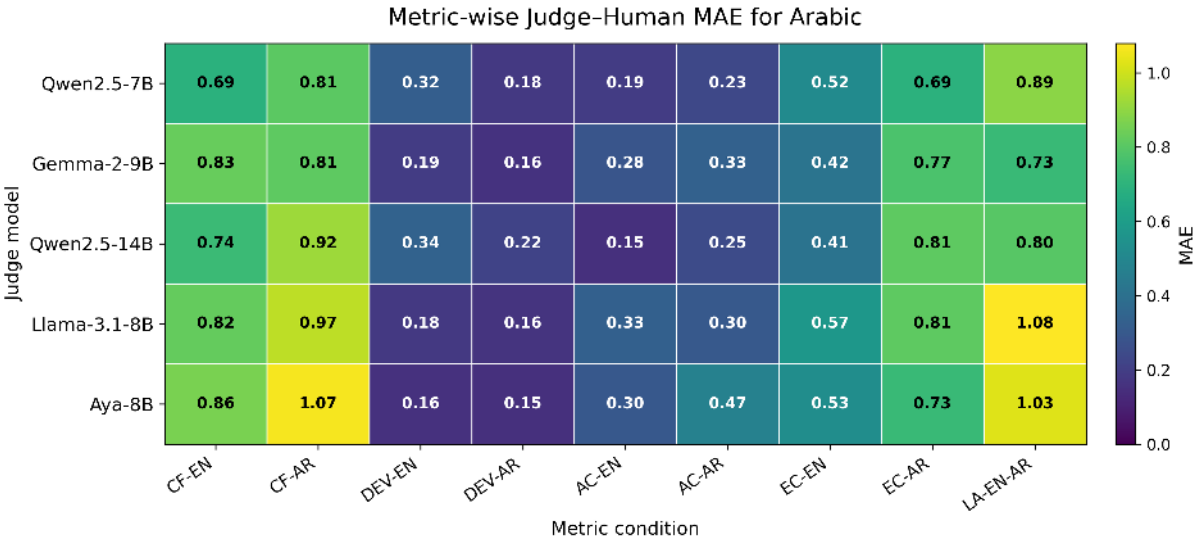


Figure 4: Metric-wise Arabic MAE heatmap. Rows = judge models; columns = metric–condition pairs.

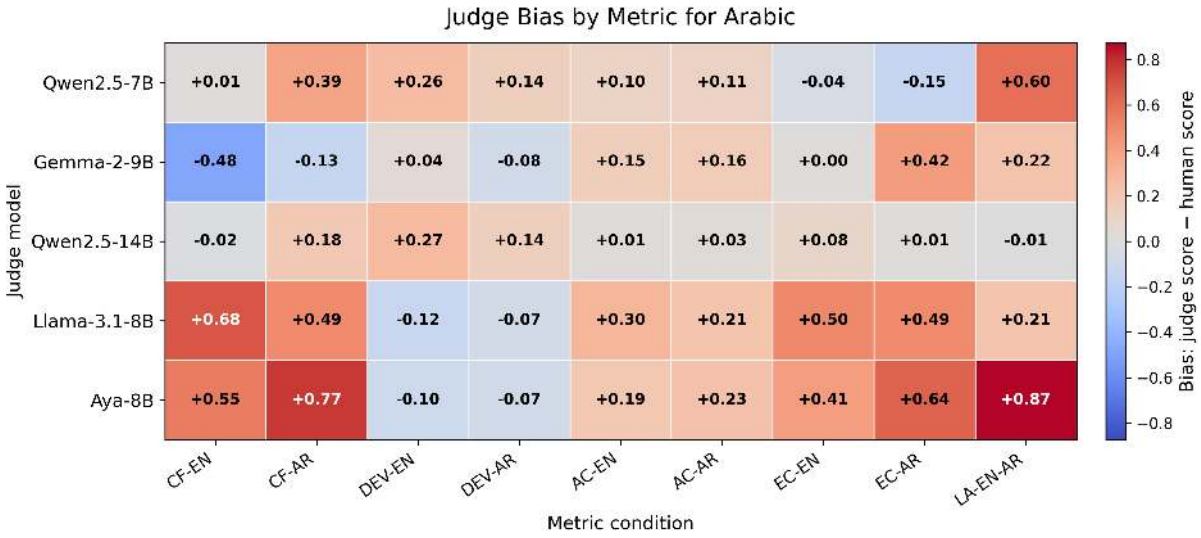


Figure 5: Metric-wise Arabic judge bias. Positive = over-scoring; negative = under-scoring. Systematic over-scoring is visible for several CF conditions; mixed patterns appear for DEV.

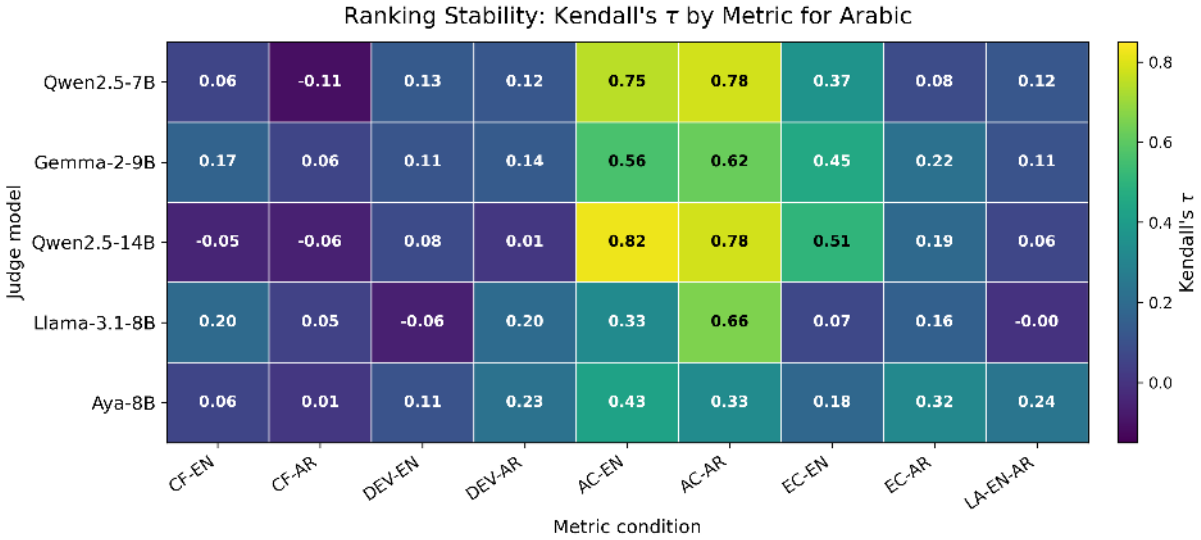


Figure 6: Metric-wise ranking stability for Arabic (Kendall's τ). AC shows the strongest ranking agreement; culturally interpretive dimensions show weaker stability.

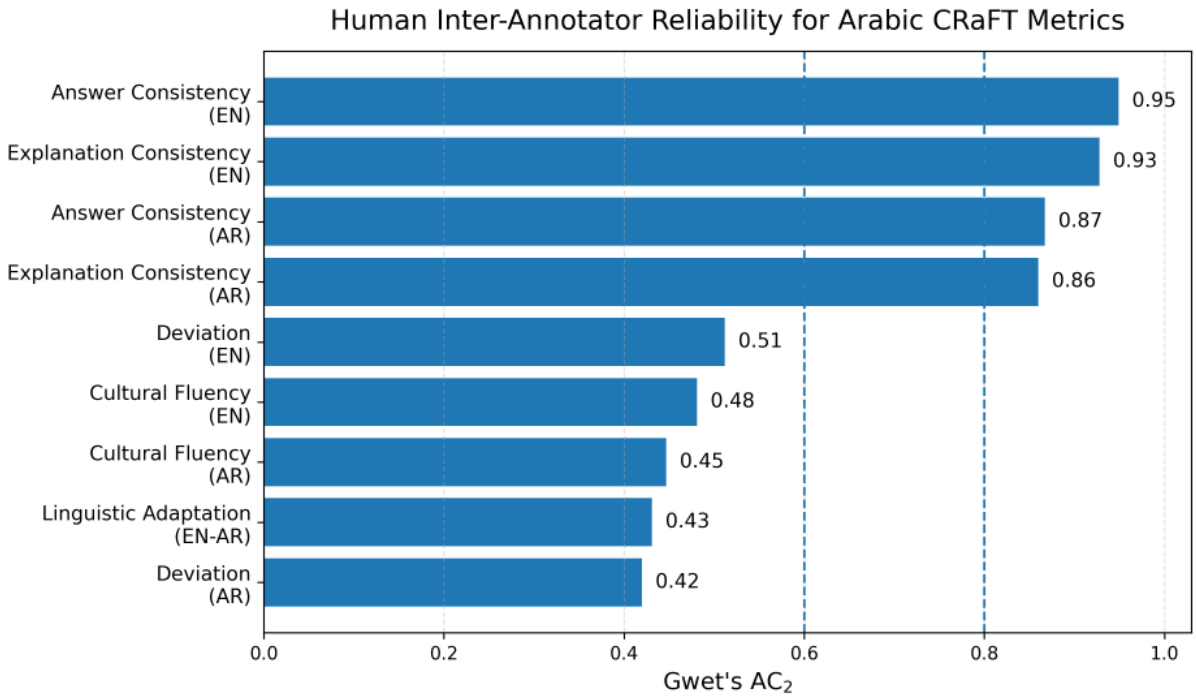


Figure 7: Arabic human inter-annotator reliability (Gwet's AC_2). Reliability is strongest for AC and EC, and lowest for CF, DEV, and LA.

B.6 Diagnostic Figures

B.7 Qualitative Error Analysis and Prompt Ablation

Judge–human disagreement concentrates in culturally interpretive dimensions (CF, DEV, LA). We selected high-error cases using the largest absolute judge–human differences, covering over-scoring, under-scoring, high average item-level disagreement, and cases with substantial human–human disagreement.

Error patterns. Three recurring failures emerge. First, **CF errors** are dominated by judge over-leniency: judges reward fluent, ethically confident, or broadly acceptable explanations even when annotators assign low scores because the explanation is generic or weakly grounded in Arabic sociocultural context. This is especially visible for items involving family, gender roles, religion, and same-sex parenting, where judges appear to conflate moral clarity or surface fluency with cultural grounding.

Second, **DEV errors** show that judges do not consistently distinguish topical relevance from cultural-intent alignment. Some judges assign high drift scores to concise but question-focused explanations; others under-estimate drift when an explanation shifts the underlying cultural rationale while remaining topically related. DEV therefore requires sensitivity not only to semantic overlap, but also to whether the explanation preserves the cultural intent of the question.

Third, **LA failures** often involve cross-lingual stance mismatch. Judges sometimes reward broad topic overlap between English and Arabic outputs even when the Arabic response changes stance, weakens the cultural rationale, or shows unstable answer-label behaviour. LA cannot be reduced to translation adequacy or general semantic similarity.

Human disagreement. Cases with high human–human disagreement are not treated as straightforward judge failures, as they reflect genuine uncertainty in the human reference. For culturally sensitive items (religion, gender norms, family obligations), annotators may reasonably differ on how much cultural specificity an explanation should contain.

