DEPART: DEcomposing PARiTy across Multilingual LLMs

Manan Uppadhyay[♡], **Prashant Kodali**[♡], **Pranjal Chitale**[♡],
Reshma Ramaprasad[♡], **Himanshu Beniwal**[♡♣*], **Sunayana Sitaram**[♡]
[♡]Microsoft Research India, [♣]IIT Gandhinagar
{t-muppadhyay, sunayana.sitaram}@microsoft.com

Abstract

Multilingual Large Language Models (mLLMs) leaderboards report per-language accuracy but rarely explain why disparities emerge, leaving systemic biases unattributed and offering practitioners no actionable levers. We first establish that these gaps are systematic rather than artifacts of sampling noise via distribution-free Friedman and Kruskal–Wallis tests, then introduce a two-step Bayesian hierarchical framework that decomposes multilingual performance variance into interpretable components. First, isolating the variance attributable to language identity, we show that observable language features (script, family, typological distance) explain $R^2_{\text{ling}} = 79\%$ of this variance on understanding tasks and 92% on reasoning, with a model's internal representational similarity to English emerging as the dominant predictor across both task buckets. Second, decomposing the full (model$\times$benchmark$\times$language) cube, we find that NLU and reasoning have fundamentally divergent variance profiles: model identity dominates understanding (66.7% of variance), whereas the benchmark$\times$model interaction dominates reasoning (46.3%). Together these results recast multilingual evaluation from passive performance mapping into an explainable, diagnostic framework with concrete levers for targeting the root drivers of language disparity.

1 Introduction

Large Language Models (LLMs) are deployed worldwide, yet their performance varies sharply across the languages they serve, producing systematically worse outputs for speakers of low-resource languages (Choudhury and Deshpande, 2021; Khanuja et al., 2023). The field measures this gap using per-language tables, occasional Gini coefficients, and English-vs-X deltas, but does not yet explain it. Rather than just measuring the size of the cross-language gap, this work addresses a deeper question: how much of the gap can be predicted before running a benchmark, versus how much remains as an irreducible per-checkpoint residual? We answer this using observable language properties (script, family, typological distance, resource class) and cheap model-conditional probes (tokenizer fertility, internal representation similarity to English). The answer determines where intervention is most consequential: feature-predictable disparity points to data, tokenization, and alignment choices made before final evaluation, while a large residual would point to checkpoint-level idiosyncrasies that only retraining can address.

Existing work addresses this question partially: fairness frameworks (Choudhury and Deshpande, 2021; Khanuja et al., 2023) reduce disparity to one-dimensional summary statistics (Gini, max–min); the closest methodological precedent (Hu et al., 2025) fits an additive mixed-effects model that ranks languages by difficulty but cannot, by construction, identify model$\times$language interactions or admit linguistic features as covariates; intrinsic-similarity rankings (Li et al., 2025) and proxy-LM forecasts (Anugraha et al., 2025) contribute individual signals without a fairness-style decomposition (see Section 2). What is missing is a single hierarchical decomposition that jointly attributes observed disparity to feature-predictable and residual components, with calibrated uncertainty on both.

We evaluate seven open-weight LLMs ($\sim$3.5B–122B parameters) on 15 multilingual benchmarks covering 63 languages, partitioned into multilingual-understanding (9 benchmarks, 63 languages) and reasoning (6 benchmarks, 62 languages) task buckets. We fit a Bayesian hierarchical model that jointly estimates the contribution of the model, the benchmark, the language, their interactions, and observable language-level and model$\times$language features: resource class (Joshi

*Work done during internship at MSR India.

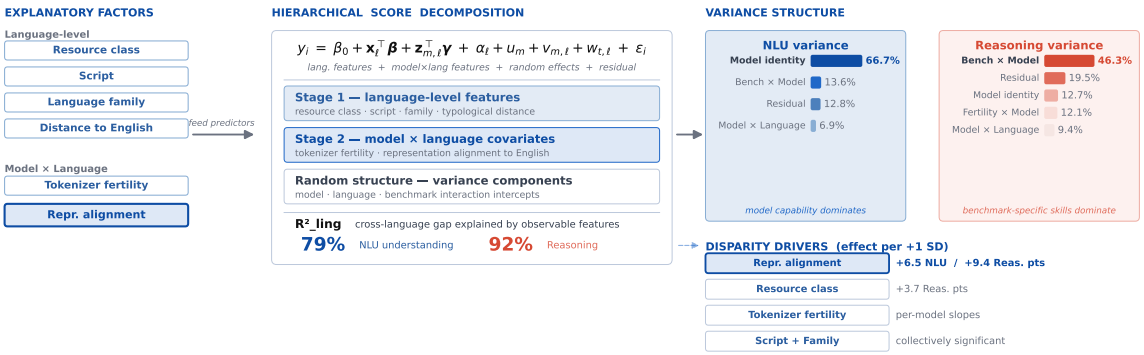

Figure 1: Structural factors dominate cross-lingual performance gaps in multilingual LLMs.

et al., 2020), lang2vec syntactic, phonological, and geographic distance to English (Littell et al., 2017), script, family, tokenizer fertility, and the alignment between a model's representation of a language and its representation of English. The analysis proceeds in two steps: distribution-free tests first verify that observed gaps are systematic, after which the hierarchical model decomposes them into feature-predictable and residual components.

The following are the key contributions of this work.

- **Non-parametric validation of widespread language disparity.** Distribution-free tests verify that observed cross-lingual gaps are systematic rather than sampling artifacts: a Friedman test rejects equal language ranks across all 18 benchmarks, and a Dunn pairwise tests rejects performance equality across resource tiers, confirming that performance drops are structured by resource level.
- **A joint hierarchical decomposition with calibrated uncertainty.** We fit a Bayesian hierarchical model that estimates the cross-language gap (i.e. ± 8–9 accuracy points) and the share explained by observable features, summarised by R^2_{ling}, the proportional reduction in σ^2_α once those features are added. Observable language features alone recover 0.79-0.92 R^2, showing that most cross-language variation is predictable in advance of evaluation. Across both task buckets, internal representation alignment to English is the single most dominant predictor (a one-SD increase is worth $+6.5$ accuracy points on understanding and $+9.4$ on reasoning), offering a benchmark-free proxy computable from model geometry alone.
- **Variance decomposition across the full design.** Model identity dominates understanding (66.7% of variance), whereas benchmark$\times$model interaction dominates reasoning (46.3%, larger than the main model effect), so a single multilingual-reasoning aggregate is misleading and per-benchmark reporting is warranted.

2 Related Works

Multilingual Evaluation Benchmarks. A growing body of multilingual evaluation benchmarks/suites for NLU tasks (XNLI, MEGA, MEGAVERSE, BUFFET) (Conneau et al., 2018; Ahuja et al., 2023, 2024; Asai et al., 2024) have established per-language reporting as the default, with recent benchmarks like MMLU-ProX (Xuan et al., 2025) and Global MMLU (Singh et al., 2025) pushing the language coverage across knowledge / reasoning-intensive tasks, reading comprehension (Bandarkar et al., 2024), and pan-cultural settings like Include (Romanou et al., 2025). These have also been complemented by region-specific suites like MILU (Verma et al., 2025), Iroko Bench (Adelani et al., 2025), IndoMMLU (Koto et al., 2023), SEA-HELM (Susanto et al., 2025) for Indic, African, and South East Asian languages, respectively. While several of these works supplement per-language tables with *bivariate* correlations against individual covariates like resource class (Ahuja et al., 2023; Asai et al., 2024), pretraining data share, or tokenizer fertility (Ahuja et al., 2024) - the dominant reporting format remains per-language accuracy together with an unweighted mean, with attribution typically performed one covariate at a time without controlling for con-

founders, modelling interactions, or quantifying uncertainty.

Disparity-aware Evaluation and its limits. Choudhury and Deshpande (2021) reframe multilingual model selection as a social-choice problem and advocate Rawlsian max–min selection over the implicit utilitarian means. Khanuja et al. (2023) operationalize fairness as the Gini coefficient over per-language scores within a Diversity–Equity–Inclusion framework, and Blasi et al. (2022) document the systemic global inequalities in language-technology performance. These works establish the normative case for disparity-aware evaluation, but stop at one-dimensional summaries (Gini, max–min) and do not run a confounder-controlled statistical model over the (model × benchmark × language) cube. The closest to our work is the work by Hu et al. (2025) who fit a linear mixed-effects model over MEGA datasets and a panel of pre-2024 model variants and report a Performance Realisation Ratio per language. The additive specification ($s = \mu + \alpha_\ell + \beta_t + u_m + \varepsilon$) can rank languages and tasks by difficulty but cannot, by construction, identify model×language interactions or admit linguistic features as covariates. Li et al. (2025) and Anugraha et al. (2025) contribute, respectively, an intrinsic similarity-to-English ranking and a proxy-LM forecast for unseen (language, task) cells, but neither yields a fairness decomposition with calibrated uncertainty.

Mechanisms of Multilingual Disparity. A separate body of literature has identified individual correlates of cross-lingual performance: data scale and the resource level taxonomy (Joshi et al., 2020; Blevins and Zettlemoyer, 2022); tokenizer fertility, which varies by orders of magnitude across languages (Rust et al., 2021; Ahia et al., 2023) and translates directly into evaluation inequity via inflated token budgets and length-sensitive metrics (Petrov et al., 2023); typological distance to English from URIEL/lang2vec (Littell et al., 2017); and the alignment of internal representations with English, established for multilingual encoders (Wu and Dredze, 2020; Conneau et al., 2020) and recently extended to decoder LLMs (Wendler et al., 2024; Li et al., 2025). Each correlate has been studied in isolation; none of these threads enter all four families of predictors into a single hierarchical model, reports their explanatory contributions on a common scale, or quantifies how much of the cross-language gap is structural and therefore predictable in advance of evaluation, which is exactly the gap our framework addresses (Section 5.1).

Model	Params	Arch.	Origin
QWEN3.5-4B	4B	Dense	Qwen
QWEN3.5-122B-A10B	122B (10B)	MoE	Qwen
AYA-EXPANSE-32B	32B	Dense	Cohere
TINY-AYA-GLOBAL	3.5B	Dense	Cohere (distilled)
GPT-OSS-20B	20B	MoE	GPT-OSS
SARVAM-30B	30B	MoE	Sarvam (Indic)
SARVAM-105B	105B	MoE	Sarvam (Indic)

Table 1: Models evaluated, grouped by family (Qwen , Cohere , GPT-OSS , Sarvam).

