Disentangling Question, Response, and Verifier Languages in Multilingual LLM Evaluation

Anonymous ACL submission

Abstract

Multilingual evaluation of large language models typically varies a single "prompt language" axis that collapses three logically distinct languages—of the question, of the model's response, and of the prompt that verifies it—into one, conflating question-side, response-side, and verifier-side effects. We disentangle them into a full factorial across six languages on three benchmarks spanning regime (language-agnostic vs. language-specific) and response format (open-ended, multiple-choice, short-answer), on a sixteen-model panel, reading four metrics jointly (accuracy, AUROC, ECE, Brier). Three findings emerge. First, the load-bearing language axis depends on the task regime: language-specific content is response-driven, language-agnostic content is question-driven. Second, the English-language advantage on LA tasks is primarily question-side, and accuracy and AUROC favor different configurations: $q_{\text{lang}} = \text{EN}$ helps accuracy but $r_{\text{lang}} \neq \text{EN}$ helps discrimination, so ($q_{\text{lang}} = \text{EN}$, $r_{\text{lang}} \neq \text{EN}$) gives the best calibration at a ~ 13 p.p. accuracy cost relative to all-English. Third, the native-language advantage on LS tasks is localized to the triple-match configuration where question, response, and content language coincide, but the same configuration compresses ranking discrimination—a saturation–discrimination tradeoff that mirrors the all-English configuration on LA. The verifier language contributes negligible main variance ($\eta^2 < 0.5\%$) but interacts asymmetrically with the response language across formats.

1 Introduction

Knowing when a language model's answer is correct is as important as the answer itself. $P(\text{True})$ [Kadavath et al., 2022]—asking the model to predict whether its own response is correct and reading off the probability assigned to "True"—has become a workhorse self-confidence signal for selective answering, abstention, and downstream verifier or retrieval pipelines [Tian et al., 2023, Xiong et al., 2024, Geng et al., 2024]. Its appeal is operational: no labels, no extra training, just a second forward pass. As LLMs are deployed across languages, a verifier signal that silently degrades when the question or response language changes will silently degrade selective answering with it.

Multilingual evaluation of $P(\text{True})$ is usually conducted with a single language at a time: the question, the model's response, and the verifier prompt are all rendered in the same language, varied as a single "prompt language" condition. The most rigorous recent treatment under this framing [Xue et al., 2025] reports two aggregate findings on top of an LS/LA task distinction: English dominance on language-agnostic (LA) tasks—math, general factual recall, content independent of any specific language community—and native-language dominance on language-specific (LS) tasks—culture-rooted knowledge such as French civic life or Japanese historical figures. Subsequent multilingual confidence work has inherited the one-language-at-a-time framing [Yang et al., 2024, Zhou et al., 2025].

Real multilingual deployments rarely respect this collapse: a Korean user asking in Korean may receive an English answer from a downstream translator pipeline, a multilingual chatbot may extract $P(\text{True})$ through an English verifier prompt that ships as the API default, and translated evaluation suites mix question and response languages. Each differs from the $q_{\text{lang}} = r_{\text{lang}} = v_{\text{lang}}$ matched setting on which the multilingual confidence literature was calibrated, and each is a setting where a verifier signal that silently degrades when the three languages diverge would silently degrade the downstream selective-answering decision built on top of it.

This framing conflates effects that need not covary. A real multilingual interaction involves three logically distinct languages—of the question, of

the response, and of the verifier prompt—and tying them together cannot say whether an observed effect lives on the question side, the response side, or the verifier side. It also conflates two distinct sources of multilingual difficulty: how well a model can read and generate a given language, versus how much knowledge the model has stored in it. Without a controlled contrast between LA content (where only the first matters) and LS content (where both do), the two cannot be separated.

We disentangle the three languages into a full factorial over six languages, evaluated on three benchmarks chosen to span both regime (LS vs. LA) and response format (open-ended, multiple-choice, short-answer): an LS benchmark extended from Xue et al. [2025], MMMLU [OpenAI, 2024], and TriviaQA [Joshi et al., 2017]. The panel is sixteen instruction-tuned models from three billion to seventy billion parameters, plus two closed APIs.

The decomposition yields three findings (Figure 1). First, the load-bearing language axis depends on the task regime: response-driven on LS, question-driven on LA—two different mechanisms the single-axis convention conflates. Second, the English-language advantage on LA tasks is primarily question-side (17 p.p. across q_{lang} vs. 2.4 p.p. across r_{lang}), but accuracy and calibration favor different axes: $q_{\text{lang}} = \text{EN}$ helps accuracy while $r_{\text{lang}} \neq \text{EN}$ helps AUROC and ECE, so ($q_{\text{lang}} = \text{EN}$, $r_{\text{lang}} \neq \text{EN}$) trades $\sim$ 13 p.p. of accuracy relative to all-English for the best calibration. Third, the native-language advantage on LS tasks is localized to the triple-match configuration, which is also where AUROC is lowest—a saturation–discrimination tradeoff mirroring the all-English configuration on LA. Two further effects worth flagging: Hindi inverts the LS pattern on accuracy (AUROC and ECE barely change), and pairing the verifier with the response language improves calibration on open-ended LSQA but worsens it on short-answer TriviaQA, separating ranking from absolute-calibration use cases.

2 Related Work

Multilingual calibration. Multilingual $P(\text{True})$ and related self-confidence methods have been evaluated by varying question, response, and verifier languages jointly as a single "prompt language" [Xue et al., 2025, Yang et al., 2024, Zhou et al., 2025]. Xue et al. [2025] introduce a language-agnostic / language-specific task distinction and contribute the LSQA benchmark we extend; we adopt that task distinction but leave the single-axis design behind. To our knowledge no prior multilingual evaluation study disentangles the three language axes, scales across a multi-family panel of this size, or contrasts language-specific with both multiple-choice and short-answer language-agnostic benchmarks.

Self-confidence and $P(\text{True})$. The $P(\text{True})$ probe was introduced by Kadavath et al. [2022], who found that larger models are well-calibrated on multiple-choice and true/false formats but that AUROC and ECE can move independently. Subsequent work explored verbalized confidence [Lin et al., 2022, Tian et al., 2023, Xiong et al., 2024] and reasoning-mode calibration [Yoon et al., 2025, Bakman et al., 2025]. Most closed-source calibration studies rely on verbalized confidence due to limited logit access; we instead apply unified first-token-logit $P(\text{True})$ across open-weight and OpenAI closed models using the `top_logprobs=20` field.

Cross-lingual knowledge. Wendler et al. [2024] show multilingual LLMs operate through an English-anchored representational pivot; Qi et al. [2023] and Wang et al. [2025] document cross-lingual factual inconsistency. By varying the question, response, and verifier languages independently rather than jointly, we can decompose such cross-lingual inconsistency into question-side and response-side components and isolate which axis it lives on.

3 Experimental Setup

3.1 Benchmarks

We use three benchmarks: one language-specific dataset (LSQA), and two language-agnostic datasets that differ in response format (multiple-choice MMMLU, short-answer TriviaQA).

LSQA (LS, open-ended). LSQA is a language-specific benchmark covering culture-rooted question–answer pairs across six languages $\mathcal{L} = \{\text{KO}, \text{HI}, \text{EN}, \text{ZH}, \text{JA}, \text{FR}\}$, 160 items per language across four cultural categories (Appendix B). Each item carries gold (question, answer) pairs in all six languages. We extend LSQA [Xue et al., 2025] from its original four shared languages (English, French, Chinese, Japanese) with Korean and Hindi, producing the parallel translations via Claude Opus 4.7 [Anthropic, 2026] with a content-preservation prompt template.

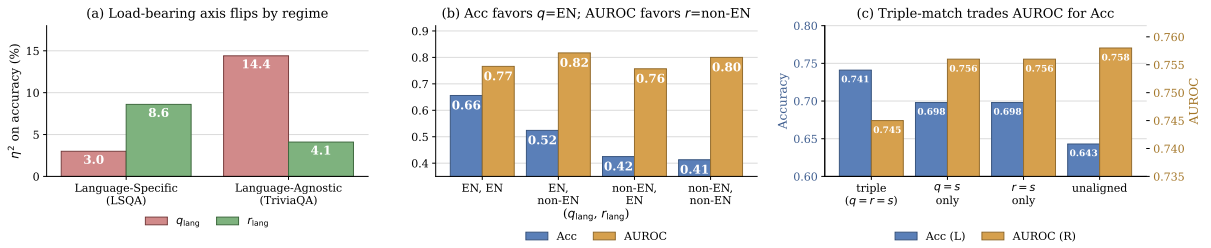


Figure 1: Three main findings of the paper. **(a)** The load-bearing language axis flips between regimes: on language-specific content (LSQA), the response axis carries the variance ($\eta^2 = 8.6\%$ vs. 3.0%); on language-agnostic content (TriviaQA), the question axis does (14.4% vs. 4.1%). **(b)** On TriviaQA, accuracy and AUROC favor different configurations: accuracy is highest at $q_{\text{lang}} = \text{EN}$, AUROC is highest at $r_{\text{lang}} \neq \text{EN}$. The ($q_{\text{lang}} = \text{EN}$, $r_{\text{lang}} \neq \text{EN}$) cell trades ~ 13 p.p. of accuracy (vs. all-English) for the best calibration on both AUROC and ECE. **(c)** On LSQA, the triple-match $q_{\text{lang}} = r_{\text{lang}} =$ subset configuration maximizes accuracy but minimizes AUROC—a saturation–discrimination tradeoff mirroring (b).

MMMLU (LA, MCQ) is a six-language restriction of OpenAI's multilingual MMLU release [OpenAI, 2024, Hendrycks et al., 2021], sampled to a matched 960-item budget stratified by MMLU subject (STEM, Humanities, Social Sciences, Other). The MMLU subset axis is a subject-difficulty axis, not a cultural-region axis: STEM items are not native to any language community, so the native-tone analysis does not transfer.

TriviaQA (LA, short-answer) [Joshi et al., 2017] disambiguates LA-regime effects from MCQ-format artifacts. We draw 960 items from `trivia_qa rc.nocontext` (validation, $\leq$ 4-token answers) and translate into the same six languages via Claude Opus 4.7 [Anthropic, 2026]. TriviaQA is single-corpus (no subset axis), and answers are free-text 1–4 word strings rather than MCQ letters.

Terminology. A $(q_{\text{lang}}, r_{\text{lang}})$ pair is *matched* if $q_{\text{lang}} = r_{\text{lang}}$ and *cross-language* otherwise. On LSQA, a configuration is *native-tone* if $q_{\text{lang}} = r_{\text{lang}} =$ subset, *question-aligned* or *response-aligned* if only q_{lang} or only r_{lang} equals the subset, and *unaligned* otherwise. The four types tile the per-subset $Q \times R$ grid ($1 + 5 + 5 + 25 = 36$).

3.2 $Q \times R \times V$ Factorial

We vary three orthogonal axes per item: the language of the question (q_{lang}, 6 levels), the language the model is instructed to respond in (r_{lang}, 6 levels), and the language of the verifier prompt used to extract $P(\text{True})$ (v_{lang}, 6 levels). The factorial yields 36 $(q_{\text{lang}}, r_{\text{lang}})$ configurations per item, multiplied by 6 verifier languages, for 216 verifier-conditioned configurations per item per model: 34,560 unique (item, q_{lang}, r_{lang}) generations per model on each dataset. A single response can be verified under any of the six v_{lang}, yielding six $P(\text{True})$ values per (item, q_{lang}, r_{lang}) configuration while the correctness label remains fixed.

The semantic of $q_{\text{lang}} = r_{\text{lang}}$ differs across datasets. On LSQA the response is open-ended free text, so $q_{\text{lang}} = r_{\text{lang}}$ means question and full answer share a language. On MMMLU the response is a letter followed by a short r_{lang} justification; $q_{\text{lang}} = r_{\text{lang}}$ corresponds to the question and the justification text sharing a language. On TriviaQA the answer is a 1–4 word free-text string.

3.3 Unified $P(\text{True})$ Extraction

For each (item, q_{lang}, r_{lang}) triple, the model produces a free-form response $\hat{a}$. We then prompt the same model with a verifier template translated into v_{lang} that asks whether $\hat{a}$ is correct, and compute $P(\text{True}) = \sigma(\ell_{\text{True}} - \ell_{\text{False}})$ from first-token logits on a fixed set of English surface variants (full template in Appendix C). Candidate target tokens are kept in English across all v_{lang}; translating them would conflate language-specific morphology with calibration discrimination.

For open-weight models, $\ell_{\text{True}}, \ell_{\text{False}}$ come from a single forward pass on the verifier prompt. For OpenAI models we use `logprobs=true, top_logprobs=20` and recover the same quantity from the returned top-20 token logprobs; the True/False candidates appear in the top-20 on 99.7% of GPT-4o configurations and 99.4% on GPT-4o-mini. We pin snapshots to `gpt-4o-2024-11-20` and `gpt-4o-mini-2024-07-18`. Reasoning-mode models (Qwen3 with `<think>`) require disabling thinking during verification; we document this

pitfall in Appendix P.

3.4 Correctness Labeling and Metrics

We label each response correct or incorrect using LLM-as-judge (GPT-4o-mini [OpenAI, 2023]) with a structured rubric (Appendix C) [Zheng et al., 2023, Bavaresco et al., 2025, Hada et al., 2024]. Token-overlap heuristics produced systematic errors on multilingual outputs (morphological attachments, romanization variants, script differences; Appendix C), so we use a single LLM judge throughout. Per-configuration metrics are accuracy (higher is better), AUROC of $P(\text{True})$ over correct/incorrect labels (discrimination; higher is better), 15-bin top-label ECE (calibration error; lower is better) [Guo et al., 2017], and Brier score (lower is better). For variance decomposition we use row-level Type III ANOVA on binary correctness with three main effects ($q_{\text{lang}}, r_{\text{lang}}$, subset) and three two-way interactions. We report each factor's share within *explained* variance ($\eta^2 = \text{SS}_{\text{factor}}/\text{SS}_{\text{explained}}$), since residual SS dominates total SS on binary targets (raw SS in Appendix D).

