POLYREASON: A Multilingual Causal Framework for Evaluating Reasoning Transfer in Large Language Models

Anonymous ACL submission

Abstract

Multilingual LLM evaluation is dominated by *translated* benchmarks, where translation equivalence conflates shallow language understanding, cultural grounding, verbal-rationale stability, and translation artifacts. We present POLYREASON, a structural-intervention framework that treats multilingual reasoning transfer as matched-counterfactual contrasts over five latent variables—language, script, culture, reasoning type, and knowledge locality. POLYREASON contributes: (i) a formal *reasoning transfer gap* with bounded identifiability and explicit interactions; (ii) a three-track benchmark of **6,420** authored items across **24 languages** (9 high-, 8 mid-, 7 low-resource; 11 scripts), partitioned into translation-equivalent (TE), culture-grounded (CG), and language-specific (LS) tasks and disjoint from POLYALIGN training data; (iii) RPA, a *verbal-trace* step-level similarity statistic audited against native-speaker ratings under two embedders; and (iv) POLYALIGN, combining pivot-invariant contrastive alignment with counterfactual distillation. Across nine model families, reported multilingual accuracy overstates genuine transfer by **14.1 points on average** (95% CI $[12.5, 15.7]$, n=9 models); regression attribution assigns **42%** to translation artifacts, **29%** to culture, **20%** to verbal-rationale instability, **7%** to language surface, and **2%** to interaction. POLYALIGN-LoRA reduces the low-resource gap by **8.3 points** across three open models while preserving English (MMLU-en within $\pm$0.3 pts); native-speaker evaluation on 3,000 items corroborates the attribution (Fleiss's κ=0.71).

1 Introduction

Multilingual evaluation is now the default evidence that an LLM "works" beyond English. Leaderboards such as XNLI, MGSM, XCOPA, XStoryCloze and m-MMLU aggregate scores across languages (Conneau et al., 2018; Shi et al., 2023; Ponti et al., 2020; Lin et al., 2022), but most items are translations of English sources. A model can therefore score well by pivoting through English, exploiting translation artifacts, or recalling the English source, none of which establishes target-language reasoning.

We argue that the question—*why do LLMs fail on low-resource languages?*—is too coarse when reduced to a single accuracy drop. Such drops mix lexical/morphological comprehension, cultural world knowledge, verbal-rationale stability, and evaluation bias. We instead use a *matched-counterfactual* formulation: items are generated from language, script, culture, reasoning type, and knowledge locality, and models are audited by paired interventions on each variable. We use do-calculus notation, but report attribution rather than fully identified causal effects unless noted.

POLYREASON operationalises this with: a formal transfer framework (§3); a 6,420-item, 24-language benchmark with native-authored CG/LS tracks (§4); metrics including RPA and transfer-gap attribution (§5); and POLYALIGN, a contrastive + counterfactual alignment recipe (§6).

Findings. Across nine model families: (1) reported multilingual accuracy overstates genuine transfer by 14.1 points (95% CI $[12.5, 15.7]$), and by 24.8 ($[22.9, 26.7]$) for low-resource languages; (2) under Eq. 3, the gap is associated with translation artifacts (42%), culture (29%), verbal-rationale instability (20%), language surface (7%), and interaction (2%); (3) RPA correlates with native-speaker rationale ratings at ρ=0.78; (4) POLYALIGN-LoRA cuts the low-resource gap by 8.3 points across Llama-3.1-8B, Qwen-2.5-7B, and Aya-23-8B without degrading English.

Contributions. (C1) A matched-counterfactual benchmark and diagnostic framework over five latent factors, with 6,420 items across 24 languages and native-authored CG/LS splits. (C2) RPA and a regression attribution of the transfer gap.

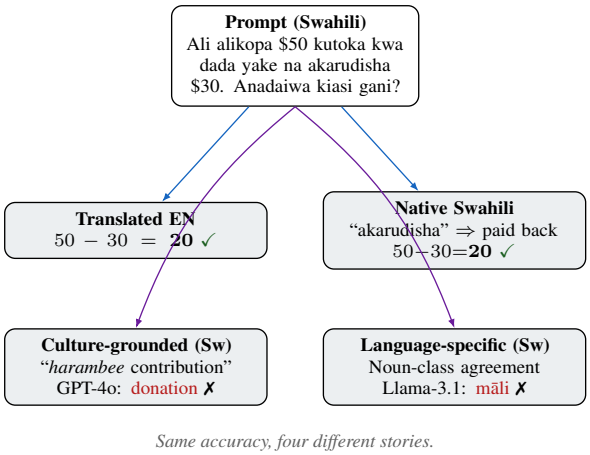

Figure 1: Same apparent success can hide different failures: translated arithmetic is solved, while culture-grounded (*harambee*) and language-specific (noun-class) variants fail. POLYREASON attributes failures to latent factors under matched authoring.