Bucket	Benchmarks	Score
NLU *(MCQ, lik, knowledge centric)*	GLOBAL-MMLU, MMLU-INDIC-ROMAN, MMLU-PROX, MMMLU, OKAPI-MMLU, MILU, INCLUDE, BOOLQ-INDIC, TRIVIAQA-INDIC-MCQ	acc $\in [0, 1]$
Reasoning *(gen. + MCQ + Infer)*	GSM8K-INDIC, MGSM (CoT math, flexible exact-match); BELEBELE, XCOPA, XSTORYCLOZE, XWINOGRAD (commonsense MCQ, likelihood)	native $[0, 1]$

Table 2: Benchmark task buckets, grouped by skill (NLU , Reasoning). Each (model, benchmark, language) cell is the benchmark-level mean of subject-level scores. NLG benchmarks are listed separately in Table 17.

3 Experimental Setup

Models As shown in Table 1, we evaluate seven open-weight multilingual LLMs spanning ∼3.5B - ∼122B parameters, covering three tokenizer families and four providers.

Benchmarks and Task Buckets Cross-language scores are drawn from LM-EVALUATION-HARNESS (Gao et al., 2024) runs against a consistent eval pipeline. Benchmarks are partitioned into three task buckets (Table 2) according to output format and scoring. The evaluation suite encompasses both N-way parallel datasets and structurally non-parallel benchmarks.

Features: Language & Model×Languages Every (model, language) cell carries two groups of predictors, listed in Appendix Table 11. The first group is purely language-level: `script` and `family` are categorical typological factors, `resource_class` is the Joshi et al. (2020) 1–5 tier,

Benchmark	k (lang)	Kendall W	sig.
BELEBELE	62	0.696	***
BOOLQ-INDIC	11	0.810	***
GLOBAL-MMLU	15	0.728	***
GSM8K-INDIC	11	0.808	***
INCLUDE	43	0.665	***
MGSM	11	0.562	***
MILU	11	0.871	***
MMLU-INDIC-ROMAN	10	0.907	***
MMLU-PROX	29	0.633	***
MMMLU	14	0.809	***
OKAPI-MMLU	34	0.729	***
TRIVIAQA-INDIC-MCQ	11	0.860	***
XCOPA	11	0.309	**
XSTORYCLOZE	11	0.578	***
XWINOGRAD	6	0.778	***

Table 3: Per-benchmark Friedman tests (languages as items, models as judges); larger Kendall's $W \in [0, 1]$ indicates stronger inter-model agreement on language rankings. Rows tinted by bucket (NLU , Reasoning); $** \; p < 0.01$, $*** p < 0.001$. Full statistics in Table 7.

and `syn_dist_en`, `phon_dist_en` are `lang2vec` distances to English. The second group is model-specific: `fertility` is tokens-per-word for the model's tokenizer, and `repr_sim_en` is the representational similarity (CKA) of the model's hidden-state geometry for the language to its geometry for English. All continuous features are standardised ($\mu = 0$, $\sigma = 1$) before entering the model.

4 Characterizing Disparity

Overview of Disparity. In this work, we evaluate N models on B multilingual benchmarks, where each benchmark $b \in \{1, \ldots, B\}$ covers a set of languages $\mathcal{L}_b$ with $|\mathcal{L}_b| = x_b$. Let $s_{n,b,\ell} \in \mathbb{R}$ denote the score of model n on benchmark b for language $\ell \in \mathcal{L}_b$. Beyond reporting raw performance, we are interested in characterizing the gap between the best- and worst-performing languages for a given model as well as more nuanced measures that summarise the full distribution of per-language scores $\{s_{n,b,\ell}\}_{\ell \in \mathcal{L}_b}$. Prior work has explored several factors contributing to such gaps, most notably the systematic under-representation of many of the world's languages in training data and benchmarks (Joshi et al., 2020; Blasi et al., 2022). While under-representation is a key driver of disparity, it is one of several confounding factors, alongside task difficulty, domain mismatch, and model capacity, that jointly shape observed cross-lingual performance gaps (Hu et al., 2023; Ahuja et al., 2023).

Resource tier	NLU ↑	Reasoning ↑	Overall ↑
1-Scraping	0.432	0.382	0.421
2-Hopefuls	0.458	0.497	0.472
3-Rising	0.536	0.527	0.533
4-Underdogs	0.542	0.667	0.604
5-Winners	**0.680**	**0.714**	**0.696**

Table 4: Mean per-language score by language resource tier, pooled across models, on the NLU and Reasoning buckets (NLU , Reasoning).

Existing Disparity Metrics. For a fixed model n and benchmark b, let $\mathbf{s} = (s_1, \ldots, s_K)$ be the vector of per-language scores with mean $\bar{s}$ and standard deviation σ_s. Prior work summarises $\mathbf{s}$ with scale-invariant dispersion indices: the coefficient of variation $\mathrm{CV} = \sigma_s / \bar{s}$ (Hu et al., 2025), the Gini coefficient $G \in [0, 1]$ adapted from welfare economics (Khanuja et al., 2023; Choudhury and Deshpande, 2021), Sen welfare $W = \bar{s}(1 - G)$ (Sen, 1976) which combines level and equity. Full definitions and per-model values are given in Table 8 (Appendix).

Shortcomings of these methods. Three limitations make these summaries inadequate as a standalone account of disparity. First, CV and Gini are scale-invariant: small numerical differences mask large absolute gaps—even the lowest-Gini model shows an average max-min gap of roughly one-third the mean. Second, most indices are insensitive to the overall performance level, so a uniformly weak model and a uniformly strong one both register "perfect parity". Third, none identify *which* factors—language, task, model, or their interactions—drive the gaps. We therefore test whether disparity is systematic (Section 4) and attribute it to interpretable factors via a Bayesian mixed-effects model (Section 5).

Is the disparity real and systematic? The summary statistics above describe *how much* scores spread across languages, but not whether that spread reflects a stable, model-independent ordering or merely sampling noise. We address this with two distribution-free tests - Kendall's W (Kendall and Smith, 1939) and Dunn's Post-Hoc (Dunn, 1964)- that together motivate the mixed-effects model of Section 5.

Rankings are consistent across models (Friedman / Kendall's W). For each benchmark, we

treat languages as k items ranked by the m=7 models and compute the Friedman statistic together with Kendall's $W = \chi^2/[m(k-1)] \in [0, 1]$ as an effect size for inter-model agreement (Tables 3 and 7), where W is the Kendall's W value; χ^2 is the Friedman test statistic value. Kendall's W coefficient assumes the value from 0 (indicating no relationship) to 1 (indicating a perfect relationship). Null hypothesis of equal language ranks is rejected in **18/18 benchmarks**, with W ranging from 0.31 (`xcopa`) to 0.91 (`mmlu_indic_roman`) and a median of ~ 0.75. This means the models show similar patterns of ranking languages from easiest to hardest, with the same languages tending to be on top and bottom across models.

Disparity tracks resource tiers (Dunn's post-hoc). We next ask whether the cross-language spread aligns with how well-represented, each language is on the web. We group languages into the five resource tiers of Joshi et al. (2020) (Class 1 *Scraping* $\rightarrow$ Class 5 *Winners*), z-normalise scores within each benchmark cell to make them comparable, and run Dunn's pairwise tests between tiers with Benjamini–Hochberg FDR correction at $\alpha = 0.05$ (Benjamini and Hochberg, 1995) Pooled across models, mean z-scores rise strictly from Class 1 to Class 5 and *all* 10 pairwise tier contrasts are significant in both NLU and reasoning. The same monotone staircase holds per model on NLU (7–10/10 contrasts significant for every model) and on reasoning for all but two models – Sarvam-30b and GPT-OSS-20B – whose tier ordering flattens at the lower end, an expected outcome when a model is near the floor across most languages. Full pairwise matrices are in Appendix G. The raw (unnormalized) means in Table 4 make the effect concrete:The average per-language score nearly doubles from Class 1 to Class 5 in both buckets.

Together, these tests establish that the cross-language gaps are neither sampling noise nor an artifact of any single model: the language ordering is stable across models and structured monotonically by resource level. This rules out the most benign explanations for disparity and motivates an explanatory model of *what* drives it, beyond any single dispersion summary.

5 What Explains Disparity?

Sources of the disparity characterised in Section 4 are entangled: a low score on a given language may reflect the language itself (its script, its typological distance from training data), the model (its tokenizer, its representation geometry), the benchmark, or simple noise. Disentangling these requires a model that respects the nested structure of the data — every score is jointly indexed by (model, benchmark, language) — and that yields a quantitative attribution rather than per-language anecdote. We address this with a Bayesian hierarchical decomposition fit independently across two task buckets: NLU (multilingual MCQ knowledge benchmarks) and Reasoning (commonsense and math MCQ benchmarks).[1]

5.1 Method in brief

We model each observed score as the sum of a benchmark effect (τ_t), a language effect (α_ℓ), a model effect (u_m), and Gaussian noise (ε):

$$\text{score}_{m,t,\ell} = \mu + \tau_t + \alpha_\ell + u_m + \varepsilon_{m,t,\ell}, \quad (1)$$

with $\alpha_\ell \sim \mathcal{N}(0, \sigma_\alpha^2)$, $u_m \sim \mathcal{N}(0, \sigma_u^2)$, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

The key quantity is σ_α, the typical magnitude of a language's deviation from the model-and-benchmark-conditional mean. It is the formal answer to "how big is the cross-language gap?"

To attribute that gap, we extend (1) in two steps. **First**, we add only language-level covariates $\mathbf{x}_\ell$ (resource class (Joshi et al., 2020), syntactic, phonological, and geographic distance to English from `lang2vec` (Littell et al., 2017), plus $\text{script}(\ell)$ and $\text{family}(\ell)$ as categorical factors). This shrinks σ_α to a residual $\sigma_\alpha^{(\text{lang})}$. **Second**, we add model–language covariates $\mathbf{z}_{m,\ell}$ (tokenizer fertility and representation similarity to English) together with the random interactions $(1 \mid \text{model:language})$, $(1 \mid \text{bench:language})$, $(1 \mid \text{bench:model})$, and a per-model random slope on fertility. We measure representational similarity to English by computing the Centered Kernel Alignment (CKA) (Kornblith et al., 2019) distance between the hidden states of English and the target language at the model's middle layer following (Wendler et al., 2024; Dumas et al., 2025), mean-pooled across tokens on the FLORES-200 dataset (Costa-Jussà et al., 2022). We measure the tokenizer fertility on FLORES-200 as well. This dataset was specifically selected because it offers expert-curated parallel sentences

[1] A third bucket, NLG (open generation; chrF on translation and summarisation), is reported in Appendix F but excluded from the main analysis: its 12-language Indic/Dravidian pool is too narrow to identify the categorical script and family effects, so the disparity-explanation metric is not interpretable.

translated uniformly across 200 languages. We provide a detailed justification for the design choices in Appendix B, and provide a glossary of all the terminology in Appendix Table 13.