For match-type contrasts, we fit a mixed-effects linear regression on each configuration's metric value with binary indicators for the four pairwise alignments ($q_{\text{lang}}{=}r_{\text{lang}}$, $r_{\text{lang}}{=}v_{\text{lang}}$, $q_{\text{lang}}{=}v_{\text{lang}}$, $q_{\text{lang}}{=}r_{\text{lang}}{=}v_{\text{lang}}$) and a random intercept per model. Each coefficient estimates the average metric change when its alignment holds, pooling across the 16 models with random intercepts absorbing between-model variance. The mixed-effects test is our primary inference target; we report a cluster bootstrap over the panel as a secondary sign-consistency check (underpowered for effect sizes of order 0.01 on AUROC/ECE). Full specification in Appendix M.

3.5 Model Panel

Sixteen instruction-tuned models across five open-weight families plus two closed APIs: Qwen2.5 (3B, 7B, 14B, 32B, 72B; Qwen et al., 2025), Qwen3-8B [Yang et al., 2025] (thinking disabled during verification), Llama-3.1-8B / 3.1-70B [Dubey et al., 2024], Llama-3.2-3B [Dubey et al., 2024], Llama-3.3-70B [Dubey et al., 2024], Gemma-2-9B [Riviere et al., 2024], Gemma-3-12B [Kamath et al., 2025], Mistral-7B-Instruct-v0.3 [Jiang et al., 2023], Aya-Expanse-8B [Üstün et al., 2024] (multilingual-tuned), and GPT-4o / GPT-4o-mini [OpenAI, 2023]. The panel spans a within-family scaling series (Qwen2.5 3B–72B), a multilingual-specialized 8B model (Aya-Expanse), and two closed APIs at the frontier. Compute details and data coverage are in Appendix A.

Model	q	r	s	$q{\times}r$	$q{\times}s$	$r{\times}s$
Qwen2.5-3B	2.2	**33.8**	17.1	15.5	15.5	16.0
Qwen2.5-7B	0.5	**37.1**	15.5	9.3	13.7	24.0
Mistral-7B	10.9	**35.1**	15.4	8.9	18.9	10.7
Qwen2.5-14B	2.3	22.2	14.6	11.4	23.2	**26.2**
Qwen2.5-32B	1.0	14.0	19.7	13.3	23.6	**28.4**
Qwen2.5-72B	0.4	7.2	22.1	3.9	24.7	**41.6**
Llama-3.1-70B	2.8	4.1	16.1	11.8	**54.6**	10.6
GPT-4o	1.9	2.2	17.4	5.9	32.3	**40.4**
GPT-4o-mini	2.4	3.2	20.5	9.0	**44.6**	20.4

Table 1: LSQA Type III ANOVA decomposition of row-level binary correctness (% of explained sum-of-squares); residual excluded. $q, r, s = q_{\text{lang}}, r_{\text{lang}}$, subset. Raw SS in Appendix D. Bold = largest factor per row.

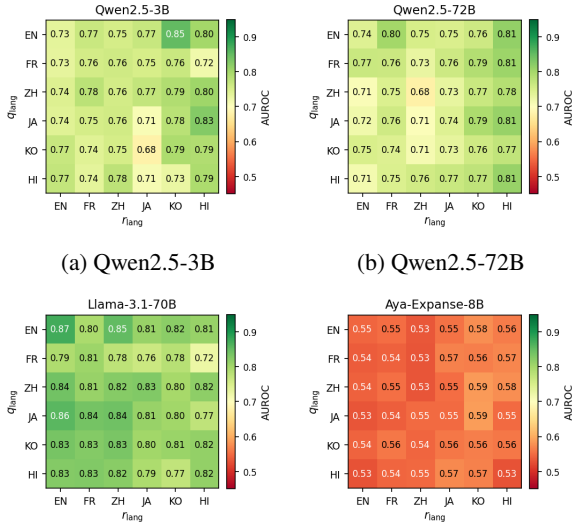
4 Results and Analysis

We organize the analysis around three questions, each examining the Q × R × V factorial from the angle of a distinct multilingual phenomenon and reading the four metrics (accuracy, AUROC, ECE, Brier) jointly. §4.1 asks which language axis carries the variance and how that shifts with task regime; §4.2 asks where the English advantage on language-agnostic tasks comes from; §4.3 asks where the native-language advantage on language-specific tasks comes from. §4.4 unifies the pairwise-alignment view including the verifier axis.

4.1 Which language axis carries the variance?

The dominant axis depends on the task regime. We decompose configuration-level binary correctness via Type III ANOVA (Table 1, LSQA; MMMLU in Appendix L) and dataset-level metric variance via η^2 (Table 2, all three datasets).

The load-bearing axis flips between regimes. On LSQA (LS, open-ended), the r_{lang} main effect dominates the smallest models (33–37% at 3B–7B), then collapses to $\leq$ 7% at 72B and is replaced by language-×-subject interactions at 40–55%. On MMMLU (LA, MCQ), the q_{lang} main effect stays large at all scales (Qwen2.5: 49.8%, 50.6%, 30.2%, 18.4% at 7/14/32/72B), while r_{lang} is essentially absent ($\leq$ 1.6% for 14B+ Qwen). The asymmetry tracks what each



(a) Qwen2.5-3B (b) Qwen2.5-72B

(c) Llama-3.1-70B (d) Aya-Expanse-8B

Figure 2: Configuration-level mean $P(\text{True})$ AUROC on LSQA, averaged across subsets, $v_{\text{lang}} = \text{EN}$. Rows are q_{lang}, columns are r_{lang}. The four panels exhibit qualitatively different structures matching the four ANOVA profiles in Table 1.

regime requires: LSQA's gold content is language-bound, so producing the right language matters (response axis); MMMLU's content is language-agnostic, so question parsing carries the signal (question axis). The single-axis $q_{\text{lang}} = r_{\text{lang}}$ convention obscures this by always coupling the two. Figure 2 shows configuration-level AUROC heatmaps over the LSQA $(q_{\text{lang}}, r_{\text{lang}})$ grid for four representative models. The dominant axis varies sharply: at small scale (Qwen2.5-3B, Aya-8B), AUROC organizes by r_{lang} (horizontal stripes); at larger scale (Qwen2.5-72B, Llama-3.1-70B), the response-axis effect attenuates and language-by-language interaction patterns emerge, matching the η^2 migration in Table 1.

Same axis-shift across metrics, with format modulation. Extending to MMMLU and TriviaQA shows the regime-driven q-vs-r split is a property of accuracy specifically. q_{lang}-on-accuracy is a regime property: both LA datasets have substantial $\eta^2(q_{\text{lang}})$ on accuracy (11.7% MMMLU, 14.4% TriviaQA) while LSQA gives only 3.0%. r_{lang}-on-AUROC is a *format* property: negligible on MCQ-LA (0.6%) and on LS-open (0.9%), but 12.4% on LA-short. MCQ pins confidence near a ceiling regardless of r_{lang}, hiding the discrimination effect (Appendix R.4); open-ended LS spreads confidence but adds cultural-knowledge variance

Table 2: Variance decomposition (η^2, % of total variance) per language axis, per metric, per dataset. The between-model variance is treated as a nuisance random effect and excluded; for reference it ranges 43.9–74.0%.

Dataset	Axis	η^2_{acc}	η^2_{AUROC}	η^2_{ECE}	η^2_{Brier}
LSQA	q	3.0	0.2	1.0	1.0
	r	8.6	0.9	2.4	2.4
	v	0.0	0.1	0.2	0.2
	subset	17.6	1.0	8.3	8.8
MMMLU	q	11.7	1.6	7.1	7.0
	r	0.3	0.6	0.1	0.1
	v	0.0	0.4	0.0	0.0
	subset	6.0	8.5	2.8	2.5
TriviaQA	q	14.4	0.7	3.1	3.4
	r	4.1	12.4	1.5	1.6
	v	0.0	0.1	0.1	0.1

that washes out the r_{lang} signal; only short-answer LA both spreads the confidence range and lets r_{lang} drive most of the variation. r_{lang}-on-accuracy is MCQ-suppressed specifically (0.3% on MMMLU vs. 8.6% on LSQA and 4.1% on TriviaQA): the MCQ letter answer is genuinely r_{lang}-invariant; free-text answers in both regimes are not.

Finally, $\eta^2(v_{\text{lang}}) < 0.5\%$ on every dataset and metric—a finding we revisit in §4.4.

4.2 Where does the English advantage come from?

The English-language advantage on LA tasks is one of the most robust findings in multilingual evaluation [Xue et al., 2025, Yang et al., 2024]. The single-axis design cannot tell whether "English" helps by being the question language, the response language, or both at once. We restrict the factorial to MMMLU (which controls for cultural-content variance) and ask how much panel-mean accuracy varies when q_{lang} is fixed and r_{lang} is varied, and vice versa.

The variation is asymmetric. Panel-mean accuracy varies by 17 p.p. across the question language but only by 2.4 p.p. across the response language; English is the best question language in 14 of 16 models and the best response language in only 5. The aggregate English-language dominance on LA tasks is thus a question-parsing phenomenon, not a response-generation one. Per-model breakdown is in Appendix K; Appendix R.3 gives a representative item where flipping q_{lang} flips the MCQ answer while flipping r_{lang} does not.

Accuracy follows q_{lang}; AUROC and ECE follow r_{lang}. The TriviaQA short-answer setting lets us push this further: by crossing $q_{\text{lang}} \in \{\text{EN}, \text{non-EN}\}$ with $r_{\text{lang}} \in \{\text{EN}, \text{non-EN}\}$, we read

off accuracy and calibration jointly for each combination (Table 3). Accuracy follows q_{lang} (+13 p.p. from non-EN to EN question), but AUROC and ECE follow r_{lang} in the opposite direction: same-language English production raises accuracy but compresses $P(\text{True})$ near the ceiling, narrowing the correct/incorrect confidence gap that AUROC depends on. Non-EN responses spread $P(\text{True})$ more, recovering discrimination on both AUROC and ECE. The (non-EN, EN) cell is worst on all three metrics—the configuration where the model is asked in a weaker language but produces a fluent English answer it cannot back up. The two best directions thus lie on different axes: $q_{\text{lang}} = \text{EN}$ for accuracy and $r_{\text{lang}} \neq \text{EN}$ for calibration. Their combination (EN, non-EN) trades ~ 13 p.p. of accuracy relative to all-English (top-left) for the best AUROC and ECE, preferred whenever the application cares about ranking by $P(\text{True})$ or absolute calibration alongside accuracy.

$(q_{\text{lang}}, r_{\text{lang}})$	Acc ↑	AUROC ↑	ECE ↓
EN, EN	**0.656**	0.766	0.213
EN, non-EN	0.524	**0.817**	**0.212**
non-EN, EN	0.425	0.757	0.316
non-EN, non-EN	0.413	0.800	0.251

Table 3: Panel-mean TriviaQA accuracy, AUROC, and ECE under the EN/non-EN split of q_{lang} and r_{lang} (averaged over v_{lang}). Accuracy is highest at (EN, EN) but AUROC and ECE are best at (EN, non-EN).

4.3 Where does the native-language advantage come from?

Prior work reports that LS tasks benefit from prompting in the subject's native language [Xue et al., 2025], but the single-prompt-language design cannot say whether the benefit comes from asking in that language, responding in it, or the joint condition. The triple-match $q_{\text{lang}} = r_{\text{lang}} =$ subset configuration that benchmarks implicitly evaluate is the natural starting point; Table 4 reports panel-mean metrics on LSQA in this configuration to establish a per-model baseline and the overall spread across the panel. MMMLU figures appear in Appendix E.

To decompose where the native-tone advantage actually originates, we partition the LSQA accuracy panel into four regimes by which of $(q_{\text{lang}}, r_{\text{lang}})$ aligns with the subset language s: triple-match ($q_{\text{lang}} = r_{\text{lang}} = s$), question-aligned only, response-aligned only, and unaligned (Table 5).

Model	Acc ↑	AUROC ↑	ECE ↓	Brier ↓
GPT-4o	0.900	0.795	0.119	0.118
Llama-3.3-70B	0.847	0.844	0.100	0.094
Llama-3.1-70B	0.778	0.829	0.137	0.131
Qwen2.5-72B	0.812	0.823	0.117	0.114
GPT-4o-mini	0.807	0.745	0.169	0.171
Gemma-3-12B	0.726	0.770	0.187	0.186
Qwen2.5-32B	0.688	0.747	0.258	0.250
Qwen3-8B	0.687	0.795	0.211	0.208
Aya-Expanse-8B	0.665	0.582	0.318	0.314
Qwen2.5-14B	0.660	0.779	0.290	0.283
Gemma-2-9B	0.629	0.751	0.260	0.258
Llama-3.1-8B	0.521	0.683	0.367	0.361
Qwen2.5-7B	0.493	0.700	0.350	0.348
Llama-3.2-3B	0.314	0.626	0.591	0.581
Mistral-7B	0.371	0.665	0.458	0.443
Qwen2.5-3B	0.371	0.652	0.526	0.516
Panel mean	0.661	0.758	0.232	0.226

Table 4: Panel-mean LSQA metrics under the triple-match configuration $q_{\text{lang}} = r_{\text{lang}} =$ subset with $v_{\text{lang}} =$ EN. The cross-model AUROC range (0.582–0.844) is tighter than the accuracy range (0.314–0.900); high accuracy and high discrimination do not always coincide.

The advantage is localized to the triple-match regime.

Every model has a positive triple-match delta, but matched configurations off the subject's native language are essentially flat (Appendix H). Question-side and response-side contributions are comparable on average but opposite within families—Qwen is response-dominant, Llama question-dominant—an asymmetry the single-prompt-language design cannot detect. The advantage shrinks but persists with scale.

