

Tokenization Sets the Measurement Scale for Cross-Lingual Diagnostics

Arij Riabi * Valentin Noël *

Devoteam

Abstract

Multilingual evaluation treats translated passages as the same input in different languages, and increasingly reads an internal diagnostic off the model while doing so. Tokenization can make that assumption false before the model answers, because the same passage occupies several times as many token positions in one language as in another, so a fixed window gives the model different amounts of text and gives the attention graph behind the diagnostic a different number of nodes. We derive a comparability criterion from the descriptors' own definitions: on a parallel corpus the tokenizer sets the graph order, so a descriptor whose attainable range grows with graph order can separate languages while measuring segmentation rather than model computation, and because those ranges are analytic the criterion can be checked before any measurement is taken. Replaying a standard analysis against the audited one changes the answer, since most of its significant associations do not survive and about one in five reverse sign. The failures concentrate in the descriptors the criterion flags in advance rather than in spectral descriptors as such, and the same split appears among interpretability descriptors that use no graph at all. Applied to high frequency energy over 60 Bebele languages and three tokenizer matched Llama models, the audit leaves the resource associated profile varying across depth and withdraws the stronger readings the raw profiles invite, among them a stable peak layer, a replicated profile shape, a stable mid-depth effect and a causal mechanism. The practical consequence is that a cross-lingual comparison should establish that its measurements are in comparable units before it reports a difference.

1 Introduction

Multilingual benchmarks often compare translated items as if each translation were the same input expressed in another language. That is the right ideal,

*These authors contributed equally.

but it is not automatically true inside a language model. Before the model reasons, the tokenizer has already decided how much text fits in the context window and how many token nodes an internal graph diagnostic will see.

The scale mismatch can be large. In our data, the same passage occupies 136 token positions in English and 1569 in Burmese. With a fixed token budget, the English example gives the model most of the passage while the Burmese example gives it only a short prefix. With a whole passage analysis, the diagnostic is computed on graphs of very different order. The central question of this paper is therefore not whether a new metric can separate languages, but whether the metric is being measured in the same units in the first place.

This input inequality is well documented. Subword vocabularies allocate several times more tokens to some languages than to others for identical content (Petrov et al., 2023; Ahia et al., 2023). Tokenizer quality explains a substantial share of monolingual performance gaps (Rust et al., 2021; Arnett and Bergen, 2025). The consequences include inflated API cost, reduced effective context, and weaker performance for low-resource languages (Ahia et al., 2023; Wu and Dredze, 2020; Blasi et al., 2022). Mitigation work changes the encoding and evaluates whole model outcomes (Limisiewicz et al., 2024; Arnett et al., 2025). What remains less clear is whether internal measurements computed from translated inputs remain comparable when tokenization density changes across languages.

Graph signal processing (Shuman et al., 2013; Ortega et al., 2018) is a useful case because it supplies both the diagnostics and a way to inspect their units. Each attention layer defines a weighted graph over tokens, and spectral descriptors summarise how residual stream signals vary over that graph. Our empirical application focuses on the high-frequency-energy ratio, HFER, while the audit applies to the broader descriptor family. These

graph-theoretic readings have already been used to study rank collapse, over-smoothing and training free diagnostics (Dong et al., 2021; Shi et al., 2022; Noël, 2026).

The useful feature is not that the descriptors are spectral. It is that their ranges are written down. HFER is a fraction of spectral mass and has the same range for any N , whereas an entropy over N modes is bounded by $\log N$. That difference turns a loose worry about tokenization into a checkable criterion. It also exposes a contrast with common residual stream baselines: effective rank has an N -dependent ceiling, while residual norm is unbounded, so the range criterion makes no promise about it.

Cross-lingual use of these descriptors is therefore not automatic. In many monolingual analyses, sequence length is a nuisance variable. In a parallel multilingual corpus, sequence length is partly a property of the tokenizer and therefore part of the measurement. Our audit first asks whether a descriptor is on a shared mathematical scale. It then asks whether the comparison changes under node refinement and how much source content remains inside a matched token window. Matching token counts fixes graph order, but it does not make the inputs equally informative: a fixed budget still covers much more of a densely encoded translation than of a heavily fragmented one.

The paper is organised around this measurement problem, and the HFER study is the application that stress tests it. We ask three questions: **RQ1:** which spectral descriptors remain comparable when token counts differ across languages? **RQ2:** after fixing graph order and using within-family permutation nulls, does the resource associated HFER profile vary across depth? **RQ3:** where do spectral descriptors sit relative to simpler residual stream baselines? Do they predict modelling quality better, and if not, what do they provide that the baselines do not?

This matters most where labelled reasoning benchmarks are unavailable, do not cover the relevant languages, or leave small base models near chance: XCOPA (Ponti et al., 2020), for instance, covers nine of the sixty languages studied here. The usual response is to build the missing supervision (Riabi et al., 2021b), which an unlabelled audit does not require.

Contributions.

- **Audit.** We give a comparability audit for cross-

lingual spectral diagnostics, built on a range based criterion, ceiling normalisation for five variable range descriptors, a refinement check and a coverage report. It imposes three constraints. Variable- N descriptors go on a shared scale, by fixed range or by division by the analytic ceiling. Matched- N probes are treated as coverage limited, with the coverage reported. Resource axis claims are tested against nulls that preserve language family.

- **Why it matters.** Replaying a standard analysis against the audited one changes the answer: most significant associations disappear, and roughly one in five change direction. The failures concentrate in the descriptors that the criterion flags before the experiment is run. The same split appears among non-graph interpretability descriptors in the Llama pair, so the problem is not specific to spectral analysis.
- **Artifact.** Code, the pre-registration log and the per-language measurements are released,¹ with every table checked against the recorded results (Appendix A).
- **Application.** We apply the audit to HFER in a tokenizer matched Llama-3.2-1B/3B pair and a Llama-3.1-8B extension, then repeat the dispersion null on two further tokenizer families. The robust finding is depth heterogeneity of the resource associated profile. Stronger readings, such as a specific peak, a stable mid-depth association or equal magnitude across model sizes, are not supported by the checks.
- **Use case.** We provide a label free audit for low-resource evaluation settings and relate it to modelling quality through bits per character and a task grounded margin. Residual stream baselines predict modelling quality better and more cheaply. The spectral descriptors answer the different question of whether, and at which depths, a resource association remains once graph order is controlled.

2 Related Work

Tokenization and cross-lingual disparity. Multilingual LMs inherit BPE, WordPiece and SentencePiece variants (Webster and Kit, 1992; Mielke et al., 2021; Gage, 1994; Sennrich et al., 2016; Schuster and Nakajima, 2012; Kudo, 2018; Kudo

¹github.com/vcnoel/tokenization-measurement-scale

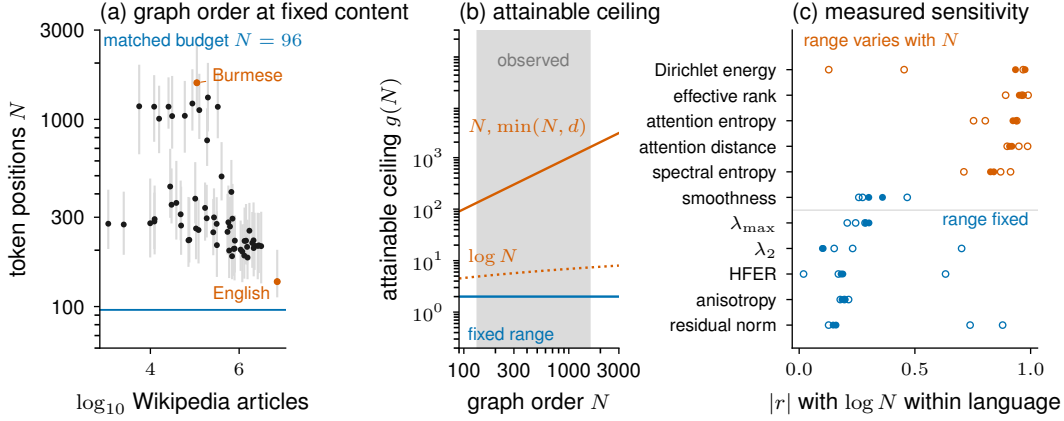


Figure 1: **The tokenizer sets the graph order, and the graph order sets what a descriptor can attain.** (a) Token positions per passage against the resource axis, 60 languages, Llama-3.2 tokenizer, content scope, mean in ink with the per language minimum to maximum range in grey, matched budget in blue. (b) Attainable ceiling $g(N)$ against graph order, with the observed range of N shaded. (c) Mean within language $|r(\text{descriptor}, \log N)|$, content scope, filled markers for the two Llama models and open markers for BLOOMZ-3B and Qwen2.5-1.5B, descriptors ordered by the Llama value and split at the criterion of Proposition 1.

and Richardson, 2018), which scale but are not typologically neutral: they over-fragment some scripts, underrepresent rare forms and are hard to adapt (Shao et al., 2018; Amrhein and Sennrich, 2021; Pruthi et al., 2019; Sun et al., 2020; Schick and Schütze, 2020; El Boukkouri et al., 2020). Consequences appear in token counts, cost, context and low-resource performance, and mitigations change the encoding (Petrov et al., 2023; Ahia et al., 2023; Rust et al., 2021; Arnett and Bergen, 2025; Wu and Dredze, 2020; Blasi et al., 2022; Limisiewicz et al., 2024; Arnett et al., 2025; Clark et al., 2022; Xue et al., 2022). Riabi et al. (2021a) evaluate the character level alternative directly, for low resource and noisy varieties where subword segmentation is least stable. Zouhar et al. (2023) give Rényi efficiency as an intrinsic quality measure, which we use as the treatment. Low-resource status is itself shaped by artefacts, sociopolitical conditions and within language variation, and transfer varies with script, typological proximity and corpus size (Hedderich et al., 2021; Nigatu et al., 2024; Bird, 2020; Hovy and Yang, 2021; Pires et al., 2019; Muller et al., 2021; Lauscher et al., 2020). Lopetegui et al. (2025) locate that variation inside a single language, across Spanish varieties, which is the scale at which a per language resource proxy stops resolving. Joshi et al. (2020) supply the taxonomy behind our resource proxy, and our outcome follows the compression view (Delétang et al., 2024).

Culture aware multilingual evaluation. Cross-cultural NLP has warned that language coverage

is not cultural coverage (Hershcovich et al., 2022): models track values weakly and misrepresent cultures in low-resource settings, and situated adaptation makes the same point from the intervention side (Bhatt and Diaz, 2024; AlKhamissi et al., 2024; Sukiennik et al., 2025; Rao et al., 2025; Bhagat et al., 2026; Havaladar et al., 2025; Cao et al., 2024; Liao et al., 2025; Pandey et al., 2026; Riabi et al., 2026). We ask a prior question: whether translated inputs induce comparable internal units at all.

Graph signal processing (Shuman et al., 2013; Ortega et al., 2018) supplies our formalism. Prior work on rank collapse, over-smoothing and anisotropy (Dong et al., 2021; Shi et al., 2022; Noël, 2026) all works within a single tokenizer (Appendix K).

3 Method

3.1 Attention graphs and spectral descriptors

Let $A^{(\ell)} \in \mathbb{R}^{N \times N}$ denote the head averaged attention matrix at layer ℓ over N content tokens, $W = \frac{1}{2}(A + A^\top)$ its symmetrisation, and D the corresponding degree matrix. We use the symmetric normalised Laplacian

$$L = I - D^{-1/2} W D^{-1/2}, \quad (1)$$

with eigendecomposition $L = U \Lambda U^\top$ and $0 = \lambda_1 \leq \dots \leq \lambda_N \leq 2$. Taking the residual stream $X^{(\ell)} \in \mathbb{R}^{N \times d}$ as the graph signal, the graph Fourier transform is $\hat{X} = U^\top X$. We consider

eleven descriptors, of which the principal one is the high-frequency energy ratio

$$\text{HFER}(\mathbf{X}) = \frac{\sum_{m > (1-\rho)N} \|\hat{\mathbf{X}}_{m,\cdot}\|_2^2}{\sum_{m=1}^N \|\hat{\mathbf{X}}_{m,\cdot}\|_2^2}, \quad (2)$$

the share of signal energy in the top ρ fraction of modes, with $\rho = 0.1$ throughout.

The eleven descriptors comprise six Laplacian quantities, two attention quantities and three residual stream baselines. The key boundary is whether a descriptor’s attainable range depends on N . Effective rank, for example, is a residual stream quantity with range $[0, \min(N, d)]$, not a graph statistic. This is why the largest association in §5.2 is a baseline result. The complete taxonomy appears in Appendix E.

3.2 A comparability criterion

Consider a parallel corpus in which passage meaning is held constant and only the tokenizer varies. A more fertile tokenizer refines the node set, dividing nodes into a larger number of nodes carrying the same text. Write $N = \phi M$, where M is a tokenizer independent measure of content length such as character count and ϕ is fertility.

