Are LLM Judges Language-Neutral? A Cross-Lingual Meta-Evaluation on Multilingual Summarization

Mohammaderfan Koupaei[1]*, Amir Esmaeili[1]*, Mahshid Khanali[1]*, Florian Mai[1,2], Patrícia Schmidtová[3]

[1]Bonn-Aachen International Center for Information Technology, University of Bonn, Germany,
[2]Lamarr Institute for Machine Learning and Artificial Intelligence, Germany
[3]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
{s32mkoup,s77aesma,s37mkhan}@uni-bonn.de schmidtova@ufal.mff.cuni.cz

Abstract

LLM-as-a-judge has emerged as a scalable proxy to human evaluation for natural language generation, yet its reliability across languages remains underexplored for open-weight judges. We conduct a systematic meta-evaluation of three open-weight judges on the WMT25 MIST Cross-Lingual Summarization subtask (X-Ling-Sum), measuring Kendall's τ against human scores across 14 languages and 4 criteria. At baseline, language-dependent gaps persist across all models and are robust under Kendall's τ, quadratic weighted kappa (QWK), and Spearman's ρ, with descriptive acc$\pm$1 gaps of 18–26pp between the best- and worst-performing languages. On a four-language subset (Arabic, Chinese, Russian, Korean), training-free interventions largely fail to reliably improve pooled rank agreement. The reliable levers are format-level and model-specific. Under the tested Egyptian-Arabic continued-pre-training configuration, no model significantly improves Arabic acc$\pm$1. Our results show that open-weight LLM judges are not language-neutral and require target-language validation before deployment in multilingual evaluation pipelines.

1 Introduction

Using a large language model (LLM) as a judge is the fastest growing evaluation trend across natural language generation (NLG) (Schmidtová et al., 2026), in some cases offering satisfactory agreement with human judgments at a fraction of the annotation cost (Zheng et al., 2023; Liu et al., 2023; Kasner et al., 2026). Proprietary judges dominate this space, but they come with significant drawbacks: high inference costs, opaque internals, and a tendency to be discontinued or silently updated after publication, undermining reproducibility. Open-weight judges offer a practical alternative, yet they remain comparatively understudied. Adoption has expanded into multilingual settings, where recruiting expert human evaluators for every language is prohibitively expensive and logistically difficult. The reliability of these judges across typologically diverse languages, however, remains largely unexplored.

Existing meta-evaluations validate LLM judges primarily on English benchmarks (Bavaresco et al., 2025; Kasner et al., 2026), where training data is abundant and morphological complexity is low. Cross-lingual summarization evaluation poses distinct challenges: languages differ substantially in script type, morphological complexity, and representation in model pretraining corpora. A judge well-calibrated on English may be systematically miscalibrated on Arabic, Korean, or Czech.

We conduct a systematic meta-evaluation of three open-weight LLM judges on the WMT25 Multilingual Instruction Shared Task (MIST) Cross-Lingual Summarization subtask (X-Ling-Sum), measuring judge agreement with human scores across 14 languages and 4 criteria. The baseline evaluation (RQ1) covers all 14 languages; the intervention experiments (RQ2–RQ4) are conducted on a four-language subset—Arabic, Chinese, Russian, and Korean—a linguistically diverse selection (Semitic, Sino-Tibetan, Slavic, Koreanic).

We organize the study around four research questions. **RQ1**: Do open-weight LLM judges agree with human scores equally across languages? **RQ2**: Can training-free prompt interventions—including prompt verbosity (the amount of detail provided to the judge), persona framing, and source-document presentation format—reliably improve pooled rank agreement? **RQ3**: Does role or strictness persona framing improve agreement or reduce scoring bias? **RQ4**: Can unlabeled continued pre-training on a low-resource language improve a judge's performance in that language?

Our contributions are as follows:

*Equal contribution.

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 1022–1040
October 17-21, 2026 ©2026 Association for Computational Linguistics

- **RQ1.** A systematic meta-evaluation of open-weight LLM judges on a 14-language cross-lingual summarization benchmark, revealing that language-dependent agreement gaps persist across all models and are robust under Kendall's τ, quadratic weighted kappa (QWK), and Spearman's ρ, with descriptive acc$\pm$1 gaps of 18–26pp between the best- and worst-performing languages.

- **RQ2.** Evidence that cheap prompt interventions largely fail to reliably improve pooled rank agreement under rank-correlation metrics on a four-language subset (Arabic, Chinese, Russian, Korean); no single source-presentation format wins across models, and the reliable levers are format-level and model-specific: joint scoring helps the Qwen judges.

- **RQ3.** Persona and cultural framing do not improve rank-agreement and can reduce it (same four-language subset); the strict-evaluator persona reduces overscoring bias rather than improving agreement.

- **RQ4.** A case study showing that, under the tested Egyptian-Arabic continued-pretraining configuration on a four-language subset (Arabic, Chinese, Russian, Korean), no model significantly improves Arabic acc$\pm$1; Gemma-4 and Qwen3-4B show significant overall acc$\pm$1 losses, while Phi-4's overall change is not significant.

2 Related Work

2.1 LLM-based Evaluation

Reference-based metrics such as BLEU and ROUGE correlate poorly with human judgments on open-ended generation (Novikova et al., 2017; Reiter, 2018), motivating many to shift to LLM-based evaluation. Zheng et al. (2023) established LLM-as-a-judge as a viable paradigm, showing GPT-4 exceeds 80% agreement with human preferences on MT-Bench. Later work applied this paradigm to the settings we build on: Liu et al. (2023) to summarization, Kim et al. (2024) to the kind of criterion-level scoring we use for our four criteria, and Kocmi and Federmann (2023) to machine translation (MT) quality estimation, evaluated with system-level pairwise accuracy and segment-level Kendall's τ – the metric family we adopt. That work found LLM judging reliable at the system level but not yet at the segment level – the item-level setting we study, and one that has been examined almost entirely in English rather than for open-weight judges across languages.

2.2 Bias and Reliability in LLM Judges

Wang et al. (2024) showed that swapping response order reverses GPT-4 verdicts at a conflict rate of 46.3%, exposing positional bias that undermines pairwise evaluation. Oi et al. (2024) demonstrated that token likelihood differences cause systematic over- and under-scoring independent of output quality. Liu et al. (2024) addressed score calibration through hierarchical criteria decomposition on English benchmarks, but the bias we document is language-dependent overscoring that varies systematically with source language and cannot be attributed to response ordering or likelihood effects. More broadly, Chehbouni et al. (2025) argue from a measurement theory perspective that enthusiasm for LLM judges has outpaced rigorous scrutiny of their validity, and that core assumptions – including their ability to act as proxies for human judgment and their scalability across settings – remain insufficiently tested. Reproducibility compounds this concern: Pattnayak and Bhatia (2026) find that model checkpoints, prompts, and decoding settings are routinely omitted, echoing the human evaluation reproducibility crisis identified by Belz et al. (2023). We therefore report full prompts, model configurations, and parse-failure rates for every experiment. Our work contributes an empirical test of one such assumption: that judge validity established in English transfers to typologically diverse languages.

2.3 Multilingual and Cross-lingual Evaluation

Clark et al. (2023) introduced SEAHORSE, a multilingual summarization evaluation benchmark covering six languages, and Gao et al. (2023) studied factuality in cross-lingual summarization, the property our faithfulness criterion measures. X-LingSum (Kocmi et al., 2025), our evaluation platform, extends this coverage to 14 languages with human scores on four criteria.

Recent work has begun to address multilingual judge reliability directly. Hada et al. (2024b) find that LLM-based evaluators degrade in calibration outside English, and METAL (Hada et al., 2024a) extends this to a multilingual meta-evaluation dataset. MM-Eval (Son et al., 2024) provides a multilingual meta-evaluation benchmark spanning

approximately 18 languages, covering judge and reward models across heterogeneous tasks including chat, reasoning, and safety. These efforts establish that judge quality varies by language, but do not study what a practitioner can do about it.

MENLO (Whitehouse et al., 2025) shows that native-like judge quality varies widely across 47 languages and that human-labeled fine-tuning narrows the gap – but this requires substantial annotation effort and, crucially, covers only naturalness, leaving criteria such as faithfulness unaddressed. Our work is complementary: rather than asking how much supervision can recover, we ask what a practitioner with no annotation budget would actually try.

3 Experimental Setup

Dataset. We evaluate on X-Ling-Sum (Kocmi et al., 2025), which provides human judgments for 11 system outputs across 14 languages, with 18 source instances per language (2,772 total). Annotators scored each output on four criteria: faithfulness, coverage, naturalness, and coherence (defined in Appendix B), using a 1–7 Likert scale. The dataset spans diverse script types (Latin, Arabic, Chinese-Japanese-Korean (CJK), Cyrillic) and resource levels, making it a demanding testbed for cross-lingual judge reliability.

Models. We evaluate three open-weight instruction-tuned judges. Qwen3-8B and Qwen3-32B (Yang et al., 2025) form a matched pair that isolates scale effects within the same model family; Gemma3-12B (Gemma Team, 2025) provides a cross-family reference point at an intermediate parameter count. All three cover the 8B–32B range relevant to practical deployment and report competitive multilingual benchmarks.

Metrics. Our primary agreement metric is Kendall's τ_b, computed over paired item-level LLM and human scores on the 1–7 Likert scale. Kendall's τ_b has served as the primary meta-evaluation metric across many editions of the WMT metrics shared task (Lavie et al., 2025) and remains a simple, well-understood, interpretable choice: it measures the proportion of correctly ordered pairs minus incorrectly ordered pairs, making it robust to the outlier sensitivity that affects Spearman's ρ and well-suited to ordinal scales with frequent ties. We use the tie-aware τ_b variant because both human and model scores contain many ties.