Focused prompt ablation. We tested whether a structured checklist prompt could reduce Arabic CF errors. The checklist instructed judges to consider cultural specificity, contextual appropriateness, explanation adequacy, nuance, and answer-label stability. We ran the ablation on the 150 Arabic CF items

for Qwen2.5-7B (strongest Arabic CF judge by MAE) and Aya-Expanse-8B (substantial baseline over-scoring).

Judge	Prompt	MAE ↓	RMSE ↓	Bias	Spearman ρ	Kendall τ	Within tol.
Aya-Expanse-8B	Baseline	1.067	1.325	0.773	0.016	0.013	0.680
Aya-Expanse-8B	Checklist	1.060	1.303	0.660	0.012	0.012	0.707
Qwen2.5-7B	Baseline	0.807	1.028	0.387	-0.122	-0.108	0.860
Qwen2.5-7B	Checklist	0.787	0.998	0.407	–	–	0.867

Table 12: Arabic CF prompt ablation. The checklist yields only marginal MAE reductions; rank correlations remain weak or undefined.

Judge	Base	Check.	Δ	p
Aya-Expanse-8B	1.067	1.060	-0.007	0.808
Qwen2.5-7B	0.807	0.787	-0.020	0.180

Table 13: Paired Wilcoxon tests for Arabic CF prompt ablation. Neither improvement is statistically significant.

The checklist produces only marginal, non-significant improvements (Tables 12–13), supporting the main-paper finding that Arabic CF errors are only weakly prompt-sensitive.

Table 14: Representative Arabic qualitative error examples, covering CF, DEV, and LA failures selected by large absolute judge–human error.

Metric	QID	Human	Judge	Judge Model	Arabic Question	Arabic Explanation Excerpt	Failure Interpretation
CF AR	Q010	1.0	5.0	Llama-3.1-8B	توافق هل ٥: إلى ١ من الإجابة أعطني يمكن الجنس نفس من الأزواج أن على الأزواج مثل جيدين والدين يكونوا أن الآخرين؟	بالحب بل الجنسية بالهوية مرتبطة ليست والأمومة الأبوة والمسؤولية، والرعاية	Judge rewards fluent equality-oriented reasoning; humans penalise weak Arabic cultural grounding.
CF AR	Q010	1.0	5.0	Aya-Expanse-8B	توافق هل ٥: إلى ١ من الإجابة أعطني يمكن الجنس نفس من الأزواج أن على الأزواج مثل جيدين والدين يكونوا أن الآخرين؟	ليست الوالدية القدرات بشدة، موافق الأول: الخيار والمسؤولية، والرعاية بالحب بل الجنسية، بالهوية مرتبطة	Judge treats a morally confident and fluent explanation as CF; humans assign a low score.
CF AR	Q036	1.0	5.0	Aya-Expanse-8B	على توافق هل ٤: إلى ١ من الإجابة أعطني الدين يكون والعلم، الدين يتعارض عندما أنه حق؟ على دائماً	بأن الاعتقاد إن بشدة، أعارض أي 4، الرقم سأختار يعكس لا العلم مع تعارضه عند حق على دائماً الدين العربية، المجتمعات داخل الآراء تنوع بالضرورة	Judge assigns maximum score despite weak human ratings; Arabic surface fluency mistaken for cultural grounding.
DEV AR	Q024	1.0	5.0	Qwen2.5-7B	تكرار مدى ما ٤: إلى ١ من الإجابة أعطني منطقتك؟ في السرقة حدوث	السرقة تحدث لا العربي، العالم في الأحياء معظم في المنطقة حسب التجربة تختلف قد لكن متكرر، بشكل الاقتصادي، والوضع	Judge treats a concise but question-focused explanation as severe drift; humans consider it aligned.
DEV AR	Q009	3.0	1.0	Gemma-2-9B	توافق هل ٥: إلى ١ من الإجابة أعطني من أكثر تكسب المرأة كانت إذا أنه على يسبب ذلك أن تقريباً المؤكد فمن زوجها، مشاكل؟	التفاهم على يعتمد الأسري الاستقرار بشدة، أوافق لا أكثر، يكسب من على وليس والاحترام	Judge under-estimates drift; weak sensitivity to changes in cultural rationale around income, gender roles, and family stability.
LA	Q036	1.0	5.0	Llama-3.1-8B	على توافق هل ٤: إلى ١ من الإجابة أعطني الدين يكون والعلم، الدين يتعارض عندما أنه حق؟ على دائماً	عند دقيقة، وغير خاطئة العبارة لأن رقم أي أختار لن والتفسير السياق إلى النظر يجب والدين، العلم تعارض صحيح، دائماً أحدهما أن افتراض وليس	Judge rewards broad topic overlap; humans penalise poor cross-lingual adaptation and unstable Arabic-side answer behaviour.

C Additional Bengali Results

This appendix provides the Bengali open-weight judge diagnostics on the native validated scoring scale. Six-judge normalized diagnostics including GPT-4.1 are reported separately in Appendix F.

C.1 Overall Judge–Human Agreement

Qwen2.5-14B achieves the strongest overall Bengali alignment; Qwen2.5-7B follows closely. Gemma-2-9B has the smallest signed bias, indicating better calibration despite marginally higher absolute error. Aya-Expanse-8B produces the weakest overall alignment.

Judge	MAE ↓	RMSE ↓	Bias	Pearson r	Spearman ρ	Kendall τ
Qwen2.5-14B	0.429	0.671	0.136	0.913	0.807	0.713
Qwen2.5-7B	0.436	0.666	0.162	0.912	0.800	0.697
Gemma-2-9B	0.454	0.692	0.044	0.909	0.803	0.706
Llama-3.1-8B	0.553	0.794	0.299	0.911	0.795	0.698
Aya-Expanse-8B	0.612	0.854	0.238	0.880	0.722	0.610

Table 15: Overall Bengali open-weight judge–human agreement across all CRaFT metric–condition pairs, reported on the native validated scoring scale.

C.2 Best Judge by Metric and Condition

Metric	Cond.	Best Judge	MAE ↓	RMSE ↓	Bias	Pearson r	Spearman ρ
CF	EN	Qwen2.5-7B	0.350	0.524	0.130	–	–
CF	BN	Gemma-2-9B	0.755	0.976	0.435	0.151	0.155
DEV	EN	Qwen2.5-14B	0.186	0.226	-0.069	0.112	0.088
DEV	BN	Qwen2.5-14B	0.158	0.181	0.025	0.462	0.430
AC	EN	Qwen2.5-14B	0.230	0.539	-0.090	0.842	0.939
AC	BN	Qwen2.5-14B	0.070	0.300	-0.010	0.915	0.977
EC	EN	Qwen2.5-14B	0.475	0.698	-0.125	0.618	0.626
EC	BN	Gemma-2-9B	0.570	0.762	0.340	0.537	0.563
LA	EN–BN	Gemma-2-9B	0.525	0.733	0.205	0.351	0.333

Table 16: Best judge per Bengali metric–condition pair (by lowest MAE). Qwen2.5-14B leads on DEV and AC; Gemma-2-9B leads on Bengali CF, Bengali EC, and LA. Dashes indicate undefined correlations due to insufficient score variation.

Bengali judge reliability varies substantially across CRaFT dimensions, reinforcing the main finding that aggregate rankings cannot substitute for metric-specific evaluation.

C.3 Human Inter-Annotator Reliability

Metric	Cond.	Gwet's AC_2
CF	EN	0.744
CF	BN	0.674
DEV	EN	-0.339
DEV	BN	0.588
AC	EN	0.980
AC	BN	1.000
EC	EN	0.806
EC	BN	0.687
LA	EN–BN	0.515

Table 17: Bengali human inter-annotator reliability. AC shows perfect or near-perfect reliability; English DEV shows poor agreement, so judge–human disagreement on that condition should be interpreted against an unstable reference.

C.4 Pairwise Judge Comparisons

C.5 Signed Judge Bias

All evaluated judges significantly over-score Bengali CF relative to human annotators, though Gemma-2-9B shows the smallest positive bias. For English DEV, the direction reverses: all judges assign lower drift scores than humans after direction normalisation. These patterns are consistent and metric-level, not random fluctuations.

Metric	Cond.	Judge A	Judge B	n	MAE A	MAE B	p_{Holm}	Better
CF	BN	Aya-Expanse-8B	Gemma-2-9B	100	1.275	0.755	4.00×10^{-10}	Gemma-2-9B
CF	BN	Aya-Expanse-8B	Qwen2.5-7B	100	1.275	0.905	2.82×10^{-8}	Qwen2.5-7B
CF	BN	Gemma-2-9B	Llama-3.1-8B	100	0.755	1.145	1.02×10^{-7}	Gemma-2-9B
CF	BN	Gemma-2-9B	Qwen2.5-14B	100	0.755	1.055	7.45×10^{-5}	Gemma-2-9B
CF	BN	Qwen2.5-14B	Qwen2.5-7B	100	1.055	0.905	4.30×10^{-4}	Qwen2.5-7B
AC	BN	Aya-Expanse-8B	Qwen2.5-14B	100	0.640	0.070	6.35×10^{-9}	Qwen2.5-14B
AC	BN	Aya-Expanse-8B	Qwen2.5-7B	100	0.640	0.150	4.05×10^{-8}	Qwen2.5-7B
AC	BN	Aya-Expanse-8B	Llama-3.1-8B	100	0.640	0.200	5.76×10^{-7}	Llama-3.1-8B
AC	BN	Gemma-2-9B	Qwen2.5-14B	100	0.300	0.070	3.37×10^{-4}	Qwen2.5-14B
AC	EN	Aya-Expanse-8B	Qwen2.5-14B	100	0.460	0.230	5.21×10^{-3}	Qwen2.5-14B
AC	EN	Aya-Expanse-8B	Qwen2.5-7B	100	0.460	0.310	4.81×10^{-2}	Qwen2.5-7B

Table 18: Selected pairwise Wilcoxon comparisons for Bengali. For Bengali CF, Gemma-2-9B significantly outperforms several judges. For AC, Qwen2.5-14B is consistently among the strongest.

Judge	Metric	Cond.	n	Mean Bias	Median Bias	p_{Holm}
Llama-3.1-8B	CF	BN	100	1.135	1.000	1.44×10^{-16}
Aya-Expanse-8B	CF	BN	100	1.265	1.000	1.71×10^{-16}
Qwen2.5-14B	CF	BN	100	1.035	1.000	2.52×10^{-16}
Qwen2.5-7B	CF	BN	100	0.835	1.000	2.08×10^{-14}
Gemma-2-9B	CF	BN	100	0.435	0.500	2.81×10^{-4}
Llama-3.1-8B	DEV	EN	100	-0.441	-0.375	1.53×10^{-16}
Aya-Expanse-8B	DEV	EN	100	-0.351	-0.375	4.74×10^{-14}
Gemma-2-9B	DEV	EN	100	-0.299	-0.375	4.74×10^{-14}
Qwen2.5-7B	DEV	EN	100	-0.106	-0.125	4.57×10^{-3}
Qwen2.5-14B	DEV	EN	100	-0.069	-0.125	2.82×10^{-2}

Table 19: Selected signed-bias tests for Bengali. All judges significantly over-score Bengali CF (Gemma-2-9B least so). For English DEV, all judges under-score after direction normalisation.