Proposition 1 (Non-identification under node set refinement). *Let f be an attention graph functional with attainable range $[0, g(N)]$ for strictly increasing g . Then on a parallel corpus a difference in f between two languages may be produced by a difference in fertility ϕ alone, with no difference in graph structure. The units of f are a function of the tokenizer, and a cross-lingual comparison of f is therefore not identified. Functionals whose attainable range is a fixed interval independent of N , in particular those defined on a fixed fraction of the spectrum, are not subject to this non-identification.*

The argument is one of units rather than of evidence. If f attains at most $g(N)$, and $N = \phi M$ with M fixed by the parallel design, then the ceiling of f is itself a function of the tokenizer, and the values of f for two languages are not expressed on a common scale, and covariate adjustment cannot recover a comparison between quantities measured on different scales. A fixed range is therefore a first requirement, not a full guarantee: it removes this range based non-identification, while leaving open ordinary tokenizer effects on the value of f .

Figure 1 shows the three facts the criterion rests on: the tokenizer sets N at fixed content, the ceiling $g(N)$ of a descriptor follows from its definition,

and the two classes separate on measured sensitivity to N in the Llama models and not in the other two architectures. Proposition 1 partitions the descriptor set before observation. Spectral entropy, effective rank, attention distance, attention entropy and Dirichlet energy have N -dependent ceilings, while HFER, λ_2 , $\lambda_{\max}$ and the smoothness index are fixed range under the normalised Laplacian. In the Llama pair the partition is empirically recovered with no overlap: variable range descriptors correlate more than twice as strongly with $\log N$ within language as fixed range ones. That recovery is a property of those models, not of the criterion, and does not hold on the two architectures of Appendix C. This is a units claim: a descriptor can be mathematically comparable and still covary with fertility.

Stability under node refinement. We therefore add a refinement invariance check. Holding a synthetic graph’s coarse structure fixed while splitting nodes shows that fixed range is not enough: HFER and λ_2 drift 114% and 83%, while $\lambda_{\max}$ drifts 0.7%. The drift is generator dependent and therefore not removable by a monotone calibration. It also runs against the reported resource association: refinement raises HFER, whereas an uncorrected artefact would flatten our positive resource slope rather than manufacture it. The sweep and generator checks are in Appendix C.

Our main results are in any case reported in the matched scope, where N is held at 96 and no refinement occurs across the comparison. For natural length comparisons of HFER or λ_2 , satisfying Proposition 1 is not sufficient, and the matched node design avoids that additional refinement dependence.

$\lambda_{\max}$ and residual norm illustrate the one directional nature of the criterion: their scales are shared across languages, but their values can still covary with fertility. Both are nearly insensitive to graph size within a language yet strongly associated with N across languages, so they satisfy the range criterion while still carrying between language variation. Comparability is necessary for cross-lingual interpretation, not sufficient.

3.3 Ceiling normalisation

For excluded descriptors, the criterion prescribes division by the analytic ceiling $g(N)$. Ceiling normalisation reduces the median discrepancy between the content and matched resource slopes

from 0.510 to 0.057, close to HFER at 0.063, and improves all ten descriptor model cells we test. Fixed range descriptors such as $\lambda_{\max}$ and residual norm are deliberately left uncorrected.

Scope of the two checks. The resulting two-stage gate is: when graph orders differ, fixed range or ceiling normalisation places descriptors on a shared scale. For natural length comparisons, refinement stability also enters the interpretation. When N is explicitly matched, the range problem disappears, but the coverage confound in §3.4 remains. The two checks disagree most clearly on effective rank: cross scope normalisation repairs its fitted resource slope, but the refinement test still shows value level drift, so we use it only in matched scope.

3.4 Token count matching and its residual confound

Each model is evaluated under two scopes. In content scope, passage meaning is fixed and N varies. In matched scope, $N = 96$ is fixed and the varying quantity is how much of the passage the budget covers. The standard reasoning is that an effect present only under content scope is a fertility artefact, whereas an effect persisting under matched scope cannot be attributed to graph size.

That reasoning is incomplete. Matching converts the graph size confound into a coverage confound: English covers 70.6% of a passage within 96 tokens, Burmese 6.1%, and coverage correlates at +0.584 with the resource axis. We control coverage in binned analyses. The per-layer profiles remain unadjusted matched scope resource profiles, testing depth heterogeneity after graph order is fixed. Because coverage is collinear with the resource axis and other controls, the binned coverage check uses a minimal specification and reports the saturated one alongside. They agree except on the 3B mid-depth cell.

4 Experimental Setup

Corpus. We use Belebele (Bandarkar et al., 2024), whose `flores_passage` field supplies FLORES-200 passages (NLLB Team, 2022, 2024). We verify item alignment by joining on (`link`, `question_number`) because the per-language files are not uniformly row aligned. The matched budget is set below English passage length, because otherwise matching discards high-resource items first. English passages span 32 to 277 tokens (median 94), so we sample 50 from the 153 passages

of at least 112 English tokens under a fixed seed.

Languages and resource axis. The inventory comprises 60 languages, 22 scripts and 24 family branches. We use Wikipedia article count as the unlabelled resource proxy from Joshi et al. (2020). It validates against both available anchors, with Spearman $\rho = 0.924$ against an a-priori resource band and $\rho = 0.890$ against the Joshi class on the 18 named languages. Cebuano, Waray and `als_Latn` are excluded for documented corpus/proxy mismatches (Appendix N).

Models, estimand and implementation. We evaluate a tokenizer matched Llama-3.2-1B/3B pair, released September 2024, and add Llama-3.1-8B, released July 2024, as a tokenizer matched extension. Tokenizer byte identity is verified over vocabulary, merges and token id sequences, so differences among profiles cannot be attributed to tokenizer vocabulary. The 8B extension changes model version, capacity and depth, so it is a matched scope generalisation check rather than a capacity only comparison. The estimand is β , the standardised per-layer slope on $\log_{10}$ article count with language clustered standard errors, since the regressor is constant within language and the effective sample is 60. Special tokens are removed before graph construction. The 1B/3B content scope sweeps cap graphs at $N = 1024$, truncating 13.2% of content scope rows and no matched scope rows. The truncation is concentrated by fertility rather than spread evenly, and the 8B run is matched scope only at $N = 512$.

5 Results

5.1 Depth heterogeneity of the resource associated HFER profile

The main empirical question is whether the resource association looks the same at every layer once graph order is fixed. It does not. The resource axis slope on HFER ranges from -0.34 to $+0.68$ across depth in the 1B, -0.65 to $+0.69$ in the 3B, and -0.61 to $+0.77$ in the 8B. Free permutation recreates about a fifth of that spread, whereas within-family relabelling recreates 58 to 71% of it (Table 1). A standardised slope of 0.4 means that one order of magnitude more Wikipedia text is associated with a 0.4 standard deviation change in high-frequency energy. In raw units mid-depth HFER runs from 0.043 in Xhosa to 0.060 in Polish across roughly three orders of magnitude of

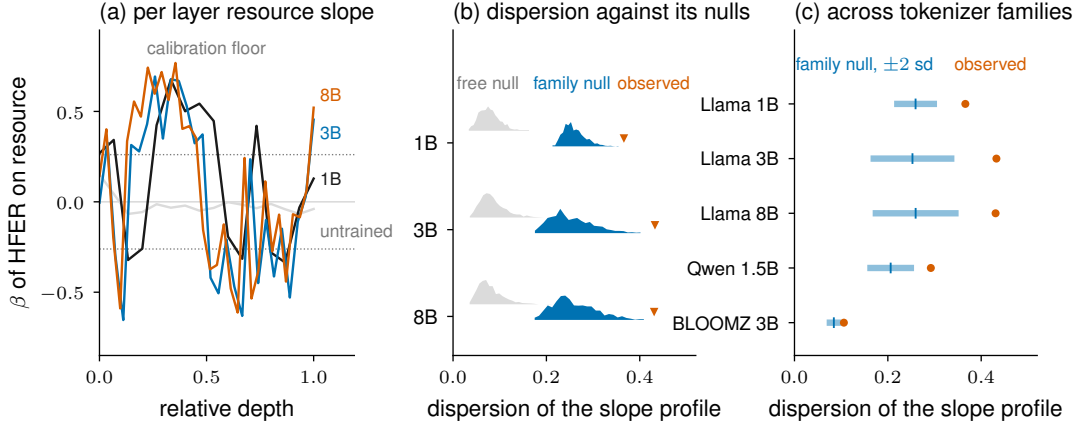


Figure 2: **The resource associated HFER profile varies across depth, and its dispersion clears a family preserving null in five models.** (a) Unadjusted matched scope per-layer slope against relative depth, the untrained Llama-3.2-1B control in grey, dotted lines at the ± 0.261 calibration floor that permuted labels reach at the 95th percentile. (b) Dispersion of the per-layer slope vector against 1000 free and 1000 within-family permutation draws, densities scaled to a common height, observed value as a marker. (c) Observed dispersion against the within-family null mean and two null standard deviations, for the three Llama models and the two further tokenizer families.

resource availability.

Main result. The unadjusted matched scope resource association is heterogeneous across depth (Figure 2). Dispersion of the per-layer slope vector exceeds all 1000 family preserving null draws in all three models ($p_{\text{perm}} = .001$, Table 1). We report dispersion rather than a maximum because the profiles are multimodal and neighbouring layers are dependent.

The 8B extension preserves and narrows this conclusion. Its profile dispersion is 0.431, lies 3.72 null standard deviations beyond the within-family null, has descriptive $I^2 = 95.8\%$, and is 60% family reproducible. The standardised margin over the family null decreases across the 1B/3B/8B sequence (4.63, 4.00, 3.72), but the gap between the observed value and the null remains comparable (0.107, 0.180, 0.172). What does not persist is the mid-band association: the slope falls from $+0.395$ to $+0.108$ to -0.015 , is inside the 8B family null ($p_{\text{perm}} = .653$), and shrinks under central windows of 3, 5, and 7 layers. The cross model finding is depth heterogeneity, while the apparent mid-band association is model specific.

Free and family preserving nulls. Free permutation breaks family structure as well as resource level. Holding family fixed raises the reproduced share from about 20% to between 58 and 71% for dispersion, puts the 3B and 8B mid-band slopes inside their family nulls, and sharply narrows the claim. Script is the weaker stratum.

Table 1: **Dispersion of the per-layer HFER slope against nulls of increasing severity.** Matched scope, unadjusted per-layer resource slopes, 1000 draws per null. The repro. column is the null mean as a fraction of the observed value, and shaded rows bound the claim.

Null	Statistic	obs.	null mean $\pm$ sd	p_{perm}	repro.
<i>Llama-3.2-1B</i>					
free	dispersion (SD)	0.366	0.0782 $\pm$ 0.0224	.001	21%
family	dispersion (SD)	0.366	0.2594 $\pm$ 0.0230	.001	71%
family	mid-band slope	0.395	0.3335 $\pm$ 0.0250	.012	84%
descriptive I^2					94.9%
<i>Llama-3.2-3B</i>					
free	dispersion (SD)	0.433	0.0845 $\pm$ 0.0270	.001	20%
family	dispersion (SD)	0.433	0.2528 $\pm$ 0.0449	.001	58%
family	mid-band slope	0.108	0.0727 $\pm$ 0.0224	.063	68%
descriptive I^2					96.3%
<i>Llama-3.1-8B</i>					
free	dispersion (SD)	0.431	0.0796 $\pm$ 0.0266	.001	18%
family	dispersion (SD)	0.431	0.2595 $\pm$ 0.0460	.001	60%
family	mid-band slope	-0.015	-0.0043 $\pm$ 0.0244	.653	28%
descriptive I^2					95.8%

Further readings checked by the nulls. Binned coverage controlled analyses support the 1B/3B mid-depth association but not a stable mid-band claim, which the 8B extension removes. The untrained Llama-3.2-1B control does not reproduce the association: its mid-band slope is -0.028 against $+0.395$ in the trained 1B. Peak position is not identified: the 1B and 3B profiles peak at relative depth 0.364, but within-family relabelling reproduces that peak in 95% of draws and the agreement does not survive reinterpolation. Depth binning obscures layer specific variation by cancelling slopes of opposite sign.

Table 2: **Three descriptors carry both legs of the tokenizer efficiency to modelling quality path.** Shaded rows have both a and b significant after correction. Smoothness and λ_2 are shown because they are natural spectral comparators. ab_{beh} repeats the decomposition with the character normalised comprehension margin, and $\dagger$ marks a row that does not clear correction in that behavioural family.

Mediator	a	b	ab [95% BCa]	ab_{beh}
eff. rank	+0.646	-0.377	-0.244 [-0.478, -0.072]	+0.218 $\dagger$
λ_{max}	-0.629	-0.215	+0.135 [+0.037, +0.245]	-0.130
HFER	+0.145	-0.248	-0.036 [-0.077, -0.007]	+0.024
smoothness	-0.196	-0.151	+0.030 [+0.001, +0.100]	-0.024 $\dagger$
λ_2	+0.014	-0.226	-0.003 [-0.117, +0.074]	+0.018 $\dagger$

5.2 Association with language modelling quality

Belebele accuracy is not a usable outcome for base models at this scale: chance is 0.250, Llama-3.2-1B attains 0.288 ± 0.037 in English, and only 14 of 60 languages exceed chance significantly. We therefore use bits per character as the primary modelling quality proxy and validate it against the character normalised log-likelihood margin between the correct option and its strongest distractor. The margin is task grounded and correlates with bits per character at $r = -0.92$ after partialling out the resource axis. Per-token NLL is not used because it rewards the fertility under study (Delétang et al., 2024). Bits per character is recoverable from the sweep’s own NLL and validates against direct measurement at $r = 0.9992$.