We also report acc$\pm$1 (the proportion of judgments falling within one point of the human score) as an auxiliary metric. It captures near-misses that τ_b penalises equally with large errors, but it should be interpreted with caution: on heavily skewed score distributions, a trivial constant predictor can attain a high acc$\pm$1 score, making it unsuitable as a standalone measure of judge quality. We therefore use acc$\pm$1 for paired comparisons within a fixed item set. Cross-language acc$\pm$1 differences are descriptive and are checked against Kendall's τ_b, QWK, and Spearman's ρ.

A recent alternative, Pairwise Difference Pearson (PDP) (DiIanni and Deutsch, 2025), computes a Pearson correlation over pairwise score differences rather than raw scores, and was proposed to recover the magnitude information that τ_b discards. We do not adopt it here for two reasons. First, PDP was developed and validated on Multidimensional Quality Metrics (MQM) annotations, which are continuous error-weighted scores; our judgments are single-rater 7-point Likert ratings, on which pairwise differences take few distinct values and ties are frequent. Second, a Pearson correlation over score differences presumes interval-scale spacing, which ordinal Likert responses do not guarantee, and it inherits Pearson's sensitivity to outliers. We therefore retain τ_b as our primary metric and treat acc$\pm$1 as an auxiliary paired statistic. Extending magnitude-aware meta-evaluation of this kind to ordinal judge ratings is a promising direction we leave to future work.

We track scoring bias separately as mean score difference (LLM minus human). Each judgment was generated once; if a model's output could not be parsed, the judgment was recorded as a failure and excluded from agreement and bias computation (failure rates below 0.3% except Qwen3-32B strict_evaluator at 5.9%, Appendix G). Agreement and bias are micro-averaged over all valid scored instances per reported condition.

Statistical testing. Agreement and bias estimates are accompanied by 95% confidence intervals from a cluster bootstrap over source documents (2,000 resamples). Comparisons between conditions are paired, recomputing both conditions on the same resampled documents. Because each language contributes only 18 source documents, per-language intervals are wide and should be read as approximate; claim-relevant paired comparisons appear in Appendix F.

Prompt design. Our evaluation prompts draw on two sources. The baseline criteria definitions

and rubrics are adapted from the WMT25 MIST shared task annotation guidelines (Kocmi et al., 2025). The intervention prompts (verbosity variants, persona framings, and source-presentation conditions) are modeled on conventions observed in LLM-as-a-judge practice, where such design choices are typically inherited from prior work or adopted from common practitioner patterns rather than empirically optimized. Following the taxonomy of Calò et al. (2026), our prompt design is grounded in the prior-research heuristic: we reuse established formulations and take inspiration from what practicioners use for their judges. Full prompt texts are given in Appendix D.

Infrastructure and implementation. Serving stack, decoding, exact model identifiers, token limits, and score-parsing details are given in Appendix A.

4 RQ1: Baseline Judge Reliability

We measured baseline agreement between each LLM judge and human annotators across all 14 languages under zero-shot, separate per-criterion prompting.

Overall agreement. Table 1 reports baseline Kendall's τ for all three judges. Gemma3-12B reaches 0.283, Qwen3-8B 0.282, and Qwen3-32B 0.271 – a narrow spread suggesting that neither model family nor scale is a reliable predictor of agreement. The acc$\pm$1 pattern tells the same story: the 32B model falls below the 8B model, confirming that larger scale within the same family does not guarantee higher agreement.

Language gap. Agreement varies substantially by language within every model (Figure 1). The gap between best- and worst-performing languages ranges from 18pp (Gemma3-12B) to 26pp (Qwen3-32B), with Qwen3-8B in between at 24pp (Table 1). These gaps remain well above zero even accounting for the small per-language sample size (full intervals in Appendix F). The worst-performing language differs by model (Czech for Qwen3-8B, Korean for Gemma3-12B, Arabic for Qwen3-32B), but these identities are metric-dependent and should not be read as stable findings.

The language gap is robust to metric choice. Because acc$\pm$1 is sensitive to the human score distribution, which varies by language, we verify that the gap holds under complementary metrics.

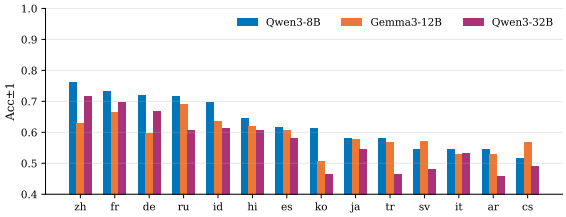


Figure 1: acc$\pm$1 per language and model under RQ1 baseline conditions.

Kendall's τ, QWK, and Spearman's ρ all show substantial best-to-worst variation within every judge (Table 14). The magnitude of cross-lingual variation is therefore not solely an artifact of the acc$\pm$1 label distribution.

Systematic bias. Bias direction differs across model families (Table 1): both Qwen3 models overscore while Gemma3-12B remains near-neutral. Qwen3-32B pairs the highest bias with the lowest overall agreement, suggesting that overscoring and accuracy are distinct failure modes that do not trade off against each other.

Criterion breakdown. Agreement differs across the four criteria, and the ordering is model-dependent. For the Qwen3 judges, coherence is easiest and naturalness hardest – Qwen3-8B reaches 71.2% on coherence but only 55.4% on naturalness. Gemma3-12B departs from this pattern, scoring highest on naturalness (66.6%) and lowest on faithfulness (49.3%), indicating that criterion difficulty is not uniform across judges.

Score distribution. Both Qwen3 models concentrate predicted scores near the top of the 1–7 scale regardless of language or criterion, consistent with their positive overscoring bias; Gemma3-12B distributes scores more broadly.

5 RQ2: Training-Free Interventions

5.1 Source Presentation

X-Ling-Sum source instances are massively multilingual – a single instance bundles reviews in two to seven languages – so the original human evaluation translated every source to English before annotation, since annotators could not read them all. Crucially, only the source inputs were translated; the system outputs being evaluated were never translated. LLM judges are not bound by the same constraint: they can be shown the raw multilingual *Original*, the *English* translation used

Model	τ	acc$\pm$1	Best lang. (acc$\pm$1)	Worst lang. (acc$\pm$1)	Gap	Bias
Qwen3-8B	0.282	**62.9%**	zh (**76.0%**)	cs (**51.6%**)	24.4pp	+0.598
Gemma3-12B	**0.283**	59.2%	ru (69.1%)	ko (50.8%)	**18.3pp**	**−0.136**
Qwen3-32B	0.271	56.6%	zh (71.7%)	ar (45.7%)	26.0pp	+0.758

Table 1: Baseline judge performance across 14 languages. τ is Kendall's τ_b; gap is best−worst language acc$\pm$1. Bias is the mean of LLM scores minus human scores. **Bold**: best per column; underline: second best.

Model	English	Original	Both
Qwen3-8B	0.375 (71.1)	0.342 (68.5)	**0.394** (69.5)
Gemma3-12B	**0.342** (63.6)	0.313 (63.3)	0.317 (61.3)
Qwen3-32B	0.351 (67.6)	0.346 (70.5)	**0.376** (71.2)

Table 2: Kendall τ (acc$\pm$1 % in parentheses) by source presentation format **Bold**: best τ per row.

Model	Short	Medium	Long
Qwen3-8B	0.331 (55.8)	**0.350** (70.1)	0.284 (67.9)
Gemma3-12B	0.181 (63.3)	0.334 (64.2)	**0.368** (56.7)
Qwen3-32B	0.328 (67.7)	**0.354** (67.8)	0.346 (60.5)

Table 3: Kendall τ (acc$\pm$1 % in parentheses) by prompt verbosity. **Bold**: best τ per row.

Model	None	Basic	Detailed
Qwen3-8B	**0.379** (71.5)	0.362 (73.4)	0.351 (73.5)
Gemma3-12B	0.344 (64.1)	0.323 (66.8)	**0.364** (71.8)
Qwen3-32B	0.362 (68.5)	0.347 (68.1)	**0.370** (69.5)

Table 4: Kendall τ (acc$\pm$1 % in parentheses) by generic evaluator persona detail. **Bold**: best τ per row.

by human annotators, or *Both*. We ask whether reproducing the human-evaluation format (English) helps or hurts relative to the alternatives. Experiments are restricted to Arabic, Chinese, Russian, and Korean as a typologically diverse representative sample.

No single presentation format wins across models. Table 2 shows that the best format is model-dependent: Gemma3-12B peaks on English (0.342), while both Qwen models peak on Both (Qwen3-8B: 0.394, Qwen3-32B: 0.376). The overall advantage of Both over English is not statistically significant for either Qwen model (Appendix F). For Qwen3-32B on Arabic, the largest numerical changes favour access to the original source (Original−English +0.051, Both−English +0.059).

English-only was a conservative default in the four-language subset: no model peaks on the raw Original alone, and the human-evaluation format (English only) is never the worst option. Adding the original was model-dependent and should be validated for the target model.

5.2 Prompt Verbosity and Detail of Guidelines

All experiments in this section cover Arabic, Chinese, Russian, and Korean. The three verbosity conditions differ in prompt length: Short (77 words including placeholders), Medium (134 words), and Long (268 words); full prompt texts are given in Appendix D.2.

Verbosity effects are mixed and model-specific. Table 3 shows no consistent pattern across models. For Qwen3-8B, medium outperforms long ($\Delta\tau = +0.066$, $p < 0.001$); for Qwen3-32B the differences are small and only the medium-short gap reaches significance ($+0.026$, $p = 0.038$). Gemma3-12B shows the largest sensitivity: short prompts yield a markedly low τ of 0.181 – well below its medium (0.334) and long (0.368) conditions – suggesting it requires more context to produce stable judgments, though we do not have sufficient detail to explain this further.