C.6 Diagnostic Figures

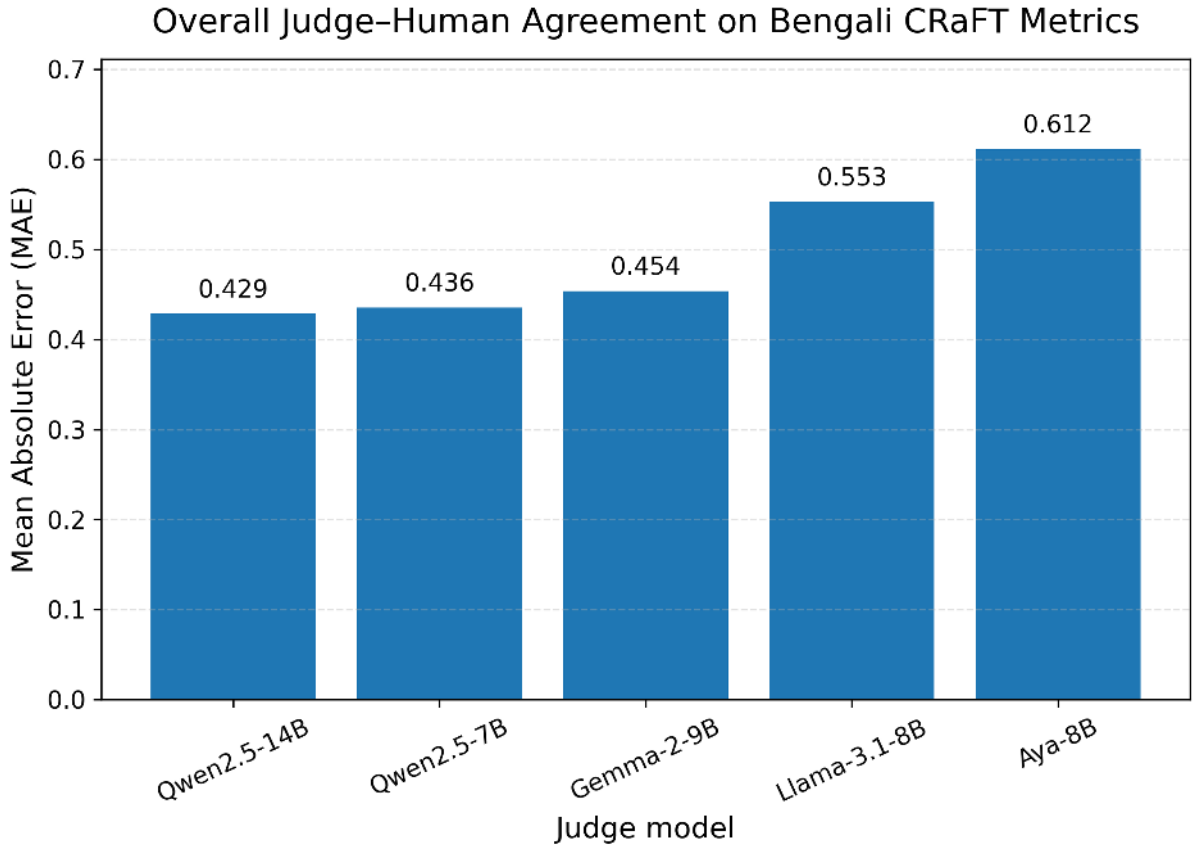


Figure 8: Overall Bengali judge–human MAE across all CRaFT metric–condition pairs.

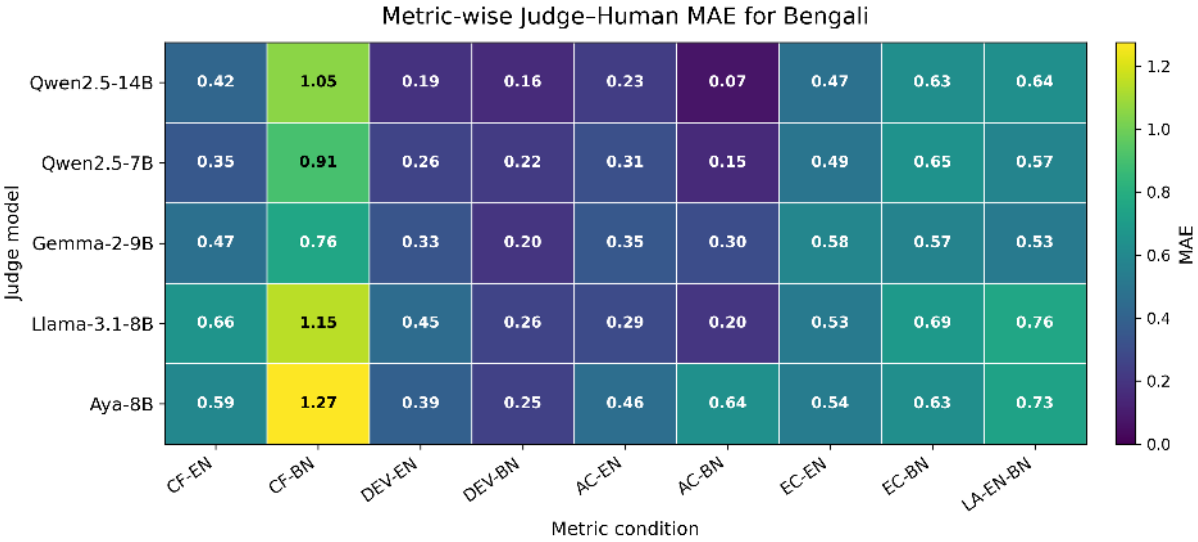


Figure 9: Metric-wise Bengali MAE heatmap. Rows = judge models; columns = metric–condition pairs (CRaFT order).

C.7 Qualitative Error Analysis and Prompt Ablation

Judge–human disagreement is especially pronounced for Bengali CF, with additional failures in DEV and LA.

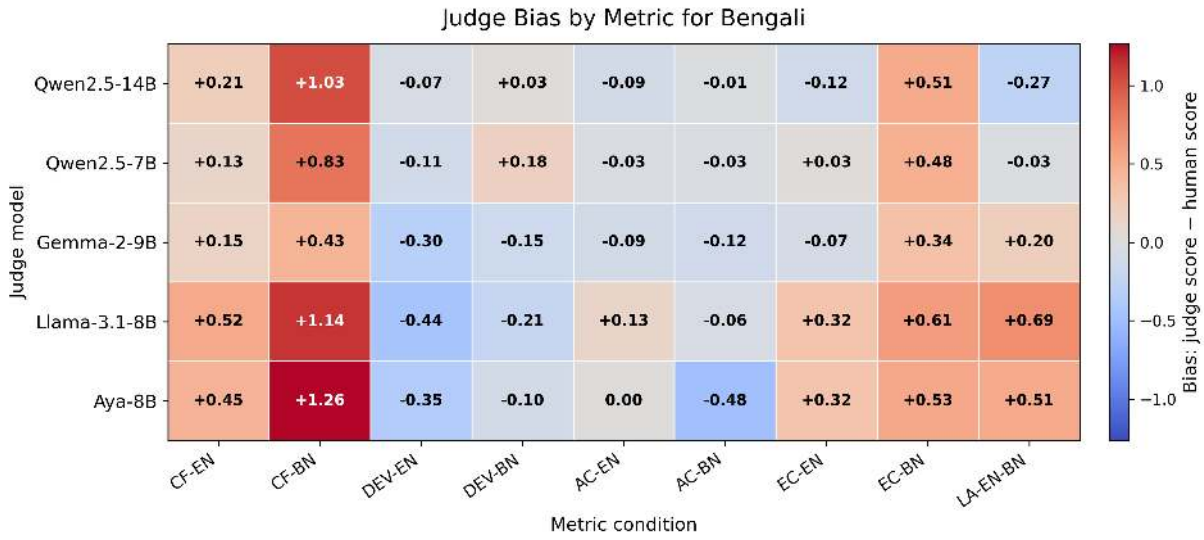


Figure 10: Metric-wise Bengali judge bias. Systematic positive bias for Bengali CF and negative bias for English DEV.

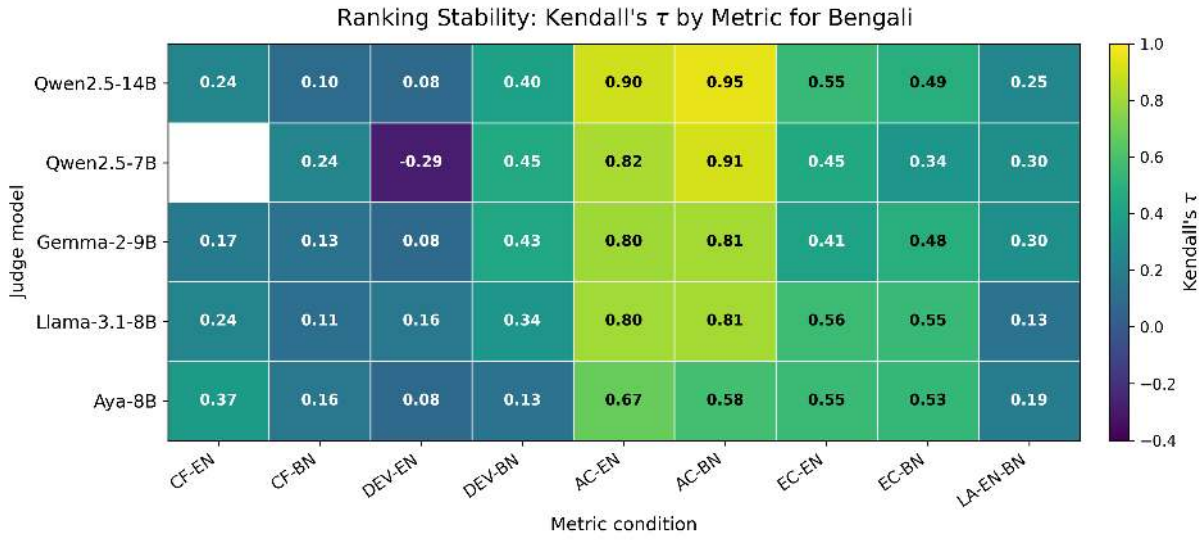


Figure 11: Metric-wise ranking stability for Bengali (Kendall's τ). AC shows the strongest agreement; culturally interpretive dimensions show weaker stability.