The decomposition is associative, not causal. The exposure is Rényi tokenizer efficiency (Zouhar et al., 2023), the candidate mediators are matched scope descriptors, and the outcome is bits per character. For layer resolved descriptors we use a fixed mid-depth summary rather than selecting a peak layer. Inference is clustered by language, with multiplicity correction on the a and b legs and descriptive BCa intervals for ab . For truncated items, the likelihood denominator follows the scored prefix. Dropping them would correlate with the resource axis and remove every Burmese item.

Three descriptors carry both legs after multiplicity correction: residual stream effective rank, λ_{max} , and HFER (Table 2). The behavioural margin preserves the ordering, but the largest association belongs to the residual stream baseline: effective rank carries $ab = -0.244$ against HFER’s -0.036 and requires no graph or eigendecomposition. λ_{max} enters with the opposite sign to the total effect,

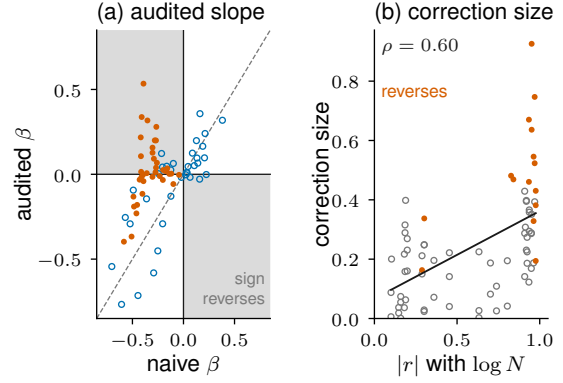


Figure 3: **The audit moves the conclusion furthest where the criterion says it will.** The 73 associations a naive analysis reports as significant, over Llama-3.2-1B/3B and BLOOMZ-3B, three depth bands, eleven descriptors. (a) Audited against naive slope, filled markers for descriptors whose attainable range varies with N and open markers for fixed range ones, the identity dashed, the two shaded quadrants holding the 15 cells whose sign reverses. (b) Correction size $|\beta_{\text{naive}} - \beta_{\text{audited}}|$ against the descriptor’s measured within language sensitivity to graph order, reversing cells marked, least squares fit in ink, Spearman $\rho = 0.60$ ($p = 2 \times 10^{-8}$).

so we report suppression rather than a mediated proportion.

5.3 Effect of the audit on reported associations

An audit is worth running only if skipping it changes the answer. We therefore compare, descriptor by descriptor, the conclusion a standard analysis reaches against the conclusion the audited analysis reaches. The *naive* arm is unaudited analysis: whole passages, raw descriptors, a regression of the descriptor on the resource axis. The *audited* arm applies the criterion: matched graph order, ceiling normalisation where the range is variable, coverage controlled, language clustered standard errors. A cell is one (model, descriptor, depth band), over the three models for which both scopes are available.

Of the 73 associations the naive analysis reports as significant, 42.5% survive. The rest divide into 37.0% that disappear and 20.5% that reverse sign: on those, a standard analysis would report a real association in the wrong direction.

The more useful result is that the criterion says in advance which conclusions are at risk (Table 3 and Figure 3). Reversals concentrate in the descriptors whose attainable range depends on N : 34.2% of their naive conclusions reverse, against 5.7% of

fate of the naive conclusion	attainable range	
	variable	fixed
survives (same sign, significant)	23.7%	62.9%
reverses sign	34.2%	5.7%
disappears	42.1%	31.4%
does not survive	76.3%	37.1%
naive conclusions (n)	38	35

Table 3: **The criterion screens conclusions before they are drawn.** Fate of every association a naive analysis would report as significant at $\alpha = .05$ without a within language control, split by whether the criterion flags the descriptor in advance. Flagged descriptors reverse sign six times as often (13/38 against 2/35, Fisher exact $p = .003$) and fail to survive twice as often ($p < .001$). Llama-3.2-1B/3B and BLOOMZ-3B, three depth bands, eleven descriptors. BLOOMZ’s own partition fails (Table 4), but dropping it widens the contrast to 42.3% against 7.1% ($p = .004$). Cells are not independent and the test is indicative (Appendix C).

fixed range ones, and three quarters of them fail in some way. This is what distinguishes a criterion from a post-hoc account of which results happened to move. Because the descriptors are defined analytically on the spectrum, the partition is fixed before any measurement is taken, and it is that derivation, not the empirical outcome, that does the screening.

5.4 Robustness and negative results

Eleven broad falsification analyses and additional pre-specified adversarial checks withdrew the late-depth, stable mid-band, peak location and profile shape claims while preserving depth heterogeneity under the family aware dispersion null (Appendix L). Four negative results further constrain the paper: residual stream baselines outperform spectra for prediction ($R^2 = 0.799/0.773$ versus $0.683/0.741$), Heaps’ exponent is not scale-free ($r = -0.96$ with fertility), tokenizer features are nearly one-dimensional (PC1=81.4%), and only dispersion generalises across model sizes. Equality of the 1B/3B mid-depth slopes is rejected ($p = .021$), sign concordance does not exceed its null ($p = .66$), and peak position is reproduced by within-family relabelling in 95% of draws.

Across tokenizer families. The Llama ladder holds the tokenizer fixed by construction, which identifies the capacity comparison but leaves open whether the depth heterogeneity result is a property of one segmentation. We therefore repeat the family aware dispersion null on two further vocab-

ularies: BLOOMZ-3B (ROOTS, ALiBi, instruction tuned, November 2022) and Qwen-2.5-1.5B (RoPE, base weights, September 2024). Dispersion clears the null in both: Qwen at 0.292 against 0.206 ± 0.025 and BLOOMZ at 0.106 against 0.084 ± 0.008 , margins of 3.44σ and 2.79σ against 4.63 , 4.00 and 3.72 for the triple. The direction replicates in three independent tokenizer families, the magnitude does not: BLOOMZ’s dispersion is roughly a quarter of the Llama values with 80% reproduced by within-family relabelling, a weak replication that also confounds vocabulary with positional encoding and instruction tuning.

6 Discussion

Contribution of the audit. A low-resource language can appear worse for at least three different reasons: the model may reason less reliably in that language, the translation may expose less of the passage within the same token budget, or the internal diagnostic may be measured on a different scale. The first is a task claim. The latter two are measurement claims, and our contribution is to make them explicit before downstream score gaps are interpreted.

Descriptor scope. The descriptors separate prediction from depth resolved diagnosis. For prediction, residual stream baselines are stronger and cheaper: anisotropy better predicts language resource level, and effective rank carries the largest association with modelling quality. Spectral descriptors answer a different question, revealing whether and how a resource association varies across layers and graph frequencies once graph order is controlled, which is why the association decomposition and the depth profile point to different lessons. For variable range descriptors, ceiling normalisation repairs matched scale comparability. For fixed range descriptors, no divisor is warranted.

Scope of the criterion beyond spectral descriptors. It would be easy to read this paper as a catalogue of ways graph descriptors fail, and that reading is wrong on the evidence. The criterion partitions descriptors by whether their attainable range depends on N , and that partition cuts across the graph boundary rather than along it. Tested as a classifier against within language sensitivity to graph size, it separates the two groups with no overlap in the Llama pair, though that recovery is a property of those models and does not hold on the archi-

tures of Appendix C. It keeps separating them when the graph descriptors are removed: among non-graph quantities alone the variable range ones sit at $|r| \geq 0.75$ against ≤ 0.38 for fixed range ones. Three of the descriptors it flags are standard interpretability tools with no eigendecomposition.

The audited population is therefore not those computing spectra on attention graphs, but anyone comparing an N -dependent internal measurement across unevenly tokenized languages. What graph signal processing supplies is what makes such a rule possible at all: descriptors with closed-form definitions, whose ceilings can be derived rather than estimated, and whose repair, division by that ceiling, follows from the same definition. A quantity introduced without an analytic range admits no equivalent check, which is why the residual stream baselines that outperform spectra at prediction cannot say whether their own cross-lingual comparison is meaningful. The spectral formalism is the instrument here, not the defendant.

Effect of the family preserving null. The family aware null is not a cosmetic robustness check. In this sample, resource level and family structure are strongly entangled, so a free permutation tests against no language level signal rather than against no resource axis signal. Preserving family reproduces most of the observed profile dispersion and almost all of the apparent peak position. The 8B extension makes the same point from the model side: dispersion persists when depth and capacity change, the mid-band association does not. The claim is therefore phrased as depth heterogeneous and resource associated, not as a stable fragmentation mechanism.

Cost and low-resource applicability. A depth profile for one language costs a forward pass over 50 parallel passages and one eigendecomposition per layer, with no labels, fine-tuning or benchmark. This matters where benchmarks have little headroom: base models sit at the Belebele chance floor in two thirds of the languages measured. The immediate use is triage before a multilingual benchmark is run or read. If a fixed token budget covers only a small share of a language’s passage, or a diagnostic fails the range criterion, an apparent per-language difference has a measurement component before any task claim is made. The audit reports fertility and coverage, checks descriptor range and refinement behaviour, and uses family aware nulls, documenting whether the comparison attached to

an accuracy number is in comparable units.

7 Conclusion

Tokenization is part of the measurement apparatus rather than a harmless preprocessing step. It fixes a translated item’s token budget, content window and attention graph node set before any reasoning occurs, so an internal quantity used to compare languages can lose comparability before a score gap is read as a model quality gap. The audit imposes the three constraints announced in §1. Put variable- N descriptors on a shared scale, by fixed range or by division by the analytic ceiling. Treat matched- N probes as coverage limited and report the coverage. Test resource axis claims against nulls that preserve language family. Applied to 73 associations a standard analysis reports as significant, those constraints retain 42.5% and reverse the sign of one in five, and the failures fall on the descriptors the criterion names in advance. Applied to HFER across three tokenizer matched Llama models, they leave one finding standing: the resource associated profile varies across depth, and its dispersion exceeds every within-family permutation draw in each model.

Limitations

One tokenizer matched Llama model set. The cleanest part of the design is also its boundary. The three Llama runs use a byte identical tokenizer, so differences among their profiles cannot be blamed on vocabulary changes. The price is that the strongest evidence comes from one tokenizer family. The 8B extension shows that depth heterogeneity persists beyond the 3.2-1B/3B pair, but model version, depth, capacity and possibly training configuration vary together, and the 8B run is matched scope only. It is therefore a generalisation check, not a capacity only experiment.

The cross tokenizer evidence is deliberately more modest. Dispersion clears the family aware null under the BLOOM/ROOTS and Qwen vocabularies, but BLOOMZ’s dispersion is roughly a quarter of the Llama values with 80% of it family reproducible, and BLOOMZ also changes positional encoding and instruction tuning. The mathematical range criterion should transfer because it depends on a descriptor’s definition, but the empirical length signature can change with architecture specific positional priors or activation scales (Appendix C). Full cross tokenizer generalisation remains a sepa-

rate validation problem.

Residual confound in the matched token scope.

It removes graph order from the per-layer comparison but is not a generic solution to tokenization inequality. Holding N fixed changes the semantic object being measured: the high density language contributes more of the original passage, while the fragmented language contributes a shorter prefix. This is why the coverage control changes the binned conclusions and why the per-layer profiles are reported as unadjusted matched scope associations. English covers 70.6% of a passage within 96 tokens where Burmese covers 6.1%, and coverage adjustment reduces that imbalance without reconstructing content absent from a fragmented prefix. The design answers the matched node question: whether the resource associated HFER profile varies across depth conditional on the same number of graph nodes, and not the content scope question of what happens at each language’s natural token length. We report no coverage adjusted per-layer permutation profile, so Table 1 supports depth heterogeneity for the matched scope profile while coverage controlled claims are restricted to the binned summaries.

Observational design. The analysis is observational. Resource level is a potential common upstream factor associated with fertility, geometry and modelling quality. It is included as a covariate, but nothing is randomised, so no causal interpretation is licensed. The depth profile reported here describes where a resource association sits, and does not identify the computation that produces it. Any statement in §6 about what an intervention would do to the profile is a prediction rather than a result.

The resource proxy is Wikipedia article count, validated against two anchors and screened for bot generated wikis. It measures unlabelled web text rather than pretraining composition, which is undisclosed for these models. Belebele accuracy is at its chance floor at this scale, so the modelling quality analysis rests on bits per character and a task grounded comprehension margin as proxy evidence for the evaluation conditions in which accuracy has little headroom. Near chance four-way arg max accuracy is sensitive to the numerical environment, so we regenerate it from stored per-item outputs rather than reconcile it with an earlier run. For the 1B and 3B content scope sweeps, graphs are capped at $N = 1024$, and §4 reports how that trun-

cation concentrates by fertility. The matched scope, where the reported effect is estimated, is unaffected. The 8B run is matched scope only at $N = 512$.