Adding more detailed guidelines does not help. Table 4 shows that adding more detailed guidelines to the baseline prompt does not significantly raise τ for any model and significantly lowers it for Qwen3-8B (detailed−none $\Delta\tau = -0.029$, $p = 0.009$); effects are not significant for the larger models. The acc$\pm$1 values rise slightly with more detail, but as established in Section 3, this reflects the skewed score distribution rather than genuine improvement in rank ordering.

Neither prompt verbosity nor generic evaluator detail provides a universal lever. Medium verbosity is the safest default – it is never the worst-performing condition across any model – but the gains from tuning either dimension are too uneven and too small to reliably improve pooled rank agree-

Model	Separate	Joint	$\Delta\tau$
Qwen3-8B	0.278 (62.3)	**0.302** (66.0)	$+0.024^{*}$
Gemma3-12B	**0.298** (60.0)	0.292 (64.5)	-0.006
Qwen3-32B	0.281 (56.8)	**0.342** (67.2)	$+0.061^{***}$

Table 5: Kendall τ (acc$\pm$1 % in parentheses): separate vs. joint criterion scoring across 14 languages. $\Delta\tau$ is the paired joint$-$separate change (2,000-resample bootstrap). **Bold**: higher τ per row. $^{*}p < .05$, $^{***}p < .001$.

ment across languages.

5.3 Format Ablation: Separate vs. Joint Criterion Scoring

Before testing prompt interventions, we establish the elicitation format used throughout RQ2 and RQ3. We compare separate per-criterion prompt calls against joint JSON elicitation of all four criteria in a single call, run across all 14 languages. Based on these results, we adopt joint elicitation as the default for all subsequent experiments.

Joint JSON scoring improves rank-agreement for the Qwen judges but not Gemma3-12B. Table 5 shows joint elicitation raises τ for Qwen3-32B ($\Delta\tau = +0.061$, $p < 0.001$) and Qwen3-8B ($+0.024$, $p = 0.025$), while the change for Gemma3-12B is not significant (-0.006, ns). The magnitude of the Qwen3-32B gain is comparable to the largest prompt intervention effects observed in RQ2. No language is consistently helped or hurt across all three models.

Criterion-level analysis. Joint gains concentrate in the Qwen judges and on the structural criteria (Table 6). Joint elicitation significantly raises τ for both Qwen models on coverage and coherence, with gains of 5–10 τ points (all $p \leq 0.001$). Gemma3-12B instead loses on coverage under joint ($\Delta\tau = -0.083$, $p = 0.002$), with no significant change elsewhere. Naturalness agreement is not significantly affected for any model.

Finding. Joint JSON elicitation improves rank-agreement for the Qwen model family and is neutral for Gemma3-12B, making it the stronger default across the three judges. We therefore adopt joint elicitation for all subsequent experiments in RQ2 and RQ3. Note that RQ1 used separate per-criterion calls, so absolute agreement values there are not directly comparable to those in later sections; all within-RQ2 and within-RQ3 comparisons are unaffected.

6 RQ3: Role and Strictness Persona Prompting

We test whether assigning a specific role (native speaker, cultural expert, linguist, or strict evaluator) affects agreement and bias, across Arabic, Chinese, Russian, and Korean on all three judges. Each role is injected as the system message of the chat-completion call (a "You are a ..." prefix delivered via the `system` role), leaving the evaluation prompt in the user turn unchanged. The role-persona baseline is a separate no-persona run from the generic-persona baseline in Section 5.2; the two were run independently and differ by up to 0.3pp.

Role personas do not improve rank-level agreement and can reduce it. Table 7 shows that no role persona raises τ significantly for any model. The pattern is most damaging for the smallest judge: for Qwen3-8B, native_speaker, cultural_expert, and linguist each significantly reduce agreement (native_speaker the largest at $\Delta\tau = -0.049$), while strict_evaluator is null. For Gemma3-12B only strict_evaluator reaches significance, in the negative direction, and no Qwen3-32B persona is significant in either direction (full deltas and p-values in Appendix F). This suggests persona sensitivity decreases with model scale, with the smallest model most vulnerable to role framing that pulls it away from its default judgment behavior.

Language-level effects. For Qwen3-8B, the largest Arabic reductions occur under the native_speaker, cultural_expert, and linguist personas, reaching $\Delta\tau = -0.054$. For the larger models, per-language persona deltas are small and rarely significant. The cultural persona framing thus appears to confuse rather than help the smallest judge when evaluating Arabic text – the opposite of its intended effect.

Bias reduction. The strict_evaluator persona reduces overscoring bias substantially for all three models (Table 8), and all three paired baseline-to-strict reductions are significant (largest for Gemma3-12B at $+0.508$; full CIs and p-values in Appendix F). For Qwen3-32B the marginal bias under strict_evaluator is near zero (-0.018), though the interval is wide owing to the 5.9% parse failure rate in that condition.

The one notable exception is the native_speaker persona for Qwen3-32B, which *worsens* bias to

	Qwen3-8B		Gemma3-12B		Qwen3-32B	
Criterion	Sep.	Joint	Sep.	Joint	Sep.	Joint
Faithfulness	0.291 (62.8)	**0.310** (75.9)	0.275 (50.8)	**0.302** (75.1)	0.264 (59.3)	**0.339** (68.8)
Coverage	0.292 (70.2)	**0.345** (74.0)	**0.276** (60.6)	0.193 (59.3)	0.302 (58.3)	**0.336** (66.7)
Coherence	0.312 (65.2)	**0.369** (60.3)	0.324 (62.6)	**0.355** (65.9)	0.250 (58.5)	**0.346** (69.8)
Naturalness	0.353 (51.0)	**0.360** (53.9)	**0.378** (66.1)	0.361 (57.8)	**0.382** (51.0)	0.366 (63.3)

Table 6: Kendall τ (acc$\pm$1 % in parentheses) by criterion and format. **Bold**: higher τ per criterion per model.

Model	Baseline	Native sp.	Cultural exp.	Linguist	Strict eval.
Qwen3-8B	**0.384** (71.4)	0.335 (70.4)	0.353 (72.0)	0.355 (72.6)	0.383 (74.8)
Gemma3-12B	0.345 (64.3)	0.339 (67.4)	**0.360** (66.8)	0.339 (66.4)	0.307 (72.5)
Qwen3-32B	0.355 (68.5)	0.364 (66.5)	0.363 (68.6)	0.364 (68.9)	**0.374** (71.5)

Table 7: Kendall τ (acc$\pm$1 % in parentheses) by model and role persona (four-language subset). **Bold**: best τ per row. Qwen3-32B strict_evaluator is computed on 2,980 of 3,168 items owing to parse failures (Appendix G); paired deltas and p-values are in Appendix F.

$+0.544$ – the largest increase over baseline of any persona-model combination (Table 8).

Parse failures under strict_evaluator. The 5.9% failure rate for Qwen3-32B strict_evaluator (188 of 3,168 judgments) is concentrated almost entirely in Chinese: 180 of 188 failures are Chinese and 8 are Korean, clustered on five Chinese source documents, three of which fail on all 44 system-by-criterion judgments. This suggests the strict_evaluator persona interacts with something specific to those documents rather than with Chinese text in general (Table 12).

Finding. Role and strictness personas do not improve rank-level agreement and can meaningfully degrade it, particularly for smaller models and in Arabic. The strict_evaluator persona is the exception on bias: it substantially reduces overscoring across all three models, at the cost of elevated parse failures for Qwen3-32B. Persona framing is therefore not a reliable path to better agreement, but may be useful as a targeted bias correction when overscoring is the primary concern.

7 RQ4: Can Unlabeled Continued Pre-training Fix a Weak Judge?

Motivation. The costly part of judge supervision is human labels; unlabeled in-domain text is cheap and available without a corporate budget. If a judge is weak on a low-resource language, is that a language-knowledge gap or a judging-skill gap? If it were only language knowledge, unsupervised continued pre-training (CPT) on in-domain dialectal text should help – and naturalness, the criterion most directly tied to what fluent language looks like, should benefit most.

Design. LoRA fine-tuning on the RQ1–RQ3 judges (8B–32B parameters) exceeded our compute budget, so we select three smaller instruction-tuned models for this experiment. We continue-pretrain these three instruction-tuned judges – Gemma-4-E2B-it, Qwen3-4B-Instruct-2507, and Phi-4-mini-instruct – with low-rank adaptation (LoRA) on an unstructured raw-text corpus of $\sim$3,800 Egyptian Arabic product reviews (Jumia), chat templates removed so the causal-LM objective sees only raw dialectal text (hyperparameters in Table 10). We re-evaluate on the four-language RQ2/RQ3 subset and compare each model before vs. after fine-tuning, paired by source document with a 2,000-resample cluster bootstrap. Chinese, Russian, and Korean are evaluated as a regression check to determine whether Arabic-targeted CPT reduces performance in languages that were not included in the CPT training. Some degradation in these languages would not be surprising; the more important finding is that Arabic itself does not improve after the Arabic-targeted CPT.

No model improves overall acc$\pm$1; the significant overall changes are negative. Table 9 shows significant overall acc$\pm$1 losses for Gemma-4 (-15.9pp) and Qwen3-4B (-3.0pp), while Phi-4's overall change is not significant. Crucially,

Model	Baseline	Native sp.	Cultural exp.	Linguist	Strict eval.
Qwen3-8B	+0.812	+0.800	+0.739	+0.672	**+0.477**
Gemma3-12B	+0.991	+0.855	+0.880	+0.874	**+0.484**
Qwen3-32B	+0.367	+0.544	+0.381	+0.401	**−0.018**

Table 8: Mean scoring bias (LLM minus human score) by model and persona. **Bold**: lowest absolute bias per row.