Putting both extensions together, each observation i (a triple $(m(i), t(i), \ell(i))$ of model, benchmark, and language) is modelled as

$$y_i = \underbrace{\beta_0 + \tau_{t(i)} + \mathbf{x}_i^\top \boldsymbol{\beta}}_{\text{fixed effects}} + \underbrace{\varepsilon_i}_{\text{residual}} + \underbrace{\alpha_{\ell(i)} + u_{m(i)} + v_{m(i),\ell(i)} + w_{t(i),\ell(i)}}_{\text{random effects}}, \quad (2)$$

where $\tau_{t(i)}$ is the benchmark fixed effect carried over from (1), $\mathbf{x}_i$ stacks the language- and model–language covariates introduced above ($\mathbf{x}_\ell$ and $\mathbf{z}_{m,\ell}$), α_ℓ, u_m, $v_{m,\ell}$, $w_{t,\ell}$ are zero-mean Gaussian random intercepts with their own variance components $\sigma_\alpha^2, \sigma_u^2, \sigma_v^2, \sigma_w^2$, and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. We place weakly informative priors on $\boldsymbol{\beta}$ and half-normal priors on the variance components, and perform full posterior inference, yielding credible intervals on every coefficient and variance component reported below.

Our disparity-explanation metric is the proportional reduction in between-language variance contributed by the language covariates:

$$R^2_{\text{ling}} = 1 - \frac{\left[\sigma_\alpha^{(\text{lang})}\right]^2}{\sigma_\alpha^2}. \quad (3)$$

This is a standard proportional-reduction-in-variance measure in multilevel modelling; we use the R^2_{ling} notation to emphasise that it quantifies linguistic-feature explanatory share. All fits use NUTS via `numpyro` with 5000 tuning iterations and 2000 post-warmup draws per chain across four chains.

5.2 The language gap and what explains it

Without any structured features, languages differ by $\sigma_\alpha = 0.078$ on NLU and 0.088 on Reasoning around the model-and-benchmark-conditional mean indicating typical deviations of roughly $\pm$8–9 accuracy points (Table 5).

Adding just the language covariates plus script and family categoricals shrinks σ_α to 0.035 on NLU and 0.021 on Reasoning, giving $R^2_{\text{ling}} = 0.79$ $[0.57, 0.93]$ on NLU and $R^2_{\text{ling}} = 0.92$ $[0.73, 0.99]$ on Reasoning (90% HDIs; Table 5). The bulk of the cross-language gap on MCQ benchmarks is recoverable from features that can be computed without any inference on the model itself.

Bucket	$\sigma_\alpha \downarrow$	$R^2_{\text{ling}} \uparrow$
NLU	0.078 $[0.064, 0.092]$	$\mathbf{0.79}$ $[0.57, 0.93]$
Reasoning	0.088 $[0.072, 0.107]$	$\mathbf{0.92}$ $[0.73, 0.99]$

Table 5: Headline disparity-explanation results per bucket. σ_α is the SD of the language random intercept; R^2_{ling} is its proportional reduction once linguistic features are added. Brackets show 90% HDIs.

Which single predictor carries the most weight? R^2_{ling} above is computed from the language-features fit alone. The full model additionally enters two *model-conditional* covariates: tokenizer fertility and representation similarity to English (`repr_sim_en`) which lets us rank all six standardised continuous predictors (four language-only, two model-conditional) against each other on a common scale. Only one is credibly positive in both buckets: `repr_sim_en`, with $\beta = +0.065$ $[0.051, 0.079]$ on NLU and $+0.094$ $[0.074, 0.116]$ on Reasoning (90% HDIs). A one-SD increase in alignment is worth 6–9 accuracy points.[2]

Resource level (`resource_class`) also has a credibly positive effect on Reasoning ($\beta = +0.037$, $[0.002, 0.07]$). The other three typological distances (syntactic, phonological, geographic) and tokenizer fertility have credible intervals that overlap zero on their own. This does *not* mean these features carry no signal but rather, they are collinear with the categorical script and family terms and with the per-language random intercept, which soak up the shared variance. The high joint R^2_{ling} confirms the linguistic block as a whole is informative.

5.3 What else drives score variation?

R^2_{ling} answers "how much of the language gap do language features explain?". A second, related question is "what *other* sources of variation in the score data should we care about, once we account for everything we can?" Table 6 reports the full-model decomposition: per component and per bucket, the share of total variance carried by that component and its standard deviation in raw score units (a component with $\sigma = 0.10$ means typical deviations of $\pm$10 accuracy points along that axis). Columns sum to 100% of total score variance within each bucket.

[2] The same effect holds on NLG (Appendix F, $\beta = +0.053$ $[0.025, 0.080]$), making this the single most stable cross-bucket finding, and a quantity that can be computed from a model alone without running any multilingual benchmark.

Component	NLU		Reasoning	
	%	σ	%	σ
σ_{model}	**66.7**	0.149	12.7	0.057
$\sigma_{\text{bench}\times\text{model}}$	13.6	0.061	**46.3**	0.119
$\sigma_{\text{fert slope}\mid\text{model}}$	1.4	0.015	12.1	0.058
$\sigma_{\text{model}\times\text{language}}$	6.9	0.044	9.4	0.053
$\sigma_{\text{bench}\times\text{language}}$	3.6	0.031	4.3	0.036
σ_{language} (residual)	1.5	0.017	1.4	0.018
σ (per-row noise)	6.3	0.042	13.8	0.065

Table 6: Variance decomposition of the full model, per bucket. "%" is each component's share of total variance (columns sum to 100 within bucket); σ is the posterior-mean SD in raw score units. Bold marks the largest component per column.

Language Residual: The language-intercept SD in this full-model decomposition ($\sigma_{\text{language}} \approx 0.017$) is smaller than the language-features residual reported in Table 5 (0.021–0.035). The two answer different questions: R^2_{ling} measures the shrinkage of σ_α when language features are added to the baseline (Eq. 3); the full model additionally includes (model:language) and (bench:language) random interactions, which absorb cell-specific deviations that were previously charged to language identity, shrinking what remains in the language intercept further.

NLU and Reasoning have qualitatively different variance structures. On NLU, model identity dominates: 66.7% of variance, $\sigma_{\text{model}} = 0.149$, meaning models differ by about 15 accuracy points and a model strong on `mmlu` tends to be strong on `include` too. On Reasoning the ordering inverts. bench × model carries 46.3% ($\sigma = 0.119$) and is *larger* than the main model effect ($\sigma_{\text{model}} = 0.057$): the typical model-by-benchmark interaction ($\sim$ 12 accuracy points) exceeds the typical model-by-grand-mean deviation ($\sim$ 6 points). A model's strength does not pool across reasoning benchmarks — math, commonsense, and coreference probe genuinely different capabilities. A single "multilingual reasoning" aggregate is misleading; per-benchmark reporting is required.

Tokenizer fertility matters on Reasoning, but heterogeneously across models. The pooled fixed effect of `fertility_z` has an HDI that just crosses zero (Appendix D), but the per-model random slope carries 12.1% of Reasoning variance ($\sigma = 0.058$). Per-model fertility coefficients differ by $\sim \pm 6$ accuracy points around the pooled mean, so averaging over models hides a real effect. The same component is small on NLU (1.4%), consistent with MCQ knowledge benchmarks being less tokenization-sensitive than chain-of-thought reasoning.

Model x Language persists at 7–9% in both buckets. With $\sigma \approx 0.044$–0.053, some model–language pairings have idiosyncratic gains or losses of ~ 5 accuracy points that none of our measured features explain — almost certainly traceable to pretraining-data specifics we cannot observe.

Non-parametric cross-check. As a sanity check on the feature ranking above, a gradient-boosted tree fit to the same per-cell frame reproduces the Stage-3 ordering — representation similarity to English dominant, fertility and syntactic distance secondary, phonological distance and resource class marginal — and recovers held-out language accuracies within $\sim$ 4 (NLU) and $\sim$ 7 (Reasoning) accuracy points (Appendix E).

6 Discussion and Implications

Most of the cross-language gap is structural. Conditional on the model and the benchmark, languages differ by about ±8–9 accuracy points. Observable language features – resource class, typological distances, script, family, and representation alignment to English – explain 79% of that gap on understanding and 92% on reasoning. This has a practical consequence: an absolute gain on a low-resource language that does not also reduce σ_α after controlling for these features is, with high probability, attributable to structural factors (Say tokenizer fertility and increased representational similarity) rather than a real improvement in multilingual capability. Reporting σ_α and R^2_{ling} alongside per-language scores would make this distinction visible by default; current Gini and max–min summaries cannot. The structure of the Bayesian decomposition attributes is already visible non-parametrically via the statistical tests we perform.

Representation alignment to English is the most robust single lever. The six standardized predictors entered together, only the similarity between a model's hidden-state geometry for a language and for English (`repr_sim_en`) is credibly positive in every bucket: +0.065 on NLU, +0.094 on reasoning, and +0.053 on generation (Appendix F). A one-standard-deviation gain is worth 6–9 accuracy

points. Two properties make this useful. First, the predictor is a property of the (model, language) pair, not of the language alone, which is why its sign and magnitude carry across three task families scored with three different metrics – consistent with prior evidence that decoder LLMs route multilingual computation through English (Wendler et al., 2024; Li et al., 2025). Second, it is computable from a model alone, with no multilingual benchmark required, making it the cheapest principled proxy available for ranking checkpoints by expected multilingual headroom. The coefficient is correlational, but its cross-bucket stability is stronger evidence than any single-task association.