Triple-match maximizes accuracy but worsens discrimination. Extending the comparison to AUROC and ECE (Table 6) surfaces the same saturation–discrimination tradeoff that (EN, EN) shows on TriviaQA. The triple-match configuration that maximizes accuracy (0.741) and minimizes ECE (0.190) also has the lowest AUROC (0.745); the fully unaligned configuration that gives the lowest accuracy (0.643) gives the highest AUROC (0.758). Aligning the response language with the subject's native language buys accuracy and absolute calibration at the cost of ranking discrimination, mirroring the (EN, EN) cell on TriviaQA: same-language production—whether "native" on LS content or "English" on LA content—compresses $P(\text{True})$ near the ceiling.

Model	no-match	+Q-only	+R-only	+native
GPT-4o	0.900	+0.018	+0.033	+0.032
GPT-4o-mini	0.807	+0.062	+0.054	+0.076
Qwen2.5-72B	0.812	+0.016	+0.039	+0.060
Llama-3.1-70B	0.778	+0.057	+0.033	+0.063
Qwen2.5-32B	0.688	+0.051	+0.069	+0.099
Qwen2.5-14B	0.639	+0.055	+0.069	+0.113
Qwen2.5-7B	0.493	+0.069	+0.076	+0.120
Qwen2.5-3B	0.371	+0.068	+0.070	+0.125
Llama-3.2-3B	0.314	+0.089	+0.068	+0.206
Panel mean	0.643	+0.055	+0.054	+0.098

Table 5: No-match baseline accuracy and advantage over baseline for the three aligned configurations, averaged within subset then across subsets. "Q-only" fixes the question in the subject's native language and the response in another; "R-only" is the symmetric configuration; "native" aligns all three. Full panel in Appendix G.

Alignment with subset	Acc ↑	AUROC ↑	ECE ↓
$q_{\text{lang}} = r_{\text{lang}} = s$	**0.741**	0.745	**0.190**
$q_{\text{lang}} = s, r_{\text{lang}} \neq s$	0.698	0.756	0.214
$q_{\text{lang}} \neq s, r_{\text{lang}} = s$	0.698	0.756	0.198
$q_{\text{lang}} \neq s, r_{\text{lang}} \neq s$	0.643	**0.758**	0.228

Table 6: Panel-mean LSQA metrics under the four alignment regimes with the subset language s (averaged over v_{lang} and over the six subsets). Triple-match maximizes accuracy and absolute calibration but minimizes ranking discrimination.

Two per-subset restrictions: Hindi inversion, Korean outlier. On Hindi, the triple-match underperforms on accuracy: $q_{\text{lang}} = \text{hi}$ with $r_{\text{lang}} \neq \text{hi}$ reaches 0.723 panel-mean accuracy, against 0.669 for triple-match and 0.644 for $r_{\text{lang}} = \text{hi}$ alone—a Hindi-production bottleneck specific to the panel's open-weight tokenizers, detectable only because we included a typologically distant non-Latin-script language. AUROC and ECE differences across the four regimes stay within 0.011 and 0.008, so the inversion is accuracy-specific (Appendices I, R.1). On ECE, the native-tone advantage is strongly English-anchored and weakens or reverses on non-Latin-script subjects; Korean is an independent calibration outlier across all four regimes (Appendix J). A representative item is in Appendix R.2.

4.4 Pairwise alignments and the verifier axis

The previous two subsections analyzed the English and native configurations separately. We now jointly estimate the four pairwise alignments with a mixed-effects regression (Table 7) and use the result to revisit the verifier-language axis v_{lang}, which

Table 7: Match-type mixed-effects coefficients Δ on three datasets, with random intercept per model. "A" = Δ_{AUROC}, "E" = Δ_{ECE}. Significance: *** $p<0.001$, ** $p<0.01$, * $p<0.05$, ns $p\geq0.05$.

	LSQA		MMMLU		TriviaQA	
	A	E	A	E	A	E
$\mathbb{1}_{q_{\text{lang}}=r_{\text{lang}}}$	$-.006^{***}$	$-.005^{*}$	$-.009^{***}$	$-.006^{***}$	$-.001^{\text{ns}}$	$-.003^{\text{ns}}$
$\mathbb{1}_{r_{\text{lang}}=v_{\text{lang}}}$	$+.008^{***}$	$-.012^{***}$	$+.000^{\text{ns}}$	$-.003^{\text{ns}}$	$+.006^{**}$	$+.010^{**}$
$\mathbb{1}_{q_{\text{lang}}=v_{\text{lang}}}$	$+.003^{*}$	$-.005^{*}$	$+.001^{\text{ns}}$	$-.004^{*}$	$+.001^{\text{ns}}$	$-.004^{\text{ns}}$
$\mathbb{1}_{q=r=v}$	$-.007^{\text{ns}}$	$+.014^{*}$	$+.001^{\text{ns}}$	$+.002^{\text{ns}}$	$-.005^{\text{ns}}$	$-.016^{\text{ns}}$

carries $\eta^2 < 0.5\%$ as a main effect (§4.1) but appears in three of the four pairwise predictors. An alignment improves calibration only when it raises AUROC (discrimination) and lowers ECE simultaneously.

Pairwise alignment results. On LSQA, $r_{\text{lang}}{=}v_{\text{lang}}$ raises AUROC and lowers ECE, the only pairing in the three datasets where both move in the favorable direction. On MMMLU, $q_{\text{lang}}{=}r_{\text{lang}}$ lowers ECE but also lowers AUROC—a within-language saturation effect, not a calibration improvement. On TriviaQA, $r_{\text{lang}}{=}v_{\text{lang}}$ raises AUROC but raises ECE, trading discrimination for over-confidence. The asymmetry coheres with the regime/format finding above: r_{lang} is the load-bearing axis on LSQA, so aligning the verifier with it is what calibration responds to; q_{lang} is the load-bearing axis on the LA datasets, but no q_{lang}-involving alignment translates that accuracy importance into a clean calibration gain. The full three-language alignment $q_{\text{lang}}{=}r_{\text{lang}}{=}v_{\text{lang}}$ is sub-additive on LSQA ECE—pairwise coefficients sum to -0.022 but the three-language term adds $+0.014$—reflecting that fully locking the model into one language couples calibration to whatever language-specific bias the model has.

Verifier axis: robust marginally, asymmetric in interaction. Prior multilingual confidence work [Xue et al., 2025, Yang et al., 2024, Zhou et al., 2025] implicitly fixes $v_{\text{lang}} = r_{\text{lang}}$, leaving the question of v_{lang}'s marginal effect unanswered. Two complementary results address it. First, v_{lang} has little *marginal* effect: an ablation (Appendix Q) confirms a panel-mean AUROC range of 0.041 on LSQA and 0.042 on MMMLU across the six v_{lang} slices (16/16 models), so applications ranking by $P(\text{True})$ can treat v_{lang} as a free knob without translating the verifier prompt. ECE shifts are larger (0.007–0.150 across models), so appli-

cations reporting absolute calibration should pin v_{lang} for reproducibility. Second, the interactions in Table 7 reveal a regime-dependent sign-flip: $r_{\text{lang}}{=}v_{\text{lang}}$ improves ECE on open-ended LSQA ($\Delta_{\text{ECE}} = -0.012$, $p < 10^{-3}$) but worsens it on short-answer TriviaQA ($+0.010$, $p < 0.01$): the same-language verifier on a short, confidently produced answer worsens calibration rather than improving it. The effect localizes to weaker multilinguals (Appendix O).

5 Discussion

The factorial design surfaces structure that the single-axis framing cannot. On LS content the load-bearing axis is the response language: the native-language advantage [Xue et al., 2025] resolves into a single triple-match configuration where question, response, and content language all coincide, with off-native matched configurations essentially flat. The operational prescription sharpens from "answer in the question's language" to "answer in the language whose pretraining corpus contains the relevant content," consistent with the English-anchored pivot of Wendler et al. [2024] and the cross-lingual factual inconsistency of Qi et al. [2023], Wang et al. [2025]. On LA content the load-bearing axis switches to the question side [Xue et al., 2025, Yang et al., 2024]: asking in English is decisive, the response language is free.

Reading accuracy and calibration jointly exposes a saturation–discrimination tradeoff visible across both regimes: the configurations that maximize accuracy (triple-match on LS, all-English on LA) also minimize ranking discrimination, because same-language production compresses $P(\text{True})$ near the ceiling. This tradeoff appears along the axis that "produces" the answer in each regime (r_{lang} on LS, the joint $(q_{\text{lang}}, r_{\text{lang}})$ alignment on LA). The strong positive accuracy–AUROC coupling visible on MMMLU collapses to near zero on TriviaQA, marking it as a format property of MCQ rather than an LA-regime property—a separation a single-LA-dataset analysis cannot make.

The design choices behind these findings were necessary rather than ornamental. Adding Hindi to the original LSQA pool surfaced the only counterexample to native-language dominance in the panel, suggesting benchmarks aimed at testing linguistic-dominance claims need at least one typologically distant non-Latin-script language. The multi-family scaling panel similarly converts regime statements into developmental trajectories: variance migrates from language main effects to language-by-subject interactions at frontier scale, on *different* axes between LSQA and MMMLU.

Practical guidance. Four prescriptions follow. *(i) Match the prompting axis to the task regime*—response language on LS, question language on LA; matching all axes by default wastes the freedom available on the inactive one. *(ii) English is a question-side resource for accuracy*, not a recipe to use everywhere; on LA tasks, pairing $q_{\text{lang}} = \text{EN}$ with a non-EN response trades ~ 13 p.p. of accuracy for the best calibration on both AUROC and ECE. *(iii) On LA, the two metrics favor different axes.* $q_{\text{lang}} = \text{EN}$ helps accuracy, $r_{\text{lang}} \neq \text{EN}$ helps calibration; their combination ($q_{\text{lang}} = \text{EN}$, $r_{\text{lang}} \neq \text{EN}$) is the calibration-best configuration but not the accuracy-best one. Applications must decide which metric to prioritize. *(iv) On LS, triple-match is accuracy-optimal but discrimination-worst*; partial alignment buys little accuracy. Applications needing absolute correctness should triple-match; applications ranking by $P(\text{True})$ should not.

6 Conclusion

The single "prompt language" axis along which multilingual LLM behavior is conventionally evaluated hides most of the structure that emerges when question, response, and verifier languages are varied independently and accuracy is read together with calibration. A full factorial over the three axes, combined with a benchmark grid that separates regime from response format, yields three findings. The load-bearing language axis depends on task regime—response-driven on LS, question-driven on LA. The English-language advantage on LA tasks is primarily question-side, and accuracy and calibration favor different axes: $q_{\text{lang}} = \text{EN}$ helps accuracy, $r_{\text{lang}} \neq \text{EN}$ helps calibration, and ($q_{\text{lang}} = \text{EN}$, $r_{\text{lang}} \neq \text{EN}$) gives the best calibration at a ~ 13 p.p. accuracy cost relative to all-English. The native-language advantage on LS tasks is localized to the triple-match configuration, but the same configuration also depresses ranking discrimination, mirroring the saturation–discrimination tradeoff that the all-English configuration exhibits on LA content. Together, these findings make multilingual evaluation a four-metric problem that a single-axis design cannot resolve.

Limitations

The two language-agnostic datasets do not exhaust the LA response-format space. We do not cover long-form open-ended LA (GSM8K-style multi-step reasoning), and cannot say whether the TriviaQA short-answer conclusions extend to longer free-form LA outputs. TriviaQA is also single-corpus, so $Q \times$ subset interactions are not estimable on it. The MMMLU subset axis is subject-difficulty, not cultural-region, so cross-regime numerical comparisons of subset effects between MMMLU and LSQA should be interpreted with that asymmetry in mind.

LSQA was produced by a single LLM translator (Claude Opus 4.7) with content-preservation spot-checks but no human inter-rater verification, and correctness labels come from a single LLM judge (GPT-4o-mini). The judge-reliability literature [Zheng et al., 2023, Bavaresco et al., 2025, Hada et al., 2024] suggests this is acceptable for relative comparisons but warrants human validation for absolute claims. The verifier-language ablation translates only the instructional prose around the `True`/`False` target tokens, leaving fully translated target morphology for future work. Our OpenAI snapshot pins (`gpt-4o-2024-11-20`, `gpt-4o-mini-2024-07-18`) may drift as default aliases rotate.

References

Anthropic. 2026. Claude opus 4.7. https://www.anthropic.com/news/claude-opus-4-7.

Yavuz Faruk Bakman, Duygu Nur Yaldiz, Sungmin Kang, Tuo Zhang, Baturalp Buyukates, Salman Avestimehr, and Sai Praneeth Karimireddy. 2025. Reconsidering LLM uncertainty estimation methods in the wild. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29531–29556, Vienna, Austria. Association for Computational Linguistics.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. 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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 516 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. 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 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *roceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1321–1330. PMLR.

Rishav Hada, Varun Gumma, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. 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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Gemma Team Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ram'e, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gael Liu, and 191 others. 2025. Gemma 3 technical report. *ArXiv*, abs/2503.19786.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. In *Transactions on Machine Learning Research (TMLR)*.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

OpenAI. 2024. Multilingual massive multitask language understanding (mmmlu). Dataset available at `https://huggingface.co/datasets/openai/MMMLU`.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

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, Junyang Lin, and 24 others. 2025. Qwen2.5 Technical Report. *arXiv preprint*. ArXiv:2412.15115 [cs].

Gemma Team Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L'eonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram'e, Johan Ferret, Peter Liu, Pouya Dehghani Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 176 others. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *Preprint*, arXiv:2402.07827.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze. 2025. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5075–5094, Vienna, Austria. Association for Computational Linguistics.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *Proceedings of the 12th International Conference on Learning Representations*.

Boyang Xue, Hongru Wang, Rui Wang, Sheng Wang, Zezhong Wang, Yiming Du, Bin Liang, Wenxuan Zhang, and Kam-Fai Wong. 2025. MlingConf: A comprehensive study of multilingual confidence estimation on large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2535–2556, Vienna, Austria. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2505.09388*.

Yahan Yang, Soham Dan, Dan Roth, and Insup Lee. 2024. On the calibration of multilingual question answering LLMs. *arXiv preprint arXiv:2311.08669*.