Model	Overall	ar	zh	ru	ko
Gemma-4	-15.9^{***}	-7.5^{*}	-25.4^{***}	-19.9^{***}	-10.6^{*}
Qwen3-4B	-3.0^{***}	-2.4	-4.5^{*}	-3.2^{**}	-2.0
Phi-4	-1.2	$+0.6$	-5.8	-3.3	$+3.7$

Table 9: acc±1 change (pp) after Egyptian Arabic CPT, paired bootstrap. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

no model significantly improves Arabic acc±1 under the tested configuration: Phi-4 is essentially unchanged, Qwen3-4B shows a small non-significant loss, and Gemma-4 degrades significantly (−7.5pp).

Naturalness is spared; structural criteria absorb the damage. Fine-tuning on dialectal text should, if anything, improve naturalness judgment – it gives the model its best chance to learn what natural language looks like in the target variety. Instead, naturalness acc±1 is not significantly harmed for any model, while the acc±1 losses concentrate on the structural criteria (faithfulness, coverage, coherence), most severely for Gemma-4 (up to −22.3pp; per-criterion changes in Table 15). The criterion that unlabeled dialectal exposure should help most is the one least affected.

Scoring disposition becomes harsher. Fine-tuning shifts scoring bias significantly harsher for Gemma-4 (the largest, −0.587) and Phi-4 – both begin penalizing summaries more strictly than human annotators – while Qwen3-4B's bias is unchanged (deltas, CIs, and p-values in Table 15). For Phi-4, the dominant story is this bias shift rather than accuracy: its acc±1 change is null, but its scoring disposition has moved substantially.

Figure 2 visualizes the bias shift: Gemma-4 moves sharply toward harsher scoring on every criterion, Phi-4 follows the same direction more moderately, and Qwen3-4B is essentially unchanged. A full per-language, per-criterion acc±1 breakdown is given in Figure 3 (Appendix), confirming that no model shows a consistent Arabic gain and that Gemma-4 degrades broadly, with Chinese and Russian absorbing the largest losses.

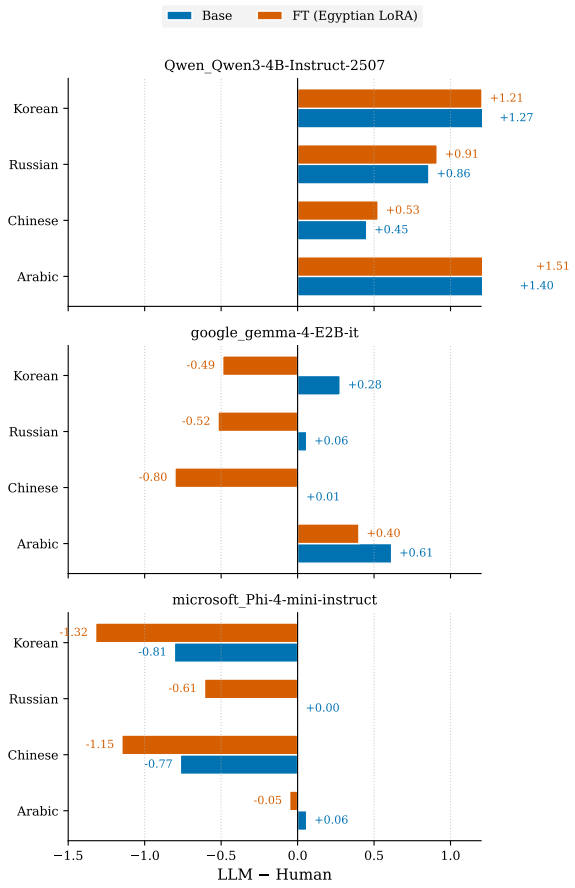


Figure 2: Per-criterion scoring bias (mean score − human gold) before and after Egyptian Arabic CPT. Fine-tuning shifts Gemma-4 and Phi-4 toward harsher scoring across all criteria, while Qwen3-4B is largely unchanged. Per-language acc±1 heatmaps are in Figure 3 (Appendix).

8 Discussion

Bias and agreement are independent failure modes. Interventions affecting scoring bias do not correspondingly affect rank-level agreement, and vice versa. The strict_evaluator persona substantially reduces overscoring while leaving τ unchanged or worsening it; joint JSON elicitation improves τ for the Qwen judges without shifting bias. A practitioner diagnosing a judge as unreliable therefore needs to identify which failure mode is present before reaching for a fix: a judge that overscores but ranks correctly needs a different in-

tervention from one that scores near human means but scrambles the ranking.

The language gap is not straightforwardly predicted by data availability. A natural hypothesis is that judges perform poorly on low-resource languages because those languages are underrepresented in pretraining. Our results complicate this picture: the worst-performing language differs across models (cs for Qwen3-8B, ko for Gemma3-12B, ar for Qwen3-32B), and the RQ4 continued pre-training experiment, which adds target-language exposure, fails to improve Arabic agreement. These observations do not rule out data volume as a contributing factor – the three models have different training mixtures, so model-specific weaknesses are expected regardless – but they suggest that the gap is not reducible to pretraining coverage alone.

The agreement metric problem. We lack consensus on what metric to use when measuring judge-human agreement. Existing papers variously report Pearson correlation, Spearman's ρ, accuracy and its variants, making cross-paper comparison effectively impossible – and as we show, these metrics can tell contradictory stories on the same data (Table 13). The MT community has addressed this through the WMT Metrics shared task (Lavie et al., 2025), but NLG evaluation has no equivalent and its needs differ: it has more diverse criteria and annotation types. We call for the community to treat the choice of agreement metric as a research question in its own right.

The structural validation gap. To contextualize our findings, we performed a survey of 31 recent multilingual NLG papers, out of which 19 actually employ LLM judges. [1] Only six validate their LLM judge in all evaluated languages – and all six are MT papers reusing pre-existing benchmarks where human scores are available by default. Four papers validate only on a subset of languages, six omit validation entirely, and three do not report their validation procedure clearly enough to assess. Papers conducting new human evaluation face a real cost to full validation, and our results show it is worth paying: agreement gaps of 18–26pp between best and worst languages within a single judge are large enough that unvalidated scores may be substantially misleading. We are not calling for every paper to conduct a full meta-evaluation; at minimum, authors should manually inspect a handful of judge outputs per language, particularly in languages they do not themselves read.

Practical recommendations. In our four-language subset, English-only source presentation was at least as good as including the original for all three judges; adding the original helped in some model–language combinations but hurt in others. Joint JSON elicitation improved τ for the two Qwen3 models we tested; medium verbosity was a reasonable default across conditions. This suggests overly long and elaborate prompts are not likely to perform better. If overscoring bias is the primary concern, the strict_evaluator persona reduced it for all three RQ3 judges, but did not improve rank agreement. Under the tested Egyptian Arabic CPT configuration, no model significantly improved Arabic acc$\pm$1, so these results do not support this particular CPT setup as a substitute for labeled supervision. None of the training-free prompt interventions we tested reliably improved pooled rank agreement across languages, suggesting that prompt-level adjustments alone may not be sufficient to close the cross-lingual gap.

9 Conclusion

We evaluated three open-weight LLM judges on the X-Ling-Sum benchmark across 14 languages and four evaluation criteria. Our central finding is that these judges are not language-neutral: agreement with human scores varies by 18 to 26 percentage points between the best- and worst-performing languages within every model, and this gap is robust across three correlation metrics. Training-free prompt interventions, including verbosity tuning, persona framing, and source-presentation format, do not reliably close this gap; the effective levers are format-level and model-specific. Continued pre-training on unlabeled dialectal text does not improve judge quality in the target language under the configuration we tested. These results point to a structural limitation that practitioners cannot work around with just prompt engineering alone: multilingual deployment of LLM judges requires per-language validation, and the NLG community would benefit from standardized agreement metrics to make such validation comparable across studies.

[1] Papers identified via keyword search using ACL Anthology 2020–2025 and categorized manually; per-paper classifications are listed in Table 11.

Limitations

Language coverage. RQ2 and RQ3 cover only Arabic, Chinese, Russian, and Korean—a linguistically diverse subset (Semitic, Sino-Tibetan, Slavic, Koreanic) selected under compute constraints. Whether the prompt interventions generalize to the remaining ten X-Ling-Sum languages is not established, and cross-lingual transfer of optimal configurations should not be assumed.

Statistical power. Each language contributes only 18 source documents, yielding wide per-language confidence intervals even under a cluster bootstrap; unpaired differences below roughly 2pp are not resolvable, though paired within-document comparisons have greater power.

Annotation source. X-Ling-Sum assigns each summary to a single annotator: three annotator slots per language partition the 306 rated items into 102 each, with no item rated more than once. Inter-annotator agreement is therefore not computable for any language, since no two annotators judge the same item. Annotator identifiers are language-local, so the total number of distinct annotators across the 14 languages lies between 3 and 42 and cannot be recovered from the released data. Kocmi et al. (2025) report neither annotator qualification criteria nor agreement statistics for this subtask. A judge–human gap may therefore reflect annotation noise rather than judge error, and this ambiguity is sharpest in languages where annotator expertise is least verifiable. We treat human scores as the reference standard throughout, as the benchmark intends, but this assumption is untested per language.

Metric. Kendall's τ_b is our primary agreement metric, but it is still computed on averaged ordinal scores with many ties and should not be interpreted as a complete model-quality measure. acc$\pm$1 is lenient and, on this top-skewed label distribution, is exceeded by a constant predictor; we use it only as an auxiliary relative comparison. Absolute acc$\pm$1 values should not be interpreted as judge quality.

Experimental design. Persona instructions were delivered in English regardless of target language, which may understate any benefit of culturally-oriented personas. Verbosity, source presentation, and persona conditions were evaluated sequentially rather than jointly optimized, so interaction effects between them remain uncharacterized.

The separate and joint conditions differ in prompt content as well as call structure (rubric length and anchor detail), so the joint-vs-separate effect cannot be attributed to output format alone; a content-matched ablation is left to future work.

The Egyptian Arabic training corpus may not match the register of the Arabic evaluation texts, so RQ4 may reflect cross-register transfer rather than same-variety adaptation.