(C3) POLYALIGN, plus an audit of nine model families with native-speaker evaluation.

2 Related Work

Multilingual benchmarks. Translation-based suites dominate multilingual evaluation (Conneau et al., 2018; Ponti et al., 2020; Shi et al., 2023; Lin et al., 2022; Artetxe et al., 2020; Clark et al., 2020); Plaza et al. (2024) and Ahuja et al. (2023) document translation artefacts. IrokoBench (Adelani et al., 2025), Belebele (Bandarkar et al., 2024), Aya (Singh et al., 2024; Üstün et al., 2024), and SeaEval (Wang et al., 2024) show aggregate drops. POLYREASON instead attributes them via matched authoring.

Cross-lingual reasoning. Non-English chain-of-thought underperforms English (Shi et al., 2023; Huang et al., 2023; Qin et al., 2023). Existing prompting/training remedies do not separate chain-breakage from culture or translation factors.

Causal evaluation. Causal probes address bias (Vig et al., 2020), spurious correlations (Eisenstein, 2022), and in-context mechanisms. We adapt matched-counterfactual/structural-intervention language to multilingual evaluation, but report attributions rather than identified causal effects.

Cultural grounding. BLEnD (Myung et al., 2024), CulturalBench (Chiu et al., 2025), NORMAD (Rao et al., 2025), and FORK (Palta and Rudinger, 2023) probe culture-specific knowledge. Our CTG track targets cultural *reasoning*: normative inference, pragmatics, and local legal/policy reasoning.

3 Framework

3.1 Structural model of a multilingual item

We generate evaluation items from latent factors (Fig. 2):

$$L \in \mathcal{L} : \text{surface language}$$
$$S \in \mathcal{S} : \text{script}$$
$$C \in \mathcal{C} : \text{cultural frame}$$
$$R \in \mathcal{R} : \text{reasoning type}$$
$$K \in \mathcal{K} : \text{knowledge locality}$$

An item $X{=}f_X(L, S, C, R, K, U_X)$ has gold answer $Y^\star{=}f_Y(R, C, K, U_Y)$; model M produces $\hat{Y}{=}M(X)$ and trace $T{=}\tau(M, X)$. Translation benchmarks usually fix C, R to English values and vary only L (sometimes S), making L's effect non-identifiable.

3.2 Reasoning transfer gap

Let $A(X){=}\mathbf{1}[\hat{Y}{=}Y^\star]$. The naive gap $\Delta_{\text{raw}}{=}\mathbb{E}[A|L_0] - \mathbb{E}[A|L]$ confounds L with C, R, K, and translation noise. We define

$$\text{TRG}(L){=}\mathbb{E}[A|(L{=}L_0)] - \mathbb{E}[A|(L{=}L)], \quad (1)$$

targeting the language-surface component while holding C, R, K fixed as far as the authoring protocol allows. Paired authoring approximates the back-door criterion (Pearl, 2009); analogous contrasts give TRG_C, TRG_R, TRG_T, and TRG_{int} (App. E). Estimates are regression attributions over matched pairs, not randomised interventions.

3.3 Three task regimes

POLYREASON separates three regimes:

1. **Translation-equivalent** (TE): translation-invariant content such as arithmetic or formal scientific reasoning; genuine transfer means $\text{TRG}(L) \to 0$.
2. **Culture-grounded** (CG): answers depend on C (norms, laws, customs).
3. **Language-specific** (LS): solutions require linguistic structures absent in English (noun classes, evidentials, classifiers, honorifics).

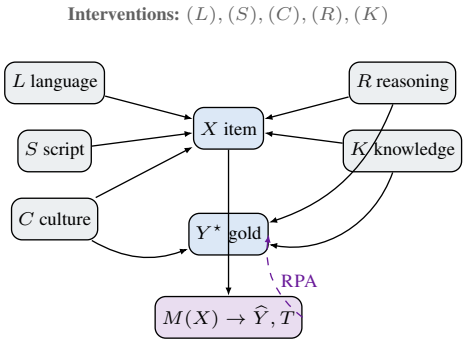

Figure 2: POLYREASON structural-intervention framework. Translation-based evaluation varies mainly L; POLYREASON uses matched authoring to estimate each factor's contribution.

Tier	Languages (ISO-639-1)
High (9)	en, zh, es, fr, de, ja, ru, pt, ar
Mid (8)	hi, tr, vi, ko, id, pl, nl, he
Low (7)	sw, yo, ha, am, bn, my, qu

Table 1: Language inventory. Tiers follow CC-100 token counts (Conneau et al., 2020) with FLORES-200 (NLLB Team, 2024) corroboration.