Dongkeun Yoon, Seungone Kim, Sohee Yang, Sunkyoung Kim, Soyeon Kim, Yongil Kim, Eunbi Choi, Yireun Kim, and Minjoon Seo. 2025. Reasoning models better express their confidence. In *Advances in Neural Information Processing Systems 38*.

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 *NeurIPS Datasets and Benchmarks Track*.

Ej Zhou, Caiqi Zhang, Tiancheng Hu, Chengzu Li, Nigel Collier, Ivan Vulić, and Anna Korhonen. 2025. Beyond the final layer: Intermediate representations for better multilingual calibration in large language models. *arXiv preprint arXiv:2510.03136*.

Appendix

A Compute and Data Coverage

Open-weight inference ran on a node with eight NVIDIA B200 GPUs (180 GB HBM3e each). Closed-API cost was ∼$320. Each dataset has 960 items per (model, q_{lang}, r_{lang}, v_{lang}) configuration. LSQA is stratified into six 160-item subsets (one per source language) and MMMLU into four 240-item subjects; TriviaQA has no subset axis. GPT-4o-mini lost 51 rows ($\sim 0.15\%$) to rate-limiting, affecting subset- and family-level conclusions negligibly.

B LSQA Construction Details

LSQA contains 160 items per subset $\times$ 6 subsets $= 960$ unique items. Each item carries: (a) a gold question and gold answer in the native (source) language of the subset; (b) translations of the question and acceptable-answer surface forms into the other five languages via a high-capacity LLM translator using a content-preservation prompt template that disallows paraphrase of named entities and dates; (c) one of four category labels (history, geography, social/culture, celebrity), with 40 items per category per subset.

Items are seeded from the cultural-knowledge region of each language community. Korean items concern Korean history, geography, celebrities, and social practices; French items concern French content; and so on. This deliberate cultural grounding is what makes the $r_{\text{lang}} \times$ subset interaction meaningful: asking a Korean-cultural question in French to a model that responds in Hindi is a stress test for cross-cultural cross-lingual transfer that a parallel-translation benchmark cannot replicate.

C LLM Judge Prompt and Rubric

The judge prompt instructs GPT-4o-mini to compare a model response $\hat{a}$ against gold answer a^* for a question q and return a single token ("correct" or "incorrect"). Tolerance rules explicitly cover: (a) cross-lingual paraphrase (Korean 이성계 vs. English "Yi Seong-gye" vs. Chinese 李成桂); (b) Sino-Korean / Sino-Japanese name variants; (c) Hindi diacritic differences (किताब vs. क़िताब); (d) trailing punctuation, particles, and articles; (e) partial-but-precise answers where the gold expects a longer form (e.g. "Paris" for "the capital of France"); (f) currency, unit, and ordinal-cardinal equivalences. Negative tolerance rules explicitly disallow: (a) hedging that produces multiple candidates with no commitment; (b) refusals or capability disclaimers; (c) near-misses that share semantic field but identify the wrong entity ("Goryeo dynasty" for "Joseon dynasty"—both correct historical Korean dynasties, but identifying the wrong one).

D Full ANOVA Tables

The raw Type III ANOVA decomposition uses one (model, factor) row per benchmark with columns for sum of squares, degrees of freedom, F-statistic, p-value, and percent of total sum of squares. The residual factor has df $= N - p$ where N is the configuration count and p is the parameter count; residual SS varies from 87.5% (GPT-4o) to 93.5% (Aya-Expanse-8B) of total SS—a property of binary ANOVA noted above. Within-explained percentages (Table 1) are computed as $\text{SS}_f / \sum_{f' \neq \text{Residual}} \text{SS}_{f'}$ for each factor f.

All main effects and interactions reach $p < 10^{-3}$ for every model on the row-level Type III ANOVA, reflecting the very large effective sample size at the row level ($N \approx 34{,}500$ rows per model). Statistical significance is therefore not the question of interest; the question of interest is which factor explains the largest share of explainable variance, which Table 1 answers.

E MMMLU Panel Results

Table 8 reports the corresponding panel summary on the MMMLU LA-regime restriction. The within-panel ordering is broadly consistent with LSQA but with two notable inversions: GPT-4o overtakes GPT-4o-mini on AUROC (0.853 vs. 0.831, reversed from LS), and Llama-3.1-70B leads all open-weight models on AUROC at 0.827.

Model	Acc	AUROC	ECE	Brier
GPT-4o	**0.821**	**0.853**	**0.145**	**0.141**
Qwen2.5-72B	0.756	0.719	0.195	0.191
Llama-3.3-70B	0.750	0.693	0.209	0.203
Llama-3.1-70B	0.730	0.827	0.159	0.168
Qwen2.5-32B	0.723	0.767	0.232	0.229
GPT-4o-mini	0.722	0.831	0.213	0.208
Qwen2.5-14B	0.669	0.751	0.244	0.239
Gemma-2-9B	0.634	0.646	0.327	0.322
Qwen3-8B	0.631	0.719	0.269	0.267
Gemma-3-12B	0.619	0.681	0.292	0.288
Qwen2.5-7B	0.583	0.674	0.315	0.311
Aya-Expanse-8B	0.495	0.520	0.496	0.492
Qwen2.5-3B	0.483	0.662	0.348	0.348
Llama-3.1-8B	0.482	0.696	0.280	0.304
Llama-3.2-3B	0.422	0.618	0.319	0.340
Mistral-7B	0.383	0.559	0.580	0.572
Panel mean	0.619	0.701	0.289	0.279

Table 8: 16-model panel summary on MMMLU LA ($v_{\text{lang}} = \text{EN}$), sorted by accuracy.

F LSQA Subset Accuracy by Model

Model	EN	FR	HI	JA	ZH	KO	EN–KO
GPT-4o	.961	.919	.927	.938	.881	.822	.139
GPT-4o-mini	.916	.866	.866	.855	.826	.620	.296
Qwen3-8B	.921	.870	.845	.849	.857	.612	.309
Qwen2.5-72B	.900	.865	.854	.846	.836	.626	.274
Llama-3.1-70B	.894	.844	.832	.810	.770	.603	.291
Llama-3.3-70B	.896	.832	.823	.799	.774	.601	.295
Gemma-3-12B	.869	.774	.761	.776	.733	.534	.335
Gemma-2-9B	.836	.727	.712	.712	.681	.459	.377
Aya-Expanse-8B	.858	.746	.694	.639	.626	.428	.430
Qwen2.5-32B	.835	.747	.709	.738	.736	.477	.358
Qwen2.5-14B	.796	.709	.631	.676	.705	.442	.354
Qwen2.5-7B	.656	.540	.492	.540	.585	.285	.371
Llama-3.1-8B	.727	.603	.573	.494	.504	.319	.408
Qwen2.5-3B	.519	.397	.352	.416	.482	.198	.321
Mistral-7B	.524	.419	.375	.362	.336	.209	.315
Llama-3.2-3B	.526	.381	.375	.324	.285	.159	.367

Table 9: Per-subset accuracy on LSQA and EN–KO gap. Korean is the lowest-accuracy subset for $16/16$ models, English the highest. The EN–KO gap is largest on history (42.7 p.p. panel-mean), smallest on social/culture (18.9 p.p.).

G Native-Tone Results: Full Panel and Subset Breakdown

Model	no-match	+Q-only	+R-only	+native
GPT-4o	0.900	+0.018	+0.033	+0.032
Qwen3-8B	0.820	+0.029	+0.006	+0.019
GPT-4o-mini	0.807	+0.062	+0.054	+0.076
Qwen2.5-72B	0.812	+0.016	+0.039	+0.060
Llama-3.1-70B	0.778	+0.057	+0.033	+0.063
Llama-3.3-70B	0.774	+0.053	+0.027	+0.066
Gemma-3-12B	0.724	+0.042	+0.061	+0.096
Qwen2.5-32B	0.688	+0.051	+0.069	+0.099
Gemma-2-9B	0.663	+0.066	+0.092	+0.111
Aya-Expanse-8B	0.649	+0.039	+0.057	+0.089
Qwen2.5-14B	0.639	+0.055	+0.069	+0.113
Llama-3.1-8B	0.507	+0.106	+0.077	+0.174
Qwen2.5-7B	0.493	+0.069	+0.076	+0.120
Qwen2.5-3B	0.371	+0.068	+0.070	+0.125
Mistral-7B	0.354	+0.060	+0.038	+0.120
Llama-3.2-3B	0.314	+0.089	+0.068	+0.206
Panel mean	0.643	+0.055	+0.054	+0.098

Table 10: Full 16-model native-tone panel (extends Table 5). Family-conditional asymmetry: Qwen2.5 R-only > Q-only at every scale; Llama Q-only > R-only at every scale.

H Decomposition by Alignment Regime

Regime	Definition	Configurations	Δ_{acc}
A (native-tone)	$q = r = s$	1	+0.102
B (bare diagonal off-native)	$q = r \neq s$	5	+0.018
C (Q-only match)	$q = s, r \neq s$	5	+0.059
D (R-only match)	$r = s, q \neq s$	5	+0.058
E (no match)	$q \neq s, r \neq s, q \neq r$	20	baseline

Table 11: Five-regime decomposition of the $Q \times R$ grid by subject-language matching. Native-tone (A) dominates; bare-diagonal off-native (B) is essentially flat. Q-only (C) and R-only (D) are nearly equal and roughly additive ($C + D \approx A$).

I Native-Tone Advantage by Subset

On ECE rather than accuracy, panel-mean $\Delta_{\text{ECE}}^{\text{native-no-match}}$ is -0.075 on EN (large improvement), -0.027 on FR, -0.013 on ZH, and effectively zero or slightly positive on JA, KO, HI. The "native-tone calibration advantage" on LS therefore narrows to an EN-anchored effect with a residual mild FR/ZH effect.

Subset	native–no-match	models pos.
EN	$+0.182$	$16/16$
FR	$+0.167$	$16/16$
ZH	$+0.129$	$16/16$
JA	$+0.060$	$13/16$
KO	$+0.055$	$11/16$
HI	-0.004	$9/16$

Table 12: Per-subset native-tone accuracy advantage. The Hindi subset inverts: native-tone 0.669 vs. Q-only ($q_{\text{lang}} = \text{HI}, r_{\text{lang}} \neq \text{HI}$) 0.723, suggesting a Hindi-production bottleneck.

J Korean LSQA Subset Calibration Outlier

On regime A (native-tone) the LSQA panel-mean Δ_{ECE} across the six cultural subsets clusters tightly at {en -0.043, fr -0.015, ja -0.014, zh -0.011, hi $+0.004$}: five of six subsets show native-tone improves or does not worsen calibration. The Korean subset breaks this pattern: $\Delta_{\text{ECE}}^{\text{KO},A} = +0.033$, eight times the next-largest positive deviation (Hindi $+0.004$). The KO outlier persists across LSQA regimes: all four $\Delta_{\text{ECE}}^{\text{KO},R}$ values for $R \in \{A, B, C, D\}$ are positive ($+0.033, +0.040, +0.002, +0.017$), while every other LSQA subset shows uniformly negative or near-zero Δ_{ECE} across regimes. The KO subset also inverts the typical native-subject accuracy pattern: $q_{\text{lang}}{=}r_{\text{lang}}{=}\text{KO}$ on the KO subset gives panel-mean 0.50 vs. 0.64 on non-KO subsets. We do not see analogous outliers in MMMLU's four subject categories ($\Delta_{\text{ECE}} \in [-0.014, -0.006]$). The KO LSQA subset warrants a follow-up audit (item difficulty, translation friction, or domain mismatch with Korean training distributions).

K English Question-Side Dominance: Detailed Breakdown

Language ℓ	panel-mean acc at $q_{\text{lang}}{=}\ell$	at $r_{\text{lang}}{=}\ell$
EN	**0.709**	**0.634**
FR	0.643	0.614
ZH	0.631	0.622
JA	0.598	0.623
KO	0.594	0.611
HI	0.539	0.610
Range	0.170	0.024
EN best in	$14/16$	$5/16$

Table 13: Q-axis vs. R-axis English dominance on MMMLU LA, panel-mean. Question-side English helps a lot ($+17$ p.p. range); response-side English helps barely ($+2.4$ p.p.).

L ANOVA Decomposition on MMMLU

Model	q	r	s	$q{\times}r$	$q{\times}s$	$r{\times}s$
Llama-3.2-3B	**35.1**	11.2	16.6	13.9	18.5	4.7
Qwen2.5-3B	14.1	**25.6**	11.5	12.1	20.8	15.9
Mistral-7B	**42.3**	3.8	25.2	3.3	18.4	7.0
Qwen2.5-7B	**49.8**	0.4	20.8	2.6	24.8	1.6
Llama-3.1-8B	**29.8**	10.2	23.4	12.0	20.1	4.4
Qwen3-8B	**32.6**	5.4	22.9	13.2	20.3	5.5
Aya-Expanse-8B	19.3	1.5	**44.2**	2.6	28.3	4.1
Gemma-2-9B	27.0	0.8	21.6	4.7	**42.8**	3.1
Gemma-3-12B	20.3	7.3	9.2	**25.5**	28.2	9.5
Qwen2.5-14B	**50.6**	0.9	11.2	10.0	23.1	4.3
Qwen2.5-32B	30.2	0.2	11.3	7.2	**43.7**	7.4
Qwen2.5-72B	18.4	1.3	23.4	6.3	**30.7**	19.9
Llama-3.1-70B	**38.4**	1.6	11.0	12.8	19.5	16.7
Llama-3.3-70B	**32.7**	0.1	8.4	4.7	26.2	27.8
GPT-4o-mini	32.7	0.6	18.4	4.2	**37.0**	7.0
GPT-4o	18.5	2.5	**27.9**	7.5	27.6	15.9

Table 14: LA-regime ANOVA decomposition (% within explained variance) on MMMLU. The q_{lang} main effect is the single largest factor in $10/16$ models; the r_{lang} main effect is essentially absent ($< 2\%$) at $\geq$ 7B scale.