NLU and reasoning are not the same kind of problem. On NLU, model identity carries two-thirds of the total variance and benchmark–model interaction is modest: a model strong on one knowledge benchmark is strong on the next. On reasoning, the main model effect drops to about a third of its NLU magnitude, and the benchmark–model component carries nearly half of the total variance, larger than the main model effect itself. Pooling commonsense, math, and coreference into a single "multilingual reasoning" score therefore averages over differences that exceed the differences between models. Per-benchmark reporting is required, and disagreement across reasoning benchmarks should be read as a signal about the underlying tasks, not as measurement noise.

Fixed-effect-only specifications hide model-dependent predictors. Tokenizer fertility is a good example. Its *pooled* coefficient on Reasoning is statistically null, averaged across all models; fertility looks like it does not matter. But once we allow each model its own fertility slope, that random slope accounts for 12.1% of total variance, with individual models' sensitivities spread roughly ± 6 accuracy points around the pooled mean. In other words, fertility hurts some models in some languages a lot, and others barely at all, averaging cancels these opposing effects out to zero. A fixed-effects-only analysis would have wrongly concluded that fertility is inert; the random slope reveals that *which* languages a tokenizer over-fragments determines *which* languages a given model fails on, and the magnitude of this effect is comparable to the main "model" effect on Reasoning. This matches prior tokenizer findings (Rust et al., 2021; Ahia et al., 2023), and argues for routinely reporting random slopes whenever an evaluation panel mixes models with different tokenizer families.

The model–language residual sets the ceiling for feature-based interventions. A $\text{model} \times \text{language}$ component of 7–9% ($\sigma \approx 0.044$–0.053) persists in both buckets after every measured covariate. These are model-specific language preferences of about ± 5 accuracy points that none of the linguistic features, the alignment probe, or the fertility slope account for. They almost certainly trace to pre-training data decisions – mixture ratios, deduplication, quality filters – that current model releases do not disclose. This residual bounds what any language-feature-only intervention can close, and it is the natural next target for the framework (Blasi et al., 2022; Blevins and Zettlemoyer, 2022).

How this disparity can be addressed Two of the variance components our framework isolates correspond to actionable levers. The first is representation alignment to English, raised either by cross-lingual alignment objectives at pretraining or by alignment fine-tuning afterward. The second is the residual model–language component – the 7–9% gap no observable language feature explains – which our results trace to in-language pretraining data: mixture and quality.

7 Conclusion

Our work reframes multilingual LLM evaluation from a passive mapping of performance gaps into an explainable, diagnostic framework: decomposing observed variance through a hierarchical model shows that most of the cross-lingual disparity is structural, attributable to predictable linguistic properties and internal representation geometries. Reporting σ_α alongside per-language accuracy attributes any reported gain to the component that actually moved; a model's internal representation alignment to English serves as a benchmark-free proxy for expected multilingual capacity during pretraining or alignment, and the residual model–language component sets the ceiling for feature-based interventions, pointing to pretraining-data mixture and curation as the next lever.

Limitations

Our analysis is observational. The seven-checkpoint panel is large enough to identify the variance components reported in Table 6 but small

enough that the per-model random slopes are aggressively shrunken by the prior; more checkpoints would tighten both. The Gaussian likelihood on raw accuracy is an approximation near the bounded scale. The NLG bucket admits the same specification, but its twelve-language Indic and Dravidian pool under-identifies the script and family categoricals; we therefore restrict the NLG result to the single `repr_sim_en` coefficient in Appendix F and do not report a bucket-level R^2_{ling} there. Finally, all predictor coefficients are correlational: the cross-bucket consistency of `repr_sim_en` is the strongest regularity we identify, but causal claims require controlled training-side experiments that we do not run here.

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 (Volume 1: Long Papers)*, pages 2732–2757, Albuquerque, New Mexico. Association for Computational Linguistics.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.

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, Singapore. Association for Computational Linguistics.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks. 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 2598–2637, Mexico City, Mexico. Association for Computational Linguistics.

David Anugraha, Genta Indra Winata, Chenyue Li, Patrick Amadeus Irawan, and En-Shiun Annie Lee. 2025. ProxyLM: Predicting language model performance on multilingual tasks via proxy models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1981–2011, Albuquerque, New Mexico. Association for Computational Linguistics.

Akari Asai, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2024. BUFFET: Benchmarking large language models for few-shot cross-lingual transfer. 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 1771–1800, Mexico City, Mexico. Association for Computational Linguistics.

R.H. Baayen, D.J. Davidson, and D.M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412. Special Issue: Emerging Data Analysis.

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 (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Monojit Choudhury and Amit Deshpande. 2021. How linguistically fair are multilingual pre-trained language models? In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12710–12718.

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, Online. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2025. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31822–31841, Vienna, Austria. Association for Computational Linguistics.

Olive Jean Dunn. 1964. Multiple comparisons using rank sums. *Technometrics*, 6(3):241–252.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Songbo Hu, Ivan Vulić, and Anna Korhonen. 2025. Quantifying language disparities in multilingual large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4003–4018, Suzhou, China. Association for Computational Linguistics.

Songbo Hu, Han Zhou, Moy Yuan, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Anna Korhonen, and Ivan Vulić. 2023. A systematic study of performance disparities in multilingual task-oriented dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6825–6851, Singapore. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Maurice G Kendall and B Babington Smith. 1939. The problem of m rankings. *The annals of mathematical statistics*, 10(3):275–287.

Simran Khanuja, Sebastian Ruder, and Partha Talukdar. 2023. Evaluating the diversity, equity, and inclusion of NLP technology: A case study for Indian languages. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1763–1777, Dubrovnik, Croatia. Association for Computational Linguistics.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374, Singapore. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2025. Language ranker: A metric for quantifying llm performance across high and low-resource languages. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):28186–28194.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL

and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. AfroBench: How good are large language models on African languages? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19048–19095, Vienna, Austria. Association for Computational Linguistics.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, Sharad Duwal, and 38 others. 2025. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations*.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Amartya Sen. 1976. Real national income. *The Review of Economic Studies*, 43(1):19–39.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Yosephine Susanto, Adithya Venkatadri Hulagadri, Jann Railey Montalan, Jian Gang Ngui, Xianbin Yong, Wei Qi Leong, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Yifan Mai, and William Chandra Tjhi. 2025. SEA-HELM: Southeast Asian holistic evaluation of language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12308–12336, Vienna, Austria. Association for Computational Linguistics.

Dennis Ulmer, Elisa Bassignana, Max Müller-Eberstein, Daniel Varab, Mike Zhang, Rob van der Goot, Christian Hardmeier, and Barbara Plank. 2022. Experimental standards for deep learning in natural language processing research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2673–2692, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2025. MILU: A multi-task Indic language understanding benchmark. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10076–10132, Albuquerque, New Mexico. 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.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. Association for Computational Linguistics.

A Appendix

Table 8 reports per-(model, bucket) disparity summaries — mean score, Gini, Sen welfare, coefficient of variation, and relative and z-scaled max–min gaps — aggregated across benchmarks within

each bucket, providing the full numerical backing for the dispersion claims discussed in Section 4.

Table 9 breaks the same scores down by resource tier and model, showing that the monotonic Class 1→Class 5 staircase recurs for every checkpoint.

Table 10 gives the Kruskal–Wallis omnibus statistics and tier-mean z-scores that motivate the Bayesian decomposition in Section 5; the corresponding full pairwise Dunn matrices appear in Appendix G.

B Design choices: covariates, random structure, and probe corpus

The methodology paragraph in Section 5.1 compresses several independent design choices into a single line. We expand each here, stating for each what we use, why, what alternatives we considered, and where the choice is standard in the literature.

B.1 Why hierarchical Bayesian modeling

Our data has the structure of a crossed panel: every (model, benchmark, language) cell yields one accuracy observation, and the three index sets are non-nested – the same language appears under every model and every benchmark, the same model appears under every benchmark and every language. Three features of this setting drive the modeling choice.

First, the quantity of interest is not a single point estimate but a *variance decomposition*: how much of the cross-language dispersion is attributable to the language itself, to its interaction with a specific model, and with its interaction with a specific benchmark. Random-effect intercepts give each of these components a single scalar (σ_α, $\sigma_{\text{model:lang}}$, $\sigma_{\text{bench:lang}}$) on the accuracy scale, which is exactly the quantity the headline R^2_{ling} statistic compares across model stages. An OLS regression with the same covariates plus a per-cell fixed effect, or an ANOVA-style sums-of-squares decomposition, would report variance *explained* rather than variance *components*, and neither shrinks sparsely-observed cells.

Second, several of the crossings are sparsely observed: many language–benchmark cells have a single trial. Maximum- likelihood mixed-effect estimators (e.g. `lme4`, `glmmTMB`) frequently fail to converge under the full random-effect structure at this sparsity, and a fully fixed-effect specification overfits the smallest cells. Hierarchical Bayesian estimation with weakly informative priors regularises these toward their group mean and returns a coherent joint posterior over the variance components, which is what the credibility intervals on R^2_{ling} are computed from; a frequentist bootstrap on the same components does not yield this joint distribution.

Third, the framework is the standard analytic tool for crossed language/item/subject panels in psycholinguistics (Baayen et al., 2008), has been explicitly recommended for LLM-evaluation panels by Ulmer et al. (2022), and is used in this form for multilingual benchmark analysis by (Khanuja et al., 2023; Ahuja et al., 2023; Ojo et al., 2025). We follow their specification.

C Full model specification

All fits share a common formula template, with the right-hand side extended in two steps from the baseline. We write the model in `bambi`/`lme4` formula syntax; the three stages are shown together in Figure 2.

Priors. We use `bambi`'s defaults throughout: weakly-informative Normal priors on fixed effects (centred at 0 with prior SD scaled to the predictor), Half-Normal priors on random-effect SDs, and Half-StudentT on the residual SD. Standardised continuous predictors carry a $\mathcal{N}(0, 1)$ prior on the standardised scale.

Sampling. NUTS via `numpyro` (JAX-backed), four chains with `chain_method="vectorized"` on a single A100, 5000 warmup and 2000 post-warmup draws per chain, target acceptance 0.99. Convergence is assessed by $\hat{R}$ and bulk effective sample size.