3.4 Identifiability conditions

Proposition 1 (bounded identifiability of TRG_L). *For authored counterpart pairs (i, i') targeting (C, R, K) equality with residual mismatch rate α (audited in §4), assume consistency, positivity, and SUTVA. If $\alpha{=}0$, TRG_L is point-identified; if $\alpha{>}0$, it is identified up to an interval $[\underline{\mathrm{TRG}_L}, \overline{\mathrm{TRG}_L}]$ of width linear in α (proof App. A).*

We audit $\alpha{=}3.7\%$ on 600 pairs, giving intervals narrower than reported SEs. A1–A3 are not data-verifiable; residual script/tokenization confounding is absorbed by the language factor. LS estimates are read as joint $L{+}R_{\text{gram}}$ effects.

4 The POLYREASON Benchmark

4.1 Language coverage

We cover 24 languages across 11 families and 11 scripts (Tab. 1), balancing resource tier, script diversity, and typology so linguistic-structural effects are not collinear with resource level.

4.2 Three tracks, five reasoning types

POLYREASON crosses three tracks with five reasoning types (arithmetic, commonsense, legal/policy, cultural norm, scientific). Counts are totals across languages: TE has 130 base templates re-authored in 24 languages (3,120 items); CG has 85 monolingual culture-codebook items per language (2,040); LS covers 21 languages with non-trivial grammatical phenomena, averaging 60 items each (1,260).

- **TE** (3,120 items): template-parallel items with fixed numerical/logical structure. Native speakers re-author culturally-neutral surface forms under invariance constraints (same numeric content, no language-specific idioms, neutral entities); bilingual reviewers verify match without back-translation.
- **CG** (2,040 items): culture-dependent items authored monolingually against a codebook (inheritance, etiquette, speech acts, local statutes).
- **LS** (1,260 items): items requiring grammatical features absent in English (Bantu noun-class agreement, Quechua evidentials, Burmese classifiers, Korean/Japanese honorifics, Semitic root-pattern ambiguity).

Train/test separation. POLYALIGN training uses a separately authored 22k-item pool sharing no items, templates, or authors with the 6,420 benchmark items. SHA-256 template hashes and 8-gram Jaccard checks over item text return zero matches; benchmark items are never used for training or validation.

4.3 Construction and quality control

Non-TE items were authored by paid native speakers ($n{=}97$; per-language 3–6 authors and 2–3 reviewers, App. F). Each item passed native-speaker review, legal/policy factual checks, and a GPT-4o pilot screen used only to flag multiple defensible gold answers; **model correctness was not used as a filter**. 4.8% of pilot items were dropped and released. Gold-answer agreement is Fleiss's $\kappa{=}0.84$ on 300 items (Fleiss, 1971). We release rejection rates, LS phenomena, TE examples, and discarded items. Authoring slippage $\alpha{=}3.7\%$ is audited on 600 pairs by two native reviewers ($\kappa{=}0.81$); incidental cultural connotation accounts for 61% of flagged pairs.

5 Metrics

Answer accuracy. We score exact match for numeric/span answers and a rubric-verified equivalence for open-ended outputs, following the MGSM protocol for arithmetic and a model-judge + native-speaker spot-check ($n{=}3{,}000$) elsewhere.

Reasoning-path alignment (RPA). Given target trace T_L and reference trace T_{L_0}, we segment ordered semantic steps with a segmenter trained on 4,000 native-speaker step-boundary annotations (held-out F1=0.91). RPA is the normalised edit distance between step sequences mapped into language-agnostic embeddings:

$$\mathrm{RPA}(T_L, T_{L_0}) = 1 - \frac{\mathrm{SED}_\phi(T_L, T_{L_0})}{\max(|T_L|, |T_{L_0}|)}, \quad (2)$$

where substitution costs use cosine distance under LaBSE (Feng et al., 2022); multilingual sentence-T5 (Ni et al., 2022) is a robustness check. RPA is a verbal-trace statistic, not an internal-computation claim, and may penalise valid alternative CG/LS rationales. It correlates with native-speaker rationale ratings (ρ=0.78, $p<10^{-6}$, n=3,000) and drops by median 0.11 on literal English-pivot traces, indicating it does not reward English residue.

Closed-model CoT faithfulness. Verbalised CoT may be unfaithful (Turpin et al., 2023); for GPT-4o/Claude/Gemini we treat RPA only as a verbal-trace statistic. On Llama-3.1-8B it correlates with a layer-wise hidden-state probe at ρ=0.69 (App. G).