M Match-Type Mixed-Effects Specification

We fit, separately on each dataset $d \in \{\text{LSQA}, \text{MMMLU}, \text{TriviaQA}\}$ and for each calibration metric $y \in \{\text{AUROC}, \text{ECE}\}$, the mixed-effects linear model

$$\begin{aligned} y_i = \mu &+ \beta_{\text{qr}} \mathbb{1}[q_{\text{lang}}{=}r_{\text{lang}}] \\ &+ \beta_{\text{rv}} \mathbb{1}[r_{\text{lang}}{=}v_{\text{lang}}] \\ &+ \beta_{\text{qv}} \mathbb{1}[q_{\text{lang}}{=}v_{\text{lang}}] \\ &+ \beta_{\text{tri}} \mathbb{1}[q_{\text{lang}}{=}r_{\text{lang}}{=}v_{\text{lang}}] \\ &+ u_{m(i)} + \varepsilon_i, \end{aligned}$$

where i ranges over 34,560 configurations per dataset, $u_m \sim \mathcal{N}(0, \sigma_m^2)$ absorbs baseline calibration differences between models, and β_{tri} measures deviation from additivity when all three matches coincide. The random intercept on model exploits within-model variance (which dwarfs across-model variance for these contrasts), the standard approach to the "model is a clusterable nuisance" setting in multi-system NLP evaluation. Cluster bootstraps over $n_{\text{model}} = 16$ are underpowered for the small effects (≤ 0.02 on AUROC/ECE) we estimate; the mixed-effects test is the primary inference target.

N Additional Cross-Dataset Findings

Six further findings emerge from the three-dataset comparison but are not central to the main text.

The r_{lang}=EN trap zone on TriviaQA. On TriviaQA only, the r_{lang}=EN slice combines panel-mean accuracy of 0.66 (high) with AUROC of 0.62 (lowest by r_{lang}); other r_{lang} slices have lower accuracy (0.45–0.55) but higher AUROC (0.73–0.78). The model answers TriviaQA well in English but cannot tell when it is wrong; the same model in non-English r_{lang} answers worse but knows which answers are uncertain. This trap zone is invisible on LSQA (cultural content makes r_{lang}=EN context-mismatched) and on MMMLU (MCQ ceiling pins all r_{lang} slices at high AUROC). It is a short-answer LA artifact.

Non-English response boosts AUROC on TriviaQA. Panel-mean TriviaQA AUROC at $r_{\text{lang}} \neq$ EN is 0.755, vs. 0.620 at $r_{\text{lang}} =$ EN—a $+0.135$ AUROC lift from not answering in English. This is the calibration counterpart of the trap zone above.

Verifier plasticity is inverse to format constraint. The fraction of configuration-level AUROC variance attributable to v_{lang}-conditional interactions ($q{\times}v + r{\times}v + v{\times}s$) is largest on TriviaQA ($\sim$ 8% of explained variance), smaller on LSQA ($\sim 4\%$), and smallest on MMMLU ($\sim 1\%$). Looser response format leaves more room for the verifier-language choice to perturb calibration.

Family-conditional saturation is dataset-invariant. Gemma is consistently negative on $\Delta^{\text{A}}_{\text{AUROC}}$ across all three datasets; GPT-4o consistently positive; Llama-3.1 vs. 3.3 flips sign within-family at fixed scale on all three datasets. The direction of native-tone saturation effect is therefore a training-recipe property, not a dataset property.

Cross-dataset configuration typology. Mapping each model's configuration-level $r(\text{acc}, \text{AUROC})$ rank to its position on each of the three datasets gives a 3-D configuration typology that clusters models into four operating regimes: well-calibrated generalists (GPT-4o, Llama-3.1-70B), MCQ-saturated (Qwen2.5-72B), multilingual-tuned-but-flat (Aya-Expanse), and calibration floor (Mistral-7B-v0.3).

Non-EN mediation. Non-EN r_{lang} choices mediate the q_{lang}=EN accuracy advantage on TriviaQA: the $r_{\text{lang}} =$ EN trap zone occupies 40% of the expected accuracy advantage, but $r_{\text{lang}} \neq$ EN configurations recover most of it on AUROC.

O Match-Type Decomposition by Model (TriviaQA)

Per-model TriviaQA r_{lang}=v_{lang} ECE coefficients (positive = worsens calibration): Mistral-7B-v0.3 $+0.042$, Llama-3.1-8B $+0.028$, Gemma-2-9B $+0.028$, Aya-Expanse-8B $+0.022$, Qwen2.5-3B $+0.018$, Llama-3.2-3B $+0.017$, Qwen2.5-7B $+0.015$, Qwen3-8B $+0.011$, Qwen2.5-14B $+0.008$, Llama-3.3-70B $+0.005$, Gemma-3-12B $+0.003$, Llama-3.1-70B $+0.001$, GPT-4o-mini -0.001, Qwen2.5-32B -0.003, Qwen2.5-72B -0.006, GPT-4o -0.008. The calibration-worsening effect is monotonically correlated with model weakness: the seven smallest/weakest models drive the panel-mean positive coefficient, while the four frontier models show null or negative effects.

P Reasoning-Mode $P(\text{True})$ Extraction (Qwen3-8B)

Qwen3-8B by default opens its response with a `<think>...</think>` block before emitting the final verdict. When we apply the verifier prompt of §3.3 and read the first-token logits at the boundary, we capture the first token of the reasoning trace—not the True/False verdict. This produces a systematically degraded $P(\text{True})$ distribution.

Setting	median $P(\text{True})$	ECE	AUROC	Brier
thinking enabled (broken)	0.069	0.744	0.677	0.692
thinking disabled (fixed)	1.000	0.167	0.798	0.159

Table 15: Qwen3-8B $P(\text{True})$ on LSQA ($v_{\text{lang}} =$ EN) with reasoning mode enabled vs. disabled. Enabled: the first token after the verifier-prompt boundary is the start of the reasoning trace, not the verdict, producing systematic underconfidence and catastrophic miscalibration (median $P(\text{True}) = 0.069$ on a panel where accuracy is ~ 0.7). Disabled: the model emits a verdict directly, producing the opposite tendency—median $P(\text{True}) = 1.000$ is overconfident in absolute terms, but AUROC and ECE both recover to panel-typical values (ECE comparable to Llama-3.1-70B at 0.166), indicating $P(\text{True})$ still discriminates correct from incorrect answers despite the high median.

Recommendation. For reasoning-mode models, either (a) disable thinking during verification, or (b) extract logits at the post-thinking position (after

the closing `</think>` token). Naive first-token extraction at the verifier boundary should be avoided. All Qwen3-8B numbers in the main paper use `enable_thinking=False`; the thinking-enabled run was discarded after this pitfall was identified.

Q Verifier-Language Robustness Ablation

The main body uses the $v_{\text{lang}} = \text{EN}$ slice as the default for two reasons: it lets readers compare directly with prior single-prompt-language multilingual calibration work, and it is the most common deployment setting for English-language pipelines. This appendix asks how much the main findings depend on that choice. Concretely, for each model we report the panel-mean AUROC range across the six v_{lang} slices (LS: LSQA; LA: MMMLU), where a small range means swapping the verifier language does not change the model's relative discrimination ability, and the qualitative claims of the main body would survive evaluating on a different v_{lang} slice.

AUROC across v_{lang}: LS regime (LSQA)

Model	AUROC min	AUROC max	Range
GPT-4o-mini	0.876	0.889	**0.013**
Qwen2.5-72B	0.742	0.762	0.020
Llama-3.1-70B	0.799	0.821	0.022
Llama-3.2-3B	0.747	0.771	0.025
Qwen2.5-14B	0.779	0.804	0.025
Qwen2.5-7B	0.769	0.795	0.026
Llama-3.1-8B	0.783	0.809	0.027
GPT-4o	0.794	0.829	0.035
Qwen2.5-32B	0.757	0.799	0.042
Gemma-3-12B	0.700	0.742	0.042
Llama-3.3-70B	0.717	0.762	0.045
Gemma-2-9B	0.709	0.759	0.049
Aya-Expanse-8B	0.533	0.582	0.050
Qwen2.5-3B	0.720	0.788	0.068
Mistral-7B	0.604	0.685	0.080
Qwen3-8B	0.723	0.812	**0.089**
Panel mean			0.041

Table 16: LS-regime v_{lang} AUROC range. Panel-mean range 0.041.

AUROC across v_{lang}: LA regime (MMMLU)

Dataset-invariance. Panel-mean AUROC range across v_{lang} is 0.041 on LS and 0.042 on LA, agreeing to within 0.001. The model-level ranking—stable models (GPT-4o, GPT-4o-mini, Llama-3.1-70B) at the top, Qwen3-8B at the bottom—is preserved across regimes. The v_{lang} stability of

Model	AUROC min	AUROC max	Range
GPT-4o	0.844	0.853	**0.009**
Aya-Expanse-8B	0.517	0.528	0.012
GPT-4o-mini	0.817	0.831	0.014
Llama-3.1-8B	0.683	0.697	0.015
Mistral-7B	0.548	0.568	0.020
Llama-3.1-70B	0.798	0.827	0.029
Qwen2.5-7B	0.652	0.690	0.038
Llama-3.2-3B	0.587	0.625	0.038
Qwen2.5-72B	0.677	0.719	0.042
Qwen2.5-3B	0.618	0.662	0.044
Gemma-2-9B	0.594	0.646	0.052
Gemma-3-12B	0.669	0.724	0.055
Qwen2.5-32B	0.710	0.767	0.057
Qwen2.5-14B	0.686	0.751	0.065
Llama-3.3-70B	0.678	0.747	0.069
Qwen3-8B	0.615	0.735	**0.120**
Panel mean			0.042

Table 17: LA-regime v_{lang} AUROC range. Panel-mean range 0.042, nearly identical to the LS-regime panel mean.

$P(\text{True})$ discrimination is therefore a model property, not a content property, exactly as predicted.

ECE across v_{lang} (less robust in both regimes)

Per-model ECE range across v_{lang} varies from 0.007 (GPT-4o-mini, LS) to 0.150 (Qwen3-8B, LS) on LS; comparable spreads obtain on LA. v_{lang} *shifts confidence levels* more than it shifts ranking quality, in both regimes. Implication: $P(\text{True})$-based selective prediction (ranking) is robust to v_{lang} choice in either regime; confidence-thresholding applications (absolute calibration) should pin v_{lang} for reproducibility.

Robustness of main-text findings to v_{lang}

The accuracy-based findings of §4 are *exactly* v_{lang}-invariant by construction: correctness labels are determined during response generation and do not depend on v_{lang}. We verified this directly: the panel-mean LS native-tone advantage $\Delta_{\text{acc}}^{\text{native–no-match}}$ computed separately on each of the six v_{lang} slices equals $+0.0981$ to within numerical precision ($< 10^{-4}$ across the six values). The same invariance holds for all accuracy-based LA findings (English question-side dominance, diagonal advantage absence, panel-mean accuracy ordering).

The AUROC- and ECE-based findings of §4 are v_{lang}-sensitive in magnitude but not in sign. For the native-tone analysis (§4.3): $\Delta_{\text{AUROC}}^{\text{native–no-match}}$ ranges $[-0.016, -0.007]$ across the six v_{lang} (panel mean -0.012), and $\Delta_{\text{ECE}}^{\text{native–no-match}}$ ranges

$[-0.052, -0.025]$ (panel mean -0.041). Both deltas agree on sign across all six v_{lang}. The ANOVA decompositions (Tables 1 and 14) were computed on the $v_{\text{lang}} = \text{EN}$ slice; computing them on any other v_{lang} slice yields the same model-level ranking of dominant factors modulo small magnitude shifts.

Hindi verifier prompts are the worst v_{lang} in 6/16 models on LS

A mild systematic effect, possibly attributable to Hindi tokenizer fragmentation interacting with the English-target-token recovery described in §3.3 (English worst in $4/16$; other languages in 1–3). We flag the observation but do not over-claim from a $6/16$ split. The same pattern is weaker but present on LA.

R Qualitative Examples

This appendix presents one concrete item per structural claim in the main paper. Every (question, model response, $P(\text{True})$) triple is copied verbatim from the logged generations at $v_{\text{lang}} = \text{EN}$; nothing is synthesized.

R.1 Hindi production bottleneck

On an LSQA history item ("Who led the 1857 revolution?"), Qwen2.5-32B-Instruct produces a fluent but fabricated answer when asked to respond in Hindi, yet answers correctly when allowed to respond in English—at *the same* verifier confidence. The model's parsing of the Hindi question is intact; the bottleneck is on the production side.

q	r	Question (truncated)	Model response (truncated)	$P(\text{True})$	Correct?
hi	hi	1857 की क्रांति का नेतृत्व किसने किया?	तनसेन राणा, बख्त करन, नन्द लाल बोस, कन्हैयालाल ...इनमें से कोई एक केंद्रीय नेता नहीं था।	0.961	✗
hi	en	1857 की क्रांति का नेतृत्व किसने किया?	The 1857 Indian Rebellion was led by several leaders including Mangal Pandey, Bahadur Shah Zafar, Nana Saheb, and Rani Lakshmibai.	0.961	✓

Table 18: Qwen2.5-32B-Instruct on an LSQA Hindi history item. Gold answer: *Mangal Pandey, Rani Lakshmibai, and many others led the 1857 uprising*. The two configurations share the same Hindi question and the same verifier confidence, but only the English-output configuration produces a factually grounded answer.