Reproducibility Statement

Every number in this paper resolves to a result file produced by a released script, and a single entry point re-runs the table checkers that compare the manuscript against those files. Appendix A describes the artifact, Appendix F the models, estimand and implementation departures from package defaults, and Appendix U the permutation nulls with their seed and draw count. The pre-registration log records each adversarial check with the decision rule fixed before the corresponding run, including the checks that withdrew a claim (Table 9), and no entry in it has been removed.

Ethics Statement

This work examines a disparity that disadvantages speakers of low-resource languages, and should not be used to rank languages or speakers, or read as endorsing any per-language quality score. The principal ethical implication is cautionary: an attention graph statistic that appears to measure per-language model quality may instead restate tokenizer fertility, and deploying such a statistic as a quality gate would encode segmentation inequality as a measured deficiency of the language. All data used is publicly released, no human subjects or personal data are involved, and all experiments were conducted on a single consumer GPU.

References

- Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah A. 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*, Singapore.
- Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. [Investigating cultural alignment of large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand.
- Chantal Amrhein and Rico Sennrich. 2021. [How suitable are subword segmentation strategies for translating non-concatenative morphology?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 689–705, Punta Cana, Dominican Republic.

- Catherine Arnett and Benjamin K. Bergen. 2025. [Why do language models perform worse for morphologically complex languages?](#) In *Proceedings of the 31st International Conference on Computational Linguistics*, Abu Dhabi, UAE.
- Catherine Arnett, Tyler A. Chang, Stella Biderman, and Benjamin K. Bergen. 2025. [Explaining and mitigating crosslingual tokenizer inequities](#). In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*.
- Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabisa. 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)*.
- Kirti Bhagat, Shaily Bhatt, Athul Velagapudi, Aditya Vashistha, Shachi Dave, and Danish Pruthi. 2026. [TALES: A taxonomy and analysis of cultural representations in LLM-generated stories](#). In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*.
- Shaily Bhatt and Fernando Diaz. 2024. [Extrinsic evaluation of cultural competence in large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16055–16074, Miami, Florida, USA.
- Steven Bird. 2020. [Decolonising speech and language technology](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online).
- 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.
- Yong Cao, Yova Kementchedjheva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024. [Cultural adaptation of recipes](#). *Transactions of the Association for Computational Linguistics*, 12:80–99.
- Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. [CANINE: Pre-training an efficient tokenization-free encoder for language representation](#). *Transactions of the Association for Computational Linguistics*, 10:73–91.
- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. [What does BERT look at? an analysis of BERT’s attention](#). In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.
- Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, Marcus Hutter, and Joel Veness. 2024. [Language modeling is compression](#). In *International Conference on Learning Representations (ICLR)*.
- Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. 2021. [Attention is not all you need: Pure attention loses rank doubly exponentially with depth](#). In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*.
- Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Junichi Tsujii. 2020. [CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6903–6915, Barcelona, Spain (Online).
- Kawin Ethayarajh. 2019. [How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65.
- Philip Gage. 1994. [A new algorithm for data compression](#). *C Users Journal*, 12(2):23–38.
- Shreya Havaldar, Sunny Rai, Young-Min Cho, and Lyle Ungar. 2025. [Culturally-aware conversations: A framework and benchmark for LLMs](#). *arXiv preprint arXiv:2510.11563*.
- Michael A. Hedderich, Lukas Lange, Heike Adel, Jan-nik Strötgen, and Dietrich Klakow. 2021. [A survey on recent approaches for natural language processing in low-resource scenarios](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568.
- Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Pi-queras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. [Challenges and strategies in cross-cultural NLP](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland.
- Dirk Hovy and Diyi Yang. 2021. [The importance of modeling social factors of language: Theory and practice](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602.

- 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.
- Taku Kudo. 2018. [Subword regularization: Improving neural network translation models with multiple subword candidates](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia.
- Taku Kudo and John Richardson. 2018. [SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium.
- 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*, pages 4483–4499.
- Xixian Liao, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari, Javier García Gilabert, Miguel Claramunt Argote, Ella Bohman, and Maite Melero. 2025. [Culture-aware machine translation: The case study of low-resource language pair Catalan–Chinese](#). In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 150–161, Geneva, Switzerland. European Association for Machine Translation.
- Tomasz Limisiewicz, Terra Blevins, Hila Gonen, Orevaoghene Ahia, and Luke Zettlemoyer. 2024. [MYTE: Morphology-driven byte encoding for better and fairer multilingual language modeling](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15059–15076, Bangkok, Thailand.
- Javier A. Lopetegui, Arij Riabi, and Djamé Seddah. 2025. [Common ground, diverse roots: The difficulty of classifying common examples in Spanish varieties](#). In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*.
- Sabrina J. Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, and Samson Tan. 2021. [Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP](#). *arXiv preprint arXiv:2112.10508*.
- Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. [When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462.
- Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. [The Zeno’s paradox of ‘low-resource’ languages](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17753–17774, Miami, Florida, USA.
- NLLB Team. 2022. [No language left behind: Scaling human-centered machine translation](#). *arXiv preprint arXiv:2207.04672*.
- NLLB Team. 2024. [Scaling neural machine translation to 200 languages](#). *Nature*.
- Valentin Noël. 2025. [A graph signal processing framework for hallucination detection in large language models](#). *arXiv preprint arXiv:2510.19117*.
- Valentin Noël. 2026. [Geometry of reason: Spectral signatures of valid mathematical reasoning](#). *arXiv preprint arXiv:2601.00791*.
- Antonio Ortega, Pascal Frossard, Jelena Kovačević, José M. F. Moura, and Pierre Vanderghenst. 2018. [Graph signal processing: Overview, challenges, and applications](#). *Proceedings of the IEEE*, 106(5):808–828.
- Saurabh Kumar Pandey, Sougata Saha, and Monojit Choudhury. 2026. [To generate or discriminate? methodological considerations for measuring cultural alignment in LLMs](#). *arXiv preprint arXiv:2601.02858*.
- Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. [Language model tokenizers introduce unfairness between languages](#). In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. [How multilingual is multilingual BERT?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy.
- Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. [XCOPA: A multilingual dataset for causal common-sense reasoning](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376.
- Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. 2019. [Combating adversarial misspellings with robust word recognition](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5582–5591, Florence, Italy.
- 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 (Volume 1: Long Papers)*, pages 2373–2403, Albuquerque, New Mexico.

- Arij Riabi, Benoît Sagot, and Djamé Seddah. 2021a. [Can character-based language models improve downstream task performances in low-resource and noisy language scenarios?](#) In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*.
- Arij Riabi, Sougata Saha, and Monojit Choudhury. 2026. [The anthropology of food: How NLP can help us unravel the food cultures of the world.](#) In *Proceedings of the First Workshop on Multilingual Multicultural Evaluation*, pages 76–98, Rabat, Morocco.
- Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. 2021b. [Synthetic data augmentation for zero-shot cross-lingual question answering.](#) In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.
- 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*.
- Timo Schick and Hinrich Schütze. 2020. [Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking.](#) In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8766–8774.
- Mike Schuster and Kaisuke Nakajima. 2012. [Japanese and Korean voice search.](#) In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5149–5152.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Neural machine translation of rare words with subword units.](#) In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany.
- Yan Shao, Christian Hardmeier, and Joakim Nivre. 2018. [Universal word segmentation: Implementation and interpretation.](#) *Transactions of the Association for Computational Linguistics*, 6:421–435.
- Han Shi, Jiahui Gao, Hang Xu, Xiaodan Liang, Zhen-guo Li, Lingpeng Kong, Stephen M. S. Lee, and James T. Kwok. 2022. [Revisiting over-smoothing in BERT from the perspective of graph.](#) In *International Conference on Learning Representations (ICLR)*.
- David I. Shuman, Sunil K. Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. 2013. [The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains.](#) *IEEE Signal Processing Magazine*, 30(3):83–98.
- Nicholas Sukiennik, Chen Gao, Fengli Xu, and Yong Li. 2025. [An evaluation of cultural value alignment in LLM.](#) *arXiv preprint arXiv:2504.08863*.
- Lichao Sun, Kazuma Hashimoto, Wenpeng Yin, Akari Asai, Jia Li, Philip Yu, and Caiming Xiong. 2020. [Adv-BERT: BERT is not robust on misspellings! generating nature adversarial samples on BERT.](#) *arXiv preprint arXiv:2003.04985*.
- Jonathan J. Webster and Chunyu Kit. 1992. [Tokenization as the initial phase in NLP.](#) In *COLING 1992 Volume 4: The 14th International Conference on 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)*.
- 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*.
- Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. [Efficient streaming language models with attention sinks.](#) In *International Conference on Learning Representations (ICLR)*.
- Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. [ByT5: Towards a token-free future with pre-trained byte-to-byte models.](#) *Transactions of the Association for Computational Linguistics*, 10:291–306.
- Vilém Zouhar, Clara Meister, Juan Luis Gastaldi, Li Du, Mrinmaya Sachan, and Ryan Cotterell. 2023. [Tokenization and the noiseless channel.](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Appendix contents

A	Reproducibility and artifact	14
B	Further implications	14
C	Cross architecture check	14
D	Depth nulls in full	15
E	Descriptor taxonomy	18
F	Models, estimand and implementation	18
G	Descriptor partition and ceiling normalisation	19
H	Mediation path table and full discussion	19
I	Negative results in full	20
J	Two validation criteria in full	21
K	Extended related work	21
L	Robustness analyses in full	22
M	Coverage collinearity	23
N	Corpus screening	23
O	Refinement, generators and the cutoff fraction	23
P	Truncation censoring of the modelling quality outcome	23
Q	Identifiability of the peak position	24
R	Stability of the maximum statistic	24
S	Node set refinement sweep	24
T	Full mediation table	25
U	Permutation nulls	26

A Reproducibility and artifact

Code, the pre-registration log and the derived measurements are released as a public artifact.² It contains the extraction and analysis pipeline, a single entry point that runs every table checker, and a tidy table of 244,800 per-language, per-layer, per-descriptor values covering 60 languages, 17 recorded series, of which the eleven descriptors of §3.1 are analysed here, and ten model by scope cells, which is sufficient to recompute every number reported here. The raw per-item sweep outputs are not shipped, being 92 MB. The regeneration command and its measured cost are documented in the repository. Every table in this paper is verified against the recorded results by a checker that fails on a changed value, a removed mark or a deleted row, and the adversarial sweep that exercises those checkers is included.

B Further implications

Where the criterion applies outside the experiments reported in the body.

As a training signal. A depth resolved geometric quantity is a natural candidate for a dense auxiliary objective, and algebraic connectivity is the quantity most often proposed. Our results do not support that choice: λ_2 carries neither leg of the path to modelling quality (Table 2), so maximising it would optimise a quantity with no established relationship to language modelling performance.

²github.com/vcnoel/tokenization-measurement-scale

Of the three descriptors that do carry both legs, $\lambda_{\max}$ enters with the sign of a suppressor and effective rank is a residual stream rather than a graph quantity, which leaves HFER as the only spectral candidate. Even there we would treat it as an objective only after the adaptation experiment above succeeds, since a diagnostic optimised as a target ceases to function as a diagnostic.

Scope beyond the cross-lingual setting. Proposition 1 concerns node set refinement rather than languages as such. Any comparison of attention graph descriptors between sequence populations of systematically different length inherits the same non-identification: prose against code, short against long prompts, models with different vocabularies compared to one another, or a single model before and after a tokenizer change. In each case the criterion and the normalisation apply unchanged and remain checkable from a definition prior to measurement. This is directly relevant to spectral descriptors deployed as runtime monitors (Noël, 2026), since a threshold calibrated on one language or one input population is not transferable to another whose sequences are systematically longer unless the descriptor satisfies Proposition 1. Bits per character is similarly transferable as an outcome: free of floor effects, comparable across scripts, and recoverable from quantities standard evaluation harnesses already record.

C Cross architecture check

The 73 cells behind Table 3 are nested within model and depth band, and the two Llama models share a tokenizer, so the effective number of independent units is well below 73 and the Fisher exact test overstates it. The association is reported as indicative rather than as a clustered inference. BLOOMZ-3B supplies 19 of those cells and is a model whose empirical partition fails, so the criterion is scored there on a model where its own length signature does not hold.

The comparability criterion in Proposition 1 is architecture independent because it depends only on a descriptor’s mathematical range. The empirical length signature it predicts is not. We swept BLOOMZ-3B (November 2022) and Qwen2.5-1.5B (September 2024) with the same pipeline, choosing one non-RoPE and one RoPE model outside the Llama family. In both, the Llama partition between descriptors whose range varies with N and fixed range ones fails at the depths where the Llama

Table 4: **The empirical partition is architecture specific.** Lowest within language $r(\cdot, \log N)$ among descriptors whose range varies with N , against the highest $|r|$ among fixed range ones, content scope, shown as the range over the three depth bands. Correlations are within language throughout: the across language form absorbs between language fertility, which is the confound under study.