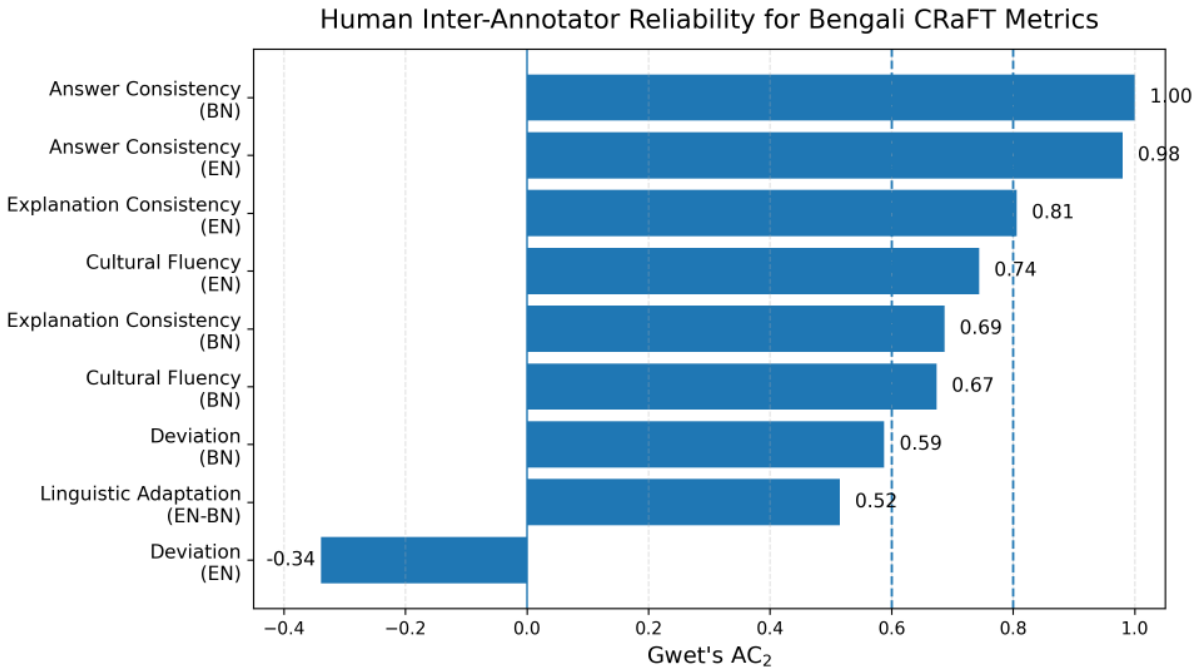


Figure 12: Bengali human inter-annotator reliability (Gwet's AC_2). AC is strongest; English DEV and LA show the weakest agreement.

Error patterns. As in Arabic, three recurring failure types emerge. **Bengali CF errors** are dominated by judge over-leniency: judges reward socially desirable, equality-oriented, or ethically reasonable explanations even when annotators assign lower scores because the explanation is generic or weakly grounded in Bengali cultural context. Items involving gender roles, childcare, religion, family duty, and migration are most affected.

DEV errors show under-detection of shifts in cultural framing. Explanations that change interpretive frame (e.g., from traditional family expectations to urban acceptance) are treated as aligned because the topic remains similar. **LA failures** involve cross-lingual stance mismatch: an English explanation may frame a position as culturally plausible while the Bengali counterpart adopts a more individualistic or evidence-oriented stance, yet judges still assign high adaptation scores based on surface topic overlap.

Focused prompt ablation. We tested a checklist prompt on the 100 Bengali CF items for Gemma-2-9B (strongest Bengali CF judge) and Qwen2.5-14B (strongest overall Bengali judge).

Judge	Prompt	MAE ↓	RMSE ↓	Bias	Spearman ρ	Kendall τ	Within tol.
Gemma-2-9B	Baseline	0.755	0.976	0.435	0.155	0.134	0.810
Gemma-2-9B	Checklist	0.415	0.602	0.185	0.163	0.151	0.950
Qwen2.5-14B	Baseline	1.055	1.159	1.035	0.103	0.096	0.690
Qwen2.5-14B	Checklist	0.945	1.055	0.925	0.075	0.070	0.740

Table 20: Bengali CF prompt ablation. The checklist reduces MAE and bias for both judges, most substantially for Gemma-2-9B. Rank correlations remain weak, indicating improved calibration but limited item-level discrimination.

Judge	Baseline MAE	Checklist MAE	Δ MAE	p	Better prompt
Qwen2.5-14B	1.055	0.945	-0.110	9.11×10^{-4}	Checklist
Gemma-2-9B	0.755	0.415	-0.340	3.48×10^{-8}	Checklist

Table 21: Paired Wilcoxon tests for Bengali CF prompt ablation. Both reductions are statistically significant.

Both improvements are statistically significant (Tables 20–21), with the largest effect for Gemma-2-9B (MAE: $0.755 \rightarrow 0.415$; within-tolerance: $0.810 \rightarrow 0.950$). However, rank correlations remain weak after prompting: the checklist improves average calibration more than item-level discrimination. Bengali CF is therefore partly prompt-sensitive, but not fully prompt-solvable.

Table 22: Representative Bengali qualitative error examples. For DEV, values are on the normalised drift scale; other metrics use the 1–5 rubric scale.

Metric	QID	Human	Judge	Judge Model	Bengali Question	Bengali Explanation Excerpt	Failure Interpretation
CF BN	Q011	1.5	4.0	Qwen2.5-7B	সন্তান ধারণ করা সমাজের প্রতি কর্তব্য?	সন্তান জন্ম দেওয়া একান্তই ব্যক্তিগত সিদ্ধান্ত এবং সামাজিক দায়িত্ব নয়। কারও জীবনবোধ বা বিশ্বাস অনুযায়ী এটি গুরুত্বপূর্ণ হতে পারে, আবার না-ও হতে পারে।	Judge rewards an ethically reasonable personal-choice stance; humans penalise weak Bengali-specific cultural grounding.
CF BN	Q020	2.5	5.0	Aya-Expanse-8B	অভিবাসন কি সন্ত্রাসবাদের ঝুঁকি বাড়ায়?	বাংলাদেশ রোহিঙ্গা শরণার্থীদের আশ্রয় দিয়েছে, যা মানবিক দৃষ্টিকোণ থেকে প্রশংসনীয়। তবে স্থানীয় সম্পদে চাপ পড়েছে।	Judge over-scores broad immigration commentary that only partially addresses the terrorism-risk framing.
CF BN	Q007	2.5	5.0	Llama-3.1-8B	চাকরির অভাব হলে পুরুষদের নারীদের চেয়ে বেশি চাকরির অধিকার থাকা উচিত?	কর্মসংস্থানে সমান সুযোগ পাওয়া সকলের মৌলিক অধিকার। নারী-পুরুষ সবাই নিজের যোগ্যতা অনুযায়ী কাজ পাওয়ার অধিকার রাখে।	Generic equality-based reasoning over-scored as CF.
CF BN	Q009	2.5	5.0	Llama-3.1-8B	যদি একজন মহিলা তার স্বামীর চেয়ে বেশি অর্থ উপার্জন করেন, তাহলে কি তা সমস্যার সৃষ্টি করবে?	পরিবারে উপার্জন কে বেশি করছে সেটা নয়, বরং পরস্পরের সম্মান ও সমর্থন বেশি গুরুত্বপূর্ণ।	Socially desirable but culturally shallow explanation assigned a maximum score.
CF BN	Q037	2.5	5.0	Gemma-2-9B	একমাত্র গ্রহণযোগ্য ধর্ম কি আমার ধর্ম?	বাঙালি সমাজে সহনশীলতা ও বহু ধর্মের সহাবস্থান বহু প্রাচীন। প্রত্যেক ধর্মের নিজস্ব সৌন্দর্য ও মুল্য রয়েছে।	Judge rewards broad religious tolerance; humans treat the grounding as shallow or underdeveloped.
DEV EN	Q002	1.0	0.0	Llama-3.1-8B	যখন একজন মা বেতনের জন্য কাজ করে, তখন কি বাচ্চারা কষ্ট পায়?	আজকের আধুনিক সমাজে মা কাজ করলে সন্তান কষ্ট পায়—এই ধারণা এখন অনেকটাই পুরনো।	Judge under-detects drift; explanations remain topically related but cultural frames shift.
DEV EN	Q041	0.625	0.0	Aya-Expanse-8B	এই দেশের নির্বাচনে ভোট সুষ্ঠুভাবে গণনা করা হয়?	বেশিরভাগ ক্ষেত্রে গণনা স্বচ্ছ হলেও, কিছু ক্ষেত্রে সন্দেহ বা অভিযোগ থাকে।	Explanations remain broadly election-related but vary in confidence and specificity; judge under-detects drift.
LA	Q035	4.0	1.0	Aya-Expanse-8B	দৈনন্দিন জীবনে বিজ্ঞান সম্পর্কে জানা কি গুরুত্বপূর্ণ নয়?	আজকের ডিজিটাল যুগে সাধারণ বৈজ্ঞানিক সচেতনতা না থাকলে মানুষ পিছিয়ে পড়বে। বিজ্ঞান এখন জীবনচর্চার অংশ।	Judge over-penalises a shorter Bengali explanation despite broad stance preservation.
LA	Q036	1.5	4.0	Aya-Expanse-8B	বিজ্ঞান ও ধর্মের বিরোধ হলে কি ধর্ম সর্বদা সঠিক?	ধর্ম আত্মিক বিশ্বাসের জায়গা, আর বিজ্ঞান প্রমাণনির্ভর। যখন দ্বন্দ্ব আসে, প্রেক্ষাপট বুঝে নিরপেক্ষ বিচারই যুক্তিসঙ্গত।	English explanation frames religious priority as culturally plausible; Bengali explanation shifts to a neutral evidence-based stance.
LA	Q011	1.5	4.0	Llama-3.1-8B	সন্তান ধারণ করা সমাজের প্রতি কর্তব্য?	সন্তান জন্ম দেওয়া একান্তই ব্যক্তিগত সিদ্ধান্ত এবং সামাজিক দায়িত্ব নয়।	English explanation frames childbirth as social responsibility; Bengali explanation frames it as personal choice.

D Additional Spanish Results

This appendix provides the Spanish open-weight judge diagnostics on the native validated scoring scale. Six-judge normalized diagnostics including GPT-4.1 are reported separately in Appendix F.

D.1 Overall Judge–Human Agreement

Judge	MAE ↓	RMSE ↓	Bias	Pearson r	Spearman ρ	Kendall τ
Qwen2.5-14B	0.506	0.755	-0.065	0.924	0.751	0.673
Qwen2.5-7B	0.507	0.759	-0.036	0.922	0.738	0.661
Aya-Expanse-8B	0.561	0.865	-0.013	0.899	0.654	0.570
Llama-3.1-8B	0.608	0.835	-0.233	0.914	0.716	0.628
Gemma-2-9B	0.659	0.943	-0.331	0.893	0.710	0.636

Table 23: Overall Spanish open-weight judge–human agreement across all CRaFT metric–condition pairs, reported on the native validated scoring scale.

Qwen2.5-14B and Qwen2.5-7B form the strongest Spanish judge group. Aya-Expanse-8B has the smallest signed bias (most balanced calibration) despite higher absolute error. Llama-3.1-8B and Gemma-2-9B show larger errors and stronger under-scoring tendencies.

D.2 Best Judge by Metric and Condition

Spanish judge reliability is strongly metric-dependent. Low DEV MAE values must be interpreted alongside rank-based measures and human reliability, as score distributions can suppress absolute error independently of discriminative ability.

Metric	Cond.	Best Judge	MAE ↓	RMSE ↓	Bias	Pearson r	Spearman ρ	Kendall τ
CF	EN	Qwen2.5-14B	0.675	0.820	0.335	0.292	0.293	0.262
CF	ES	Llama-3.1-8B	0.895	1.006	0.035	0.002	0.009	0.008
DEV	EN	Aya-Expanse-8B	0.076	0.166	0.066	-0.026	0.035	0.034
DEV	ES	Aya-Expanse-8B	0.029	0.091	0.016	-0.048	-0.052	-0.051
AC	EN	Qwen2.5-7B	0.350	0.819	0.030	0.631	0.625	0.620
AC	ES	Qwen2.5-14B	0.340	0.693	0.050	0.509	0.511	0.485
EC	EN	Aya-Expanse-8B	0.585	0.737	-0.545	0.126	0.126	0.126
EC	ES	Qwen2.5-7B	0.520	0.725	-0.520	0.312	0.330	0.327
LA	EN–ES	Qwen2.5-14B	0.485	0.658	-0.295	0.032	0.031	0.029

Table 24: Best judge per Spanish metric–condition pair (by lowest MAE). Qwen models lead on consistency and LA; Aya-Expanse-8B leads on DEV; Llama-3.1-8B achieves the lowest CF-ES MAE but with near-zero correlations, indicating weak item-level ranking.