C.1 Language-level covariates

Resource class. We use the six-tier taxonomy (0–5) of Joshi et al. (2020), treated as a standardised continuous covariate. It is a single scalar that captures the NLP-resource situation of a language (availability of labelled data, raw text, and tools) more directly than raw demographics. Among the alternatives considered – log token count from CommonCrawl, mC4, or OSCAR; Wikipedia article count; and Ethnologue speaker population – each conflates the quantity of interest with web-crawl artefacts, encyclopaedic activity, or user demand respectively. The resource taxonomy is the de facto resource proxy in multilingual NLP and

Benchmark	k (lang)	m	χ^2	p	Kendall W	sig
BELEBELE	62	7	297.1	3.4×10^{-61}	0.696	***
BOOLQ-INDIC	11	7	56.7	2.1×10^{-10}	0.810	***
GLOBAL-MMLU	15	7	71.4	2.1×10^{-13}	0.728	***
GSM8K-INDIC	11	7	56.6	2.3×10^{-10}	0.808	***
INCLUDE	43	7	195.6	1.6×10^{-39}	0.665	***
MGSM	11	7	39.3	6.1×10^{-7}	0.562	***
MILU	11	7	60.9	2.9×10^{-11}	0.871	***
MMLU-INDIC-ROMAN	10	7	57.1	1.7×10^{-10}	0.907	***
MMLU-PROX	29	7	124.1	2.2×10^{-24}	0.633	***
MMMLU	14	7	73.7	7.3×10^{-14}	0.809	***
OKAPI-MMLU	34	7	168.4	1.0×10^{-33}	0.729	***
TRIVIAQA-INDIC-MCQ	11	7	60.2	4.0×10^{-11}	0.860	***
XCOPA	11	7	21.6	1.0×10^{-3}	0.309	**
XSTORYCLOZE	11	7	40.5	3.7×10^{-7}	0.578	***
XWINOGRAD	6	7	27.2	1.3×10^{-4}	0.778	***

Table 7: Per-benchmark Friedman test (languages as items, m=7 models as judges) for the NLU and Reasoning buckets. Kendall's $W = \chi^2/[m(k-1)] \in [0,1]$ is an effect-size analog. Rows tinted by bucket (NLU , Reasoning).

Model	**Bucket**	$\mu\uparrow$	Gini$\downarrow$	$W_{\text{Sen}}\uparrow$	CV$\downarrow$	$\Delta_{\text{rel}}\downarrow$	$\Delta_z\downarrow$
QWEN3.5-122B-A10B	NLU	**0.741**	**0.064**	**0.695**	**0.099**	**0.363**	**3.581**
	Reasoning	0.647	0.119	0.602	0.290	1.030	3.638
QWEN3.5-4B	NLU	0.525	0.118	0.464	0.197	0.664	3.422
	Reasoning	0.567	0.134	0.506	0.273	0.890	3.075
AYA-EXPANSE-32B	NLU	0.562	0.112	0.502	0.180	0.599	3.338
	Reasoning	**0.677**	0.120	0.598	0.226	0.684	**2.986**
GPT-OSS-20B	NLU	0.440	0.134	0.391	0.185	0.721	3.753
	Reasoning	0.638	0.080	0.589	0.158	0.548	3.316
SARVAM-105B	NLU	0.591	0.108	0.539	0.171	0.630	3.705
	Reasoning	0.648	**0.075**	**0.599**	**0.143**	**0.481**	3.165
SARVAM-30B	NLU	0.354	0.110	0.325	0.153	0.640	3.939
	Reasoning	0.379	0.131	0.345	0.247	0.754	3.213
TINY-AYA-GLOBAL	NLU	0.421	0.134	0.368	0.213	0.712	3.363
	Reasoning	0.525	0.143	0.464	0.278	0.882	3.026

Table 8: Per-(model, bucket) language disparity, unweighted means across benchmarks in the bucket (NLU , Reasoning). For each (model, benchmark) cell, metrics are computed over the per-language scores $\{x_1, \ldots, x_L\}$ and then averaged across benchmarks within the bucket. $\mu = \frac{1}{L}\sum_i x_i$ is the mean per-language score (higher is better). **Gini** $= \frac{\sum_i \sum_j |x_i - x_j|}{2L^2\mu} \in [0,1]$ measures relative inequality across languages (lower is more equal). $W_{\text{Sen}} = \mu\,(1 - \text{Gini})$ is Sen's welfare index, an equity-discounted mean (higher is better). **CV** $= \sigma/\mu$ is the coefficient of variation, a scale-free spread (lower is better). $\Delta_{\text{rel}} = (\max_i x_i - \min_i x_i)/\mu$ is the best-vs.-worst gap normalised by the mean (lower is better). $\Delta_z = (\max_i x_i - \min_i x_i)/\sigma$ is the same gap expressed in standard deviations (lower is better). Arrows show direction of improvement; bold marks the best value per column within each bucket.

Model	1-Scraping↑	2-Hopefuls↑	3-Rising↑	4-Underdogs↑	5-Winners↑
QWEN3.5-122B-A10B	**0.513**	**0.620**	**0.684**	**0.793**	0.806
QWEN3.5-4B	0.361	0.413	0.498	0.669	0.735
AYA-EXPANSE-32B	0.473	0.482	0.646	0.694	**0.808**
GPT-OSS-20B	0.459	0.500	0.510	0.558	0.680
SARVAM-105B	0.514	0.600	0.610	0.617	0.747
SARVAM-30B	0.244	0.355	0.320	0.350	0.448
TINY-AYA-GLOBAL	0.288	0.371	0.453	0.551	0.658

Table 9: Mean per-language score by resource tier for each model, averaged across the NLU and Reasoning buckets. Construction matches Table 4. Bold marks the best model per tier.

Bucket	H	p	C1	C2	C3	C4 / C5
NLU	1243.89	4.9×10^{-268}	-1.02	-0.40	-0.01	$+0.21$ / $+0.75$
Reasoning	78.69	3.3×10^{-16}	-0.78	-0.32	-0.04	$+0.24$ / $+0.81$

Table 10: Kruskal–Wallis omnibus across the five resource tiers on z-normalised scores (within each (benchmark, task_id) cell), with tier-mean z-scores from Class 1 to Class 5. Dunn pairwise tests with BH–FDR correction: all 10 contrasts significant in both buckets (weakest C1 vs. C2 in Reasoning at p=0.014). The same staircase recurs per model.

Stage 1 — Baseline (used to estimate the raw cross-language gap σ_α).

```
score ~ 1 + C(task) + (1|language) + (1|model)
```

Stage 2 — Language-features fit (used to compute R^2_{ling} via Equation 3).

```
score ~ 1 + joshi_class_z + syn_dist_en_z + phon_dist_en_z + geo_dist_en_z
          + C(script) + C(family) + C(task)
          + (1|language) + (1|model)
```

Stage 3 — Full model (used for all coefficient and variance-share tables in the main text).

```
score ~ 1 + joshi_class_z + syn_dist_en_z + phon_dist_en_z + geo_dist_en_z
          + C(script) + C(family) + C(task)
          + fertility_z + repr_sim_en_z
          + (1|language) + (1|model)
          + (1|model:language) + (1|task:language) + (1|task:model)
          + (0 + fertility_z|model)
```

Figure 2: The three nested model specifications fit in this work, in `bambi`/`lme4` formula syntax. Stage 2 adds structured language-level covariates on top of the Stage 1 baseline; Stage 3 adds model-conditional covariates (`fertility_z`, `repr_sim_en_z`), two-way random interactions, and a per-model random slope on fertility.

Feature	Source	Type
Language-level features		
`script`	Hardcoded ISO-1 → script map	Categorical (15 levels)
`family`	Hardcoded ISO-1 → family map	Categorical (17 levels)
`resource_cla`	Joshi et al. (2020) resource class	Integer 1–5 → std.
`syn_dist_en`	`lang2vec` syntactic distance to English	Float → std.
`phon_dist_en`	`lang2vec` phonological distance to English	Float → std.
Model × language features		
`fertility`	`compute_fertility` — tokens-per-word for the model's tokenizer on a language sample	Float → std.
`repr_sim_en`	`compute_repr_sim` — cosine similarity of hidden-state geometry for a language to English	Float → std.

Table 11: Features used per language and model, grouped by scope (language-level, model × language). All continuous features are standardized ($\mu = 0$, $\sigma = 1$) before use.

is used directly in (Khanuja et al., 2023; Ojo et al., 2025; Ahuja et al., 2023) among others.

Typological distances to English. We use the syntactic, phonological, and geographic distances from `lang2vec`/URIEL (Littell et al., 2017), each standardised before entering the regression. These are the three least mutually redundant distance families in URIEL: syntactic distance summarises word-order and morphology, phonological distance summarises sound inventory (relevant to subword segmentation), and geographic distance captures areal/contact effects orthogonal to genealogy. We exclude the URIEL *genetic* distance because it duplicates the categorical `family` factor we include separately. We considered raw WALS features, Glottolog family identifiers only, but rejected them due to lack of a continuous distance, and incomplete language coverage respectively. URIEL/lang2vec is the dominant cross-lingual distance toolkit in multilingual transfer (Lauscher et al., 2020).

Script and family as categorical factors. We include $\mathrm{script}(\ell)$ and $\mathrm{family}(\ell)$ as unordered categorical factors with a coarse grouping (7 scripts; 8 families on our panel). Categorical coding is the natural choice as scripts (Latin, Cyrillic, Devanagari, Arabic, CJK, Ge'ez, Brahmic-other) and families (Indo-European, Sino-Tibetan, Afro-Asiatic, Niger-Congo, Dravidian, Turkic, Austronesian, Uralic) admit no defensible ordering. Despite the URIEL distances already encoding typology, we keep both: script captures engineering-level effects (subword vocabulary coverage, romanisation availability) that no continuous distance reflects, and family captures pretraining-data clustering, since languages of the same family typically co-occur in web crawls. We considered a binary Latin / non-Latin coding and a per-script random intercept; the former collapses Devanagari, Arabic, and CJK into one bucket and hides script-specific effects we observe empirically, while the latter is not data-efficient at our 7-script panel.

C.2 Model–language covariates

Tokenizer fertility. We define fertility as the mean number of subword tokens per whitespace-separated word, computed on FLORES-200 (see corpus discussion below). Fertility is the most widely used tokenizer-quality proxy in multilingual NLP and has documented downstream effects on both accuracy and cost (Rust et al., 2021; Ahia et al., 2023; Petrov et al., 2023). We considered bytes-per-token and the tokenizer parity ratio (Petrov et al., 2023), and the inverse characters-per-token compression rate; we chose fertility because it preserves the subword-fragmentation signal that affects context utilisation, avoids the reference-language dependency of a parity ratio, and follows the definition used by (Rust et al., 2021; Ahia et al., 2023).