Ethics Statement

This study uses only publicly available data and models. Human evaluation scores were taken directly from X-Ling-Sum (Kocmi et al., 2025); no human annotators were recruited or compensated by the authors. All three LLM judges are open-weight models released for research use. No personal data was collected or processed at any stage of this study.

Data Availability Statement

X-Ling-Sum and the pretrained model checkpoints used in this study are publicly available from their original sources. The prompts, analysis code, generated judgments, bootstrap outputs, and continued-pre-training configuration will be released upon publication, subject to the redistribution terms of the underlying Egyptian Arabic corpus.

Acknowledgments

This work was partially supported by the Lamarr Institute and b-it center, co-funded by the European Union (ERC, NG-NLG, 101039303) and by Charles University projects GAUK 252986 and SVV project 260 821.

References

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan,

Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Mingqi Gao, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, and 23 others. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Eduardo Calò, David M. Howcroft, Leo Leppänen, Saad Mahamood, Simon Mille, Patrícia Schmidtová, and Emiel van Miltenburg. 2026. Justify your prompts! *Computational Linguistics*, pages 1–13.

Khaoula Chehbouni, Mohammed Haddou, Jackie Chi Kit Cheung, and Golnoosh Farnadi. 2025. Neither valid nor reliable? investigating the use of llms as judges. *Preprint*, arXiv:2508.18076.

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur Parikh. 2023. SEAHORSE: A multilingual, multifaceted dataset for summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9397–9413, Singapore. Association for Computational Linguistics.

Colten DiIanni and Daniel Deutsch. 2025. Don't sweat the small stuff: Segment-level meta-evaluation based on pairwise difference correlation. *Preprint*, arXiv:2509.25546.

Mingqi Gao, Wenqing Wang, Xiaojun Wan, and Yuemei Xu. 2023. Evaluating factuality in cross-lingual summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12415–12431, Toronto, Canada. Association for Computational Linguistics.

Gemma Team. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Rishav Hada, Varun Gumma, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024a. METAL: Towards multilingual meta-evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2280–2298, Mexico City, Mexico. Association for Computational Linguistics.

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2024b. Are large language model-based evaluators the solution to scaling up multilingual evaluation? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1051–1070, St. Julian's, Malta. Association for Computational Linguistics.

Zdeněk Kasner, Vilém Zouhar, Patrícia Schmidtová, Ivan Kartáč, Kristýna Onderková, Ondrej Platek, Dimitra Gkatzia, Saad Mahamood, Ondrej Dusek, and Simone Balloccu. 2026. LLMs as span annotators: A comparative study of LLMs and humans. In *Proceedings of the First Workshop on Multilingual Multicultural Evaluation*, pages 1–22, Rabat, Morocco. Association for Computational Linguistics.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Tom Kocmi, Sweta Agrawal, Ekaterina Artemova, Eleftherios Avramidis, Eleftheria Briakou, Pinzhen Chen, Marzieh Fadaee, Markus Freitag, Roman Grundkiewicz, Yupeng Hou, Philipp Koehn, Julia Kreutzer, Saab Mansour, Stefano Perrella, Lorenzo Proietti, Parker Riley, Eduardo Sánchez, Patricia Schmidtova, Mariya Shmatova, and Vilém Zouhar. 2025. Findings of the WMT25 multilingual instruction shared task: Persistent hurdles in reasoning, generation, and evaluation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 414–435, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. Association for Computational Linguistics.

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.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024. HD-eval: Aligning large language model evaluators through hierarchical criteria decomposition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7641–7660, Bangkok, Thailand. Association for Computational Linguistics.

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.

Masanari Oi, Masahiro Kaneko, Ryuto Koike, Mengsay Loem, and Naoaki Okazaki. 2024. Likelihood-based mitigation of evaluation bias in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3237–3245, Bangkok, Thailand. Association for Computational Linguistics.

Priyaranjan Pattnayak and Apoorv Bhatia. 2026. ReproEvalCard: A reporting standard for reproducible evaluation of LLM pipelines. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–249, San Diego, California, United States. Association for Computational Linguistics.

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

Patrícia Schmidtová, Saad Mahamood, and Ondřej Dušek. 2026. Never truly out of fashion: A retrospective look at evaluation in NLG. In *Proceedings of the 1st Symposium on Natural Language Generation Evaluations*, pages 63–72, Aberdeen, United Kingdom. Association for Computational Linguistics.

Guijin Son, Dongkeun Yoon, Juyoung Suk, Javier Aula-Blasco, Mano Aslan, Vu Trong Kim, Shayekh Bin Islam, Jaume Prats-Cristià, Lucía Tormo-Bañuelos, and Seungone Kim. 2024. MM-Eval: A multilingual meta-evaluation benchmark for LLM-as-a-judge and reward models. *Preprint*, arXiv:2410.17578.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Chenxi Whitehouse, Sebastian Ruder, Tony Lin, Oksana Kurylo, Haruka Takagi, Janice Lam, Nicolò Busetto, Denise Diaz, and Francisco Guzmán. 2025. MENLO: From preferences to proficiency – evaluating and modeling native-like quality across 47 languages. *arXiv preprint arXiv:2509.26601*, arXiv:2509.26601.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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, pages 46595–46623.

A Infrastructure and Implementation Details

Models were served via vLLM on NVIDIA A40 and A100 GPUs on an academic HPC cluster. All inference used greedy decoding (temperature 0.0) to ensure reproducibility. Exact model identifiers were `Qwen/Qwen3-8B`, `Qwen/Qwen3-32B`, and `google/gemma-3-12b-it`, served via vLLM 0.11.2 with tensor parallelism 2. Maximum output tokens were set to 20 for separate-criterion calls and 120 for joint JSON calls. Scores were extracted by parsing the first integer in the range 1–7 from the model response; joint JSON responses were parsed as JSON objects with integer fields. Responses that could not be parsed were excluded from agreement and bias computation; per-condition failure counts are in Appendix G.

B Evaluation Criteria

Human annotators and LLM judges scored each summary on four criteria on a 1–7 scale (Kocmi et al., 2025); the full rubrics appear in Appendix D.

Faithfulness whether all information in the summary is supported by the source reviews, with no contradicted or unverifiable claims.

Coverage whether the summary captures the key points of the source reviews.

Naturalness whether the summary reads as fluent, natural text in the target language, appropriate in tone and formality.

Coherence whether the summary is logically sound, well-structured, and internally consistent.

C LoRA Fine-tuning Configuration

Table 10 lists the configuration used for the Egyptian Arabic dialectal adaptation experiments in Section 7. The adaptation used unsupervised continued pre-training (causal LM objective) via LoRA on a raw-text corpus of approximately 3,800 Egyptian Arabic Jumia product reviews, with all chat templates removed during training.

Hyperparameter	Value
Base models	Qwen3-4B-Instruct-2507, gemma-4-E2B-it, Phi-4-mini-instruct
Corpus	Jumia Egyptian Arabic reviews (~3,800)
Corpus format	Raw text, chat templates removed
Objective	Causal LM continued pre-training (unsupervised)
Adaptation	LoRA, bf16 (not quantized)
LoRA rank (r)	16
LoRA alpha	32
LoRA dropout	0.05
Target modules	all-linear
Learning rate	2e-4
Epochs	3
Batch size	4 per device (grad. accum. 4, effective 16)
Max sequence length	512
LR scheduler	Cosine, warmup ratio 0.05
Weight decay	0.01
Optimizer	adamw_torch_fused
Gradient checkpointing	Enabled
Seed	42
Hardware	1× NVIDIA A100

Table 10: LoRA fine-tuning configuration for this case study's dialectal adaptation experiments.

D Prompt Templates

Prompts below are copied verbatim from the inference code. Braced tokens ({language}, {input}, {response}, {input_block}) are placeholders filled at runtime; {input} is the English translation of the source reviews and {response} the summary. Output-format braces shown as {...} are the rendered single-brace form of the doubled braces in the source. All calls use greedy decoding. Message roles are noted per experiment.

D.1 RQ1 Baseline: Zero-shot, Separate Calls

Each criterion is scored in its own call as a single user message (no system message).

```
[faithfulness]
Score the summary generated by a system based on a set of
      reviews in {language} on a likert scale from 1 to 7.
      Evaluate whether all information in the summary can be
      traced back to the reviews. Treat the reviews as the
      source of truth and do not consider any external
      information. The quality levels associated with
      numerical scores are provided below:
7: All of the information in the summary is fully supported
      by the reviews and no meaning was changed.
5: Most information is supported, but a small part of the
      summary contains information that either contradicts or
      cannot be verified by the reviews.
3: More than half of the information in the summary either
      contradicts or cannot be verified by the reviews.
1: The summary is fully made up of information that either
      contradicts or cannot be verified by the reviews.
Return only the score and nothing else.

Reviews: {input}
Summary: {response}

Score (1-7):

[coverage]
Score the summary generated by a system based on a set of
      reviews in {language} on a likert scale from 1 to 7.
      Read the reviews and identify the most important points,
       then evaluate whether these key points are covered by
      the summary. The quality levels associated with
      numerical scores are provided below:
7: The summary covers all key points.
5: The summary covers about two thirds of the key points.
3: The summary covers about a third of the key points.
1: The summary does not cover any of the key points mentioned
       in the reviews.
Return only the score and nothing else.

Reviews: {input}
Summary: {response}

Score (1-7):

[naturalness]
Score the summary generated by a system based on a set of
      reviews in {language} on a likert scale from 1 to 7.
      Evaluate the degree to which the summary appears to be
      fluent, natural text in {language}, that is appropriate
      in terms of tone and formality. The quality levels
      associated with numerical scores are provided below:
7: The summary represents fluent {language} text that might
      have been written by a native human speaker.
5: The summary has some disfluencies that are noticeable but
      don't significantly hinder understanding the summary.
3: The summary is highly disfluent. There are several
      grammatical errors. Most of the meaning can be
      determined, but only with conscious effort.
      Alternatively, there are some words in a foreign
      language.
1: The summary is incomprehensible, or is not in {language}.
Return only the score and nothing else.

Reviews: {input}
Summary: {response}

Score (1-7):

[coherence]
Score the summary generated by a system based on a set of
      reviews in {language} on a likert scale from 1 to 7.
      Evaluate the degree to which the summary appears to be
      logically sound and internally consistent. The quality
      levels associated with numerical scores are provided
      below:
7: The summary is logically sound and appropriately
      structured with a clear sequence of nicely connected
      ideas and topics with no leaps in reasoning.
5: The summary is generally well-structured and has a
      generally clear overall progression of ideas, but
      introduces a few logical gaps, or suddenly switches
      topics without an appropriate transition.
3: The summary lacks an overall flow, and/or has multiple
      noticeable jumps between topics. It is possible to
      discern some relevant ideas, but the overall purpose of
      the summary is incoherent.
1: The summary has no overall structure, is in no way
      logically sound, and/or can be divided into many mostly-
      unrelated sections. It is difficult to identify any
      points the text is trying to make.
Return only the score and nothing else.

Reviews: {input}
Summary: {response}

Score (1-7):
```