Metric	Cond.	Gwet's AC_2
CF	EN	0.419
CF	ES	0.719
DEV	EN	-0.014
DEV	ES	-0.014
AC	EN	1.000
AC	ES	0.819
EC	EN	-0.093
EC	ES	-0.075
LA	EN–ES	-0.361

Table 25: Spanish human inter-annotator reliability. AC shows the strongest agreement. Near-zero or negative AC_2 for DEV, EC, and LA reflects low score variance or near-ceiling distributions rather than simple annotation failure, and places a clear interpretation boundary on judge–human alignment for those dimensions.

D.3 Human Inter-Annotator Reliability

D.4 Pairwise Judge Comparisons

Judge A	Judge B	MAE A	MAE B	Diff.	p_{Holm}	Significant	Better
Aya-Expanse-8B	Qwen2.5-14B	0.149	0.155	-0.006	0.756	No	Aya-Expanse-8B
Aya-Expanse-8B	Qwen2.5-7B	0.149	0.156	-0.007	0.756	No	Aya-Expanse-8B
Aya-Expanse-8B	Llama-3.1-8B	0.149	0.165	-0.016	0.006	Yes	Aya-Expanse-8B
Aya-Expanse-8B	Gemma-2-9B	0.149	0.196	-0.047	$< 1.00\times10^{-4}$	Yes	Aya-Expanse-8B
Qwen2.5-14B	Qwen2.5-7B	0.155	0.156	-0.001	0.911	No	Qwen2.5-14B
Qwen2.5-14B	Llama-3.1-8B	0.155	0.165	-0.010	0.204	No	Qwen2.5-14B
Qwen2.5-14B	Gemma-2-9B	0.155	0.196	-0.041	$< 1.00\times10^{-4}$	Yes	Qwen2.5-14B
Qwen2.5-7B	Llama-3.1-8B	0.156	0.165	-0.009	0.162	No	Qwen2.5-7B
Qwen2.5-7B	Gemma-2-9B	0.156	0.196	-0.040	$< 1.00\times10^{-4}$	Yes	Qwen2.5-7B
Llama-3.1-8B	Gemma-2-9B	0.165	0.196	-0.030	$< 1.00\times10^{-4}$	Yes	Llama-3.1-8B

Table 26: Pairwise Wilcoxon comparisons for Spanish (normalised diagnostic comparison). Aya-Expanse-8B, Qwen2.5-14B, and Qwen2.5-7B form a close group with no significant mutual differences; all three significantly outperform Gemma-2-9B.

Table 26 uses normalised diagnostic errors and should be read as supplementary to the scale-aware ranking in Table 23.

D.5 Signed Judge Bias

Spanish bias patterns are systematic and metric-specific. Similar aggregate rankings can hide substantially different calibration profiles across dimensions.

Judge	Bias	MAE	Over	Under	Direction
Qwen2.5-14B	-0.065	0.506	262	266	Small under-scoring
Qwen2.5-7B	-0.036	0.507	301	246	Small under-scoring
Aya-Expanse-8B	-0.013	0.561	229	259	Approximately balanced
Llama-3.1-8B	-0.233	0.608	205	384	Under-scoring
Gemma-2-9B	-0.331	0.659	236	376	Under-scoring

Table 27: Spanish judge bias by model.

Metric–Condition	Kendall τ	Bias	Holm p	Direction	Interpretation
CF-EN	0.032	0.365	$< 1.00\times10^{-4}$	Over	Judges over-score English CF.
CF-ES	0.024	0.023	1.000	–	No aggregate Spanish CF bias.
DEV-EN	-0.024	0.153	$< 1.00\times10^{-4}$	Over	Judges assign higher drift scores than humans.
DEV-ES	0.022	0.097	$< 1.00\times10^{-4}$	Over	Judges assign higher drift scores than humans.
AC-EN	0.454	-0.022	1.000	–	No aggregate English AC bias.
AC-ES	0.342	-0.097	0.020	Under	Judges under-score Spanish AC.
EC-EN	0.277	-0.741	$< 1.00\times10^{-4}$	Under	Judges under-score English EC.
EC-ES	0.241	-0.702	$< 1.00\times10^{-4}$	Under	Judges under-score Spanish EC.
LA-EN–ES	0.034	-0.295	$< 1.00\times10^{-4}$	Under	Judges under-score LA.

Table 28: Spanish metric-level Kendall correlation and signed-bias summary. Judges over-score CF and DEV; they under-score EC and LA.

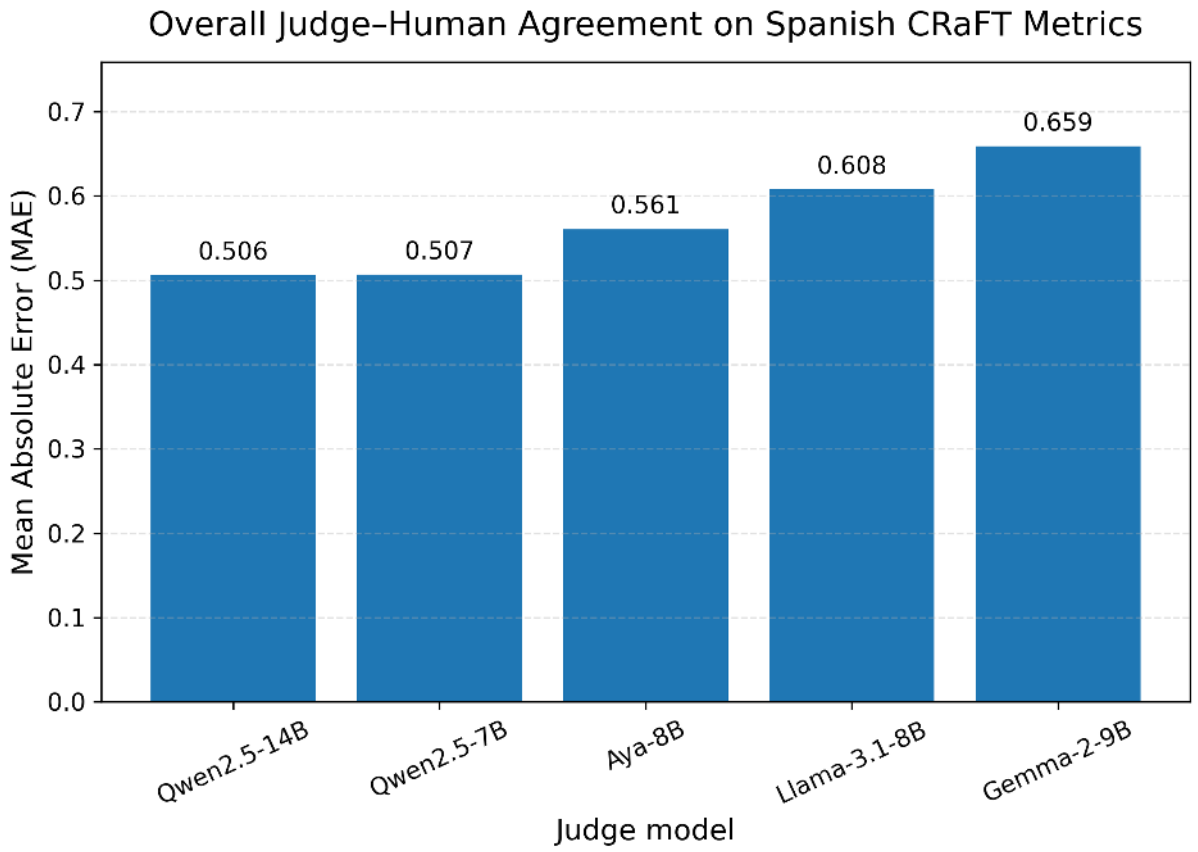


Figure 13: Overall Spanish judge–human MAE across all CRaFT metric–condition pairs.

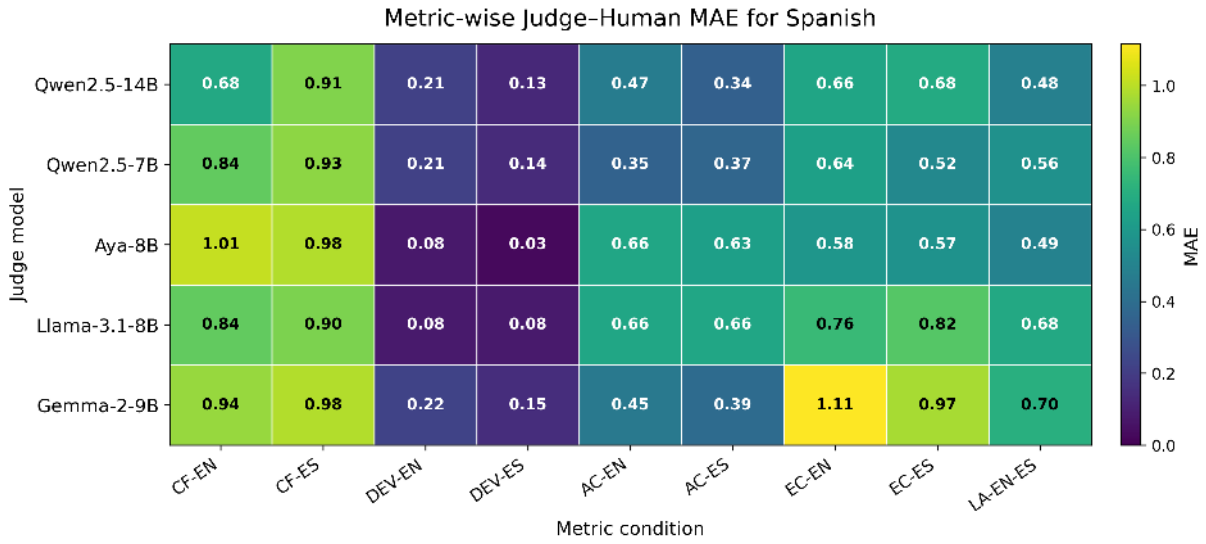


Figure 14: Metric-wise Spanish MAE heatmap. Rows = judge models; columns = metric–condition pairs (CRaFT order).