Representation similarity to English (`repr_sim_en`). We compute Centered Kernel Alignment (Kornblith et al., 2019) between the hidden-state matrices of an English FLORES-200 sentence and its translation in the target language, at the model's middle layer ($\lfloor L/2 \rfloor$ for an L-layer model), mean-pooled across content tokens. Cross-lingual hidden-state comparison requires a metric invariant to orthogonal rotation and isotropic scaling of the representation space and well-defined between matrices of possibly different dimension; CKA is the dominant tool that meets all three requirements and is standard in LLM-internals analyses (Kornblith et al., 2019; Wendler et al., 2024; Dumas et al., 2025; Li et al., 2025). We considered mean cosine similarity over paired tokens, SVCCA (Raghu et al., 2017), Representational Similarity Analysis,

and Procrustes alignment; we rejected them respectively because they are not rotation-invariant, require matched dimensions and are unstable at our sample sizes, offer only rank-level invariance, and depend on a paired token-level alignment that is ill-defined when source and target tokenisations differ. We summarise at the middle layer because decoder-only LLMs exhibit a U-shaped language-specificity profile – early layers dominated by surface token form, late layers by output-language reformatting, and mid-depth the most language-agnostic (Wendler et al., 2024; Dumas et al., 2025; Li et al., 2025). We mean-pool because the source and target tokenisations differ token-by-token, ruling out positional alignment; mean-pooling is the standard choice in CKA-on-LLM analyses, whereas last-token pooling is unstable on short FLORES-200 sentences.

C.3 Random-effect structure

Three two-way interactions, no three-way. We include $(1 \mid \text{model:language})$, $(1 \mid \text{bench:language})$, and $(1 \mid \text{bench:model})$ as random intercepts. The three-way $(\text{model:bench:language})$ interaction is omitted because with one observation per $(\text{model}, \text{benchmark}, \text{language})$ cell it is unidentified from the residual ε and would simply re-label noise. We use random rather than fixed interactions because a panel of 7 models, $\sim$ 18 benchmarks, and 54 languages yields hundreds of pairwise interactions: a fixed-effect specification would overfit and produce unstable per-cell estimates, whereas random-effect shrinkage regularises toward zero where data is sparse. Crossed-random-effects models of this form are standard in psycholinguistics (Baayen et al., 2008) and are explicitly recommended for LLM-evaluation panels by Ulmer et al. (2022).

Per-model random slope on fertility. We add $(0 + \texttt{fertility\_z} \mid \text{model})$ – one fertility coefficient per model – but no analogous slopes for the other covariates. Prior work (Rust et al., 2021; Ahia et al., 2023) documents that tokenizers from different families penalise different languages, so the per-model fertility effect is expected to vary; a pooled fixed effect averages these opposing slopes toward zero, which is what we observe (Section 5.3). The other covariates are properties of the language alone ($\mathbf{x}_\ell$), so per-model slopes would imply "each model has its own typology", which has no mechanistic justification and did not improve leave-one-language-out predictive accuracy in pilot fits (within ± 1 ELPD-SE). A `fertility_z` $\times$ `model_id` fixed interaction is equivalent to a random slope with a flat prior and is less stable at our sample size.

C.4 Measurement corpus: FLORES-200

We compute both fertility and `repr_sim_en` on the FLORES-200 `devtest` split (Costa-jussà et al., 2024). A parallel corpus is required because cross-lingual hidden-state comparison must hold semantic content fixed – otherwise CKA conflates language-geometry differences with content differences – and fertility is comparable across languages only when measured on the same underlying content. We chose FLORES-200 because it is professionally human-translated across 200 languages, fully covers all 54 languages in our model panel without imputation, is domain-consistent (Wikipedia news/encyclopedia), and is the standard parallel evaluation corpus in multilingual LLM work (Costa-jussà et al., 2024). We considered Tatoeba, OPUS / mC4 parallel subsets, and the NLLB seed corpus, but each is weaker on coverage, translation quality, or per-language size, the last of which would inflate CKA estimator variance.

D Per-feature coefficients of the full model

The main text (Section 5.2) reports the headline finding that `repr_sim_en` is the only continuous predictor credibly non-null in both buckets. This appendix presents the full per-feature picture numerically (Table 12), and discusses why the four non-named predictors are individually attenuated despite the high joint R^2_{ling}.

Resource level (`resource_class`) is credible on Reasoning and borderline on NLU. Resource class is credibly positive on Reasoning ($+0.037$ $[0.002, 0.07]$) and just-crosses-zero on NLU ($+0.022$ $[-0.005, 0.05]$). The full-model coefficient understates the marginal effect: in the language-features-only fit (without `repr_sim_en`), β_{resource} rises to $+0.038$ on NLU and $+0.057$ on Reasoning, both then credibly positive. The reduction in the full model is expected — resource level is itself a coarse predictor of representation alignment, and the two compete for shared variance.

Why the four typology / fertility predictors are not individually credible. Syntactic, phonolog-

ical, and geographic distance to English, together with tokenizer fertility, all have HDIs that span zero in both substantive buckets. This is the typical pattern when correlated predictors are entered jointly into a mixed model: `syn_dist_en` and `phon_dist_en` are themselves correlated with each other and with the categorical $\mathrm{family}(\ell)$ term, while `geo_dist_en` (an areal proxy) overlaps with both `syn_dist_en` and the script categoricals. The language random intercept further absorbs residual language-level structure. The pattern is collinear shrinkage of individual coefficients, not absence of feature signal: the joint R^2_{ling} remains high because the features are *collectively* explanatory.

Fertility behaves differently again. Its pooled fixed effect is washed out because its sign and magnitude vary by model, but the random per-model slope $(0 + \texttt{fertility\_z} \mid \mathrm{model})$ carries 12.1% of Reasoning variance. Fertility *does* matter, but model-specifically rather than universally. We caution against reading the borderline pooled HDIs as "the feature is uninformative": for the typology predictors the right reading is collective rather than individual explanatory power; for fertility it is the random slope, not the fixed slope, that carries the signal.

Predictor	**NLU**	**Reasoning**
`resource_class_z`	$+0.022^{\dagger}$	$+0.037^{\star}$
	$[-0.005, +0.050]$	$[+0.002, +0.070]$
`syn_dist_en_z`	-0.012	$+0.025$
	$[-0.045, +0.020]$	$[-0.015, +0.070]$
`phon_dist_en_z`	$+0.009$	$+0.007$
	$[-0.004, +0.020]$	$[-0.012, +0.030]$
`geo_dist_en_z`	-0.011	-0.007
	$[-0.041, +0.020]$	$[-0.044, +0.030]$
`fertility_z`	$-0.021^{\dagger}$	$-0.051^{\dagger}$
	$[-0.050, +0.005]$	$[-0.112, +0.007]$
`repr_sim_en_z`	$+0.065^{\star}$	$+0.094^{\star}$
	$[+0.051, +0.079]$	$[+0.074, +0.116]$

Table 12: Posterior mean and 90% highest density interval (HDI) of each standardised continuous predictor in the full model for the NLU and Reasoning buckets. For each predictor, the first row reports the posterior mean and the second row reports the corresponding 90% HDI in brackets. Markers: $\star$ = HDI excludes zero; † = HDI just crosses zero at one boundary.

Symbol / name	Formula term / source	Definition
Indices and Observation Unit		
m	Model identifier (7 models).	One of the seven evaluated LLMs (e.g., QWEN-3.5-122B, AYA-EXPANSE-32B, ...; see Table 1).
ℓ	Language identifier.	ISO-1 code. Per-bucket pools contain 63 languages for NLU, 62 for Reasoning, and 12 for NLG, restricted to languages with both `fertility` and `repr_sim_en` available.
t	Benchmark or task identifier.	After benchmark-as-task aggregation, this equals the benchmark name. Translation direction is preserved as a separate task in NLG. Counts: 9 for NLU, 6 for Reasoning, 5 for NLG.
$c = (m, t, \ell)$	Observation cell.	A single aggregated row in the fit frame, expanded to $(m, t, \ell,$ direction$)$ for NLG translation.
y_c, `score`	Outcome in $[0, 1]$.	Multiple-choice accuracy (NLU and most Reasoning tasks), `exact_match`, `flexible-extract` (GSM variants), or `chrF`/100 (NLG).
Language-Level Features (functions of ℓ only)		
`resource_class`, `resource_class_z`	Integer 1–5; z-scored.	Resource class from Joshi et al. (2020), standardised within each bucket fit frame.
`syn_dist_en`, `syn_dist_en_z`	`lang2vec` Hamming distance on `syntax_knn`.	Syntactic distance to English from URIEL (Littell et al., 2017), standardised.
`phon_dist_en`, `phon_dist_en_z`	`lang2vec` Hamming distance on `phonology_knn`.	Phonological distance to English.
`geo_dist_en`, `geo_dist_en_z`	Great-circle distance between `lang2vec` geographic centroids.	Areal/geographic distance to English used in this run.
script(ℓ)	`C(script)` categorical.	Writing system. The reference level is absorbed into β_0. NLU has approximately 18 levels; NLG has 10, of which 9 are singletons.
family(ℓ)	`C(family)` categorical.	Language family. NLU has approximately 19 levels; NLG has 2 (Indo-Aryan and Dravidian).
$\mathbf{x}_\ell$	Vector.	Stack of the four standardised numeric language features together with the script and family contrasts.
Model $\times$ Language Features (functions of (m, ℓ))		
`fertility`, `fertility_z`	Tokens per word for m's tokeniser on language ℓ.	Tokeniser inefficiency for ℓ, standardised within bucket. Paired models share tokenisers (Qwen-122B/4B; sarvam-105b/30b).
`repr_sim_en`, `repr_sim_en_z`	CKA similarity (Kornblith et al., 2019) between m's middle-layer mean-pooled hidden states for ℓ and English.	Representation alignment to English; the single most consistent predictor across all three buckets.
$\mathbf{z}_{m,\ell}$	Vector.	Stack of `fertility_z` and `repr_sim_en_z` evaluated at (m, ℓ).
Fixed Effects (population-level coefficients)		
β_0	`Intercept`, 1.	Grand mean at the reference levels of all categorical variables.
τ_t	`C(task)`, one per benchmark.	Benchmark fixed effect, carried over from Eq. 1.
γ_{resource}	`resource_class_z`.	Effect of one standard deviation on the resource axis.
$\gamma_{\text{syn}}, \gamma_{\text{phon}}, \gamma_{\text{geo}}$	`syn_/phon_/geo_dist_en_z`.	Effects of typological and areal distance to English.
γ_{fert}	`fertility_z`.	Pooled fertility effect; per-model deviations are carried by r_m below.
γ_{rs}	`repr_sim_en_z`.	Representation-alignment effect; the 90% HDI excludes zero in every bucket.
$\boldsymbol{\beta}_s^{\text{script}}, \boldsymbol{\beta}_f^{\text{family}}$	`C(script)`, `C(family)`.	Per-level contrasts against the reference script or family.
$\boldsymbol{\beta}, \mathbf{x}_i$	Stacked.	The term $\mathbf{x}_i^\top \boldsymbol{\beta}$ in Eq. 2 bundles all of the above continuous and categorical fixed effects for cell i.
Random Effects (group-level intercepts and slopes)		
α_ℓ	`(1|language)`.	Language random intercept capturing residual language signal beyond features and categoricals, with $\alpha_\ell \sim \mathcal{N}(0, \sigma_\alpha^2)$.
u_m	`(1|model)`.	Model random intercept, with $u_m \sim \mathcal{N}(0, \sigma_u^2)$.
$v_{m,\ell}$	`(1|model:language)`.	Per-(model, language) idiosyncratic boost or deficit beyond the main effects.