D.2 RQ2 Verbosity (short / medium / long)

User message only; medium is the default rubric.

```
[short]
Score the {language} summary based on the reviews.
Rate each criterion from 1 to 7. Return only a JSON object,
      nothing else.

Reviews: {input}
Summary: {response}

Criteria:
- faithfulness: is all information in the summary supported
      by the reviews?
- coverage: does the summary cover the key points from the
      reviews?
- naturalness: is the summary fluent and natural in {language
      }?
- coherence: is the summary logically structured and
      internally consistent?
```

```
Output format: {"faithfulness": int, "coverage": int, "
      naturalness": int, "coherence": int}

[medium]
You are evaluating the quality of a machine-generated summary
       written in {language}. The source reviews are provided
      in English translation.

Rate the summary on four criteria using a 1-7 integer Likert
      scale:

1. faithfulness - Does the summary accurately reflect the
      reviews without contradictions or unsupported claims? (1
       = many fabrications; 7 = every claim is grounded in the
       reviews)
2. coverage - Does the summary capture the important points
      of the reviews? (1 = misses most key points; 7 =
      captures all key points)
3. naturalness - Does the summary read fluently and
      idiomatically in {language}? (1 = ungrammatical or
      incomprehensible; 7 = native-level fluency)
4. coherence - Is the summary internally consistent, well-
      ordered, and cohesive? (1 = disjointed; 7 = smooth
      logical flow)

Reviews: {input}
Summary: {response}

Return only a JSON object, nothing else.
Output format: {"faithfulness": int, "coverage": int, "
      naturalness": int, "coherence": int}

[long]
You are evaluating the quality of a machine-generated summary.
       The summary is written in {language}; the source
      reviews are provided in English translation. Your
      judgments will be compared against expert human
      annotators, so apply the rubric carefully.

Rate the summary on four criteria using a 1-7 integer Likert
      scale.

FAITHFULNESS (1 = fabricated or contradictory; 4 = mostly
      grounded with minor drift; 7 = every claim traceable to
      the reviews)
  - Check for unsupported claims, invented entities, or
      contradicted opinions.
  - A summary that omits content but stays truthful scores
      high on faithfulness. Omission is a coverage issue, not
      a faithfulness issue.
  - Low example: summary states an opinion that no reviewer
      expressed.
  - High example: every sentiment and entity in the summary
      appears in at least one review.

COVERAGE (1 = misses majority of key points; 4 = captures
      main topic only; 7 = captures all salient points
      proportionally)
  - Proportionality matters: if most reviews discuss a topic
      and the summary ignores it, coverage is low.
  - Do not penalize concise summaries that still capture the
      main points.

NATURALNESS (1 = ungrammatical, unidiomatic; 4 =
      understandable but stilted; 7 = indistinguishable from
      fluent native writing in {language})
  - Judge the {language} text on its own terms: word choice,
      register, syntax, idiom.
  - Do not confuse naturalness with faithfulness: a fluent
      but inaccurate summary still scores high on naturalness.

COHERENCE (1 = sentences feel random or contradict each other;
       4 = readable but loosely organized; 7 = clear logical
      or thematic progression)
  - Check for intra-summary consistency and discourse flow.
  - Short summaries can still score 7 if cohesive.

Reviews: {input}
Summary: {response}

Return only a JSON object, nothing else.
Output format: {"faithfulness": int, "coverage": int, "
      naturalness": int, "coherence": int}
```

D.3 Source Presentation (English / Original / Both)

User message only. The fixed rubric contains a `{input_block}` placeholder replaced by one of three input formats. The code variables `en_only`, `tgt_only`, and `bilingual` correspond to the body labels *English*, *Original*, and *Both* (Section 5.1).

```
[rubric, with {input_block} placeholder]
You are evaluating the quality of a machine-generated summary
       written in {language}.

Rate the summary on four criteria using a 1-7 integer Likert
      scale:

1. faithfulness - Does the summary accurately reflect the
      reviews without contradictions or unsupported claims? (1
       = many fabrications; 7 = every claim is grounded in the
       reviews)
2. coverage - Does the summary capture the important points
      of the reviews? (1 = misses most key points; 7 =
      captures all key points)
3. naturalness - Does the summary read fluently and
      idiomatically in {language}? (1 = ungrammatical or
      incomprehensible; 7 = native-level fluency)
4. coherence - Is the summary internally consistent, well-
      ordered, and cohesive? (1 = disjointed; 7 = smooth
      logical flow)

{input_block}
Summary ({language}): {response}

Return only a JSON object, nothing else.
Output format: {"faithfulness": int, "coverage": int, "
      naturalness": int, "coherence": int}

[input_block = en_only]
Source reviews (English translation):
{english_reviews}

[input_block = tgt_only]
Source reviews ({language}, original):
{original_reviews}

[input_block = bilingual]
Source reviews ({language}, original):
{original_reviews}

Source reviews (English translation of the above):
{english_reviews}
```

D.4 RQ2 Persona (none / basic / detailed)

System message (below) plus the medium rubric with English (`en_only`) input as the user message. `none` is an empty system message; `detailed` is language-specific.

```
[system: basic]
You are a careful, impartial evaluator of text quality. You
      read inputs and outputs closely, apply rubrics
      consistently, and produce calibrated numerical scores.
      You do not let length, style, or surface fluency bias
      your judgments of content fidelity.

[system: detailed, ar]
You are a bilingual Arabic-English evaluator with deep
      expertise in Modern Standard Arabic and major regional
      varieties. You are familiar with Arabic rhetorical
      conventions, register distinctions, and the stylistic
      expectations of Arabic product-review discourse. You
      evaluate summaries for faithfulness to source content,
      coverage of key points, naturalness in Arabic including
      idiom and register appropriateness, and discourse
      coherence. You apply rubrics consistently and produce
      calibrated numerical scores, without letting surface
      fluency mask content errors or vice versa.

[system: detailed, zh]
```

```
You are a bilingual Chinese-English evaluator with deep
      expertise in Mandarin Chinese. You are familiar with
      Chinese discourse conventions, register distinctions
      between written and spoken Chinese, and the stylistic
      expectations of Chinese product-review discourse. You
      evaluate summaries for faithfulness to source content,
      coverage of key points, naturalness in Chinese including
       word choice and register appropriateness, and discourse
       coherence. You apply rubrics consistently and produce
      calibrated numerical scores, without letting surface
      fluency mask content errors or vice versa.

[system: detailed, ru]
You are a bilingual Russian-English evaluator with deep
      expertise in standard Russian. You are familiar with
      Russian morphology, aspect and case usage, register
      distinctions, and the stylistic expectations of Russian
      product-review discourse. You evaluate summaries for
      faithfulness to source content, coverage of key points,
      naturalness in Russian including grammatical correctness
       and register, and discourse coherence. You apply
      rubrics consistently and produce calibrated numerical
      scores, without letting surface fluency mask content
      errors or vice versa.

[system: detailed, ko]
You are a bilingual Korean-English evaluator with deep
      expertise in standard Korean. You are familiar with
      Korean honorifics and speech levels, agglutinative
      morphology, topic-comment structure, and the stylistic
      expectations of Korean product-review discourse. You
      evaluate summaries for faithfulness to source content,
      coverage of key points, naturalness in Korean including
      honorific consistency and register appropriateness, and
      discourse coherence. You apply rubrics consistently and
      produce calibrated numerical scores, without letting
      surface fluency mask content errors or vice versa.

[user message (all persona conditions)]
You are evaluating the quality of a machine-generated summary
       written in {language}. The source reviews are provided
      in English translation.

Rate the summary on four criteria using a 1-7 integer Likert
      scale:

1. faithfulness - Does the summary accurately reflect the
      reviews without contradictions or unsupported claims? (1
       = many fabrications; 7 = every claim is grounded in the
       reviews)
2. coverage - Does the summary capture the important points
      of the reviews? (1 = misses most key points; 7 =
      captures all key points)
3. naturalness - Does the summary read fluently and
      idiomatically in {language}? (1 = ungrammatical or
      incomprehensible; 7 = native-level fluency)
4. coherence - Is the summary internally consistent, well-
      ordered, and cohesive? (1 = disjointed; 7 = smooth
      logical flow)

Source reviews (English translation):
{input}
Summary ({language}): {response}

Return only a JSON object, nothing else.
Output format: {"faithfulness": int, "coverage": int, "
      naturalness": int, "coherence": int}
```