Model	lo N -dep.	hi fixed	overlap
Llama-3.2-1B	0.679 to 0.888	0.268 to 0.332	none
Llama-3.2-3B	0.718 to 0.890	0.166 to 0.288	none
Qwen2.5-1.5B	0.074 to 0.161	0.858 to 0.879	all
BLOOMZ-3B	0.161 to 0.581	0.740 to 0.826	all

pair separates by a factor of two or more (Table 4). The failures are informative, not replications: each has a model specific explanation unrelated to tokenizer vocabulary.

Under BLOOM, the main culprit is ALiBi: a pre-softmax distance penalty makes locality a monotone function of sequence length. Passing the content free ALiBi attention matrix through the same pipeline gives $r(\text{HFER}, \log N) = -0.798$ from position alone, against -0.632 observed and $+0.144$ for Llama. The sign is robust to the choice of graph signal but too small in magnitude over the within language token count range, so the signature is substantially positional and does not carry tokenizer information.

Under Qwen, the issue is residual stream magnitude. Its HFER is flat ($r = -0.045$), even though its fertility correlates more strongly with Llama’s tokenizer ($+0.907$) than BLOOM’s does ($+0.573$). In both Qwen and Llama, $\text{tr}(X^\top X)$ is dominated by a single token, but that token’s cross passage variance differs by roughly twenty fold, burying the length signal in one model and not the other. A third 1B base model screened under the same pipeline, OLMo-2-1B (April 2025), places 3.0% of $\text{tr}(X^\top X)$ in its largest token against Qwen’s 99.3% and Llama’s 97.5%, so this is a claim about Qwen rather than about model scale. The implication is methodological: the range criterion transfers, but the empirical length signature remains model specific.

D Depth nulls in full

The narration supporting Figure 2 and §5.1: the calibration bound on per-layer slopes, why script is a weaker stratum than family, the multimodality of the profiles, and the untrained control’s limitations.

Three depth summaries and their names.

These are easily confused, so we fix terminology. The **binned mid-depth slope** is the fit over the whole middle third of the network ($+0.315/ +0.323$ in Table 10). The **per-layer mid-depth slope** is the mean of the per-layer fits over relative depth $[0.33, 0.66)$ ($+0.395$ for the 1B, $+0.108$ for the 3B). This is the quantity all permutation tests below use. The **peak layer slope** is the single largest per-layer value ($+0.60$ and $+0.67$, both at relative depth ≈ 0.36). These are three different estimands and none is a rounding of another.

Free permutation. Permuting the resource axis across all 60 languages and refitting the entire profile (1000 draws, Appendix U) places the observed per-layer mid-depth slope 5.8 (1B) and 3.1 (3B) null standard deviations out. With the finite permutation correction, this gives $p_{\text{perm}} = .0010$ and $.0020$ respectively.

This test is weaker than it first appears, and the next paragraph is the one that matters. A free permutation scrambles every language level structure simultaneously, so it tests the observed effect against “no language level signal whatsoever”, not against “no resource effect”. Language family and script are not independent of resource level, since the high-resource languages are overwhelmingly Indo-European and Latin script, so a family or script signal would clear this null exactly as a resource effect does.

Stratified permutation. We therefore permute the resource axis within strata, leaving the stratum’s structure intact in every draw (Table 16). Under within-script permutation the effect survives in both models, and the 1B null mean is -0.036 : the within-script null reproduces substantially less of the observed slope than the within-family null in this sample. Under within-family permutation the picture changes materially. The null mean rises to $+0.333$ against an observed $+0.395$ for the 1B, so 84% of the observed slope is reproduced by relabelling within family branches (68% for the 3B). The free null margin above becomes a smaller 2.5 null-SD margin ($p = .012$) for the 1B, and for the 3B the effect falls inside the null at the conventional threshold (1.6 null SDs, $p = .063$).

Untrained control. The permutation tests above ask whether the labels carry signal. A different question is whether the model does: the profile might be a consequence of depth, head count,

causal masking and per-language token counts acting on any network of this shape. We therefore ran the entire pipeline on a randomly initialised Llama-3.2-1B (same configuration, same tokenizer, same corpus, both scopes, 60 languages), whose mean NLL is 12.19 against a uniform prediction floor of $\log |V| = 11.76$ and a trained value of 2.66. Because the tokenizer is untouched, this control holds the per-language token counts exactly fixed.

The resource association is not reproduced by this untrained control. The per-layer mid-depth slope is -0.028 against $+0.395$ in the trained model of identical architecture, and the peak layer statistic is 0.151, below the 0.261 that permuted labels reach at the 95th percentile, so the untrained model’s strongest layer is weaker than this pipeline’s noise floor. Sign concordance with the trained profile is 0.500, exactly chance.

The control has a limit, which we state. An untrained transformer has near-uniform attention: row-wise attention entropy is 3.374 nats against 3.598 for an exactly uniform causal row at $N = 96$, so the untrained graph carries only 6% less entropy than the maximum. The null result therefore bounds the association to a model whose attention is not uniform, and does not separate that condition from learned computation, which an all-ones attention matrix would also fail to separate. Two observations bear on this. The untrained graph is not degenerate on the descriptor of interest: HFER has cross language standard deviation 0.0062 at mid depth, coefficient of variation 0.084 against the trained model’s 0.083, so the descriptor has the same relative room to move in both models and the null is not a floor effect: the variance exists, it simply is not organised by resource level. We note in passing that HFER is higher in the untrained model (0.074 vs 0.055), so training reduces high-frequency graph energy, which is consistent with the over-smoothing literature (Shi et al., 2022). The sharper version of this control would use an early training checkpoint. None is publicly released for Llama-3.2-1B, so we report the limitation rather than resolve it.

Dispersion of the per-layer profile. The striking fact in Figure 2 is that both models peak at relative depth 0.364 despite having 16 and 28 layers. We tested it. The cross model peak agreement is reproduced by 11.9% of free permutations ($p = .12$): profiles are smooth by construction, the grid has 12 points, and coincidental agreement is common.

Under within-family permutation the peak lands within one grid step of 0.364 in 94.8% of draws, and exact cross model agreement occurs in 45.5% ($p = .46$). For peak position, the family null is also five times tighter than the free null (± 0.091 vs ± 0.323) and centred on the observed value: within family, the peak reliably lands where we observed it whatever the labels. Rerunning on grids of 8, 12 and 16 points settles it: the two models’ peaks coincide at 12 and 16 but lie four cells apart at 8 (Appendix Q). A fact that flips with an arbitrary interpolation choice is not a finding. We therefore withdraw the peak position claim and report peak position as description.

The profile shape resolves an apparent tension. Each has competing near-equal maxima: (0.33, $+0.676$) and (0.47, $+0.543$) for the 1B, and for the 3B (0.26, $+0.693$) and (0.33, $+0.678$), two peaks differing by 2.2% and sitting 0.074 apart in relative depth. A gridding choice decides which wins. That is why the maximum value is stable to ± 0.019 under resampling while its location moves half the network: a multimodal profile has a well-determined maximum and an ill-determined argmax.

What survives is the dispersion of the per-layer profile, with amplitude and range corroborating the same quantity rather than confirming it independently (they correlate 0.57 to 0.87 across draws). SD sits 4.6, 4.0, and 3.72 null SDs over the within-family null across the 1B/3B/8B sequence. In the 8B run, observed SD is 0.431 against a free null of 0.0796 ± 0.0266 and a family null of 0.2595 ± 0.0460 . No permutation draw reaches it, and the family null reproduces 60% of the observed dispersion. A random-effects decomposition against the clustered per-layer standard errors gives $I^2 = 94.9\%$, 96.3%, and 95.8%: estimation noise accounts for little of the between layer variance. The peak layer statistic is 0.676 and 0.693 against free null 95th percentiles of 0.261 and 0.288. It clears every stratified null in both models at $p \leq .002$ (Table 16), and the untrained control puts it at 0.151, below the noise floor. Its sampling variability is small next to the effect’s margin over that null (bootstrap SD 0.019 against a margin of 0.190, or 10%, Appendix R).

We report the family reproducible share for what we keep, not only for what we withdraw, because the alternative is an uneven disclosure standard. For the 1B model, the within-family null mean for amplitude is 0.486 against an observed 0.676, so 72%

of that candidate statistic is reproduced by within-family relabelling, against 84% for the mid-band slope and 95% for peak position. The dispersion statistic we keep is itself 58 to 71% family reproducible in the 1B/3B pair and 60% in the 8B run. Surviving statistics are substantially family aligned. What distinguishes them from the withdrawn ones is that they still clear their nulls, not that they are free of phylogeny.

Depth heterogeneity. The claim that does not presume a position is that the effect is not constant across depth. The dispersion figures given above are that claim. The dispersion of the per-layer slope vector is outside both nulls across the tokenizer matched Llama set: for the 1B, SD is 0.366 against a free null of 0.078 ± 0.022 and a family null of 0.259 ± 0.023 ($p = .001$, 71% reproducible). For the 3B, the family null margin is 4.0 null SDs. For the 8B, SD is 0.431 against a family null of 0.2595 ± 0.0460 ($p = .001$, 60% reproducible).

We lead with SD because it is the strongest of the candidates, not because it reads best. In the 1B/3B pair, margins over the family null in null SDs are 4.6 (SD), 4.1 (amplitude) and 3.6 (range) for the 1B, and 4.0, 3.3 and 4.2 for the 3B: SD is the only one in the top two of both models, and it carries the lowest family reproducible share. It is also the appropriate choice given the profiles are multimodal: with two maxima 2.2% apart in the 3B, a maximum is a coin flip between two layers, whereas a dispersion measure is well defined. These three are in any case not independent confirmations: all are dispersion measures of the same per-layer vector, correlating 0.57 to 0.87 across permutation draws.

Netting out estimation noise directly gives the same answer without a null. A random-effects decomposition of the per-layer slopes against their clustered standard errors returns $\tau = 0.349$ and $I^2 = 94.9\%$ for the 1B, and $\tau = 0.405$, $I^2 = 96.3\%$ for the 3B, and the 8B gives $I^2 = 95.8\%$. About 95% of the observed between layer variance in the resource slope is not attributable to estimation noise. That τ sits close to the raw SD in the 1B/3B pair confirms the unweighted dispersion was not mostly sampling error. This is the claim the three-bin cancellation argument needs, and what supports the statement in §6 that a per-layer fit is required to detect the effect, not to locate it. The defensible claim is therefore that resource level is associated with high-frequency graph energy in a way that is consistent with tokenizer fragmenta-

tion and varies across depth, that the dispersion of that variation requires a trained model and is not reducible to family or script, and that its position in depth we cannot attribute to resource level rather than phylogeny. A large depth resolved change would describe the estimator rather than the finding, and the paper claims the dispersion instead.

Family and resource level are genuinely confounded in the world and not merely in this sample: they share 70% of their variance across our 60 languages. A within-family test is therefore close to the ceiling of what an observational design can establish here, which is a limit on what can be concluded and not a reason to prefer the free null.

The shape of the profile is a separate question. Under the same null the lag-1 autocorrelation of the per-layer sequence is inside the null in both models ($p = 0.41$, 0.37): adjacent layers are fitted on nearly the same observations, so profiles are smooth whatever the labels. Sign concordance between the two models on a common depth grid is 0.583 against a null of 0.593 ± 0.135 ($p = 0.66$), below the null mean. The sign sequence is therefore described rather than claimed as replicated structure, and individual phases are left uninterpreted. The same nulls give a calibration we apply throughout: the largest $|\beta|$ across depth under permuted labels reaches 0.261 and 0.288 at the 95th percentile, so a single layer slope below ≈ 0.29 lies inside what this pipeline produces from meaningless labels, whatever its per-layer confidence interval, which does not account for the search across layers.

Depth binning obscures the depth structure in both directions: early slopes of opposite sign fall within one bin and cancel, and the binned late-depth average for the 1B fails coverage control ($p = 0.57$) although layers 12 and 13 individually survive it. We therefore treat the per-layer fit as primary. Since three-bin depth summaries are widely used in interpretability work, the point generalises beyond this setting.

A fragment integration account would explain the sign: processing content distributed across many byte-level fragments may spend capacity on smoother, lower-frequency integration, while a language whose words survive tokenization more intact may proceed to more differentiated, hence higher-frequency, computation. We state this as motivation, not as a finding, and do not attach it to a depth, since our tests cannot identify where in the network the effect sits (Appendix Q). A direct test

of word boundary alignment in attention proved confounded with token type and is not yet interpretable. The attempt is recorded in the analysis log. We observe geometry, not the computation that produces it.

E Descriptor taxonomy

The eleven descriptors fall into three groups, and the boundaries matter for the results, so we state them explicitly. **Six require the Laplacian:** HFER, spectral entropy of the mode energy distribution, Dirichlet energy $\text{Tr}(\mathbf{X}^\top \mathbf{L} \mathbf{X})$, the smoothness index $\text{Tr}(\mathbf{X}^\top \mathbf{L} \mathbf{X}) / \text{Tr}(\mathbf{X}^\top \mathbf{X})$, λ_2 and $\lambda_{\max}$. **Two require the attention matrix but not its spectrum:** mean attention distance and row-wise attention entropy. **Three are computed from the residual stream alone, with no graph at all:** anisotropy (mean pairwise cosine of $\mathbf{X}$), mean residual norm, and *effective rank*, defined as exp of the entropy of the normalised singular-value spectrum of the centred $\mathbf{X}$.