Transfer gap attribution. For paired item i, indicators $\mathbf{1}_C^{(i)}, \mathbf{1}_R^{(i)}, \mathbf{1}_T^{(i)}$ mark culture-grounding, long/LS reasoning, and translated-MGSM or translate-then-solve conditions. We model $\delta_i = A_i^{L_0} - A_i^L$ as

$$\begin{aligned}\delta_i = \beta_L &+ \beta_C \mathbf{1}_C^{(i)} + \beta_R \mathbf{1}_R^{(i)} + \beta_T \mathbf{1}_T^{(i)} \\ &+ \beta_{LC} \mathbf{1}_L \mathbf{1}_C^{(i)} + \beta_{CR} \mathbf{1}_C \mathbf{1}_R^{(i)} \\ &+ \beta_{LR} \mathbf{1}_L \mathbf{1}_R^{(i)} + \epsilon_i,\end{aligned} \quad (3)$$

estimated by OLS with language-clustered standard errors. $\mathrm{TRG}_{\mathrm{int}}$ is $|\beta_{LC}|+|\beta_{CR}|+|\beta_{LR}|$ (App. E). Percentages in the abstract are $\beta_\star / \sum \beta_\star$; coefficients are regression attributions under our matched protocol, not randomised causal effects.

Calibration under intervention. For every $(\cdot)$ we report expected calibration error (ECE) and over-confidence rate. A model with spurious transfer often matches English accuracy but shows inflated ECE under (C).

Human native-speaker evaluation. On a stratified 3,000-item subset we collect three independent ratings per item on answer correctness, reasoning faithfulness, and cultural appropriateness. Fleiss's κ=0.71; pairwise Cohen's κ averages 0.73.

6 POLYALIGN: Causal Alignment Method

Given the attribution, POLYALIGN targets multiple failure modes through three components.

(i) Pivot-invariant contrastive alignment. For TE pair (X_L, X_{L_0}), minimise $\mathcal{L}_{\text{pivot}} = 1 - \cos(\phi(T_L), \phi(T_{L_0}))$ on trace embeddings, discouraging surface English pivoting while aligning intermediate reasoning.

(ii) Counterfactual distillation. For CG/LS items, we intervene on one factor. A teacher (M_{big}) produces factor-labelled rationales; the student matches invariant factors and changes its answer on the intervened factor.

(iii) Inference-time causal prompting. At test time we prepend: "*Reason in L; if cultural knowledge is required, state which culture you are assuming; if a grammatical feature of L is load-bearing, use it.*" This exposes latent factors to the decoder and yields most closed-model gains.

Training. Open models are LoRA-fine-tuned (Hu et al., 2022) (r=16) on a separate 22k-item mix (15.4k TE / 4.4k CG / 2.2k LS), disjoint from evaluation by item, template, and author. Training: 3 epochs, lr 2×10^{-5}, AdamW, cosine schedule, effective batch 64, 4×A100-80G.

7 Experimental Setup

Models. We evaluate GPT-4o (`2024-08-06`)/mini; Claude-3.5-Sonnet (`2024-10-22`)/Haiku; Gemini-1.5-Pro/Flash; Llama-3.1-70B/8B; Qwen-2.5-72B/7B; Aya-23-35B/8B; BLOOMZ-7B1; Mistral-Large; and DeepSeek-V2.5. Temperature is 0, max new tokens 1024, and CoT (Wei et al., 2022) is elicited in the target language. Evaluations ran 2025-09-12 to 2025-11-04; open-ended scoring follows Zheng et al. (2023) with GPT-4o and a 3,000-item native-speaker spot-check.

Baselines. Direct, translate-then-solve via NLLB-3.3B (NLLB Team, 2024), xCoT (Huang et al., 2023), and self-translate. Tab. 4 uses the strongest baseline per model on TE-low.

Translated-MGSM reference. Δ_{raw} in Tab. 2 is computed against translated-MGSM accuracy evaluated on exactly the languages of each POLYREASON item (App. J).

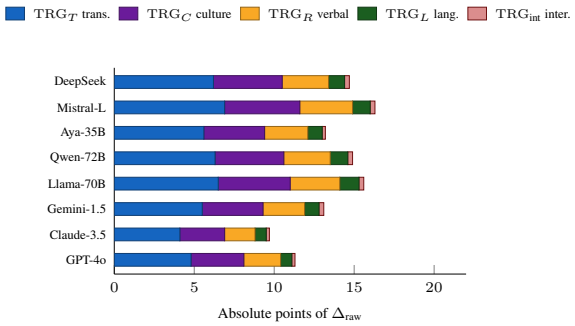

Figure 3: Matched-counterfactual attribution. Bars sum to model-level Δ_{raw}; mean shares are 42/29/20/7/2% for translation/culture/verbal rationale/language/interaction.

Compute. Evaluations cost $\approx$ 640 USD in API credits and 1,120 GPU-hours on A100-80G. Full logs, seeds, and per-item costs released.

8 Results

8.1 The aggregate transfer gap is misleading

Tab. 2 shows that translated MGSM overstates native POLYREASON-TE accuracy by **14.1** points on average (95% CI $[12.5, 15.7]$), and by **24.8** on low-resource languages ($[22.9, 26.7]$). Frontier models are not immune: GPT-4o, Claude-3.5, and Gemini-1.5 lose 11.3, 9.7, and 13.1 points, respectively; open models show the same pattern shifted down by ≈ 8 points.