R.2 Matched configuration alone is insufficient

On an LSQA Korean history item ("When did the 4.19 Revolution take place?"), Qwen2.5-32B-Instruct produces three informative configurations. The Korean matched configuration ($q{=}r{=}\text{KO}$) answers correctly with full verifier confidence; the English matched configuration off the native subject swaps the event entirely for the 1775 American Revolutionary War and the verifier registers it at the floor; a cross-language configuration with $q{=}\text{KO}$, $r{=}\text{JA}$ recovers the correct year (the model actually does switch its response into Japanese) and again carries full verifier confidence. The useful calibration signal touches the subset language; it is not localized to the matched configuration in general.

q	r	Question	Model response	$P(\text{True})$	Correct?
ko	ko	4.19 혁명이 일어난 해는 언제인가?	4.19 혁명이 일어난 해는 1960년입니다.	1.000	✓
en	en	When did the 4.19 Revolution take place?	The April 19 Revolution took place on April 19, 1775, marking the beginning of the American Revolutionary War.	0.076	✗
ko	ja	4.19 혁명이 일어난 해는 언제인가?	4.19革命が起こった年は1960年です	1.000	✓

Table 19: Qwen2.5-32B-Instruct on an LSQA Korean history item. Gold answer: *The April 19 Revolution was a student-led pro-democracy movement in South Korea in 1960.* The English matched configuration does not merely lose confidence: it swaps the historical event entirely, treating "April 19 Revolution" as a US event. The triple-match configuration and the ($q{=}\text{KO}, r{=}\text{JA}$) cross-language configuration both carry full verifier confidence and the correct year; the off-native matched configuration is at the floor.

R.3 English helps for asking, not for answering

On an MMMLU astronomy item ("Why isn't Mars warm despite its CO_2 atmosphere?"), Qwen2.5-7B-Instruct picks the correct MCQ letter in both English-question configurations and the wrong one in both Hindi-question configurations—independent of the response language. Flipping the response language has no effect on either the selected letter or on $P(\text{True})$; flipping the question language flips both.

q	r	Letter	Model response (truncated)	Correct?
en	en	D	D) The atmosphere on Mars is too thin to trap a significant amount of heat. …	✓
en	hi	D	D) अंतरिक्ष मंडल मARS पर अत्यधिक खारिज है जिससे ग्रीनहाउस प्रभाव …	✓
hi	en	A	A) Mars does not have sufficient internal heat to maintain the greenhouse effect. …	✗
hi	hi	A	A) मंगल पर ग्रीनहाउस प्रभाव को बनाए रखने के लिए पर्याप्त आंतरिक गर्मी नहीं है	✗

Table 20: Qwen2.5-7B-Instruct on an MMMLU astronomy item. Gold letter: D. The two English-question configurations (top) are correct; the two Hindi-question configurations (bottom) are wrong. The response language has no measurable effect on correctness.

R.4 The MCQ confidence ceiling

Qwen2.5-72B-Instruct at $q = r = v_{\text{lang}} = \text{EN}$ on two biology-recall items the model answers correctly. The MMMLU item asks the model to identify a detritivore from four labeled options; $P(\text{True})$ saturates at the ceiling. The TriviaQA item asks the same model to recall the animal sense of "lupine" in free-form; $P(\text{True})$ drops to 0.350 even though the answer ("Wolf") is also correct. The two items are matched in topical domain and factual difficulty (a vocabulary-to-organism mapping); only the surface response format differs.

Dataset	Topic	Question (truncated)	Model response	$P(\text{True})$	Correct?
MMMLU	biology	Which of the following is an example of a detritivore? A) Cactus B) Algae C) Bat D) Fungus	D) Fungus	1.000	✓
TriviaQA	biology	Lupine relates to which animal?	Wolf	0.350	✓

Table 21: Qwen2.5-72B-Instruct on two matched biology vocabulary items, $q = r = v_{\text{lang}} = \text{EN}$. The model answers both items correctly with comparable factual difficulty, but $P(\text{True})$ differs by 0.65. The MCQ format puts the response at a ceiling while the open-ended format does not. This is the format-driven ceiling effect discussed in §4.1; the higher $P(\text{True})$ on MMMLU is not evidence that the model knows more about detritivores than about Latin etymology.

S Raw Data: LSQA

Table 22: LSQA — ACC ↑ per $(q_{\text{lang}}, r_{\text{lang}}, v_{\text{lang}})$. Panel mean.

q_{lang}	r_{lang}	**EN**	FR	ZH	JA	KO	HI
EN	EN	0.839	0.839	0.839	0.839	0.839	0.839
	FR	0.786	0.786	0.786	0.786	0.786	0.786
	ZH	0.741	0.741	0.741	0.741	0.741	0.741
	JA	0.716	0.716	0.716	0.716	0.716	0.717
	KO	0.668	0.668	0.668	0.668	0.668	0.668
	HI	0.592	0.592	0.592	0.592	0.592	0.592
FR	EN	0.805	0.805	0.805	0.805	0.805	0.805
	FR	0.775	0.775	0.775	0.775	0.775	0.775
	ZH	0.712	0.712	0.712	0.712	0.712	0.712
	JA	0.687	0.687	0.687	0.687	0.687	0.687
	KO	0.625	0.625	0.625	0.625	0.625	0.625
	HI	0.566	0.566	0.566	0.566	0.566	0.566
ZH	EN	0.741	0.741	0.741	0.741	0.741	0.741
	FR	0.707	0.707	0.707	0.707	0.707	0.707
	ZH	0.683	0.683	0.683	0.683	0.683	0.683
	JA	0.681	0.681	0.681	0.681	0.681	0.681
	KO	0.641	0.641	0.641	0.641	0.641	0.641
	HI	0.592	0.592	0.592	0.592	0.592	0.592
JA	EN	0.723	0.723	0.723	0.723	0.723	0.723
	FR	0.686	0.686	0.686	0.686	0.686	0.686
	ZH	0.670	0.670	0.670	0.670	0.670	0.670
	JA	0.639	0.639	0.639	0.639	0.639	0.639
	KO	0.606	0.606	0.606	0.606	0.606	0.606
	HI	0.564	0.564	0.564	0.564	0.564	0.564
KO	EN	0.692	0.692	0.692	0.692	0.692	0.692
	FR	0.644	0.644	0.644	0.644	0.644	0.644
	ZH	0.646	0.646	0.646	0.646	0.646	0.646
	JA	0.624	0.624	0.624	0.624	0.624	0.624
	KO	0.567	0.567	0.567	0.567	0.567	0.567
	HI	0.521	0.521	0.521	0.521	0.521	0.521
HI	EN	0.692	0.692	0.692	0.692	0.692	0.692
	FR	0.641	0.641	0.641	0.641	0.641	0.641
	ZH	0.635	0.635	0.635	0.635	0.635	0.635
	JA	0.608	0.608	0.608	0.608	0.608	0.608
	KO	0.559	0.559	0.559	0.559	0.559	0.559
	HI	0.529	0.529	0.529	0.529	0.529	0.529

Table 23: LSQA — AUROC ↑ per $(q_{\text{lang}}, r_{\text{lang}}, v_{\text{lang}})$. Panel mean.

q_{lang}	r_{lang}	**EN**	FR	ZH	JA	KO	HI
EN	EN	0.763	0.757	0.756	0.746	0.747	0.753
	FR	0.752	0.749	0.750	0.745	0.750	0.732
	ZH	0.763	0.755	0.756	0.745	0.755	0.735
	JA	0.767	0.766	0.767	0.759	0.766	0.752
	KO	0.788	0.779	0.784	0.778	0.788	0.766
	HI	0.779	0.775	0.763	0.767	0.774	0.764
FR	EN	0.756	0.745	0.748	0.740	0.741	0.746
	FR	0.753	0.745	0.742	0.738	0.735	0.739
	ZH	0.751	0.742	0.753	0.746	0.747	0.735
	JA	0.742	0.752	0.759	0.757	0.756	0.741
	KO	0.750	0.753	0.754	0.760	0.778	0.745
	HI	0.746	0.751	0.744	0.746	0.748	0.759
ZH	EN	0.760	0.758	0.756	0.744	0.744	0.751
	FR	0.755	0.768	0.757	0.753	0.748	0.750
	ZH	0.747	0.743	0.743	0.740	0.736	0.736
	JA	0.759	0.761	0.762	0.754	0.753	0.751
	KO	0.771	0.777	0.772	0.772	0.771	0.766
	HI	0.775	0.779	0.776	0.778	0.780	0.770
JA	EN	0.759	0.759	0.753	0.747	0.741	0.751
	FR	0.745	0.751	0.742	0.746	0.736	0.744
	ZH	0.755	0.750	0.753	0.744	0.741	0.737
	JA	0.745	0.744	0.738	0.744	0.731	0.740
	KO	0.770	0.774	0.767	0.771	0.771	0.767
	HI	0.766	0.770	0.766	0.769	0.771	0.769
KO	EN	0.766	0.762	0.758	0.755	0.757	0.752
	FR	0.742	0.766	0.747	0.751	0.757	0.752
	ZH	0.768	0.763	0.764	0.762	0.760	0.747
	JA	0.752	0.757	0.760	0.759	0.756	0.738
	KO	0.769	0.770	0.766	0.766	0.769	0.761
	HI	0.766	0.778	0.774	0.774	0.780	0.765
HI	EN	0.766	0.758	0.752	0.753	0.749	0.753
	FR	0.741	0.769	0.749	0.755	0.760	0.754
	ZH	0.752	0.761	0.771	0.761	0.762	0.748
	JA	0.752	0.767	0.770	0.778	0.777	0.759
	KO	0.737	0.756	0.754	0.760	0.777	0.756
	HI	0.766	0.769	0.756	0.759	0.757	0.761

Table 24: LSQA — ECE ↓ per $(q_{\text{lang}}, r_{\text{lang}}, v_{\text{lang}})$. Panel mean.

q_{lang}	r_{lang}	**EN**	FR	ZH	JA	KO	HI
EN	EN	0.143	0.154	0.137	0.139	0.160	0.144
	FR	0.206	0.173	0.189	0.189	0.211	0.226
	ZH	0.199	0.198	0.185	0.186	0.213	0.216
	JA	0.213	0.206	0.198	0.183	0.203	0.218
	KO	0.221	0.234	0.206	0.207	0.204	0.231
	HI	0.208	0.221	0.208	0.218	0.213	0.223
FR	EN	0.153	0.166	0.149	0.155	0.172	0.158
	FR	0.166	0.179	0.166	0.173	0.191	0.172
	ZH	0.222	0.227	0.206	0.208	0.220	0.222
	JA	0.268	0.234	0.223	0.196	0.206	0.236
	KO	0.298	0.273	0.256	0.238	0.218	0.244
	HI	0.284	0.252	0.249	0.249	0.232	0.234
ZH	EN	0.185	0.197	0.186	0.178	0.193	0.183
	FR	0.227	0.199	0.203	0.200	0.218	0.211
	ZH	0.227	0.233	0.226	0.212	0.233	0.224
	JA	0.230	0.217	0.222	0.208	0.222	0.212
	KO	0.246	0.237	0.234	0.224	0.235	0.228
	HI	0.258	0.240	0.231	0.217	0.226	0.236
JA	EN	0.187	0.187	0.196	0.181	0.192	0.176
	FR	0.232	0.206	0.225	0.204	0.224	0.206
	ZH	0.231	0.236	0.232	0.221	0.235	0.223
	JA	0.246	0.239	0.255	0.233	0.255	0.225
	KO	0.247	0.231	0.250	0.226	0.235	0.221
	HI	0.256	0.232	0.249	0.214	0.223	0.229
KO	EN	0.206	0.196	0.216	0.192	0.198	0.191
	FR	0.256	0.211	0.242	0.224	0.216	0.223
	ZH	0.243	0.235	0.242	0.225	0.235	0.237
	JA	0.247	0.227	0.254	0.221	0.227	0.227
	KO	0.264	0.244	0.277	0.249	0.257	0.238
	HI	0.250	0.226	0.248	0.224	0.215	0.246
HI	EN	0.196	0.206	0.200	0.196	0.205	0.201
	FR	0.261	0.212	0.240	0.237	0.227	0.231
	ZH	0.253	0.232	0.225	0.230	0.240	0.240
	JA	0.253	0.225	0.241	0.221	0.229	0.226
	KO	0.299	0.257	0.266	0.254	0.248	0.249
	HI	0.261	0.255	0.275	0.265	0.288	0.258

Table 25: LSQA — Brier ↓ per $(q_{\text{lang}}, r_{\text{lang}}, v_{\text{lang}})$. Panel mean.

q_{lang}	r_{lang}	**EN**	FR	ZH	JA	KO	HI
EN	EN	0.140	0.148	0.134	0.134	0.151	0.140
	FR	0.203	0.170	0.185	0.183	0.205	0.217
	ZH	0.195	0.197	0.183	0.183	0.207	0.216
	JA	0.209	0.205	0.197	0.182	0.198	0.216
	KO	0.217	0.231	0.206	0.205	0.199	0.230
	HI	0.206	0.219	0.210	0.214	0.209	0.221
FR	EN	0.151	0.163	0.149	0.151	0.166	0.156
	FR	0.166	0.177	0.166	0.169	0.185	0.171
	ZH	0.218	0.225	0.202	0.206	0.215	0.224
	JA	0.260	0.232	0.222	0.195	0.204	0.232
	KO	0.290	0.267	0.253	0.236	0.214	0.244
	HI	0.276	0.247	0.246	0.244	0.225	0.234
ZH	EN	0.181	0.194	0.182	0.177	0.189	0.183
	FR	0.223	0.197	0.201	0.199	0.213	0.210
	ZH	0.221	0.230	0.220	0.211	0.227	0.223
	JA	0.224	0.215	0.215	0.206	0.218	0.214
	KO	0.241	0.235	0.229	0.221	0.229	0.229
	HI	0.251	0.234	0.228	0.214	0.221	0.234
JA	EN	0.182	0.186	0.189	0.178	0.188	0.178
	FR	0.227	0.203	0.221	0.202	0.221	0.206
	ZH	0.224	0.233	0.225	0.217	0.229	0.224
	JA	0.238	0.235	0.248	0.229	0.250	0.226
	KO	0.241	0.229	0.245	0.224	0.230	0.223
	HI	0.249	0.227	0.245	0.214	0.221	0.229
KO	EN	0.198	0.196	0.208	0.191	0.196	0.193
	FR	0.250	0.209	0.238	0.223	0.213	0.224
	ZH	0.235	0.232	0.235	0.224	0.231	0.237
	JA	0.239	0.224	0.244	0.219	0.224	0.228
	KO	0.252	0.239	0.265	0.243	0.253	0.237
	HI	0.244	0.222	0.243	0.222	0.212	0.243
HI	EN	0.193	0.202	0.198	0.194	0.202	0.199
	FR	0.255	0.210	0.236	0.233	0.223	0.227
	ZH	0.250	0.230	0.221	0.228	0.234	0.239
	JA	0.247	0.222	0.235	0.219	0.225	0.226
	KO	0.291	0.253	0.261	0.250	0.242	0.245
	HI	0.254	0.250	0.270	0.259	0.281	0.254

Table 26: LSQA per-subset ACC ↑ under v=en, columns are alignment regimes (q=r=s; q-only; r-only; neither). Panel-mean over 16 models.