Effective rank is grouped with the residual stream baselines and not with the spectral family. This is worth stating because its name and its $[0, \min(N, d)]$ range invite the opposite reading, and because it is the descriptor that carries the largest mediated effect in §5.2: that result is a result about a baseline, and would be much less interesting if effective rank were a graph quantity. Its ceiling is $\min(N, d)$ rather than N , since the singular-value spectrum of an $N \times d$ matrix has at most $\min(N, d)$ nonzero entries. In our setting $d = 2048$ and $N \leq 1024$, so N binds and the distinction does not affect the normalisation of §3.3. It does affect the synthetic sweep of §3.2, where the signal dimension is chosen larger than the largest refined node count so the ceiling under test is the intended one.

The criterion evaluated as a classifier. Table 5 scores the range criterion against a purely empirical target: sensitivity to graph size measured within a language, where resourcedness is constant so that nothing but graph size sensitivity remains. Correlations are computed within each (language, layer) and then averaged. Conditioning on layer is necessary rather than fastidious, because every descriptor has a strong depth profile that is unrelated to passage length, and pooling layers attenuates all of these correlations toward zero, enough, in the case of Dirichlet energy, to invert the verdict.

The partition is clean, and it survives deleting the

graph descriptors: among the five residual stream and attention quantities alone, the variable range ones stay above 0.75 and the fixed range ones below 0.39. BLOOMZ is the least clean column, with several fixed range descriptors reaching $|r| \approx 0.4$ to 0.6. This is the same model whose positional encoding and instruction tuning confound the tokenizer comparison elsewhere in the paper.

descriptor	range	1B	3B	BLM	Qwen
<i>range varies with N</i>					
effective rank [†]	$\min(N, d)$	.94	.96	.95	.86
attn. distance [†]	N	.87	.88	.92	.97
attn. entropy [†]	$\log N$	.86	.87	.64	.64
Dirichlet energy	N	.98	.97	.43	.18
spectral entropy	$\log N$	.78	.78	.78	.49
<i>range fixed</i>					
HFER	$[0, 1]$	.02	.03	-.38	-.01
λ_2	$[0, 2]$	-.10	-.02	-.59	-.20
$\lambda_{\max}$	$[0, 2]$	.10	.08	-.16	-.12
smoothness	$[0, 2]$	.21	.19	-.38	.36
anisotropy [†]	$[-1, 1]$	-.01	-.01	-.11	-.04
residual norm [†]	unbounded	-.07	-.08	-.60	-.79

Table 5: **The range criterion works off the graph.** Within language $r(\text{descriptor}, \log N)$, computed within each (language, layer) and averaged over languages, content scope, 60 languages. [†] marks descriptors requiring no eigendecomposition. Residual norm is unbounded rather than fixed range, so the criterion makes no prediction for it. It is shown for reference and is indeed the least well-behaved row.

F Models, estimand and implementation

The model set, the estimated quantity and the two departures from the spectral package’s defaults, in full.

Models and estimand. We evaluate a tokenizer matched Llama-3.2-1B/3B pair, released September 2024, and add Llama-3.1-8B, released July 2024, as a tokenizer matched extension. Tokenizer byte identity is verified rather than assumed, including vocabulary, merges, and token id sequences, so fertility, script and content are held exactly fixed across the three models. The 8B extension differs in model version, capacity, depth, and potentially training configuration. It is a generalisation check rather than a capacity only comparison, and is matched scope only because its extraction cap is $N = 512$ rather than 1024. Differences among the three model profiles cannot be attributed to differences in tokenizer vocabulary. We report β , the standardised slope of a per-layer descriptor on

$\log_{10}$ article count, with language clustered standard errors. The regressor is constant within a language, so the effective sample size for a between language slope is 60 rather than the row count of a cell. Clustering leaves point estimates largely unchanged and widens intervals by approximately an order of magnitude. This also motivates allocating budget to languages rather than items, since doubling the language count approximately halves the intervals whereas tripling items per language does not.

Implementation. Spectral quantities are computed with a published GSP package under two deliberate departures from its defaults: the symmetric normalised Laplacian of (1), since the random-walk operator is not symmetric and energy ratios on its eigenvectors are not well defined, and a dense eigensolver, since the default sparse solver retains 50 modes, which renders a high-frequency energy ratio uninterpretable at these graph orders. Special tokens are removed before graph construction. Graphs are capped at $N = 1024$ in the 1B and 3B content scope sweeps, truncating 13.2% of content scope rows and no matched scope rows. The truncation falls on 12 high fertility languages and is total for Burmese. The 8B run is matched scope only at $N = 512$.

G Descriptor partition and ceiling normalisation

The two tables supporting §3.2 and §3.3.

Resource level is constant within a language, so every association in Table 6 is graph size sensitivity alone. HFER, set in bold as the descriptor the paper uses, sits at 0.144/0.147, on the correct side of the partition. The two axes of the taxonomy are independent. Whether a descriptor needs the Laplacian is a separate question from whether its attainable range grows with N , and the criterion applies to the range axis whatever the answer on the graph axis. Effective rank is the cell that establishes this: it is computed on the residual stream with no graph at all, yet it has an N -dependent range and the strongest graph size sensitivity in Table 6. The partition in that table accordingly cuts across both the graph and the non-graph rows.

Constant factors in the ceiling do not affect Table 7, the outcome being a standardised slope. HFER needs no correction and is listed for reference only.

Table 6: **The comparability criterion recovers the empirical partition with no overlap.** Mean within language correlation between each descriptor and $\log N$, content scope, 59 languages: Burmese is excluded because the $N = 1024$ cap fixes every one of its passages at the same length, leaving its token count constant and the correlation undefined. Shading marks the N -dependent range: in both models every shaded descriptor exceeds every unshaded one by more than a factor of two, with no overlap.

Descriptor	Range	1B	3B
GRAPH $\times$ N -DEPENDENT RANGE			
Dirichlet energy	$[0, N]$	0.935	0.976
attention entropy	$[0, \log N]$	0.924	0.937
attention distance	$[0, N]$	0.909	0.919
spectral entropy	$[0, \log N]$	0.827	0.841
NON-GRAPH $\times$ N -DEPENDENT RANGE			
effective rank	$[0, \min(N, d)]$	0.951	0.970
GRAPH $\times$ FIXED RANGE			
smoothness	$[0, 2]$	0.311	0.255
$\lambda_{\max}$	$[0, 2]$	0.302	0.268
HFER	$[0, 1]$	0.144	0.147
λ_2	$[0, 2]$	-0.081	-0.001
NON-GRAPH $\times$ FIXED RANGE			
anisotropy	$[-1, 1]$	-0.062	-0.021
residual norm	unbounded	-0.031	-0.064

Table 7: **Ceiling normalisation places the five variable range descriptors on a shared scale.** Cross scope discrepancy $|\Delta|$ before and after division by the analytic ceiling, as 1B/3B, and nearer zero is better. [†]Effective rank passes this test but fails the refinement test of Table 12.

Descriptor	Ceiling	Raw	Normalised
effective rank [†]	$\min(N, d)$	0.92/0.90	-0.07/-0.05
spectral entropy	$\log N$	0.63/0.80	-0.14/-0.05
Dirichlet energy	N	0.51/0.51	0.00/0.02
attention distance	N	0.45/0.40	-0.15/-0.06
attention entropy	$\log N$	0.22/0.25	-0.09/-0.03
<i>HFER (reference)</i>	n/a	-0.07/0.06	<i>unchanged</i>

H Mediation path table and full discussion

The decomposition of Table 2 with its estimation details, its sensitivity to an unmeasured confounder, and the reason no mediated proportion is reported.

Estimation follows §4 throughout Table 8: the treatment is Rényi efficiency (Zouhar et al., 2023) with the resource axis as a covariate, standard errors are clustered over the 60 languages, and the interval on ab is bias-corrected and accelerated over language clusters with 10,000 draws. The q values are Benjamini–Hochberg adjusted separately

Table 8: **Three descriptors carry both legs of the path, and the largest is not a spectral one.** Here a is the coefficient from tokenizer efficiency to geometry, b the coefficient from geometry to modelling quality at fixed tokenizer, and ab the indirect effect. Shaded rows, both $q < .05$, are the only survivors. ab_{beh} repeats the decomposition against a behavioural outcome, the character normalised comprehension margin on the same items, under its own Benjamini–Hochberg family over the sixteen mediators. Signs invert because a lower bits per character and a higher margin both mean better. Shading refers to the bits per character family only, and $\dagger$ marks a row whose b leg does not clear correction in the behavioural family: effective rank at $q = .069$, on a wider standard error rather than a smaller coefficient, its b being 0.329 against HFER’s 0.171. Intervals for the behavioural arm are not shown in this compact table.

Mediator	a	b	ab [95% BCa]	ab_{beh}
eff. rank	+0.646	−0.377	−0.244 [−0.478, −0.072]	+0.218 [†]
λ_{max}	−0.629	−0.215	+0.135 [+0.037, +0.245]	−0.130
HFER	+0.145	−0.248	−0.036 [−0.077, −0.007]	+0.024
smoothness	−0.196	−0.151	+0.030 [+0.001, +0.100]	−0.024 [†]
λ_2	+0.014	−0.226	−0.003 [−0.117, +0.074]	+0.018 [†]

for the a and b legs across the eleven descriptors and the five ceiling normalised forms. The BCa intervals for ab are unadjusted and descriptive. The remaining thirteen rows appear in Appendix T.

One of the three cuts against the framing that motivated the study. The largest indirect effect belongs to a baseline rather than to the spectral family: residual stream effective rank carries $ab = -0.244$, roughly seven times HFER’s -0.036 , and requires no graph, no Laplacian and no eigendecomposition. If the objective is to account for as much of the association between tokenizer efficiency and modelling quality as possible with one measured quantity, the residual stream statistic is the stronger summary and the spectral machinery is not required. Two qualifications attach and neither removes the point: effective rank carries the highest collinearity in the set (variance inflation 3.4), and it is the one descriptor for which our two validation criteria disagree, being repaired by the cross scope test and not by the stricter refinement test (§3.3). Because it is used here at fixed $N = 96$, the refinement failure does not bear on this estimate. It bars the same comparison at each language’s natural length. λ_{max} , for its part, enters with the opposite sign to the total effect, which is suppression rather than mediation, and we report no effect share for it. Its a -path (-0.629) is close to the negative of effective rank’s ($+0.646$), which a reader may read as a sign error.

It is not: the two descriptors correlate -0.47 with each other and $+0.745$ and -0.596 respectively with the treatment, so the mirroring is a property of the data and not a transcription. A more efficient tokenizer raises effective rank and lowers λ_{max} , and the two descriptors sit in different comparability classes (Table 6), so this is not one quantity counted twice. HFER’s path is the smallest of the three, though it is the only one whose sign matches the mechanism of §5.1. It is one mediator with an interpretable sign, not the mediator. That λ_2 carries neither leg is worth stating explicitly, given that algebraic connectivity is the descriptor most commonly chosen in this setting.

Two figures bound how much weight the table takes. The treatment is not collinear with the resource axis ($R^2 = 0.059$, variance inflation 1.06), so conditioning is not what produces these estimates. But the E -values for HFER’s legs are only 1.54 and 1.82: an unmeasured language level confounder of about that strength with both treatment and outcome would explain each leg away, and the space of such confounders (morphological type, orthographic depth, register, translation direction of the parallel corpus) is large.

We do not report a mediated proportion. The mediated share is ab/c , and the total effect c is not significant net of the resource axis ($c = -0.078$, $p = 0.278$). A ratio to a denominator not distinguishable from zero has arbitrary magnitude, and reporting one would present a quantitatively strong claim resting on a null result. The defensible statement is narrower: both legs are individually significant and the indirect effect is non-zero, while the total effect of the tokenizer on modelling quality is absorbed by resource level. This is a decomposition of an observational association and does not constitute causal identification.

I Negative results in full

Four negative results constrain the conclusions and are reported in full.

First, the spectral family does not outperform a non-graph baseline. Under leave-one-out cross-validation at the language level (60 languages, matched scope), the baseline family attains $R^2 = 0.799/0.773$ against 0.683/0.741 for the spectral family. Anisotropy of the residual stream, computed with no graph at all, provides the largest single surviving slope at -0.63 mid-depth, and residual stream effective rank carries the largest

path to modelling quality (Table 2). Three independent analyses thus place a non-graph quantity ahead of the spectral family, which is consistent with prior findings that representation geometry alone is highly informative (Ethayarajh, 2019).

Second, Heaps’ exponent is not scale-free. We introduced it as a principled alternative to type-token ratio, and it correlates with fertility at $r = -0.96$, more strongly than the ratio it was intended to replace. Refitting each language over an identical token range, to exclude a range artefact, increased the correlation. Fragmentation induces fragment reuse, and the two are not separable.

Third, the tokenizer feature space is effectively one-dimensional. Across five statistics the first principal component accounts for 81.4% of the variance and every feature correlates with token count at $|r|$ between 0.81 and 0.96. On a fixed parallel corpus, tokenization and fertility are not separable constructs, and additional tokenizer features will not separate them. Holding the tokenizer fixed while varying the model is the function of the Llama model set.