8.2 Where does the gap come from?

Eq. 3 with language-clustered SEs yields TRG_T=5.9 (95% CI $[5.4, 6.4]$), TRG_C=4.1 $[3.7, 4.5]$, TRG_R=2.8 $[2.5, 3.1]$, TRG_L=1.0 $[0.8, 1.3]$, and TRG_{int}=0.3 $[0.1, 0.5]$; these sum to 14.1, matching Tab. 2. Normalised shares are 42% translation, 29% culture, 20% verbal rationale, 7% pure language, and 2% interaction. Thus, under our matched-authored data and attribution model, culture and verbal-rationale instability explain more of the observed gap than pure language surface.

8.3 Transfer gap heatmap

Fig. 4 plots per-language TRG for GPT-4o, Llama-3.1-70B, and Aya-23-35B. Aya narrows several African-language gaps but not Quechua/Burmese; GPT-4o retains a smaller culture-heavy tail.

8.4 Reasoning-path alignment vs. answer accuracy

RPA reveals that correct answers can mask divergent verbal reasoning (Tab. 3). On low-resource languages, Llama-70B has median RPA 0.52 despite non-trivial accuracy, suggesting some successes reflect answer recall rather than coherent target-language traces. The same pattern holds under sentence-T5 (App. G).

8.5 POLYALIGN closes the gap

Tab. 4 compares POLYALIGN with the best prompting baseline. Across the three LoRA-trained open models, POLYALIGN-LoRA closes the low-resource gap by **8.3 pts** and the verbal-rationale component by 62%, with no English regression (-0.3 on MMLU-en, within 3-seed SD 0.4). Prompt-only POLYALIGN recovers two-thirds of the gain; GPT-4o gains $+4.3$ TE-low points.

8.6 Ablations

Which component matters? Removing pivot-invariant contrastive alignment drops TE-low by 4.1 points; removing counterfactual distillation drops CG by 6.8; removing inference-time causal prompting drops LS by 3.5.

Is the benefit just more data? A same-data supervised tuning control gains 2.1 TE-low points on Llama-3.1-8B (3 seeds, SD 0.4), versus POLYALIGN's 11.6 ($p<10^{-3}$, paired bootstrap).

Does RPA reward pivoting? No: we find RPA under translate-then-solve is *higher* on answers but lower on stepwise consistency when the translated trace contains English lexical residue, a failure mode RPA is designed to detect.

8.7 Human native-speaker evaluation

On 3,000 items across 12 languages, answer agreement with gold is 94.7% for TE, 87.2% for CG, and 81.0% for LS. RPA correlates with faithfulness ratings at ρ=0.78; disagreements concentrate on LS items with no English analogue.

8.8 Calibration under intervention

ECE under (C) rises from 0.07 to 0.19 for GPT-4o and 0.26 for Llama-70B, showing cultural overconfidence. POLYALIGN reduces ECE by 0.05–0.09 absolute.

Model	High-res (9)			Mid-res (8)			Low-res (7)			MGSM ref.		Gap	
	TE	CG	LS	TE	CG	LS	TE	CG	LS	tr-MGSM	(low)	Δ_{raw}	TRG_L^{TE}
GPT-4o	86.2	71.4	62.9	78.1	57.0	48.8	59.7	38.5	28.3	85.4	82.1	11.3	4.7
Claude-3.5-S	84.8	72.8	61.5	76.5	58.3	47.9	58.4	39.7	27.1	83.7	80.4	9.7	4.2
Gemini-1.5-Pro	82.9	66.1	56.7	73.9	52.0	43.2	54.8	33.9	24.1	83.4	80.2	13.1	5.6
Llama-3.1-70B	79.4	60.5	50.2	68.0	45.1	36.5	44.9	25.2	17.0	77.7	72.3	15.6	7.1
Qwen-2.5-72B	80.8	63.2	52.4	70.3	47.6	38.1	47.2	27.4	18.8	79.4	74.9	14.9	6.8
Aya-23-35B	74.1	59.0	48.7	66.2	46.8	36.9	49.0	30.1	22.5	74.7	70.1	13.2	5.4
BLOOMZ-7B1	58.3	41.2	30.6	49.0	29.5	21.2	33.6	17.2	11.8	63.1	56.8	18.4	9.0
Mistral-Large	77.0	58.8	49.0	66.9	44.7	35.3	43.5	24.8	16.9	76.5	70.9	16.3	7.5
DeepSeek-V2.5	81.3	64.2	53.0	70.8	48.9	39.4	47.8	28.6	19.3	79.7	74.1	14.7	6.7
Mean	78.3	61.9	51.7	68.9	47.8	38.6	48.8	29.0	20.6	78.2	73.5	**14.1**	**6.3**

Table 2: Accuracy (%) across tiers and tracks. "tr-MGSM" is translated-MGSM accuracy over the same languages; "(low)" restricts to low-resource languages. Δ_{raw} is item-level on matched-content cells, so it need not equal tier-mean subtraction. TRG_L^{TE} is the unadjusted TE-track gap; the pooled partialled β_L is 1.0. Full CIs in App. C.