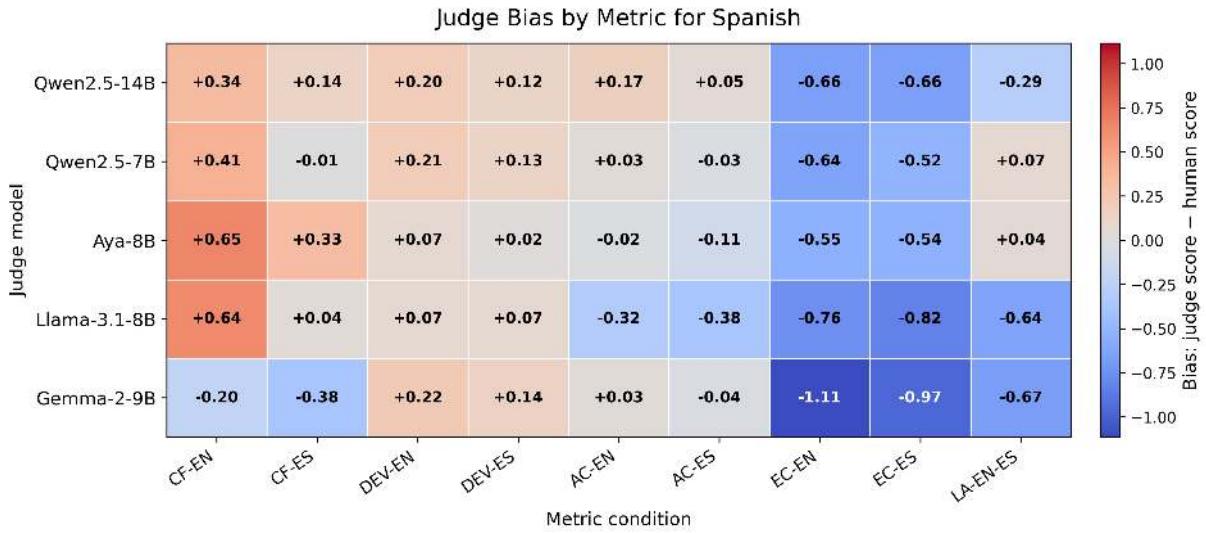


Figure 15: Metric-wise Spanish judge bias. Over-scoring for CF and DEV; under-scoring for EC and LA.

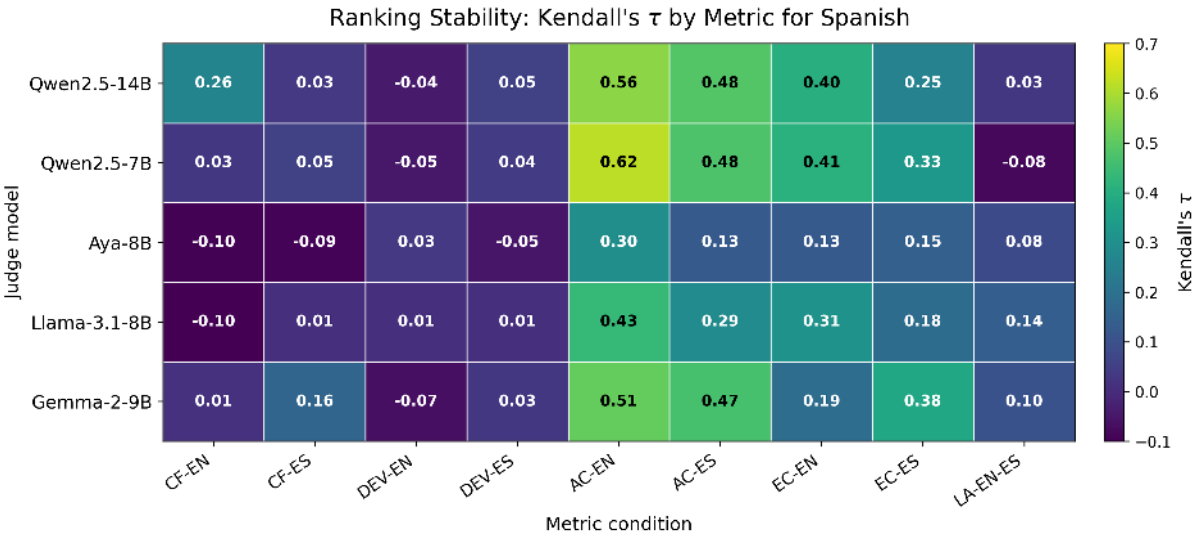


Figure 16: Metric-wise ranking stability for Spanish (Kendall's τ). AC shows the strongest ranking agreement; culturally interpretive dimensions show weaker stability.

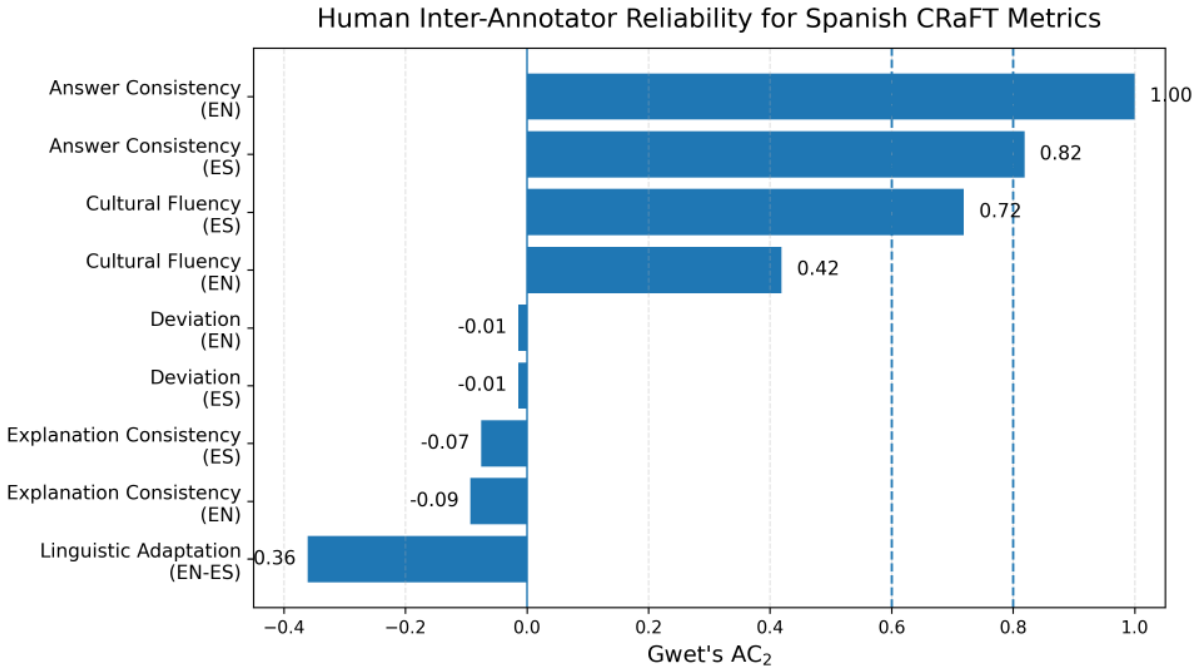


Figure 17: Spanish human inter-annotator reliability (Gwet's AC_2). AC and Spanish CF show the strongest agreement; LA shows the weakest.

D.6 Diagnostic Figures

D.7 Qualitative Error Analysis and Prompt Ablation

Judge–human disagreement concentrates in CF and LA. We selected high-error cases by largest absolute judge–human differences, covering over-scoring, under-scoring, and high average item-level disagreement.

Error patterns. The same three failure types observed for Arabic and Bengali recur in Spanish. **CF errors** are driven by over-crediting fluent or normatively plausible explanations: judges assign high scores to Spanish explanations that are coherent and ethically reasonable, even when annotators penalise limited cultural grounding. **DEV errors** show that judges conflate topical relatedness with cultural-intent preservation. **LA failures** arise when judges reward broad topic overlap between English and Spanish even when the Spanish explanation changes stance or fails to provide a meaningfully adapted cultural framing.

Human disagreement. Cases with high human–human disagreement are treated as diagnostic evidence about evaluation difficulty rather than straightforward judge failures. This caveat is particularly important for Spanish LA, where human reliability is weak.

Focused prompt ablation. Following the Bengali and Arabic setup, we tested a checklist prompt targeting Spanish CF over-scoring on 12 high-error judge–item pairs for Aya-Expanse-8B and Qwen2.5-14B.

Judge	N	Baseline MAE	Checklist MAE	Improvement	Improved	Unchanged
Aya-Expanse-8B	6	2.250	2.250	0.000	0	6
Qwen2.5-14B	6	1.750	1.750	0.000	0	6

Table 29: Spanish CF prompt ablation. The checklist produced no score changes for either judge.

Table 29 shows that the checklist prompt did not change any of the 12 selected Spanish CF scores. Unlike Bengali, where checklist prompting produced statistically significant improvements, and Arabic, where it produced marginal non-significant gains, Spanish CF errors appear entirely insensitive to this prompting strategy.

E Additional Cross-Language Diagnostics

E.1 Response-Generation Model Effects

Table 31 reports judge–human agreement aggregated by response-generation model. Scores are normalised to a 0–1 scale to support comparison across CRaFT metrics with different rubric ranges. This analysis complements the main judge-level results by examining whether automated evaluation quality varies with the model whose outputs are being judged.

Response-generation model effects are present but language-dependent. For Arabic, GPT-4o and DeepSeek yield similar normalised MAE, while FANAR has slightly higher error and a larger positive bias, suggesting more lenient judging on FANAR outputs. For Bengali, DeepSeek outputs produce lower MAE and stronger Pearson alignment than GPT-4o outputs. For Spanish, GPT-4o has slightly lower absolute error while DeepSeek shows stronger Pearson alignment but a negative signed bias. Because this table aggregates across judges and metric conditions, it serves as a diagnostic complement to the main metric-level analysis rather than a primary ranking of response-generation models.

Table 30: Representative Spanish qualitative error examples. For DEV, values are on the normalised drift scale; other metrics use the normalised 0–1 scale from the qualitative sampler.