Continued on next page.

Table (continued).

Symbol / name	Formula term / source	Definition
$w_{t,\ell}$	`(1|bench:language)`.	Per-(benchmark, language) difficulty quirk.
$s_{t,m}$	`(1|bench:model)`.	Per-(benchmark, model) capability quirk; the dominant variance share on Reasoning (46%).
r_m	`(0 + fertility_z|model)`.	Per-model random slope on fertility, allowing fertility's effect to vary across models.
ε_c	`sigma`.	Gaussian residual, with $\varepsilon_c \sim \mathcal{N}(0, \sigma^2)$.
Variance Components (standard deviations of the random effects)		
σ_α	Standard deviation of α_ℓ in Stage 3 (or any single fit).	Size of the cross-language gap. Reported as 0.078 on NLU and 0.088 on Reasoning.
$\sigma_\alpha^{(\text{lang})}$	Standard deviation of α_ℓ in Stage 2.	Residual language SD after the language covariates are added; the numerator of R^2_{ling}.
σ_u	Standard deviation of u_m.	Drives the `model` variance share.
σ_v	Standard deviation of $v_{m,\ell}$.	The `model`×`language` variance share.
σ_w	Standard deviation of $w_{t,\ell}$.	The `task`×`language` variance share.
σ_{tm}	Standard deviation of $s_{t,m}$.	The `task`×`model` variance share.
σ_{fm}	Standard deviation of r_m.	The `fertility-slope`×`model` variance share (12.1% on Reasoning).
σ	Residual standard deviation.	Within-cell noise.
Derived Quantities Reported in the Paper		
R^2_{ling}	$1 - \left[\sigma_\alpha^{(\text{lang})}\right]^2 / \sigma_\alpha^2$ (Eq. 3).	Proportional reduction in the language random-intercept variance attributable to observable language covariates. Reported as the posterior mean with a 90% HDI.
Composite Objects Appearing in Equations 1–2		
μ	Intercept in Eq. 1.	Plays the same role as β_0 in Eq. 2: the global mean.
η_c	Linear predictor.	$\mu + \tau_t + \mathbf{x}_i^\top \boldsymbol{\beta} + \alpha_\ell + u_m + v_{m,\ell} + w_{t,\ell} + s_{t,m}$ (Stage 3 full model).
Stage 1	Baseline.	Drops the structured-features block, keeping $\mu+\tau_t+\alpha_\ell+u_m+\varepsilon$. Estimates the raw σ_α.
Stage 2	Language-features fit.	Adds $\mathbf{x}_\ell$ (the four γ terms together with $\boldsymbol{\beta}^{\text{script}}$ and $\boldsymbol{\beta}^{\text{family}}$). Yields $\sigma_\alpha^{(\text{lang})}$.
Stage 3	Full model.	Adds $\mathbf{z}_{m,\ell}$, the random interactions (v, w, s), and the per-model fertility slope r_m. Used for all coefficient and variance-share tables.

Table 13: Glossary of every variable in the Bayesian hierarchical framework. Symbols and `bambi`/`lme4` formula names are aligned with Sections 5.1, 5.2 and the modelling appendix. "ℓ" indexes language, "m" model, "t" benchmark/task, and "$c = (m, t, \ell)$" a single observation cell.

E Non-parametric cross-check of the feature ranking

This appendix provides a non-parametric sanity check on the Stage-3 feature ordering reported in Section 5.3. It is not intended as a standalone imputation method or as an independent estimator: the tree is trained on the same per-cell frame and the same feature set the Bayesian model uses, and is reported only to confirm that the hierarchy of feature contributions is recoverable from a model with very different inductive biases.

Setup. We fit a gradient-boosted regressor (`XGBoost`, 600 trees, depth 4, learning rate 0.05) on the tidy frame of (model, benchmark, language) cells. Continuous features enter standardised; model, language, task, script, and language family enter as one-hot encodings. For each of 5 random seeds and each (benchmark, resource tier) combination with ≥ 2 languages, one language is held out and its seven model rows form the test set; the held-out language never appears for that benchmark in training.

Held-out error and feature ranking. Table 14 reports per-benchmark MAE/RMSE. Pooled across held-out cells the imputation lies within 3.98 ± 0.83 points (NLU) and 6.92 ± 1.34 points (Reasoning) of the true accuracy on the $[0, 1]$ scale; the per-tier breakdown (Table 15) shows graceful degradation from Class 4 to Class 1 rather than collapse on low-resource tiers. SHAP-based feature importance (Table 16, categorical one-hots summed back to their parent feature) places REPR_SIM_EN as the largest linguistic contributor in both buckets (19.9% NLU, 29.5% Reasoning), followed by fertility and syntactic distance, with phonological distance and RESOURCE class trailing. This ranking matches the Stage-3 credible-interval ordering in Section 5.3.

Caveats. Feature selection is shared with the Bayesian model, so the agreement should be read as internal consistency rather than independent corroboration. The tree is evaluated on held-out languages from *seen* benchmarks; generalisation to entirely unseen benchmarks or model families is not tested here. We report no baselines (e.g. language-mean or RESOURCE TIER-only regressors); the held-out numbers should accordingly be read as an upper bound on what the feature set affords, not as a benchmarked imputation result.

Benchmark	MAE	RMSE
NLU		
OKAPI-MMLU	0.019 ± 0.003	0.026 ± 0.008
MMLU-INDIC-ROMAN	0.021 ± 0.008	0.025 ± 0.008
MILU	0.022 ± 0.006	0.025 ± 0.006
BOOLQ-INDIC	0.030 ± 0.010	0.040 ± 0.012
MMMLU	0.031 ± 0.011	0.038 ± 0.013
GLOBAL-MMLU	0.039 ± 0.019	0.050 ± 0.025
TRIVIAQA-INDIC-MCQ	0.047 ± 0.036	0.058 ± 0.041
INCLUDE	0.053 ± 0.024	0.065 ± 0.030
MMLU-PROX	0.097 ± 0.066	0.127 ± 0.073
NLU overall	$\mathbf{0.040 \pm 0.008}$	$\mathbf{0.063 \pm 0.020}$
Reasoning		
XCOPA	0.034 ± 0.027	0.041 ± 0.033
XSTORYCLOZE	0.063 ± 0.024	0.076 ± 0.028
BELEBELE	0.065 ± 0.027	0.078 ± 0.033
XWINOGRAD	0.075 ± 0.021	0.083 ± 0.021
GSM8K-INDIC	0.084 ± 0.038	0.102 ± 0.045
MGSM	0.094 ± 0.054	0.115 ± 0.052
Reasoning overall	$\mathbf{0.069 \pm 0.013}$	$\mathbf{0.091 \pm 0.017}$

Table 14: Per-benchmark held-out imputation error. For each of 5 seeds, one language per benchmark is held out (all 7 model rows for that language form the test set); MAE/RMSE on the $[0, 1]$ scale, mean±std across seeds. Bucket shading as in main-text tables (NLU , Reasoning).

Resource tier	n	MAE	RMSE
NLU			
1-Scraping	5	0.070 ± 0.042	0.083 ± 0.048
2-Hopefuls	45	0.040 ± 0.023	0.050 ± 0.028
3-Rising	25	0.044 ± 0.031	0.054 ± 0.035
4-Underdogs	25	0.038 ± 0.029	0.048 ± 0.038
Reasoning			
1-Scraping	5	0.106 ± 0.049	0.123 ± 0.053
2-Hopefuls	25	0.097 ± 0.034	0.115 ± 0.038
3-Rising	15	0.051 ± 0.024	0.062 ± 0.025
4-Underdogs	25	0.050 ± 0.028	0.059 ± 0.029

Table 15: Held-out imputation error stratified by the held-out language's resource tier; n is the number of held-out (language, benchmark) cells across 5 seeds. Class 5 is omitted (English is the only Class 5 language and has no within-tier holdout partner). Shading as in main-text tables (NLU , Reasoning).

Feature group	NLU	Reasoning
Linguistic / model–language features		
REPR_SIM_EN (CKA)	**19.9%**	**29.5%**
FERTILITY	6.5%	4.5%
$d_{\text{en}}^{\text{syn}}$	4.6%	5.9%
JOSHI_CLASS	3.4%	2.6%
$d_{\text{en}}^{\text{phon}}$	0.5%	0.8%
Language identity (categorical)		
language	4.3%	5.1%
family	2.3%	3.8%
script	1.8%	1.7%
Model- and task-level identifiers		
model	31.7%	28.5%
task	24.9%	17.6%

Table 16: Tree feature importance: share of total SHAP magnitude per feature group (one-hot encodings summed back to parents). Shading as in main-text variance decomposition (NLU , Reasoning); REPR_SIM_EN is the largest linguistic contributor in both buckets, reproducing the Stage-3 Bayesian ordering (§5.3).