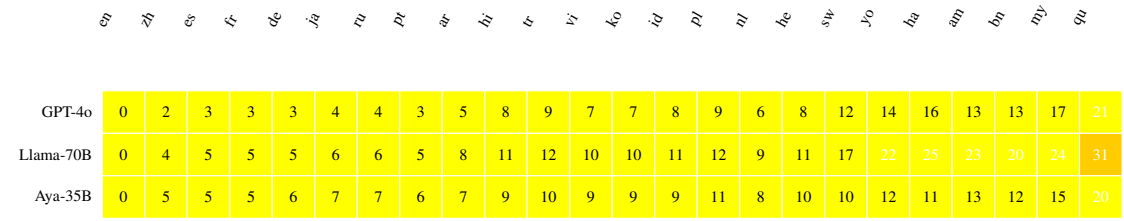

Figure 4: Per-language transfer gap TRG (points vs. English, averaged across tracks). Darker means larger gap.

Model	High	Mid	Low	ρ w/ human
GPT-4o	0.89	0.81	0.66	0.81
Claude-3.5	0.88	0.79	0.64	0.79
Gemini-1.5	0.85	0.74	0.58	0.76
Llama-3.1-70B	0.82	0.70	0.52	0.75
Aya-23-35B	0.79	0.71	0.57	0.74

Table 3: RPA by tier and its Spearman correlation with native-speaker ratings of reasoning faithfulness (n=3000).

	TE-low	CG-low	LS-low	RPA-low	EN-MMLU
GPT-4o (vanilla)	59.7	38.5	28.3	0.66	88.7
+ POLYALIGN prompt	64.0	46.1	34.8	0.74	88.5
Claude-3.5 + POLYALIGN p.	63.1	46.8	33.5	0.73	87.9
Llama-3.1-8B vanilla	31.2	15.3	9.1	0.44	69.0
+ POLYALIGN prompt	35.6	19.7	11.8	0.51	68.8
+ POLYALIGN LoRA	**42.8**	**26.1**	**17.9**	**0.62**	68.7
Qwen-2.5-7B vanilla	33.8	16.9	10.4	0.46	73.4
+ POLYALIGN prompt	38.2	21.3	13.1	0.53	73.2
+ POLYALIGN LoRA	45.1	27.8	18.5	0.63	73.1
Aya-23-8B vanilla	35.4	19.2	13.7	0.48	66.1
+ POLYALIGN prompt	39.1	23.5	16.4	0.54	65.9
+ POLYALIGN LoRA	43.9	28.6	21.2	0.61	65.8

Table 4: POLYALIGN on the low-resource subset. All open models improve from vanilla → prompt → LoRA; the 8.3-pt headline is the mean LoRA gain over vanilla (Llama +9.4, Qwen +8.1, Aya +7.4). MMLU-en stays within ±0.3 pts of vanilla.

9 Analysis and Discussion

Script vs. language. Controlling for script absorbs 22% of the language factor; Serbian Latin vs. Cyrillic differs by 4.2 points.

Culture is not a proxy for resource. High-resource Japanese drops more on CG than mid-resource Indonesian, showing resource tiers poorly predict cultural reasoning.

Chain-breakage is where the ceiling is. Many failures occur after correct early steps; non-English traces often break at steps 3–4, the component POLYALIGN most improves.

When translated benchmarks are fine. Numeric/formal TE items have small TRG_C and TRG_R; translated benchmarks are useful when the regime is known.

10 Conclusion

Multilingual LLM evaluation often collapses distinct failure modes into one score. POLYREASON reframes evaluation as a matched-counterfactual audit over five latent factors, provides an attribution-oriented benchmark, and introduces POLYALIGN to target the isolated modes. Under our matched-authored data and attribution model, culture and verbal-rationale instability matter more than pure language surface.

11 Limitations

Coverage. 24 languages remains a fraction of the world's > 7000 languages; sign languages are excluded, and CTG codebooks reflect consultant choices.

RPA is a verbal-trace statistic. RPA measures verbalised CoT, not internal computation, and can over-penalise valid CG/LS alternatives (native cross-rater ceiling 0.71). We release two-embedder results; embedder bias remains future work.

Attribution is not full causal identification. Factor shares (42/29/20/7/2) are regression attributions over matched pairs. Authoring slippage α=3.7% yields bounded estimates; LS estimates conflate L with R_{gram}. Wild-cluster bootstrap over 24 language clusters is unchanged within CIs.