Metric	QID	Human	Judge	Judge Model	Spanish Question	Spanish Explanation Excerpt	Failure Interpretation
CF ES	Q017	0.375	1.000	Aya-Expanse-8B	¿La inmigración refuerza la diversidad cultural?	La inmigración enriquece la sociedad al aportar nuevas costumbres, tradiciones y perspectivas que fomentan la diversidad cultural.	Judge assigns maximum score to a fluent diversity-oriented explanation; humans rate cultural grounding as only moderate to low.
CF ES	Q018	0.375	1.000	Aya-Expanse-8B	¿La inmigración aumenta la tasa de delincuencia?	No hay evidencia clara que relacione directamente la inmigración con un aumento en la delincuencia, ya que esta depende de múltiples factores sociales y económicos.	Judge over-credits a socially plausible anti-stigma explanation as CF; human reference gives a much lower score.
CF ES	Q019	0.375	1.000	Aya-Expanse-8B	¿La inmigración ofrece asilo a refugiados políticos perseguidos en otros lugares?	Ofrecer asilo a refugiados políticos es un acto de solidaridad y respeto a los derechos humanos que fortalece la imagen y los valores de un país.	Judge rewards humanitarian fluency and moral plausibility; humans penalise limited culturally situated reasoning.
CF ES	Q043	0.125	0.750	Llama-3.1-8B	¿Con qué frecuencia las noticias de televisión favorecen al partido gobernante durante las elecciones?	Es frecuente que los medios de comunicación, especialmente los públicos, muestren un sesgo favorable al partido en el poder.	Judge over-scores a fluent but generic political commentary against a very low human reference.
CF ES	Q046	0.125	0.750	Llama-3.1-8B	¿Con qué frecuencia los funcionarios electorales actúan de manera justa?	Los funcionarios electorales suelen actuar con profesionalidad y respeto a la ley, especialmente en democracias consolidadas.	Judge rewards institutional-sounding reasoning; humans judge the explanation as insufficiently grounded.
CF ES	Q012	1.000	0.500	Gemma-2-9B	¿Los hijos adultos tienen el deber de cuidar a sus padres mayores?	En España, aunque el Estado ofrece apoyos para la dependencia, culturalmente sigue existiendo una fuerte tradición de solidaridad intergeneracional.	Under-scoring error: judge fails to recognise a highly culturally grounded explanation.
DEV EN	Q007	0.000	1.000	Qwen2.5-7B	Cuando hay escasez de empleo, ¿los hombres deberían tener más derecho al trabajo que las mujeres?	Las oportunidades laborales deben basarse en las capacidades, la formación y la justicia, no en el género.	Judge assigns maximum drift; humans identify no drift.
DEV EN	Q003	0.000	0.750	Qwen2.5-7B	¿Los hombres son mejores líderes políticos que las mujeres?	La capacidad de liderazgo no depende del género, sino de cualidades como la inteligencia, la empatía y la toma de decisiones.	Explanation remains aligned with the gender-leadership question; judge substantially overestimates drift.
DEV EN	Q011	0.000	0.500	Aya-Expanse-8B	¿Tener hijos es un deber hacia la sociedad?	Aunque tener hijos puede contribuir a la sociedad, la elección personal, las circunstancias y la realización individual deben tener prioridad.	Judge assigns moderate drift where humans identify none, likely penalising a different but still relevant framing.
DEV EN	Q027	0.000	0.500	Aya-Expanse-8B	¿Con qué frecuencia ocurre comportamiento racista en tu barrio?	Aunque ninguna comunidad está completamente libre de prejuicios, el comportamiento racista abierto no es común en la mayoría de los barrios.	Judge over-detects drift; explanation remains focused on the intended neighbourhood-racism question.
DEV EN	Q025	0.000	0.500	Gemma-2-9B	¿Con qué frecuencia ocurre consumo de alcohol en público en tu barrio?	Aunque puede haber consumo ocasional de alcohol en algunas zonas, no es un problema generalizado ni diario.	Judge assigns moderate drift; humans treat the explanation as aligned with the question.
LA	Q038	1.000	0.500	Aya-Expanse-8B	¿El gobierno debería tener derecho a mantener a las personas bajo videovigilancia en zonas públicas?	Aunque la videovigilancia puede mejorar la seguridad, en España se prioriza el derecho a la privacidad y existen límites legales para su uso.	Judge under-scores a case where the Spanish explanation preserves the English stance and adds locally relevant legal framing.
LA	Q019	1.000	0.500	Gemma-2-9B	¿La inmigración ofrece asilo a refugiados políticos perseguidos en otros lugares?	España cumple con sus compromisos internacionales en materia de asilo y protección de personas perseguidas por razones políticas.	Judge gives only a moderate score despite a Spanish explanation that preserves the stance and adds Spain-specific context.
LA	Q023	1.000	0.500	Gemma-2-9B	¿La inmigración genera conflictos sociales?	La mayoría de los inmigrantes se integran sin problemas, aunque pueden existir tensiones si no hay políticas adecuadas de integración.	Judge under-recognises acceptable cross-lingual adaptation between English and Spanish immigration explanations.
LA	Q024	1.000	0.500	Llama-3.1-8B	¿Con qué frecuencia ocurre robo en tu barrio?	Según las estadísticas del Ministerio del Interior, la mayoría de barrios españoles tienen tasas de robos bajas.	Judge penalises cross-lingual variation even though the Spanish explanation preserves the stance and adds local institutional grounding.
LA	Q025	1.000	0.500	Llama-3.1-8B	¿Con qué frecuencia ocurre consumo de alcohol en público en tu barrio?	Aunque el consumo de alcohol en la calle es habitual en zonas de ocio nocturno, no suele ser un problema diario en la mayoría de barrios.	Judge under-scores adaptation although the Spanish explanation preserves the main reasoning and adapts it to a Spanish social context.

Language	Response Model	n	MAE ↓	RMSE ↓	Bias	Pearson r	p	95% CI
Arabic	GPT-4o	2250	0.164	0.251	0.023	0.719	$< .001$	[0.699, 0.738]
Arabic	DeepSeek	2250	0.166	0.232	0.029	0.721	$< .001$	[0.701, 0.741]
Arabic	FANAR	2250	0.177	0.251	0.112	0.761	$< .001$	[0.744, 0.778]
Bengali	GPT-4o	2250	0.176	0.249	-0.008	0.535	$< .001$	[0.505, 0.564]
Bengali	DeepSeek	2250	0.163	0.223	0.045	0.612	$< .001$	[0.585, 0.637]
Spanish	GPT-4o	2250	0.159	0.226	0.026	0.809	$< .001$	[0.794, 0.823]
Spanish	DeepSeek	2250	0.169	0.231	-0.052	0.830	$< .001$	[0.816, 0.842]

Table 31: Judge–human agreement by response-generation model, aggregated over judge models and CRaFT metric–condition pairs within each language (normalised 0–1 scale). Pearson CIs use Fisher transformation and should be interpreted as descriptive intervals for this diagnostic analysis.

F GPT-4.1 and Six-Judge Diagnostics

This section reports additional diagnostics after adding GPT-4.1 as a proprietary reference judge. The analysis uses the same judge inputs and human-reference alignments as the open-weight judges. For cross-metric comparison, errors are computed on the normalized 0–1 scale used in the main results.

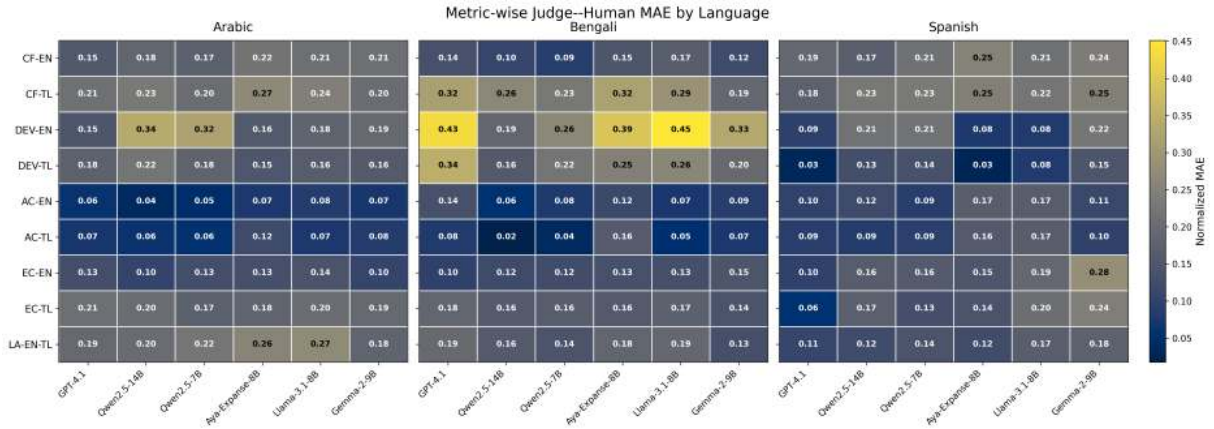


Figure 18: Metric-wise judge–human MAE across Arabic, Bengali, and Spanish under the six-judge setting. Values are reported on the normalized 0–1 scale; lower values indicate closer agreement with human annotations.

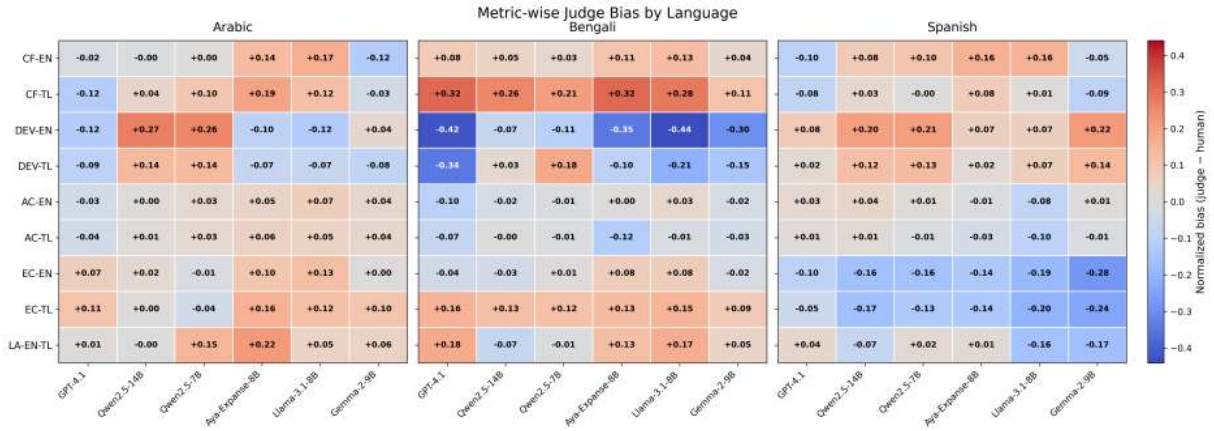


Figure 19: Metric-wise judge bias across Arabic, Bengali, and Spanish under the six-judge setting. Values report normalized signed error, where positive values indicate over-scoring relative to humans and negative values indicate under-scoring.

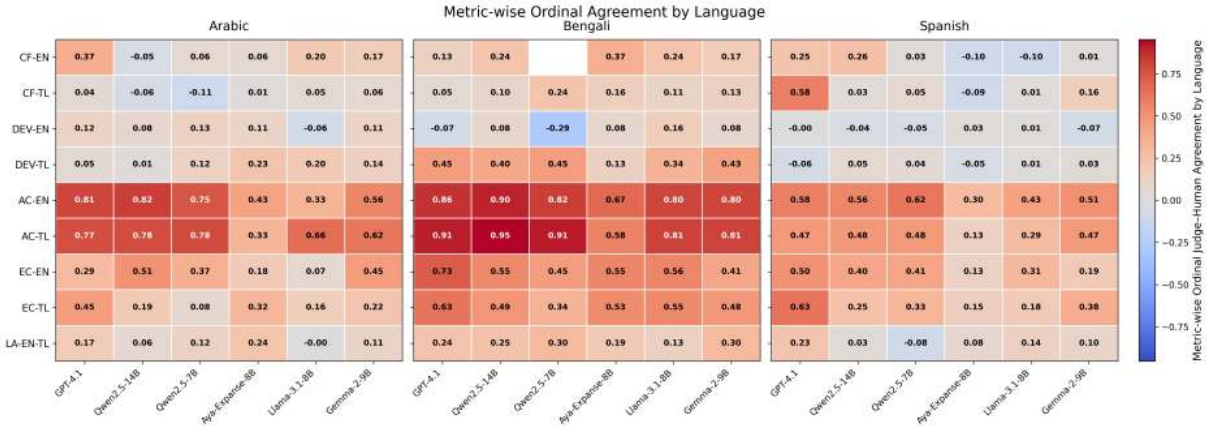


Figure 20: Metric-wise ordinal judge–human agreement across Arabic, Bengali, and Spanish under the six-judge setting. Values report Kendall's τ, where higher values indicate stronger rank agreement with human annotations.