Subset	q=r=s	q=s	r=s	none
EN	0.945	0.812	0.871	0.763
FR	0.845	0.732	0.766	0.678
ZH	0.764	0.729	0.721	0.635
JA	0.724	0.694	0.693	0.663
KO	0.502	0.501	0.490	0.447
HI	0.670	0.723	0.644	0.674

Table 27: LSQA per-subset AUROC ↑ under v=en, columns are alignment regimes (q=r=s; q-only; r-only; neither). Panel-mean over 16 models.

Subset	q=r=s	q=s	r=s	none
EN	0.768	0.759	0.771	0.759
FR	0.753	0.757	0.771	0.762
ZH	0.696	0.735	0.736	0.741
JA	0.753	0.778	0.765	0.768
KO	0.762	0.761	0.747	0.754
HI	0.776	0.751	0.769	0.766

Table 28: LSQA per-subset ECE ↓ under v=en, columns are alignment regimes (q=r=s; q-only; r-only; neither). Panel-mean over 16 models.

Subset	q=r=s	q=s	r=s	none
EN	0.090	0.189	0.106	0.222
FR	0.135	0.248	0.169	0.213
ZH	0.196	0.220	0.216	0.249
JA	0.194	0.210	0.209	0.224
KO	0.313	0.268	0.292	0.309
HI	0.203	0.239	0.210	0.226

Table 29: LSQA per-subset Brier ↓ under v=en, columns are alignment regimes (q=r=s; q-only; r-only; neither). Panel-mean over 16 models.

Subset	q=r=s	q=s	r=s	none
EN	0.084	0.185	0.105	0.216
FR	0.134	0.239	0.167	0.210
ZH	0.195	0.217	0.211	0.243
JA	0.192	0.206	0.205	0.220
KO	0.298	0.259	0.281	0.298
HI	0.198	0.233	0.207	0.220

T Raw Data: MMMLU

Table 30: MMMLU — ACC ↑ per $(q_{\text{lang}}, r_{\text{lang}}, v_{\text{lang}})$. Panel mean.

q_{lang}	r_{lang}	**EN**	FR	ZH	JA	KO	HI
EN	EN	0.730	0.730	0.730	0.730	0.730	0.730
	FR	0.705	0.706	0.705	0.705	0.705	0.705
	ZH	0.711	0.711	0.711	0.711	0.711	0.711
	JA	0.709	0.709	0.709	0.709	0.709	0.709
	KO	0.699	0.698	0.698	0.698	0.699	0.699
	HI	0.699	0.699	0.699	0.699	0.699	0.698
FR	EN	0.657	0.657	0.657	0.657	0.657	0.657
	FR	0.655	0.655	0.655	0.655	0.655	0.655
	ZH	0.641	0.641	0.641	0.641	0.641	0.641
	JA	0.640	0.640	0.640	0.640	0.640	0.640
	KO	0.632	0.632	0.632	0.632	0.632	0.632
	HI	0.630	0.630	0.630	0.630	0.630	0.630
ZH	EN	0.646	0.646	0.646	0.646	0.646	0.646
	FR	0.620	0.620	0.620	0.620	0.620	0.620
	ZH	0.646	0.646	0.646	0.646	0.646	0.646
	JA	0.636	0.636	0.636	0.636	0.636	0.636
	KO	0.619	0.619	0.619	0.619	0.619	0.619
	HI	0.621	0.621	0.621	0.621	0.621	0.621
JA	EN	0.613	0.612	0.612	0.612	0.613	0.612
	FR	0.589	0.589	0.589	0.589	0.589	0.589
	ZH	0.599	0.599	0.599	0.599	0.599	0.599
	JA	0.616	0.616	0.616	0.616	0.616	0.616
	KO	0.588	0.588	0.588	0.588	0.588	0.588
	HI	0.585	0.585	0.585	0.585	0.585	0.585
KO	EN	0.608	0.608	0.608	0.608	0.608	0.608
	FR	0.585	0.585	0.585	0.585	0.585	0.585
	ZH	0.594	0.594	0.594	0.594	0.594	0.594
	JA	0.599	0.599	0.599	0.599	0.599	0.599
	KO	0.600	0.600	0.600	0.600	0.600	0.600
	HI	0.578	0.578	0.578	0.578	0.578	0.578
HI	EN	0.550	0.550	0.549	0.550	0.550	0.550
	FR	0.529	0.529	0.529	0.529	0.529	0.529
	ZH	0.538	0.538	0.538	0.538	0.538	0.538
	JA	0.541	0.541	0.541	0.541	0.541	0.541
	KO	0.529	0.529	0.529	0.529	0.529	0.529
	HI	0.549	0.549	0.549	0.549	0.549	0.549

Table 31: MMMLU — AUROC ↑ per $(q_{\text{lang}}, r_{\text{lang}}, v_{\text{lang}})$. Panel mean.

q_{lang}	r_{lang}	**EN**	FR	ZH	JA	KO	HI
EN	EN	0.722	0.713	0.712	0.700	0.702	0.707
	FR	0.723	0.721	0.715	0.706	0.708	0.708
	ZH	0.728	0.720	0.711	0.701	0.706	0.705
	JA	0.720	0.711	0.712	0.706	0.705	0.702
	KO	0.730	0.719	0.721	0.707	0.708	0.708
	HI	0.736	0.733	0.727	0.723	0.718	0.716
FR	EN	0.706	0.703	0.701	0.689	0.686	0.685
	FR	0.706	0.701	0.697	0.687	0.682	0.684
	ZH	0.713	0.710	0.706	0.694	0.693	0.695
	JA	0.695	0.687	0.690	0.680	0.680	0.678
	KO	0.701	0.698	0.702	0.686	0.686	0.682
	HI	0.714	0.716	0.709	0.704	0.702	0.698
ZH	EN	0.707	0.698	0.696	0.689	0.687	0.688
	FR	0.719	0.708	0.706	0.695	0.697	0.696
	ZH	0.706	0.693	0.692	0.680	0.681	0.686
	JA	0.694	0.685	0.686	0.677	0.681	0.676
	KO	0.704	0.693	0.696	0.686	0.688	0.686
	HI	0.722	0.710	0.709	0.701	0.697	0.697
JA	EN	0.687	0.684	0.681	0.676	0.671	0.672
	FR	0.712	0.703	0.703	0.690	0.691	0.686
	ZH	0.692	0.685	0.682	0.675	0.668	0.675
	JA	0.679	0.669	0.670	0.666	0.663	0.660
	KO	0.682	0.677	0.679	0.667	0.667	0.666
	HI	0.698	0.692	0.696	0.684	0.682	0.679
KO	EN	0.685	0.677	0.676	0.668	0.670	0.668
	FR	0.702	0.696	0.695	0.688	0.689	0.684
	ZH	0.692	0.683	0.682	0.671	0.675	0.669
	JA	0.674	0.666	0.666	0.661	0.662	0.656
	KO	0.680	0.671	0.670	0.661	0.665	0.657
	HI	0.698	0.695	0.694	0.684	0.688	0.684
HI	EN	0.690	0.684	0.686	0.671	0.673	0.676
	FR	0.703	0.699	0.697	0.686	0.688	0.683
	ZH	0.688	0.685	0.687	0.675	0.673	0.670
	JA	0.673	0.675	0.673	0.667	0.668	0.662
	KO	0.670	0.674	0.677	0.667	0.662	0.663
	HI	0.683	0.676	0.676	0.662	0.664	0.666

Table 32: MMMLU — ECE ↓ per $(q_{\text{lang}}, r_{\text{lang}}, v_{\text{lang}})$. Panel mean.

q_{lang}	r_{lang}	**EN**	FR	ZH	JA	KO	HI
EN	EN	0.217	0.222	0.221	0.222	0.230	0.227
	FR	0.223	0.225	0.224	0.233	0.248	0.234
	ZH	0.218	0.226	0.227	0.225	0.241	0.236
	JA	0.230	0.226	0.230	0.225	0.234	0.236
	KO	0.226	0.233	0.236	0.232	0.241	0.239
	HI	0.231	0.237	0.232	0.227	0.236	0.229
FR	EN	0.261	0.267	0.270	0.275	0.274	0.278
	FR	0.256	0.266	0.264	0.271	0.270	0.268
	ZH	0.264	0.271	0.271	0.274	0.282	0.279
	JA	0.274	0.277	0.269	0.274	0.277	0.275
	KO	0.275	0.283	0.282	0.284	0.281	0.284
	HI	0.287	0.280	0.278	0.281	0.278	0.270
ZH	EN	0.280	0.290	0.281	0.284	0.284	0.291
	FR	0.278	0.289	0.282	0.286	0.293	0.286
	ZH	0.273	0.280	0.274	0.275	0.281	0.283
	JA	0.288	0.292	0.287	0.293	0.289	0.292
	KO	0.291	0.297	0.292	0.292	0.293	0.299
	HI	0.279	0.288	0.278	0.280	0.283	0.277
JA	EN	0.320	0.325	0.321	0.318	0.312	0.321
	FR	0.305	0.315	0.310	0.311	0.309	0.310
	ZH	0.318	0.318	0.319	0.308	0.309	0.312
	JA	0.310	0.313	0.313	0.309	0.302	0.311
	KO	0.322	0.329	0.329	0.319	0.318	0.328
	HI	0.315	0.318	0.320	0.312	0.307	0.307
KO	EN	0.311	0.319	0.325	0.319	0.307	0.316
	FR	0.307	0.315	0.316	0.313	0.313	0.311
	ZH	0.314	0.321	0.327	0.316	0.317	0.315
	JA	0.319	0.319	0.324	0.320	0.315	0.320
	KO	0.311	0.318	0.328	0.319	0.310	0.319
	HI	0.314	0.315	0.323	0.319	0.305	0.300
HI	EN	0.331	0.348	0.341	0.345	0.338	0.340
	FR	0.325	0.339	0.330	0.336	0.335	0.332
	ZH	0.331	0.349	0.344	0.339	0.337	0.339
	JA	0.328	0.332	0.329	0.333	0.329	0.333
	KO	0.338	0.346	0.346	0.351	0.342	0.345
	HI	0.326	0.342	0.338	0.339	0.330	0.330

Table 33: MMMLU — Brier ↓ per $(q_{\text{lang}}, r_{\text{lang}}, v_{\text{lang}})$. Panel mean.

q_{lang}	r_{lang}	**EN**	FR	ZH	JA	KO	HI
EN	EN	0.217	0.221	0.221	0.224	0.228	0.229
	FR	0.226	0.225	0.227	0.234	0.245	0.239
	ZH	0.222	0.228	0.229	0.229	0.240	0.241
	JA	0.230	0.229	0.230	0.227	0.235	0.240
	KO	0.228	0.235	0.234	0.235	0.239	0.243
	HI	0.234	0.238	0.234	0.233	0.240	0.233
FR	EN	0.262	0.266	0.268	0.275	0.274	0.279
	FR	0.259	0.265	0.264	0.272	0.272	0.274
	ZH	0.265	0.270	0.271	0.275	0.282	0.280
	JA	0.275	0.277	0.272	0.276	0.280	0.280
	KO	0.278	0.282	0.281	0.287	0.283	0.288
	HI	0.288	0.281	0.280	0.285	0.283	0.273
ZH	EN	0.279	0.287	0.282	0.285	0.285	0.293
	FR	0.279	0.286	0.283	0.288	0.293	0.291
	ZH	0.274	0.280	0.276	0.278	0.282	0.287
	JA	0.288	0.292	0.288	0.293	0.292	0.295
	KO	0.291	0.295	0.292	0.294	0.294	0.302
	HI	0.280	0.287	0.280	0.284	0.287	0.281
JA	EN	0.315	0.320	0.318	0.315	0.313	0.321
	FR	0.303	0.311	0.308	0.311	0.310	0.313
	ZH	0.315	0.315	0.316	0.310	0.311	0.316
	JA	0.309	0.312	0.312	0.309	0.306	0.314
	KO	0.322	0.326	0.327	0.322	0.319	0.331
	HI	0.313	0.315	0.317	0.311	0.310	0.310
KO	EN	0.308	0.316	0.320	0.319	0.308	0.316
	FR	0.308	0.313	0.314	0.315	0.313	0.315
	ZH	0.312	0.318	0.324	0.317	0.316	0.319
	JA	0.318	0.319	0.323	0.319	0.317	0.323
	KO	0.311	0.317	0.325	0.320	0.312	0.321
	HI	0.314	0.314	0.321	0.319	0.306	0.304
HI	EN	0.329	0.344	0.339	0.343	0.338	0.339
	FR	0.324	0.335	0.330	0.337	0.335	0.332
	ZH	0.331	0.345	0.343	0.342	0.340	0.339
	JA	0.330	0.332	0.331	0.336	0.334	0.335
	KO	0.340	0.345	0.347	0.354	0.346	0.345
	HI	0.327	0.339	0.336	0.339	0.334	0.331

Table 34: MMMLU per-subject ACC ↑ under v=en, columns are alignment regimes (q=r=s; q-only; r-only; neither). Note: MMMLU subjects are not cultural categories — same (q, r, s) construction as LSQA, for completeness. Panel-mean over 16 models.