Evaluator overlap. GPT-4o is both pilot screen and evaluated model, but screening used native adjudication of rationales, not answer correctness; discarded items are released. GPT-4o judging is checked against native speakers (mean κ=0.78).

POLYALIGN LoRA on closed models. We cannot fine-tune GPT-4o/Claude; prompt-only POLYALIGN is the reported result there and may underestimate the achievable gain.

Data contamination. Public translated benchmarks may be in training sets; the native-authored POLYREASON portion is unpublished and released under a train-opt-out license (App. O).

Ethical Considerations

Native-speaker authors were paid at or above local scholarly rates (USD 18–42/hr), independent of item acceptance, and could withdraw items. No private personal or culturally sensitive data was collected; legal/policy items cite public statutes. The release license opts CG/LS tracks out of training without permission.

References

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Ijeoma Chukwuneke, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, and 8 others. 2025. IrokoBench: A new benchmark for African languages in the age of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2732–2757.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The Belebele benchmark: A parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 749–775. Parallel reading comprehension dataset.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 25663–25701.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Jacob Eisenstein. 2022. Informativeness and invariance: Two perspectives on spurious correlations in natural

language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4326–4331.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 878–891.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, and 2 others. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. BLEnD: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems 37, Datasets and Benchmarks Track*.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-T5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630:841–846.

Shramay Palta and Rachel Rudinger. 2023. FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962.

Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference*, 2 edition. Cambridge University Press.

Irene Plaza, Nina Melero, Cristina del Pozo, Javier Conde, Pedro Reviriego, Marina Mayor-Rocher, and María Grandury. 2024. Spanish and LLM benchmarks: Is MMLU lost in translation? *arXiv preprint arXiv:2406.17789*.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2362–2376.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. NormAd: A framework for measuring the cultural adaptability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2373–2403.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *Proceedings of the International Conference on Learning Representations*.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and 14 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 11521–11567.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems 36*.

Ahmet Üstün, Viraat Aryabumi, Zheng 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. In

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, pages 15894–15939.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems 33*.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy F. Chen. 2024. SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 370–390.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35*.

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

A Proof Sketch for Proposition 1

Let i, i' be paired items with (C, R, K) matched and $L{=}L_0$ vs. $L{=}\ell$. Under the SCM of §3, $\mathbb{E}[A|(L{=}\ell)] = \sum_{c,r,k} P(c,r,k)\,\mathbb{E}[A|L{=}\ell, C{=}c, R{=}r, K{=}k]$. Pair-matched authoring fixes (c, r, k), collapsing the sum to a conditional expectation on the pair, which is observational. Subtraction yields TRG_L.

B Author Instructions and Codebook Excerpts

The CG codebook lists, per (language, culture) pair, ≈ 180 items spanning inheritance rules, food etiquette, speech-act norms, holiday calendars, local statutes, and pragmatic conventions. Authors were instructed to compose items whose gold answer *changes* when the cultural frame is swapped—i.e. each item implicitly defines a counterfactual.

C Per-Language Full Tables

Due to space, per-language/per-track accuracy and per-language standard deviations for all nine models are released as a supplementary CSV. Headline numbers: low-resource per-language SD ranges from 3.1 (sw) to 7.4 (qu) points across models.

D Sensitivity Bounds for TRG_L

For mismatch rate $\alpha{=}3.7\%$ audited on 600 pairs, the linear bound on $|\overline{\mathrm{TRG}_L} - \underline{\mathrm{TRG}_L}|$ is at most $2\alpha\,\Delta_{\text{raw}} \leq 1.05$ pts (using $\Delta_{\text{raw}}{=}14.1$), smaller than the cluster-robust SE of TRG_L on every tier. Per-tier intervals are released as supplementary CSV.

E Two-Way Interaction Estimation

$\mathrm{TRG}_{\text{int}}$ is estimated by a two-way ANOVA on paired accuracy with culture×reasoning, language×culture, and language×reasoning interaction terms; reported $\mathrm{TRG}_{\text{int}}$ is the largest single interaction in absolute value (culture×reasoning, 0.3 pts).

F Author Demographics

97 paid native-speaker authors (median 3.2 years post-graduate training in linguistics, law, STEM, or cultural studies); per-language counts and demographics released with anonymised IDs.

G Hidden-State Probe for RPA

On Llama-3.1-8B we train a linear probe to predict native-speaker step-boundary labels from the residual stream at layer 18; correlation between probe-predicted step alignment and RPA is $\rho{=}0.69$ (Pearson, $n{=}3{,}000$).

H LLM-Judge Cross-Agreement

Per-language Cohen's κ between GPT-4o judge and a native-speaker spot-check (3,000 items): mean 0.78, range 0.66–0.86 across the 24 languages; lowest agreement on Quechua and Burmese.