F NLG bucket — why it is excluded from the main text

The NLG bucket is the open-generation analogue of the analysis in the main text: 413 rows of chrF scores from FLORES and IN22-conv translation and IGB-XSum summarisation, on 12 Indic and Dravidian languages, with the same seven models. The fit produces a credibly positive coefficient on `repr_sim_en_z` ($\beta = +0.053$ $[+0.025, +0.080]$), consistent with NLU and Reasoning, but the headline disparity-explanation metric R^2_{ling} is not interpretable on this bucket.

The identifiability failure. Adding the structured language features inflates rather than shrinks σ_α ($0.014 \rightarrow 0.304$, a $22\times$ expansion), which produces a strongly negative R^2_{ling}. The mechanism is categorical sparsity: of the 12 languages in the pool, 9 of 10 scripts are singletons (Bengali, Gujarati, Kannada, Malayalam, Odia, Gurmukhi, Tamil, Telugu, Arabic) and the only two language families present are Indic (8) and Dravidian (4). With this structure, the $C(\text{script})$ and $C(\text{family})$ categorical effects are collinear with the language random intercept α_ℓ for those languages, and the model cannot identify them separately.

Diagnostic implication. The variance share of the language residual is 55.5% in NLG but this is an artefact of the same identifiability failure, not a substantive finding about Indic and Dravidian languages. Reading the NLG variance decomposition as "language matters most" would be a misinterpretation.

What would unblock NLG. The fix is not methodological: it is data coverage. Adding a single non-Indic, non-Dravidian generation benchmark (e.g. FLORES en↔es, en↔zh, or en↔ar) would bring the script-family categoricals out of singleton status and restore R^2_{ling} identifiability. Until that is done, the NLG analysis should be read as "`repr_sim_en` predicts generation quality on Indic languages" rather than as a parallel disparity decomposition.

Bucket	Benchmarks	Score
NLG *(open gen.)*	IGB-FLORES (FLORES-200 translation, both dir.), IN22-CONV-16K (IN22 conv. translation, both dir.), IGB-XSUM (cross-lingual summarisation)	chrF $\in [0, 100]$, rescaled to $[0, 1]$

Table 17: NLG benchmarks (NLG); companion to Table 2. Translation source→target directions are preserved as separate modeling tasks so that TASK:LANGUAGE interactions are identifiable.

Benchmark	k (lang)	m	χ^2	p	Kendall W	sig
IGB-FLORES	29	7	148.4	1.7×10^{-29}	0.757	***
IGB-XSUM	28	7	154.4	9.3×10^{-31}	0.817	***
IN22-CONV-16K	22	7	107.2	7.8×10^{-21}	0.729	***

Table 18: Per-benchmark Friedman test on language scores with models as judges (rows = languages, columns = models, balanced via dropna) for the NLG bucket. Kendall's $W = \chi^2/(m(k-1))$ is the $[0, 1]$ effect-size analog. Companion to Table 7.

G Full Dunn's pairwise post-hoc matrices

Each table reports Dunn's pairwise post-hoc test for one (model, bucket) combination, comparing the five resource tiers (Joshi et al., 2020): 1=Scraping, 2=Hopefuls, 3=Rising, 4=Underdogs, 5=Winners. Entries are Benjamini–Hochberg FDR-adjusted p-values for the null hypothesis that two tiers have the same distribution of within-(benchmark, task_id) z-normalised per-language scores. The matrices are symmetric and the diagonal is omitted. Significance markers follow the usual convention: *** $p < .001$, ** $p < .01$, * $p < .05$, and *ns* for $p \geq .05$; p-values below 10^{-4} are written as "$< 10^{-4}$". A significant off-diagonal cell means

the two tiers differ after correcting for all ten pairwise comparisons; an *ns* cell means the data do not separate them. Read together with the omnibus Kruskal–Wallis result in Table 10, these per-model matrices show *where* on the resource ladder each model's tier ordering is actually resolved versus where adjacent tiers collapse.

	1	2	3	4	5
1	—	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***
2	$<10^{-4}$ ***	—	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***
3	$<10^{-4}$ ***	$<10^{-4}$ ***	—	.0886 ns	$<10^{-4}$ ***
4	$<10^{-4}$ ***	$<10^{-4}$ ***	.0886 ns	—	.0001 ***
5	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***	.0001 ***	—

Table 19: Qwen3.5-122B-A10B, NLU.

	1	2	3	4	5
1	—	.0311 *	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***
2	.0311 *	—	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***
3	$<10^{-4}$ ***	$<10^{-4}$ ***	—	$<10^{-4}$ ***	$<10^{-4}$ ***
4	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***	—	.0021 **
5	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***	.0021 **	—

Table 20: Qwen3.5-4B, NLU.

	1	2	3	4	5
1	—	.0003 ***	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***
2	.0003 ***	—	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***
3	$<10^{-4}$ ***	$<10^{-4}$ ***	—	$<10^{-4}$ ***	$<10^{-4}$ ***
4	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***	—	.0328 *
5	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***	.0328 *	—

Table 21: aya-expanse-32b, NLU.

	1	2	3	4	5
1	—	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***
2	$<10^{-4}$ ***	—	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***
3	$<10^{-4}$ ***	$<10^{-4}$ ***	—	$<10^{-4}$ ***	$<10^{-4}$ ***
4	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***	—	.0011 **
5	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***	.0011 **	—

Table 22: gpt-oss-20b, NLU.

	1	2	3	4	5
1	—	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***
2	$<10^{-4}$ ***	—	.5950 ns	$<10^{-4}$ ***	$<10^{-4}$ ***
3	$<10^{-4}$ ***	.5950 ns	—	$<10^{-4}$ ***	$<10^{-4}$ ***
4	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***	—	$<10^{-4}$ ***
5	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***	—

Table 23: sarvam-105b, NLU.

	1	2	3	4	5
1	—	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***
2	$<10^{-4}$ ***	—	.3295 ns	.4224 ns	$<10^{-4}$ ***
3	$<10^{-4}$ ***	.3295 ns	—	.6598 ns	$<10^{-4}$ ***
4	$<10^{-4}$ ***	.4224 ns	.6598 ns	—	$<10^{-4}$ ***
5	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***	—

Table 24: sarvam-30b, NLU.

	1	2	3	4	5
1	—	.0862 ns	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***
2	.0862 ns	—	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***
3	$<10^{-4}$ ***	$<10^{-4}$ ***	—	$<10^{-4}$ ***	.0001 ***
4	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***	—	.0830 ns
5	$<10^{-4}$ ***	$<10^{-4}$ ***	.0001 ***	.0830 ns	—

Table 25: tiny-aya-global, NLU.

	1	2	3	4	5
1	—	.8552 ns	.0142 *	.0105 *	.0260 *
2	.8552 ns	—	$<10^{-4}$ ***	$<10^{-4}$ ***	.0105 *
3	.0142 *	$<10^{-4}$ ***	—	.8007 ns	.8007 ns
4	.0105 *	$<10^{-4}$ ***	.8007 ns	—	.8007 ns
5	.0260 *	.0105 *	.8007 ns	.8007 ns	—

Table 26: Qwen3.5-122B-A10B, Reasoning.

	1	2	3	4	5
1	—	.9191 ns	.0210 *	.0010 ***	.0010 ***
2	.9191 ns	—	$<10^{-4}$ ***	$<10^{-4}$ ***	$<10^{-4}$ ***
3	.0210 *	$<10^{-4}$ ***	—	.0408 *	.0394 *
4	.0010 ***	$<10^{-4}$ ***	.0408 *	—	.2596 ns
5	.0010 ***	$<10^{-4}$ ***	.0394 *	.2596 ns	—

Table 27: Qwen3.5-4B, Reasoning.

	1	2	3	4	5
1	—	.5501 ns	.0464 *	.0015 **	.0002 ***
2	.5501 ns	—	.0088 **	$<10^{-4}$ ***	$<10^{-4}$ ***
3	.0464 *	.0088 **	—	.0071 **	.0016 **
4	.0015 **	$<10^{-4}$ ***	.0071 **	—	.0500 ns
5	.0002 ***	$<10^{-4}$ ***	.0016 **	.0500 ns	—

Table 28: aya-expanse-32b, Reasoning.

	1	2	3	4	5
1	—	.0755 ns	.9072 ns	.2782 ns	.0240 *
2	.0755 ns	—	.0004 ***	.1282 ns	.2250 ns
3	.9072 ns	.0004 ***	—	.0240 *	.0018 **
4	.2782 ns	.1282 ns	.0240 *	—	.0431 *
5	.0240 *	.2250 ns	.0018 **	.0431 *	—

Table 29: gpt-oss-20b, Reasoning.

	1	2	3	4	5
1	—	.1318 ns	.8575 ns	.1318 ns	.0268 *
2	.1318 ns	—	.0073 **	.8575 ns	.1562 ns
3	.8575 ns	.0073 **	—	.0063 **	.0063 **
4	.1318 ns	.8575 ns	.0063 **	—	.1318 ns
5	.0268 *	.1562 ns	.0063 **	.1318 ns	—

Table 30: sarvam-105b, Reasoning.

	1	2	3	4	5
1	—	.0349 *	.4451 ns	.1495 ns	.1325 ns
2	.0349 *	—	.0027 **	.1053 ns	.8418 ns
3	.4451 ns	.0027 **	—	.1495 ns	.1495 ns
4	.1495 ns	.1053 ns	.1495 ns	—	.4253 ns
5	.1325 ns	.8418 ns	.1495 ns	.4253 ns	—

Table 31: sarvam-30b, Reasoning.

	1	2	3	4	5
1	—	.6452 ns	.0209 *	.0015 **	.0002 ***
2	.6452 ns	—	.0006 ***	$< 10^{-4}$ ***	$< 10^{-4}$ ***
3	.0209 *	.0006 ***	—	.0531 ns	.0044 **
4	.0015 **	$< 10^{-4}$ ***	.0531 ns	—	.0488 *
5	.0002 ***	$< 10^{-4}$ ***	.0044 **	.0488 *	—

Table 32: tiny-aya-global, Reasoning.

H Compute

All large language model evaluations were executed using the LM-EVALUATION-HARNESS framework across a single computational node equipped with eight NVIDIA B200 GPUs. Downstream statistical estimation and posterior inference for the nested Bayesian hierarchical framework were performed subsequently, with the sampling process requiring approximately 1.5 hours of compute time per modelling stage.