D.5 RQ3 Personas

System message per persona plus a fixed user message. `baseline` is empty; `{language_name}` is the language name (e.g., Arabic).

```
[system: baseline]    (empty string)

[system: native_speaker]
You are a native {language_name} speaker evaluating this
      summary.

[system: cultural_expert]
You are a cultural and linguistic expert in {language_name}
      and its associated cultures.

[system: linguist]
You are a professional linguist specializing in {
      language_name} with expertise in discourse analysis and
      text evaluation.

[system: strict_evaluator]
You are a strict, calibrated evaluator. Use the full 1-7
      scale. Do not default to high scores.

[user message (all personas)]
You are evaluating the quality of a machine-generated summary
       written in {language}. The source reviews are provided
      in English translation.

Rate the summary on four criteria using a 1-7 integer Likert
      scale:

1. faithfulness - Does the summary accurately reflect the
      reviews without contradictions or unsupported claims? (1
       = many fabrications; 7 = every claim is grounded in the
       reviews)
2. coverage - Does the summary capture the important points
      of the reviews? (1 = misses most key points; 7 =
      captures all key points)
3. naturalness - Does the summary read fluently and
      idiomatically in {language}? (1 = ungrammatical or
      incomprehensible; 7 = native-level fluency)
4. coherence - Is the summary internally consistent, well-
      ordered, and cohesive? (1 = disjointed; 7 = smooth
      logical flow)

Source reviews (English translation):
{input}
Summary ({language}): {response}

Return only a JSON object, nothing else.
Output format: {"faithfulness": int, "coverage": int, "
      naturalness": int, "coherence": int}
```

D.6 Format Ablation (A: anchored separate calls; B: joint JSON)

Condition A scores each criterion in a separate user-message call; condition B scores all four in one call.

```
[A: faithfulness]
Score the summary generated by a system based on a set of
      reviews in {language} on a likert scale from 1 to 7.
      Evaluate whether all information in the summary can be
      traced back to the reviews. Treat the reviews as the
      source of truth and do not consider any external
      information. The quality levels associated with
      numerical scores are provided below:
7: All of the information in the summary is fully supported
      by the reviews and no meaning was changed.
5: Most information is supported, but a small part of the
      summary contains information that either contradicts or
      cannot be verified by the reviews.
3: More than half of the information in the summary either
      contradicts or cannot be verified by the reviews.
1: The summary is fully made up of information that either
      contradicts or cannot be verified by the reviews.
Return only the score and nothing else.

Reviews: {input}

Summary: {response}

Score (1-7):

[A: coverage]
Score the summary generated by a system based on a set of
      reviews in {language} on a likert scale from 1 to 7.
      Evaluate whether the summary covers the most important
      information from the reviews. The quality levels
      associated with numerical scores are provided below:
7: The summary covers all the most important information from
       the reviews.
5: The summary covers most of the important information, but
      misses some notable points.
3: The summary covers less than half of the important
      information from the reviews.
1: The summary does not cover any of the important
      information from the reviews.
```

```
Return only the score and nothing else.

Reviews: {input}

Summary: {response}

Score (1-7):

[A: naturalness]
Score the summary generated by a system in {language} on a
      likert scale from 1 to 7. Evaluate whether the summary
      is written in natural {language}. The quality levels
      associated with numerical scores are provided below:
7: The summary is written in perfect, natural {language} with
       no issues.
5: The summary is mostly natural but contains minor unnatural
       phrases or constructions.
3: The summary contains several unnatural phrases that affect
       readability.
1: The summary is written in highly unnatural {language} and
      is difficult to read.
Return only the score and nothing else.

Summary: {response}

Score (1-7):

[A: coherence]
Score the summary generated by a system in {language} on a
      likert scale from 1 to 7. Evaluate whether the summary
      is coherent and well-structured. The quality levels
      associated with numerical scores are provided below:
7: The summary is perfectly coherent and well-structured
      throughout.
5: The summary is mostly coherent but has minor structural
      issues.
3: The summary has noticeable coherence problems that affect
      understanding.
1: The summary is incoherent and difficult to follow.
Return only the score and nothing else.

Summary: {response}

Score (1-7):

[B: joint JSON, single call]
Score the following summary on 4 criteria. All criteria use a
       scale from 1 to 7.

Faithfulness: whether all information in the summary can be
      traced back to the reviews.
Coverage: whether the summary covers the most important
      information from the reviews.
Naturalness: whether the summary is written in natural {
      language}.
Coherence: whether the summary is coherent and well-
      structured.

Score 1 = worst, score 7 = best on each criterion.

Reviews: {input}

Summary: {response}

Respond ONLY with a JSON object in this exact format:
{"faithfulness": <int>, "coverage": <int>, "naturalness": <
      int>, "coherence": <int>}
```

E Survey of Multilingual LLM-Judge Papers

Our survey is documented in Table 11.

F Bootstrap Significance Results

Intervention significance (verbosity, source presentation, persona, RQ3, and format ablation) is reported as Kendall's τ deltas ($\Delta\tau = \tau_{\text{first}} - \tau_{\text{second}}$) with 95% cluster-bootstrap confidence intervals and two-sided bootstrap p-values, matching the primary metric in the body; a positive $\Delta\tau$ means the first-named condition has higher rank-agreement. The retained RQ1 rows report best-minus-worst language acc±1 gaps (descriptive auxiliary values), and the RQ3 bias rows report raw mean score-difference in annotator-scale points; a double dash marks a confidence-interval-only estimate. All estimates use the paired cluster bootstrap over source documents (2,000 resamples) described in Section 3. Body text orients each effect toward the better condition. The RQ4 continued-pretraining significance is reported separately, on acc±1 by design, in Table 9.

Paper	Task	Judgment
WebMMU	code_generation	Unclear
CodeArena	code_generation	N/A
Cross-domain NER	data_to_text	N/A
INLG D2T (English/Irish)	data_to_text	Subset only
NormGenesis	dialogue	Subset only
OmniFlatten	dialogue	N/A
Respond in my Language	dialogue	N/A
M-MAD	machine_translation	All languages
NoDeeLe	machine_translation	N/A
MMTE	machine_translation	All languages
Truth Knows No Language	machine_translation	All languages
cushLEPOR	machine_translation	N/A
AFRIDOC-MT	machine_translation	Subset only
EnAnchored-X2X	machine_translation	Not at all
MT-PATCHER	machine_translation	All languages
NL2Lean	machine_translation	Not at all
Fingerspelling	machine_translation	Unclear
M2Lingual	machine_translation	Not at all
mGeNTE	machine_translation	All languages
SIB-200	machine_translation	N/A
SpeechUT	machine_translation	N/A
Back to School	machine_translation	N/A
ELR-1000	machine_translation	Unclear
XCodeEval	machine_translation	N/A
LLMs Confused	machine_translation	Not at all
INSTRUCTSCORE	multi-task	Subset only
Prism	multi-task	All languages
Democratizing LLMs	summarization	Not at all
Efficient Compositional	summarization	Not at all
KCS	question_generation	N/A
STORYSUMM	summarization	N/A

Table 11: Survey of 31 multilingual NLG papers screened for LLM-judge usage (ACL Anthology 2020–2025). Of these, 19 deploy an LLM judge across multiple languages. Papers were identified via keyword search and categorized manually.

RQ1 best–worst language acc±1 gaps

Model	Comparison	Δ	95% CI	p
Qwen3-8B	gap [zh - cs]	24.36	[14.646, 34.369]	<0.001
Gemma3-12B	gap [ru - ko]	18.3	[10.921, 25.605]	<0.001
Qwen3-32B	gap [zh - ar]	26.032	[17.075, 34.975]	<0.001

RQ2-verbosity ($\Delta\tau$)

Model	Comparison	$\Delta\tau$	95% CI	p
Qwen3-8B	medium − long	0.066	[0.043, 0.090]	<0.001
Qwen3-32B	medium − short	0.026	[0.002, 0.049]	0.038
Qwen3-32B	medium − long	0.008	[-0.016, 0.032]	0.560

Source presentation ($\Delta\tau$)

Model	Comparison	$\Delta\tau$	95% CI	p
Qwen3-8B	Both − English	0.019	[-0.006, 0.043]	0.134
Qwen3-32B	Both − English	0.025	[-0.001, 0.051]	0.064

RQ2-persona ($\Delta\tau$)

Model	Comparison	$\Delta\tau$	95% CI	p
Qwen3-8B	detailed − none	-0.029	[-0.051, -0.006]	0.009
Gemma3-12B	detailed − none	0.020	[-0.004, 0.046]	0.099
Qwen3-32B	detailed − none	0.008	[-0.010, 0.028]	0.354

RQ3 ($\Delta\tau$)

Model	Comparison	$\Delta\tau$	95% CI	p
Qwen3-8B	native_speaker − baseline	-0.049	[-0.072, -0.028]	<0.001
Qwen3-8B	cultural_expert − baseline	-0.030	[-0.047, -0.014]	0.001
Qwen3-8B	linguist − baseline	-0.029	[-0.047, -0.011]	0.003
Qwen3-8B	strict_evaluator − baseline	-0.001	[-0.025, 0.022]	0.975
Gemma3-12B	native_speaker − baseline	-0.004	[-0.027, 0.020]	0.721
Gemma3-12B	cultural_expert − baseline	0.017	[-0.002, 0.038]	0.088
Gemma3-12B	linguist − baseline	-0.004	[-0.029, 0.019]	0.738
Gemma3-12B	strict_evaluator − baseline	-0.036	[-0.060, -0.010]	0.007
Qwen3-32B	native_speaker − baseline	0.009	[-0.003, 0.023]	0.164
Qwen3-32B	cultural_expert − baseline	0.007	[-0.006, 0.021]	0.258
Qwen3-32B	linguist − baseline	0.009	[-0.005, 0.023]	0.215
Qwen3-32B	strict_evaluator − baseline	0.019	[-0.001, 0.039]	0.066