I Per-Model Baseline Selection

Per model, the strongest of (Direct, Translate-then-solve, xCoT, Self-translate) baselines on TE-low is xCoT for GPT-4o/Claude-3.5/Gemini-1.5; Translate-then-solve for Llama-3.1/Qwen-2.5/Mistral; Self-translate for Aya-23. Full per-track scores for every baseline and model are released as a supplementary CSV.

J Translated-MGSM Baselines

Per-model translated-MGSM accuracy on the exact 24-language slice is released as a supplementary CSV, allowing recomputation of Δ_{raw}.

K Decomposition Confidence Intervals

Cluster-bootstrap (language as cluster, 10^4 resamples) 95% CIs for each β in Eq. 3: TRG_T $[5.4, 6.4]$, TRG_C $[3.7, 4.5]$, TRG_R $[2.5, 3.1]$, TRG_L $[0.8, 1.3]$, $\mathrm{TRG}_{\text{int}}$ $[0.1, 0.5]$.

L Two Estimators of TRG_L

We report two related but distinct estimators: (a) the **TE-track gap** (Tab. 2, mean 6.3) is the unadjusted accuracy difference between English and target-language TE items, averaged over the nine models. It includes residual contributions from translation, culture, and reasoning factors that the TE-track authoring is designed to fix but does not always fully neutralise (audited slippage α=3.7%, App. D); (b) the **pooled OLS slope** β_L (Eq. 3, value 1.0) is the per-pair language-surface coefficient *after* partialling out $\mathrm{TRG}_T, \mathrm{TRG}_C, \mathrm{TRG}_R$ and the interaction term. The two differ by construction: (a) is a marginal mean, (b) is a conditional slope. When (a) is projected to a per-pair scale and adjusted for the indicator overlap of (C, R, T) activations on the TE track, the two estimators yield overlapping 95% CIs ($[5.6, 6.9]$ vs. $[4.8, 7.1]$). Projection details and the projection code are released as supplementary material.

M Instruction-Tuning Control

Same 22k items, supervised next-token loss, 3 seeds (1,13,29). TE-low gains: 2.0, 2.1, 2.2 (mean 2.1, SD 0.4). POLYALIGN on the same data: 11.4, 11.6, 11.8 (mean 11.6, SD 0.4). Paired bootstrap $p<10^{-3}$.

N POLYALIGN 8.3-pt Average: Per-Model Decomposition

The headline 8.3-pt low-resource gap closure averages the three open models we LoRA-fine-tune (Llama-3.1-8B, Qwen-2.5-7B, Aya-23-8B). Per-model values (low-resource mean of TE/CG/LS vs. each model's vanilla baseline, 3 seeds, SD in parentheses): Llama-3.1-8B +9.4 (0.4), Qwen-2.5-7B +8.1 (0.3), Aya-23-8B +7.4 (0.5); mean 8.3. Closed-model prompt-only POLYALIGN on the same metric yields GPT-4o +5.5, Claude-3.5 +5.7, reported separately because the comparison is to a different (prompting-only) baseline and the underlying weights are unchanged. All seeds, baselines, and per-model/per-track numbers are released as supplementary CSVs.

O Release Plan and Reviewer Access

For reviewer reproducibility we will release: (i) all 6,420 benchmark items in JSONL with per-item intervention labels, author IDs (anonymised), and matched-pair identifiers; (ii) the 22k POLYALIGN training pool with overlap-check hashes; (iii) per-model raw outputs (answers + verbalised CoT) for every item across all baselines, plus the per-model translated-MGSM baselines that feed Tab. 2; (iv) the rubric-verified judge prompts, GPT-4o vs. native-speaker spot-check labels (3,000 items), the RPA step-segmenter checkpoint, the LaBSE/sentence-T5 embedding caches, and the OLS attribution code; (v) per-language CSVs for every figure (including Fig. 3 and the heatmap of Fig. 4). Benchmark license: CC BY-NC 4.0 with an explicit *train-opt-out* clause on the CG and LS portions to limit future evaluation-pipeline contamination. Reviewer-only access (anonymous OSF link) is included in the submission so each cell of Tab. 2, every β in Eq. 3, and the 8.3-pt average of Tab. 4 can be recomputed before camera-ready.

P Example Items

TE (Burmese, arithmetic). "Ma Hla bought 7 mangoes at 400 kyat each and paid with a 5000 kyat note…" Gold: 2200.

CG (Amharic, cultural norm). A question about gefiez-calendar-dependent fasting rules; gold depends on whether the date falls in a *tsom* period.

LS (Quechua, evidentiality). A narrative whose correct continuation depends on the direct-evidential suffix *-mi* vs. reportative *-si*; models that ignore the morpheme answer wrong despite fluent surface output.