Fourth, the claim we make about depth is about dispersion. The magnitude and the position of the effect are not supported across model sizes. A pooled fit with a model interaction rejects equality of the mid-depth slopes (-0.124 ± 0.054 , $p = .021$), so $+0.395$ and $+0.108$ are not co-equal evidence. Sign concordance between the profiles does not exceed its permutation null ($p = .66$), and the peak position is reproduced by within-family relabelling in 95% of draws. What clears the family aware dispersion null is the dispersion of the profile across the 1B/3B/8B sequence (4.6, 4.0, and 3.72 null SDs over the family null, $I^2 \approx 95\%$), with amplitude and range corroborating the same quantity rather than confirming it independently. Reasons in Appendix D.

J Two validation criteria in full

Why the two comparability tests disagree on effective rank, and which one the paper defers to.

Cross scope agreement against refinement invariance. Cross scope agreement is not the only test of comparability available, and it is not the strictest. The refinement sweep of §3.2 asks a different question: not whether a descriptor’s slope is stable when the population of passages changes, but whether its value is stable when one fixed graph is re-expressed over a finer node set.

These come apart, and the descriptor on which they come apart is effective rank. It is repaired by the cross scope criterion, with discrepancy $0.92/0.90 \rightarrow -0.07/-0.05$, the largest improvement in Table 7, and it is not repaired by the refinement criterion, where normalisation moves drift from 0.654 to 0.793, making it worse (Table 12). Conversely HFER needs no cross scope correction ($|\Delta| = 0.063$) yet drifts 114% under refinement. No descriptor in our set passes one test and fails the other by accident: the two tests are measuring different things.

Refinement invariance is the stricter requirement, and where the two disagree we defer to it. Cross scope agreement is a property of a fitted slope across two samples, so a descriptor can achieve it by having its scale contamination absorbed into the standardisation within each scope, so that the slope agrees while the underlying values remain scale dependent. Refinement invariance is a property of the descriptor evaluated on a single graph, with the coarse structure held fixed to 6.7×10^{-16} by construction, so there is no sample and nothing for standardisation to absorb. A descriptor that fails it is reading the node set directly. The methodological consequence is stated in §6: normalised effective rank is admissible for between language comparison at fixed N , which is the matched scope where we use it, and is not admissible for comparison at each language’s natural length. Passing the cross scope test alone does not license the latter.

K Extended related work

What each cited result establishes, at the length §2 cannot carry.

Tokenization and cross-lingual disparity. Petrov et al. (2023) and Ahia et al. (2023) document that subword tokenizers allocate several times more tokens to some languages than to others for identical content, with direct consequences for context budget, latency and API pricing. Rust et al. (2021) isolate tokenizer quality from the remainder of the model and attribute to it much of the monolingual performance gap of multilingual models, and Arnett and Bergen (2025) find that the gap between agglutinative and fusional languages is reduced principally when training data is equalised after accounting for encoding efficiency. Recent work seeks to mitigate the disparity at the encoding level, through morphology driven byte encodings (Limisiewicz et al., 2024) and

through parity constraints on the merge procedure (Arnett et al., 2025). Zouhar et al. (2023) propose Rényi efficiency of the subword distribution as an intrinsic tokenizer quality measure, which we adopt as the treatment variable in §5.2. These studies characterise the inequality at the input and at the output. The present work measures its effect inside the network.

Wu and Dredze (2020) establish that multilingual BERT performs substantially worse for low-resource languages, and Blasi et al. (2022) quantify the resulting global disparity in the utility of language technology. Joshi et al. (2020) supply the resource taxonomy from which we take the unlabelled resource proxy. Our outcome measure follows the compression view of language modelling (Delétang et al., 2024).

Graph and spectral analyses of transformers. Graph signal processing (Shuman et al., 2013; Ortega et al., 2018) supplies the formalism we adopt. Within transformer analysis, Dong et al. (2021) show that pure attention converges doubly exponentially to a rank-one matrix with depth, and Shi et al. (2022) analyse over-smoothing in BERT by identifying the attention matrix with a normalised adjacency matrix. A separate line applies training free spectral descriptors of the attention graph as deployment diagnostics, for hallucination detection (Noël, 2025) and for discriminating valid from invalid mathematical reasoning (Noël, 2026). These studies establish that the descriptors we use carry signal, but all operate within a single tokenizer, where graph order varies only with content length. The present work identifies the additional condition required when the node set is itself the treatment, and shows that under it most of the descriptor set is inadmissible.

Clark et al. (2019) established the broader practice of reading attention maps as structure, and Xiao et al. (2024) document attention sinks, which motivate our reporting of N -normalised alongside raw quantities. Ethayarajh (2019) characterise the anisotropy of contextual representations, which supplies our strongest non-graph baseline. Wendler et al. (2024) show that Llama models pass through an English adjacent latent space in the middle layers. That is a depth specific claim of the kind we found we could not support for our own effect (Appendix Q), and we cite it as related in subject rather than as corroboration.

L Robustness analyses in full

The twelve verdicts behind §5.4, and the checks whose decision rules were fixed in writing before the corresponding runs.

The eleven falsification analyses. Content coverage, within language graph size sensitivity, phylogenetic non-independence with standard errors clustered on 24 family branches, leave one language out over all 60, Benjamini–Hochberg correction across all 792 fits, per-layer refitting, cross model depth alignment, refitting within the Latin script subset, substitution of an ordinal resource band and of total edit count as a second proxy, ceiling normalisation, and tokenizer statistics as competing explanations. These analyses produce twelve verdicts because depth binning yields separate mid- and late-depth claims. Eleven of the twelve verdicts preserved the relevant conclusion. The single falsification is substantive and is reported in §5.1: the late-depth association is content coverage and is withdrawn.

The pre-specified adversarial analyses. A second set of checks used decision rules fixed in writing before the corresponding runs. Several changed a conclusion we would otherwise have stated, including the within-family null and the peak position test. Each is reported where it bears rather than collected as post-hoc reassurance. Table 9 lists the named checks that are described here.

Table 9: **Named pre-specified adversarial checks.** “Changed” marks an analysis whose recorded outcome altered a conclusion we would otherwise have stated: a withdrawn claim, a narrowed one, or a headline statistic replaced. It does not mark an analysis that merely returned a smaller number than expected.

Analysis	Outcome	Changed
Untrained control	association not reproduced	no
Free permutation null	free null too permissive	yes
Node set refinement	HFER, λ_2 drift 114/83%	yes
α -path identification	$R^2 = 0.059$, VIF 1.06	no
Truncation censoring	denominator selected on treatment	yes
Model size stability	mid-band and concordance fail	yes
All mediator BH screen	three mediators survive	yes
Citation and anonymity	clean	no
Stratified nulls	84% family reproducible	yes
Mediator first PC	common factor does not mediate	no
Peak position	not identified, withdrawn	yes
ρ sweep	0.1 not worst cutoff	no
Grid resolution	peak agreement is grid sensitive	yes
Depth heterogeneity	dispersion clears nulls	no
Maximum statistic stability	bootstrap SD 0.019	no

Table 10: **Coverage control alters the estimate, and does so differently at each depth.** Standardised HFER slope on the resource axis, matched scope, 60 languages, clustered errors. *Retained* is the controlled coefficient over the uncontrolled one. Both mid-depth estimates attain $p < 10^{-3}$. The 1B late-depth estimate is not significant after control and is not claimed.

Depth	Model	Raw	+ Coverage	Retained
mid	1B	+0.315	+0.323	102%
mid	3B	+0.191	+0.087	45%
late	1B	-0.348	-0.033 ^{n.s.}	9%
late	3B	-0.257	<u>-0.113</u>	44%

M Coverage collinearity

Coverage is substantially collinear with the remaining controls, correlating at approximately 0.58 with the resource axis, with script, and with NLL. Entering it alongside the full control set yields VIF 8.6 for coverage and 5.8 for the resource axis, inflating the standard error of the latter by a factor of approximately 2.4 through adjusters that are collinear with the quantity being adjusted, and the matched design already neutralises most of them by construction. For the binned coverage check, we therefore use a minimal specification (resource axis plus coverage, $VIF \approx 1.5$) and report the saturated specification alongside it. The two agree except on the 3B mid-depth cell, where the saturated specification yields $p = 0.11$. We report the disagreement rather than select the specification that resolves it.

N Corpus screening

Two failure modes of this proxy are screened in code. Cebuano reports 6,115,373 articles at 6.1 edits per article, consistent with bot generated stubs, and would otherwise enter as the second most resourced language in the study. Waray presents the same pattern. Separately, als_Latn denotes Tosk Albanian in Belebele, whereas als.wikipedia.org reports lang=gsw and is the Alemannic German wiki. Both are excluded. A fixed edits per article threshold was evaluated first and rejected, as it also flags Hausa and Burmese, which are genuine low-resource wikis. The adopted screen instead tests whether article count and total edit count rank the inventory consistently, with a maximum rank shift of 10 of 60.

O Refinement, generators and the cutoff fraction

The effect is generator dependent, which matters for both its interpretation and its direction: HFER rises 114% under a block community graph and 103% under sink plus uniform, and falls 28% under a banded one, so no monotone correction calibrates it away. Most of the movement occurs by $k \approx 4$, inside the empirical fertility range. The λ_2 drift is almost entirely the sink generator ($0.992 \rightarrow 0.167$), with the banded case flat to four decimals, and sink plus uniform is the applicable regime, since attention graphs are sink dominated (Xiao et al., 2024).

Dependence of the drift on the cutoff fraction.

HFER is defined by a cutoff on mode index, so the drift above is a function of ρ , and $\rho = 0.1$ was chosen for consistency with the extraction pipeline rather than on any principle. Sweeping it (Table 13) shows the drift grows monotonically as ρ shrinks (1.811 at $\rho = 0.02$ against 1.140 at our 0.10), so our reported figure is 63% of the worst case and not adversarially chosen. There is no ρ at which HFER becomes refinement invariant: the drift still exceeds 0.5 at every $\rho \leq 0.20$ and is 0.425 at $\rho = 0.50$.

The artefact runs against our result, so our estimate is conservative. The applicable generators drive HFER up under refinement, and refinement corresponds to higher fertility, hence to lower resource languages. Our finding is a positive resource axis slope: higher resource, less fragmented languages carry more mid-depth high-frequency energy. An uncorrected refinement artefact would therefore raise HFER for exactly the low-resource languages, flattening or reversing the slope we report rather than manufacturing it. Were the artefact operating in the content scope, the true effect would be larger than the measured one. This holds across the sweep: block and sink drift stays positive for every $\rho \leq 0.30$ (Table 13), so the direction of the argument does not depend on the cutoff.

P Truncation censoring of the modelling quality outcome

One detail of the derivation is worth reporting as a failure mode that generalises. A passage exceeding the token cap has surprisal measured over its truncated prefix but a character count covering the whole passage. The two available derivations either drop such items or match their denominator to

the prefix. Dropping them is wrong here: which passages truncate is determined by fertility, so the exclusion correlates with the treatment (-0.33 with the resource axis). It removes items from 12 of 60 languages, over 70% from seven of them, and all items for Burmese, which would otherwise have no modelling quality measurement at all. Matching the denominator to the scored prefix retains every item. It moves no language by as much as its own between item standard deviation, but moves 20 of 22 affected cells in the same direction, so the bias removed is systematic rather than noise. Bits per character was adopted precisely to be tokenization invariant, and the sampling had reintroduced a tokenization dependence behind it.

Q Identifiability of the peak position

We initially tested whether the argmax of the per-layer profile sits at an unusual relative depth. That question has no null hypothesis, and the reasoning is recorded here because the construction recurs. An argmax always exists. Under a null that removes the effect it is approximately uniform over the grid, so with 12 points every position carries probability $\approx 1/12$ and no position can be surprising without a prior privileging one. Our free permutation arm returned 21.5% of draws within one grid step of the observed peak, against $3/12 = 25.0\%$ under exact uniformity: it reproduced the uniform prediction. The statistic can neither reject nor fail to reject.

The within-family arm is worse, because it inverts the logic. Within-family permutation preserves 84% of the effect’s magnitude, hence everything downstream of the magnitude, including where the profile peaks. That 94.8% of within-family draws peak near the observed position restates the confound between family and resource level rather than testing anything. The diagnostic tell is that, for peak position, the family null was five times tighter than the free null and centred on the observed value, the signature of a null that obscured nothing.

Pre-specifying a decision rule prevents post-hoc rescue. It does not certify a statistic. We withdraw the position claim on the grounds of $P2$ and of its grid dependence (Table 11), not on this argument, and we claim depth heterogeneity instead (§5.1), which is identified without a privileged position.

Table 11: **Cross model peak agreement is an artefact of grid resolution.** Chance exact agreement exceeds the uniform expectation at every resolution, because smooth profiles co-locate. The observed agreement holds at 12 and 16 points but not at 8, where the peaks differ by more than half the depth range.