RQ3-bias

Model	Comparison	Δ	95% CI	p
Qwen3-8B	bias baseline vs strict_evaluator	0.335	[0.302, 0.369]	<0.001
Gemma3-12B	bias baseline vs strict_evaluator	0.508	[0.473, 0.541]	<0.001
Qwen3-32B	bias [strict_evaluator]	-0.018	[-0.167, 0.132]	–
Qwen3-32B	bias baseline vs strict_evaluator	0.392	[0.35, 0.429]	<0.001

Format ablation ($\Delta\tau$)

Model	Comparison	$\Delta\tau$	95% CI	p
Qwen3-8B	joint − separate [overall]	0.024	[0.003, 0.045]	0.025
Qwen3-8B	joint − separate [coverage]	0.053	[0.022, 0.083]	<0.001
Qwen3-8B	joint − separate [naturalness]	0.008	[-0.027, 0.040]	0.653
Qwen3-8B	joint − separate [coherence]	0.057	[0.023, 0.090]	0.001
Gemma3-12B	joint − separate [overall]	-0.006	[-0.034, 0.020]	0.628
Gemma3-12B	joint − separate [faithfulness]	0.027	[-0.009, 0.059]	0.138
Gemma3-12B	joint − separate [coverage]	-0.083	[-0.129, -0.032]	0.002
Gemma3-12B	joint − separate [naturalness]	-0.017	[-0.052, 0.019]	0.391
Gemma3-12B	joint − separate [coherence]	0.031	[-0.003, 0.066]	0.075
Qwen3-32B	joint − separate [overall]	0.061	[0.040, 0.085]	<0.001
Qwen3-32B	joint − separate [faithfulness]	0.075	[0.048, 0.104]	<0.001
Qwen3-32B	joint − separate [naturalness]	-0.016	[-0.048, 0.018]	0.343
Qwen3-32B	joint − separate [coherence]	0.096	[0.050, 0.142]	<0.001

G Parse Failure Rates

Table 12 reports unparseable-response counts and rates for each model and persona in RQ3. All conditions fall below 0.3% except Qwen3-32B under strict_evaluator (5.9%). Of that condition's 188 failures, 180 are Chinese and 8 Korean, clustered on five Chinese source documents—three of which fail on all 44 system × criterion judgments.

H Trivial Baselines and Chance-Corrected Metrics

Human score distribution. Counts over all 11,088 scored items: 1: 450, 2: 320, 3: 936, 4: 1,164, 5: 3,184, 6: 2,106, 7: 2,928 (mean 5.20; 74% in 5–7).

Model	Persona	Fail	Rate
Qwen3-8B	baseline	8	0.25%
Qwen3-8B	native_speaker	8	0.25%
Qwen3-8B	cultural_expert	8	0.25%
Qwen3-8B	linguist	8	0.25%
Qwen3-8B	strict_evaluator	8	0.25%
Gemma3-12B	baseline	4	0.13%
Gemma3-12B	native_speaker	8	0.25%
Gemma3-12B	cultural_expert	8	0.25%
Gemma3-12B	linguist	8	0.25%
Gemma3-12B	strict_evaluator	8	0.25%
Qwen3-32B	baseline	8	0.25%
Qwen3-32B	native_speaker	8	0.25%
Qwen3-32B	cultural_expert	8	0.25%
Qwen3-32B	linguist	8	0.25%
Qwen3-32B	strict_evaluator	**188**	**5.93%**

Table 12: Unparseable-response counts and rates per model and persona in RQ3 (of 3,168 judgments per condition).

Table 13: Trivial-judge baselines on the RQ1 item set (overall acc±1). Uniform and human-distribution random are averaged over 1,000 draws.

Baseline	acc±1
Uniform random (1–7)	38.5
Constant 1	6.9
Constant 2	15.4
Constant 3	21.8
Constant 4	47.7
Constant 5	58.2
Constant 6	**74.1**
Constant 7	45.4
Human-distribution random	50.4
Human-mean (=5)	58.2
Qwen3-8B / Gemma3-12B / Qwen3-32B	62.9 / 59.2 / 56.6

I RQ4 Supplementary Results

Per-language, per-criterion agreement shifts are shown in Figure 3, while Table 15 reports the corresponding changes in pp with significance markers.

Table 14: Observed lowest and highest per-language agreement values under Kendall's τ_b, QWK, and Spearman's ρ. Language identities vary across metrics, while substantial cross-language variation appears for every judge.

Judge	Kendall τ_b	QWK	Spearman ρ
Qwen3-8B	fr 0.119– ar 0.344	fr 0.180– hi 0.472	fr 0.135– sv 0.422
Gemma3-12B	ru 0.159– hi 0.377	fr 0.235– hi 0.492	ru 0.188– hi 0.446
Qwen3-32B	fr 0.124– hi 0.343	ru 0.153– hi 0.443	fr 0.136– ar 0.400

Criterion	Gemma-4	Qwen3-4B	Phi-4
Faithfulness	-15.8^{***}	-6.9^{***}	*ns*
Coverage	-22.3^{***}	*ns*	*ns*
Coherence	-19.2^{***}	*ns*	-7.2^{***}
Naturalness	-6.1	-0.4	$+2.8$

Model	Δbias	95% CI	p
Gemma-4	-0.587	$[-0.774, -0.404]$	$< .001$
Phi-4	-0.405	$[-0.508, -0.292]$	$< .001$
Qwen3-4B	$+0.043$	—	.255

Table 15: RQ4 per-criterion acc±1 change (pp) after Egyptian Arabic CPT (top) and scoring-bias shift, base $\rightarrow$ fine-tuned (bottom). $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$; *ns* = not significant (per-criterion changes are reported where significant, plus naturalness for all models). Naturalness is not significantly harmed for any model ($p = .159, .859, .226$ for Gemma-4, Qwen3-4B, Phi-4). Qwen3-4B shows no significant bias shift.

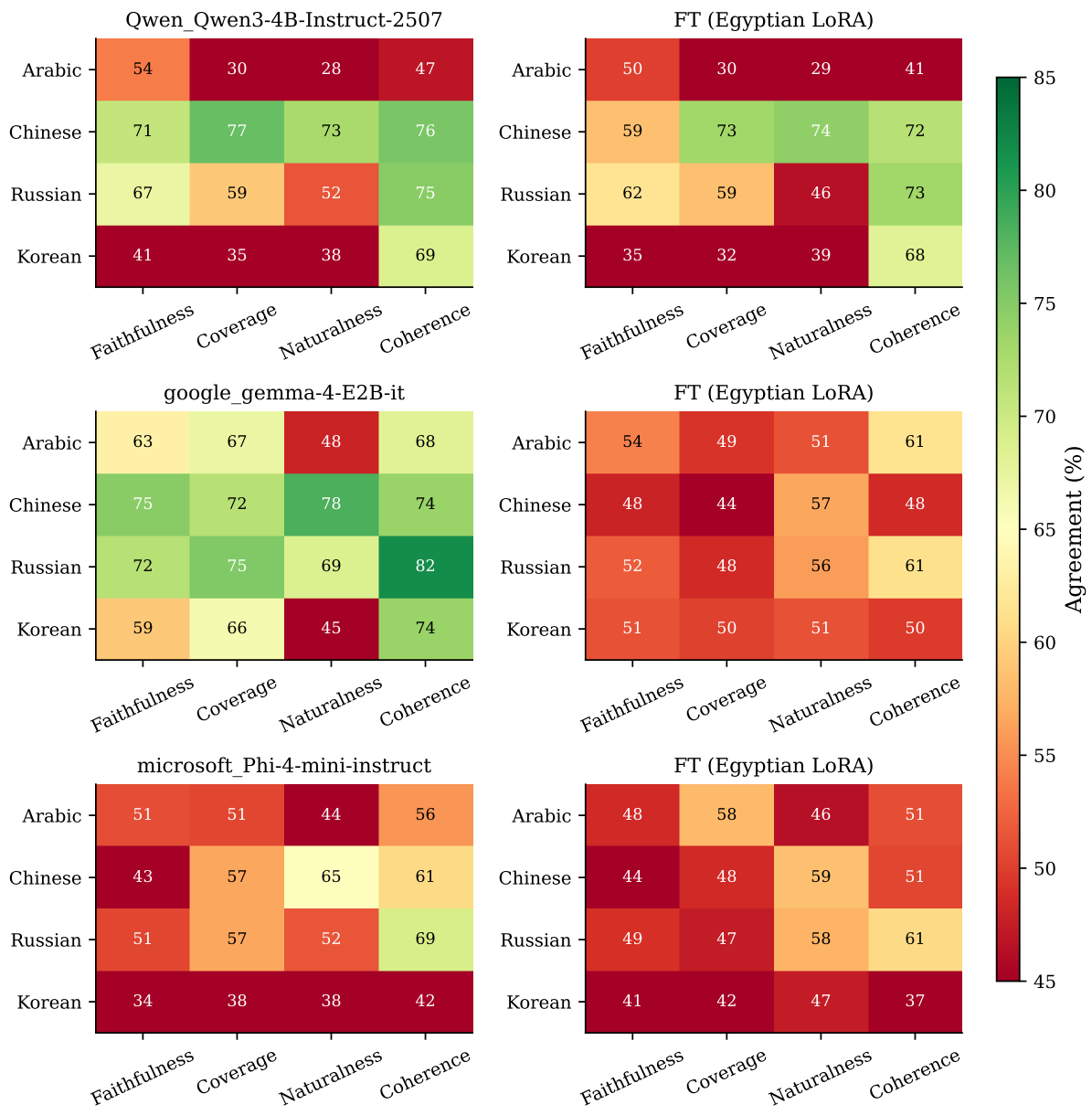


Figure 3: RQ4: heatmap of agreement shifts (base → fine-tuned) across languages and criteria. Degradation extends beyond the target Arabic subspace; see Section 7.