Subject	q=r=s	q=s	r=s	none
Humanities	0.692	0.671	0.575	0.561
Other	0.757	0.732	0.625	0.610
STEM	0.679	0.644	0.591	0.572
Social Sciences	0.792	0.771	0.666	0.651

Table 35: MMMLU per-subject AUROC ↑ under v=en, columns are alignment regimes (q=r=s; q-only; r-only; neither). Note: MMMLU subjects are not cultural categories — same (q, r, s) construction as LSQA, for completeness. Panel-mean over 16 models.

Subject	q=r=s	q=s	r=s	none
Humanities	0.687	0.707	0.661	0.672
Other	0.713	0.725	0.694	0.692
STEM	0.781	0.767	0.759	0.747
Social Sciences	0.705	0.710	0.666	0.672

Table 36: MMMLU per-subject ECE ↓ under v=en, columns are alignment regimes (q=r=s; q-only; r-only; neither). Note: MMMLU subjects are not cultural categories — same (q, r, s) construction as LSQA, for completeness. Panel-mean over 16 models.

Subject	q=r=s	q=s	r=s	none
Humanities	0.250	0.250	0.342	0.334
Other	0.209	0.211	0.301	0.302
STEM	0.225	0.244	0.277	0.286
Social Sciences	0.186	0.198	0.283	0.286

Table 37: MMMLU per-subject Brier ↓ under v=en, columns are alignment regimes (q=r=s; q-only; r-only; neither). Note: MMMLU subjects are not cultural categories — same (q, r, s) construction as LSQA, for completeness. Panel-mean over 16 models.

Subject	q=r=s	q=s	r=s	none
Humanities	0.253	0.255	0.341	0.335
Other	0.206	0.212	0.297	0.300
STEM	0.222	0.245	0.272	0.285
Social Sciences	0.187	0.199	0.284	0.288

U Raw Data: TriviaQA

Table 38: TriviaQA — ACC ↑ per $(q_{\text{lang}}, r_{\text{lang}}, v_{\text{lang}})$. Panel mean.

q_{lang}	r_{lang}	**EN**	FR	ZH	JA	KO	HI
EN	EN	0.656	0.656	0.656	0.656	0.656	0.656
	FR	0.598	0.598	0.598	0.598	0.598	0.598
	ZH	0.545	0.545	0.545	0.545	0.545	0.545
	JA	0.520	0.520	0.520	0.520	0.520	0.520
	KO	0.497	0.497	0.497	0.497	0.497	0.497
	HI	0.458	0.458	0.458	0.458	0.458	0.458
FR	EN	0.529	0.529	0.529	0.529	0.529	0.529
	FR	0.571	0.571	0.571	0.571	0.571	0.571
	ZH	0.520	0.520	0.520	0.520	0.520	0.520
	JA	0.492	0.492	0.492	0.492	0.492	0.492
	KO	0.465	0.465	0.465	0.465	0.465	0.465
	HI	0.417	0.417	0.417	0.417	0.417	0.417
ZH	EN	0.451	0.451	0.451	0.451	0.451	0.451
	FR	0.453	0.453	0.453	0.453	0.453	0.453
	ZH	0.455	0.455	0.455	0.455	0.455	0.455
	JA	0.406	0.406	0.406	0.406	0.406	0.406
	KO	0.397	0.397	0.397	0.397	0.397	0.397
	HI	0.362	0.362	0.362	0.362	0.362	0.362
JA	EN	0.386	0.386	0.386	0.386	0.386	0.386
	FR	0.401	0.401	0.401	0.401	0.401	0.401
	ZH	0.422	0.422	0.422	0.422	0.422	0.422
	JA	0.417	0.417	0.417	0.417	0.417	0.417
	KO	0.381	0.381	0.381	0.381	0.381	0.381
	HI	0.345	0.345	0.345	0.345	0.345	0.345
KO	EN	0.377	0.377	0.377	0.377	0.377	0.377
	FR	0.379	0.379	0.379	0.379	0.379	0.379
	ZH	0.394	0.394	0.394	0.394	0.394	0.394
	JA	0.384	0.384	0.384	0.384	0.384	0.384
	KO	0.367	0.367	0.367	0.367	0.367	0.367
	HI	0.328	0.328	0.328	0.328	0.328	0.328
HI	EN	0.384	0.384	0.384	0.384	0.384	0.384
	FR	0.396	0.396	0.396	0.396	0.396	0.396
	ZH	0.426	0.426	0.426	0.426	0.426	0.426
	JA	0.405	0.405	0.405	0.405	0.405	0.405
	KO	0.386	0.386	0.386	0.386	0.386	0.386
	HI	0.381	0.381	0.381	0.381	0.381	0.381

Table 39: TriviaQA — AUROC ↑ per $(q_{\text{lang}}, r_{\text{lang}}, v_{\text{lang}})$. Panel mean.

q_{lang}	r_{lang}	**EN**	FR	ZH	JA	KO	HI
EN	EN	0.773	0.768	0.766	0.760	0.768	0.762
	FR	0.791	0.803	0.789	0.791	0.786	0.783
	ZH	0.825	0.822	0.821	0.822	0.819	0.804
	JA	0.825	0.826	0.826	0.819	0.822	0.812
	KO	0.830	0.829	0.827	0.829	0.827	0.819
	HI	0.833	0.830	0.823	0.827	0.829	0.813
FR	EN	0.754	0.748	0.747	0.747	0.745	0.748
	FR	0.796	0.793	0.787	0.786	0.787	0.785
	ZH	0.803	0.807	0.810	0.811	0.807	0.788
	JA	0.806	0.810	0.819	0.815	0.808	0.799
	KO	0.810	0.816	0.819	0.822	0.823	0.813
	HI	0.820	0.829	0.827	0.829	0.825	0.819
ZH	EN	0.777	0.772	0.765	0.764	0.766	0.762
	FR	0.780	0.796	0.777	0.778	0.768	0.773
	ZH	0.807	0.805	0.797	0.795	0.796	0.791
	JA	0.816	0.821	0.813	0.808	0.811	0.804
	KO	0.822	0.821	0.817	0.820	0.821	0.811
	HI	0.815	0.821	0.819	0.820	0.817	0.811
JA	EN	0.758	0.756	0.742	0.747	0.746	0.749
	FR	0.751	0.776	0.750	0.754	0.740	0.753
	ZH	0.796	0.799	0.794	0.792	0.786	0.788
	JA	0.799	0.799	0.786	0.796	0.782	0.792
	KO	0.818	0.821	0.817	0.821	0.816	0.818
	HI	0.808	0.811	0.807	0.809	0.804	0.802
KO	EN	0.773	0.769	0.763	0.761	0.766	0.762
	FR	0.764	0.791	0.768	0.768	0.768	0.763
	ZH	0.796	0.798	0.804	0.800	0.797	0.781
	JA	0.807	0.809	0.811	0.808	0.809	0.792
	KO	0.812	0.807	0.806	0.801	0.807	0.801
	HI	0.810	0.812	0.814	0.811	0.810	0.804
HI	EN	0.760	0.756	0.754	0.755	0.755	0.744
	FR	0.753	0.781	0.755	0.757	0.752	0.750
	ZH	0.786	0.794	0.809	0.807	0.798	0.779
	JA	0.800	0.809	0.823	0.819	0.813	0.801
	KO	0.796	0.809	0.808	0.815	0.820	0.807
	HI	0.815	0.815	0.811	0.810	0.809	0.806

Table 40: TriviaQA — ECE ↓ per $(q_{\text{lang}}, r_{\text{lang}}, v_{\text{lang}})$. Panel mean.

q_{lang}	r_{lang}	**EN**	FR	ZH	JA	KO	HI
EN	EN	0.204	0.206	0.211	0.217	0.228	0.213
	FR	0.218	0.200	0.215	0.221	0.233	0.220
	ZH	0.205	0.200	0.224	0.216	0.231	0.225
	JA	0.207	0.193	0.211	0.212	0.209	0.204
	KO	0.210	0.199	0.210	0.211	0.225	0.215
	HI	0.198	0.191	0.207	0.210	0.217	0.224
FR	EN	0.274	0.262	0.286	0.277	0.287	0.266
	FR	0.216	0.205	0.226	0.224	0.234	0.221
	ZH	0.221	0.210	0.238	0.229	0.227	0.224
	JA	0.220	0.201	0.220	0.227	0.210	0.215
	KO	0.227	0.208	0.226	0.222	0.233	0.215
	HI	0.221	0.191	0.209	0.206	0.200	0.229
ZH	EN	0.307	0.287	0.313	0.298	0.303	0.291
	FR	0.270	0.254	0.278	0.260	0.267	0.260
	ZH	0.279	0.259	0.282	0.266	0.276	0.270
	JA	0.265	0.244	0.277	0.267	0.249	0.252
	KO	0.246	0.227	0.260	0.244	0.257	0.243
	HI	0.235	0.219	0.250	0.238	0.236	0.263
JA	EN	0.343	0.321	0.365	0.331	0.332	0.317
	FR	0.298	0.282	0.319	0.287	0.295	0.281
	ZH	0.279	0.251	0.301	0.260	0.265	0.258
	JA	0.278	0.253	0.305	0.258	0.273	0.254
	KO	0.233	0.215	0.270	0.231	0.249	0.227
	HI	0.229	0.211	0.258	0.227	0.228	0.253
KO	EN	0.342	0.313	0.357	0.336	0.318	0.313
	FR	0.297	0.277	0.318	0.292	0.277	0.274
	ZH	0.282	0.255	0.311	0.280	0.261	0.262
	JA	0.262	0.233	0.290	0.266	0.240	0.246
	KO	0.281	0.253	0.302	0.273	0.254	0.258
	HI	0.236	0.212	0.262	0.245	0.222	0.254
HI	EN	0.340	0.329	0.362	0.342	0.342	0.334
	FR	0.289	0.286	0.306	0.296	0.300	0.296
	ZH	0.260	0.235	0.284	0.259	0.251	0.254
	JA	0.223	0.206	0.251	0.242	0.222	0.226
	KO	0.245	0.220	0.259	0.243	0.253	0.240
	HI	0.246	0.237	0.271	0.251	0.254	0.245

Table 41: TriviaQA — Brier ↓ per $(q_{\text{lang}}, r_{\text{lang}}, v_{\text{lang}})$. Panel mean.

q_{lang}	r_{lang}	**EN**	FR	ZH	JA	KO	HI
EN	EN	0.211	0.213	0.219	0.222	0.228	0.218
	FR	0.229	0.209	0.228	0.227	0.237	0.228
	ZH	0.215	0.211	0.229	0.223	0.236	0.238
	JA	0.215	0.204	0.217	0.219	0.215	0.217
	KO	0.218	0.210	0.219	0.218	0.226	0.226
	HI	0.206	0.201	0.216	0.214	0.220	0.228
FR	EN	0.276	0.269	0.286	0.282	0.288	0.273
	FR	0.225	0.217	0.232	0.231	0.237	0.231
	ZH	0.232	0.222	0.241	0.235	0.232	0.237
	JA	0.231	0.214	0.228	0.234	0.219	0.226
	KO	0.235	0.218	0.233	0.230	0.235	0.227
	HI	0.227	0.201	0.221	0.214	0.208	0.237
ZH	EN	0.297	0.285	0.304	0.296	0.302	0.292
	FR	0.270	0.258	0.277	0.267	0.273	0.270
	ZH	0.274	0.261	0.278	0.269	0.278	0.275
	JA	0.260	0.246	0.270	0.266	0.253	0.261
	KO	0.244	0.232	0.255	0.246	0.258	0.254
	HI	0.235	0.224	0.246	0.241	0.239	0.266
JA	EN	0.328	0.312	0.351	0.321	0.329	0.312
	FR	0.294	0.279	0.313	0.287	0.297	0.283
	ZH	0.273	0.253	0.293	0.260	0.269	0.265
	JA	0.271	0.255	0.297	0.258	0.278	0.263
	KO	0.232	0.219	0.262	0.231	0.253	0.235
	HI	0.226	0.215	0.254	0.228	0.231	0.255
KO	EN	0.322	0.303	0.339	0.324	0.312	0.306
	FR	0.288	0.273	0.306	0.289	0.278	0.274
	ZH	0.271	0.255	0.298	0.274	0.261	0.264
	JA	0.256	0.237	0.281	0.263	0.243	0.252
	KO	0.268	0.249	0.289	0.266	0.254	0.259
	HI	0.230	0.215	0.254	0.243	0.223	0.255
HI	EN	0.329	0.320	0.347	0.335	0.334	0.327
	FR	0.290	0.284	0.304	0.298	0.299	0.294
	ZH	0.264	0.243	0.277	0.263	0.254	0.260
	JA	0.234	0.220	0.251	0.248	0.229	0.237
	KO	0.249	0.229	0.256	0.247	0.253	0.244
	HI	0.246	0.239	0.266	0.254	0.255	0.247

V Artifact Licenses and AI Assistance

Artifact licenses. All datasets and models used in this study are publicly released for research use, and we use them in accordance with their respective licenses. The benchmarks include LSQA (MIT-licensed companion to Xue et al., 2025), MMMLU [OpenAI, 2024] (MIT, derived from MMLU [Hendrycks et al., 2021] also MIT), and TriviaQA [Joshi et al., 2017] (Apache 2.0). The evaluated models include the Qwen family [Qwen et al., 2025, Yang et al., 2025] (Apache 2.0 for the open-weight variants), the Llama 3 family [Dubey et al., 2024] (Llama 3 Community License), Gemma 2 and 3 [Riviere et al., 2024, Kamath et al., 2025] (Gemma Terms of Use), Mistral-7B-Instruct [Jiang et al., 2023] (Apache 2.0), Aya-Expanse-8B [Üstün et al., 2024] (CC-BY-NC 4.0, used here for non-commercial research), and the GPT-4o family (accessed via the OpenAI API under its standard research terms). Our LSQA Korean and Hindi extensions, produced by prompting Claude Opus 4.7 [Anthropic, 2026] with a content-preservation template, are released under the same research-use terms as the original LSQA.

AI assistance. AI assistants (a Claude chatbot) were used for editing English prose for fluency and for debugging plotting code. All experimental design, factorial construction, statistical analyses, claims, and figures are the authors' own; AI assistants did not generate scientific content or run experiments.