Grid	observed gap	chance agree.	uniform
8 points	0.571	19.6%	12.5%
12 points	0.000	11.9%	8.3%
16 points	0.000	10.5%	6.2%

R Stability of the maximum statistic

Our headline is a maximum over layers, the most sample sensitive kind of statistic, so we bound its sampling variability with a bootstrap over the 50 parallel passages (200 draws, resampled with replacement, and each drawn passage carries all 60 languages, which is the correct resampling unit for a parallel design). The 1B gives 0.6768 ± 0.0193 around an observed 0.6757, 95% [0.636, 0.713]. The 3B gives 0.6939 ± 0.0105 around 0.6932, 95% [0.674, 0.712]. The bootstrap SD is 41% and 21% of the corresponding family null SD and the whole interval lies above the family null’s 95th percentile, so the headline is not sample driven. This measures item level variability only: it resamples the existing passage pool and cannot see variance from passages absent from it, so it is not a substitute for a fresh draw.

S Node set refinement sweep

The construction referenced in §3.2. Each node of a base graph of order $M = 64$ is split into k children with a per-parent child weight profile, giving $W_{(i,a),(j,b)} = W_{0,ij} P_{ab}^{(i)}$ with $\sum_{ab} P_{ab}^{(i)} = k$, so that coarse-graining the child blocks recovers W_0 exactly. Three base generators, $k = 1, \dots, 8$, seed 0. Signals are 768-dimensional, which exceeds the largest refined node count $Mk = 512$. Had it not, the signal dimension rather than N would bound the effective rank and the ceiling the normalisation assumes would not be the binding one.

Per-parent profiles are required rather than a plain Kronecker product. A symmetric refinement makes the refined spectrum exactly degenerate (up to 87.5% of eigenvalues in tied clusters at $k = 8$), and HFER is defined by a cutoff on mode index, which is not well posed inside a tie. Under per-parent profiles the measured degeneracy is 0.000 at every k and every generator, and the figures below

are from that regime.

Table 12: **Drift under refinement with coarse structure held fixed.** Signed endpoint drift $(v(k=8) - v(k=1)) / |v(k=1)|$ per generator, hetero regime, and **norm.** is the same statistic after ceiling normalisation, reduced by $\max |\cdot|$. HFER and λ_2 satisfy Proposition 1 and move regardless, a fixed range leaving no ceiling to divide out. Normalisation repairs four of the five failing descriptors but overshoots effective rank[†] to -0.793 .

Descriptor	raw, by generator				
	block	band	sink	norm.	flat?
HFER	+1.140	-0.279	+1.028	1.140	yes
λ_2	-0.165	-0.000	-0.832	0.832	yes
$\lambda_{\max}$	+0.000	+0.007	+0.000	0.007	yes
smoothness	+0.000	+0.000	+0.000	0.000	yes
spec. entropy	+0.269	+0.189	+0.323	0.208	no
attn. entropy	+0.369	+0.426	+0.542	0.087	no
eff. rank [†]	+0.654	+0.654	+0.654	0.793	no
energy	+7.021	+7.022	+7.020	0.003	no
attn. distance	+7.005	+7.026	+6.998	0.003	no

Two rows in Table 12 are diagnostic of the taxonomy of §3.1. Effective rank[†] drifts identically to three decimals under all three generators: the generator changes the graph and the descriptor does not notice, which is what a residual stream quantity computed without the Laplacian does by construction. It is also the one descriptor on which this test and the cross scope test of Table 7 disagree (§3.3). Energy and attention distance both drift by about 7.0 because both are pure $O(N)$ accumulators: for any $f(k) = ck$ the endpoint drift is exactly $(8c - c)/c = 7$ whatever c may be, so the agreement is arithmetic rather than shared code. Their absolute values differ by four orders of magnitude, 42611 against 3.66 at $k = 1$.

Table 13: **The refinement drift of HFER across cutoff fractions.** Signed endpoint drift, hetero regime, same construction and seeds as Table 12, and the $\rho = 0.10$ row reproduces that table exactly. Our setting is shaded.

ρ	block	band	sink	max
0.02	+1.811	+0.031	+0.849	1.811
0.05	+1.420	-0.255	+1.092	1.420
0.10	+1.140	-0.279	+1.028	1.140
0.15	+0.834	-0.403	+0.817	0.834
0.20	+0.557	-0.458	+0.617	0.617
0.30	+0.170	-0.498	+0.257	0.498
0.50	-0.028	-0.425	-0.000	0.425

T Full mediation table

The thirteen rows not shown as survivors in Table 2, same specification throughout. None carries both legs after Benjamini–Hochberg correction. Ceiling normalised forms reproduce their raw counterparts to three decimals, which is expected in the matched scope where N is fixed and the correction divides by a constant.

Table 14: **Mediators that do not survive correction.** Leg failing correction given in the last column.

Mediator	a	b	ab	fails
attn. distance	+0.166	-0.233	-0.039	a
attn. entropy	-0.122	-0.213	+0.026	a
spec. entropy	-0.064	-0.200	+0.013	a
anisotropy	-0.009	+0.293	-0.003	a
energy	+0.026	-0.149	-0.004	a
resid. norm	+0.014	-0.151	-0.002	a
smoothness	-0.196	-0.151	+0.030	b
λ_2	+0.014	-0.226	-0.003	both
eff. rank (norm.)	+0.646	-0.371	-0.240	n/a
spec. entropy (norm.)	-0.065	-0.199	+0.013	a
attn. entropy (norm.)	-0.122	-0.212	+0.026	a
energy (norm.)	+0.026	-0.148	-0.004	a
attn. distance (norm.)	+0.165	-0.231	-0.038	a

Row count, since it is checkable: eleven raw descriptors plus five ceiling normalised forms is sixteen candidate mediators. Three appear in Table 2 as survivors (HFER, $\lambda_{\max}$, effective rank). Smoothness and λ_2 appear there too, as the two non-survivors most likely to be expected, and the thirteen remaining rows are above, of which smoothness and λ_2 are repeated for completeness. Sixteen unique mediators, no omissions.

Dimensionality of the mediator set. Ten of eleven b -legs are negative and eight lie between -0.147 and -0.244 , which invites the reading that there is one factor from geometry to modelling quality and the descriptors are noisy proxies for it. If so, “three survive correction” would be one finding plus a ranking of measurement noise. We tested this directly. It is not what is happening, although the descriptors are more correlated than their count suggests.

The mediators are substantially correlated (mean $|r| = 0.385$ off-diagonal) and two pairs are near-redundant: Dirichlet energy with residual norm at $r = 0.995$, and spectral entropy with anisotropy at $r = -0.98$. But the space is not one-dimensional. Four eigenvalues of the correlation matrix exceed unity (4.19, 3.70, 1.50, 1.01), the first principal

component captures only 38.1% of mediator variance, and the first two together 71.8%.

PC1 does not mediate. Fitted as a single mediator under the identical specification, its a -leg is -0.042 ($p = .44$) and its indirect effect is -0.012 , BCa 95% $[-0.045, +0.021]$, spanning zero. The tokenizer does not move the common factor. Controlling for PC1 attenuates the three survivors' b -legs by 31 to 41% but does not absorb them (effective rank $-0.363 \rightarrow -0.250$, HFER $-0.244 \rightarrow -0.169$). By contrast PC1 of the three survivors alone does mediate ($ab = -0.214$, $[-0.467, -0.033]$), loading $+0.78$ on effective rank.

The reading we adopt: what the tokenizer moves is a specific direction in mediator space, well approximated by effective rank, and not the dominant axis of variation among the descriptors. The uniform negativity of the b -legs is real but reflects a shared outcome rather than a shared cause. Normalised effective rank survives with its raw counterpart and is counted as the same descriptor, not a fourth. Two further robustness checks: the same three descriptors survive whether correction is applied over the eleven raw descriptors or all sixteen, and refitting on the censored modelling quality outcome (dropping truncated items rather than matching the denominator, 59 languages) changes no verdict, with HFER at $ab = -0.036$ $[-0.077, -0.008]$.

U Permutation nulls

All nulls permute the resource axis across the 60 language labels and broadcast the permuted value back onto every row of a language. Covariates stay attached to their own language. Permuting rows independently would break the clustering that fixes the effective sample at 60 rather than $\sim 90,000$ and would yield a null far too tight. 1000 draws, seed 0, matched scope.

The strata behind Table 16 are uneven. Family gives 24 strata with 49/60 languages in a permutable one and 11 singletons. Script gives 22 strata with 42/60 permutable and 18 singletons. Singletons contribute no permutation, so the script null rests on less freedom than the family null despite its wider margins.

The shaded rows carry the substantive limitation. A free permutation gives a null mean of ≈ 0 for this statistic. Holding family fixed gives $+0.333$, which is 84% of the observed $+0.395$. Most of

Table 15: **Label permutation nulls for the per-layer profile.** $p = 0.0010$ is the Monte Carlo floor at 1000 draws with finite permutation correction.

Model	Statistic	obs.	null mean $\pm$ sd	p
1B	mid-depth slope	+0.395	+0.000 $\pm$ 0.068	.0010
1B	max $ \beta $	+0.676	+0.167 $\pm$ 0.050	.0010
1B	lag-1 autocorr.	+0.349	+0.295 $\pm$ 0.210	.408
3B	mid-depth slope	+0.108	+0.001 $\pm$ 0.035	.0020
3B	max $ \beta $	+0.693	+0.194 $\pm$ 0.053	.0010
3B	lag-1 autocorr.	+0.507	+0.388 $\pm$ 0.247	.366
pair	sign concordance	0.583	0.593 $\pm$ 0.135	.657

Table 16: **Stratified nulls give a much stronger family constraint than script constraint.** The resource axis is permuted only within a stratum, so that stratum's structure survives every draw. "reach" counts draws attaining the observed value.

Stratum	Model	obs.	null mean $\pm$ sd	reach	p
<i>per-layer mid-depth slope</i>					
family	1B	+0.395	+0.333 $\pm$ 0.025	11	.0120
family	3B	+0.108	+0.073 $\pm$ 0.022	62	.0629
script	1B	+0.395	-0.036 $\pm$ 0.075	0	.0010
script	3B	+0.108	+0.022 $\pm$ 0.031	2	.0030
<i>peak layer slope (max β)</i>					
family	1B	+0.676	+0.486 $\pm$ 0.047	1	.0020
family	3B	+0.693	+0.497 $\pm$ 0.060	1	.0020
script	1B	+0.676	+0.207 $\pm$ 0.046	0	.0010
script	3B	+0.693	+0.224 $\pm$ 0.057	0	.0010

what the free null credits to the resource axis is family aligned. The 3B falls inside its family null at the conventional threshold, so for that model the resource axis reading is not supported once phylogeny is held fixed.

Script and family as strata. The two stratifications behave very differently and the reason is not the singleton count. Script has more frozen singletons (18 vs 11), which if anything would pull its null toward the observed value. What matters is how much resource variance survives within strata. Family branches are resource homogeneous: $\eta^2 = 0.701$, so 70% of the variance in log Wikipedia size lies between branches, and mean within-branch variance is 0.262 against a total of 0.714. Shuffling inside a branch therefore barely changes the labels. Script groups are not: $\eta^2 = 0.279$, the largest stratum is Latin with 31 of 60 languages spanning English to Yoruba, and mean within-stratum variance is 0.735, higher than the total. Shuffling within script is close to shuffling freely, which is why that null sits near zero. This also explains why the family test is the punishing

one: family and resource level are 70% collinear at the stratum level in any natural sample of languages.

Table 17: **Peak position against its nulls.** The observed peak positions coincide at relative depth 0.364. $P2$ tests whether that coincidence exceeds chance, as the cross model gap on the 12 point grid, with an observed value of 0.000. “match” is the share of draws reproducing the observed position ($P1$) to within one grid step, or the observed gap ($P2$).

Null	Statistic	null mean $\pm$ sd	match	p
free	$P1$ (1B position)	0.403 ± 0.323	21.5%	n/a
free	$P1$ (3B position)	0.489 ± 0.330	18.2%	n/a
free	$P2$ cross model	0.314 ± 0.267	11.9%	.120
family	$P1$ (1B position)	0.407 ± 0.091	94.8%	n/a
family	$P1$ (3B position)	0.402 ± 0.152	93.9%	n/a
family	$P2$ cross model	0.082 ± 0.134	45.5%	.456

Table 17 is why §5.1 claims amplitude rather than position. Note the family null’s spread (± 0.091) is a fifth of the free null’s and its mean (0.407) sits beside the observed 0.364: within family the peak lands where we observed it whatever the labels. Note also that $P2$ does not clear even the free null: with 12 grid points and profiles that are smooth by construction, two models peaking at the same point is a 1-in-8 coincidence.

The autocorrelation rows are why the shape of the profile is left uninterpreted. Permuted labels already yield smooth profiles, because adjacent layers are fitted on nearly the same observations. The concordance row is why the two plotted models are not presented as replicating that shape. The dispersion rows in Table 1, not the mid-depth or peak position rows, are what the depth claim in §5.1 rests on.

We note one further use of these nulls. A statistic defined by reading a template off an observed profile and then testing that profile against the template is circular, and we found no non-circular construction of the “phase sequence” statistic we had originally intended. Lag-1 autocorrelation is reported in its place because it is template free.