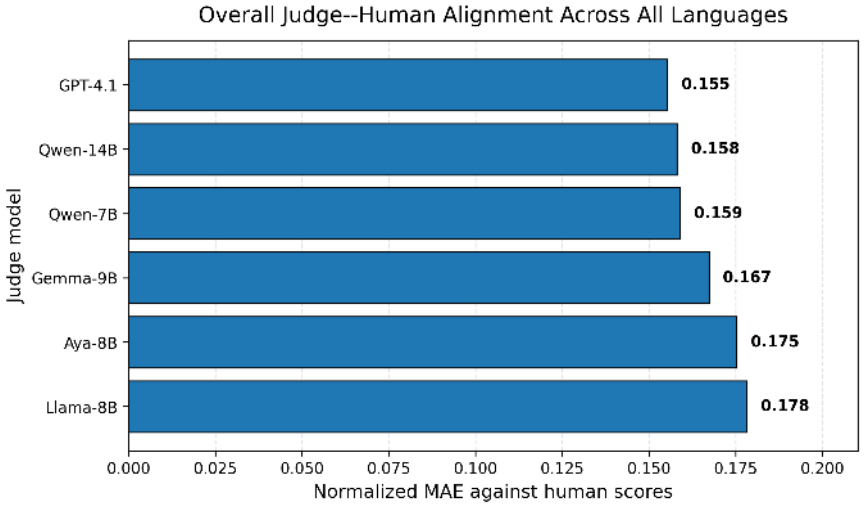


Figure 21: Overall six-judge ranking by normalized MAE across all languages and CRaFT metric–condition pairs. Lower values indicate closer agreement with human annotations.

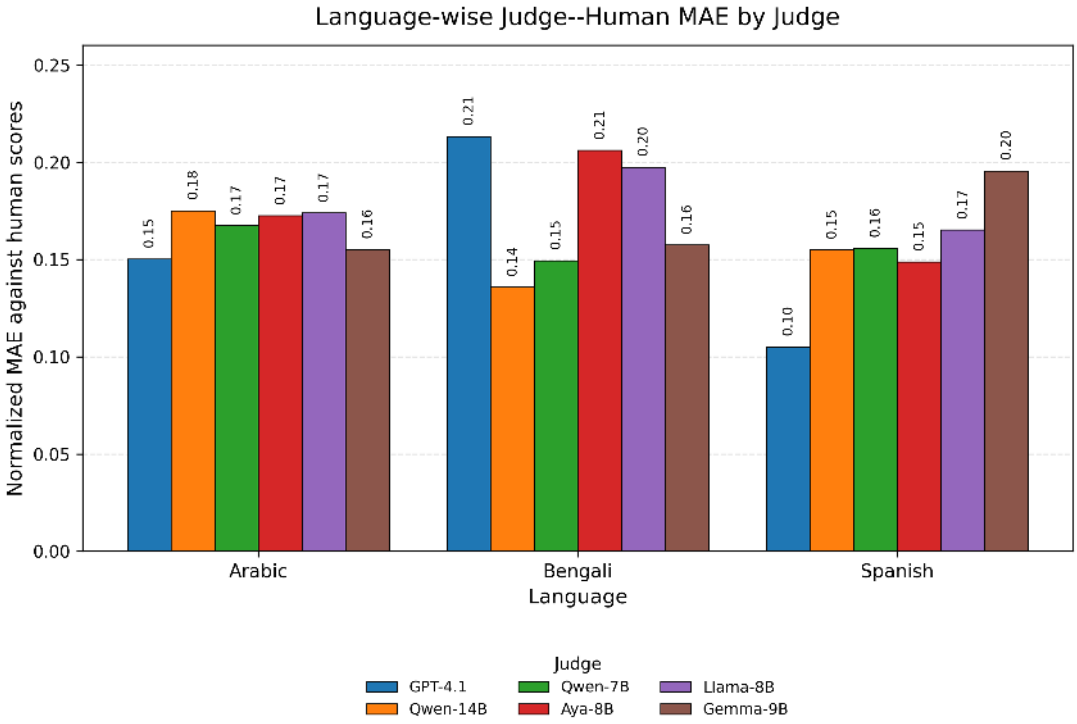


Figure 22: Language-wise judge–human MAE by judge under the six-judge setting. Values are reported on the normalized 0–1 scale.

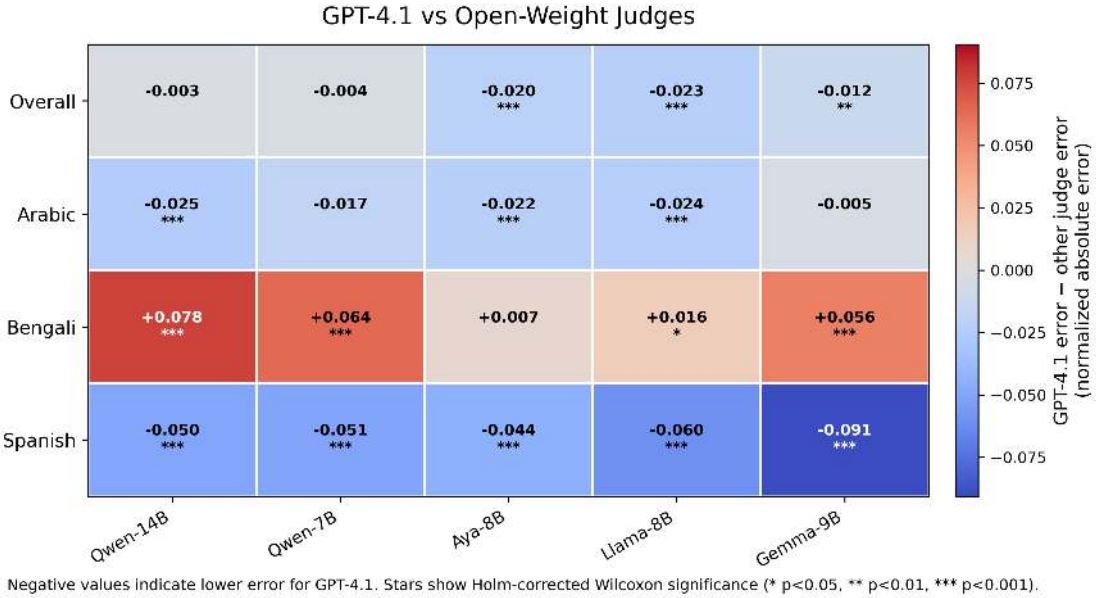


Figure 23: Pairwise comparison of GPT-4.1 against open-weight judges using Wilcoxon signed-rank tests over item-level normalized absolute errors. Each cell reports the mean error difference between GPT-4.1 and the comparison judge; negative values indicate lower error for GPT-4.1. Significance markers denote Holm-corrected p-values (* $p < .05$, ** $p < .01$, *** $p < .001$).

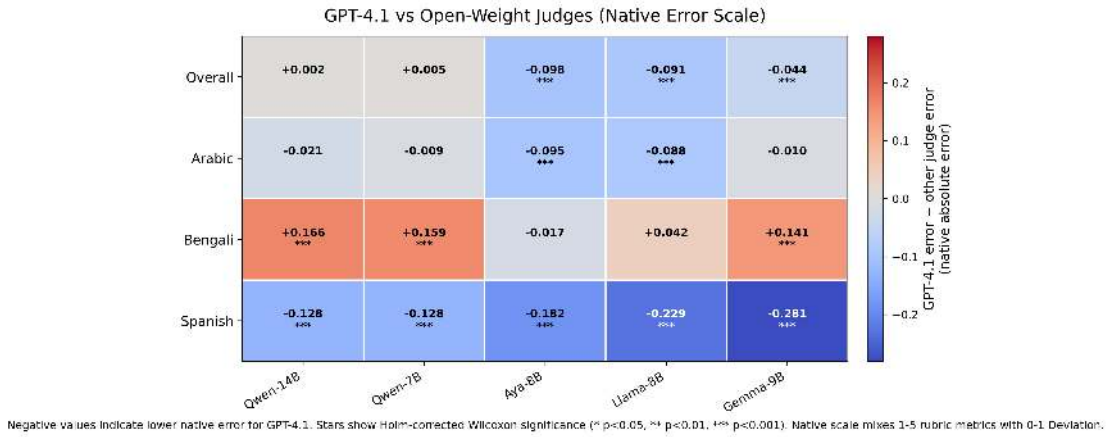


Figure 24: Native-scale pairwise comparison of GPT-4.1 against open-weight judges. Negative values indicate lower native absolute error for GPT-4.1. This diagnostic mixes 1–5 rubric metrics with 0–1 Deviation scores and is therefore reported as a supplementary robustness view rather than the primary comparison.

The normalized comparison shows that GPT-4.1 significantly outperforms all open-weight judges in Spanish and several judges in Arabic, but it is not significantly better than the strongest Qwen judges overall. Bengali shows the opposite pattern: GPT-4.1 has higher error than the strongest Qwen-based judges. The native-scale diagnostic leads to the same broad conclusion, although its magnitude should be interpreted cautiously because it combines metrics with different original score ranges.

G Additional Robustness Analyses

Following concerns about annotation ambiguity, score variance, multilingual transfer, and systematic judge bias, we conducted four additional analyses using the existing human annotations and judge outputs. These analyses are intended as robustness checks rather than causal validation of the structural–interpretive distinction.

G.1 Human-Agreement-Stratified Analysis

We first examined whether the structural–interpretive reliability gap could be explained primarily by instability in the human reference. Restricting evaluation to items with high agreement between the two human annotators preserves the overall pattern.

	Structural	**Interpretive**
MAE	**0.119**	0.194
Mean Spearman ρ	**0.560**	0.240

Table 32: Judge–human agreement on high-human-agreement items. Structural dimensions continue to show stronger alignment with human judgments than culturally interpretive dimensions.

Cultural Fluency (CF) remains difficult even within this comparatively stable subset, with MAE $=$ 0.215 and Spearman $\rho = 0.143$. Thus, poor judge–human agreement on interpretive dimensions cannot be attributed solely to items on which the human annotators themselves disagree.

G.2 Variance-Normalized Error

Because the evaluation dimensions differ in their human-score variance, we also normalize MAE by the standard deviation of the corresponding human scores.

The difference becomes smaller after normalization, indicating that score-distribution effects contribute to the observed gap, but they do not eliminate it. At the metric level, CF remains a clear outlier.

Evaluation group	MAE / Human SD
Structural	**0.615**
Interpretive	0.644

Table 33: Error normalized by human-score variability.

Metric	MAE / Human SD	Spearman ρ
Cultural Fluency	0.975	0.137

Table 34: Variance-normalized error and rank correlation for Cultural Fluency.

G.3 English vs. Target-Language Evaluation

We next use the English evaluation as a control to examine whether the difficulty of Cultural Fluency arises only from multilingual transfer.

Condition	CF MAE	Spearman ρ
English	**0.178**	**0.238**
Target language	0.237	0.028

Table 35: Cultural Fluency judge–human agreement under English and target-language evaluation.

CF remains comparatively difficult under English evaluation and degrades further in the target-language condition. This suggests that multilingual transfer contributes to the difficulty of cultural evaluation but does not fully account for it.

G.4 Direction of Cultural-Fluency Bias

Finally, we examine whether CF disagreement fluctuates randomly around the human reference or exhibits a consistent direction.

Language	Over-scoring CF	Under-scoring CF
Arabic	**4/6**	2/6
Bengali	**6/6**	0/6
Spanish	**4/6**	2/6

Table 36: Direction of mean Cultural Fluency bias across the six judge models.

The dominant tendency is CF over-scoring, particularly in Bengali where all six judges over-score relative to the human reference. This indicates that the disagreement is not purely random. However, the direction of bias alone does not establish its underlying cause, which may involve cultural knowledge, rubric interpretation, calibration, or other evaluator properties.