MULTI-SCALE INTRINSIC DIMENSION ESTIMATORS FOR TEXT EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Persistent homology dimension (PHDim) characterizes the geometry of text embeddings, but its single value combines structures at different scales. We introduce scale-resolved PHDim based on quantile-trimmed minimum spanning tree energies and prove consistent dimension recovery for every fixed quantile interval under regular sampling assumptions. In contextual token embeddings, we identify an approximate cell–skeleton geometry: repeated token types form local cells connected by a type-level skeleton. Building on this structure, we propose a three-band estimator with a fixed sampling budget for stable, informative comparisons between texts. Controlled interventions and synthetic experiments link its responses to contextual variability, lexical diversity, and the organization of longer-range geometry. Scale separation reveals a more consistent signal for AI-text detection: classical PHDim can be either higher or lower for generated texts, whereas removing the lexically dominated lower band yields a consistently nonnegative AI–human difference across the tested generators. Beyond AI-text detection, scale separation provides a more informative characterization of text complexity and defects. Different defects produce distinct three-band profiles: recognition errors predominantly raise fine and middle dimensions, repetitive boilerplate lowers middle and coarse dimensions, and disrupted semantic coherence yields a different pattern from phrase repetition. Together with scale-dependent differences in writing proficiency, these profiles enable a more granular analysis of text collections than a single dimension estimate.

1 INTRODUCTION

Language models represent each token occurrence as a vector that depends on its context. A text therefore gives rise to a cloud of points: repeated occurrences of the same token may occupy nearby locations or spread across the representation space as their contexts change. The geometry of this cloud offers a way to study how a model represents variation within a text. Intrinsic dimension summarizes the effective degrees of freedom in this cloud and has been used to study how textual properties relate to representation geometry and for AI-generated text detection (Vladislav et al., 2026; Tulchinskii et al., 2023). Unlike likelihood-based measures of token predictability, it describes how representations are organized. However, a dimension change may reflect different text properties, from lexical repetition to contextual diversity. Understanding these contributions is essential for meaningful comparisons between texts, datasets, or models.

One way to estimate intrinsic dimension is persistent homology dimension (PHDim), which uses a minimum spanning tree (MST): a graph connecting all points with the smallest total edge length and no cycles. PHDim estimates dimension from how MST energy—a sum of powers of edge lengths—grows with subsample size (Schweinhart, 2021; Tulchinskii et al., 2023). Unlike estimators restricted to a few nearest neighbours, it uses connections throughout the edge-length spectrum. However, it combines their contributions into one scaling exponent. A change in PHDim therefore does not reveal whether it arises from local contextual variation, lexical repetition, or the arrangement of more distant representations. This ambiguity matters in applications: the direction of the human–AI dimension difference can reverse across generators.

We address this problem by selecting quantile intervals of MST edge lengths and estimating a separate scaling exponent for each interval. On regular data, these estimators recover the same underlying dimension: choosing a band changes the energy constant but preserves its asymptotic exponent. On

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

structured finite samples, their differences instead reveal which parts of the geometry contribute to the observed scaling. Our goal is to interpret these profiles and establish a protocol under which they can be compared across texts.

The key empirical structure is an approximate decomposition into token-type cells and a skeleton connecting them. Occurrences of the same type cluster locally, while longer edges increasingly connect different types. In ModernBERT, the nearest neighbour has the same type for 98.9% of occurrences belonging to repeated types. Diagnostics across seven models and their layers show that this organization extends beyond a single representation, although its strength varies. The decomposition also exposes a sampling issue: as a sample grows, its numbers of occurrences and represented types need not grow proportionally, changing the balance between geometric scales. We therefore hold the sampling budget and subsample-size grid fixed within each comparison.

Controlled interventions link this geometry to linguistic changes. Fine-scale dimension tracks the spread of contextual representations within types; the middle band strongly reflects lexical occupancy; and coarse responses help distinguish fluent enrichment from disruptive variation. Synthetic clouds reproduce a substantial part of the observed response pattern, supporting a partial geometric explanation. We then use these insights to analyse AI-text detection, writing proficiency, and naturally occurring text errors. Across these settings, the central question is which property drives a dimension change, rather than whether a larger value universally indicates better text.

Our contributions are:

1. **A scale-resolved dimension estimator with theoretical guarantees.** We introduce quantile-trimmed MST energies and prove consistent dimension recovery for fixed quantile intervals under regular sampling assumptions (Section 3).
2. **A geometric explanation and a comparable measurement protocol.** We identify and test the cell–skeleton organization of token embeddings, explain its sampling dependence, and connect band responses to geometric and linguistic changes through controlled text interventions and synthetic experiments (Sections 4–5).
3. **An interpretable analysis of generation, proficiency, and text errors.** We show how lexical dependence contributes to reversals in dimension-based AI detection, identify an upper-band signal that persists after lexical matching, and demonstrate distinct profiles across proficiency groups and error sources (Sections 6–7).

2 Related Work

Intrinsic Dimension in Neural Networks. Intrinsic dimension is widely used to analyze neural networks, from their optimization landscapes to their learned representations. Li et al. (2018) and Aghajanyan et al. (2021) show that neural networks can be trained within random low-dimensional parameter subspaces, inspiring LoRA (Hu et al., 2022). Recent studies connect ID of the models representation space with scaling laws (Havrilla & Liao, 2024), training saturation (Razzhigaev et al., 2024), entropy and loss (Viswanathan et al., 2025). ID is used for architecture search (He et al., 2023), memorization (Arnold, 2025), and compositional semantics (Lee et al., 2025). Local ID (LID) patterns reflect training dynamics and overfitting (Ruppik et al., 2025), and differ between in-context learning and fine-tuning (Janapati & Ji, 2025). However, measured ID can also reflect properties of the estimator and embedding geometry. Quemy (2026) identify a hub of low-norm token embeddings that inflates nearest-neighbor ID estimates, including TwoNN. Removing the hub or normalizing embedding rows substantially reduces these estimates. In their Pythia experiments, hub removal also eliminates the reported increase in token-embedding ID with model size. These findings qualify interpretations of ID scaling in token-embedding tables.

Data-Centric View of Intrinsic Dimension. Narayanan & Mitter (2010) connect dataset ID to sample complexity, empirically confirmed on images (Pope et al., 2021). Tsukagoshi & Sasano (2025) show that task-specific embeddings can be aggressively reduced. Yin et al. (2024) use Local ID (LID) for hallucination detection: higher LID indicates fabricated content. Other work explores the separability of hallucinated and truthful samples in the representation space (Song et al., 2026).

Persistent Homology Dimension (PHD) in Neural Networks. PHD combines local and global properties of point clouds (Schweinhart, 2021). Birdal et al. (2021) link it to generalization in neural networks, while Tulchinskii et al. (2023) show that early GPT-generated texts have lower PHD than human-written ones, enabling detection; this gap shrinks for newer models (Kuznetsov et al., 2024). Kushnareva et al. (2024) track ID along the text to detect where human writing turns into machine generation. Recent work applies topological descriptors to LLM representations across layers and training (Gardinazzi et al., 2025; Birdal et al., 2021).

PHDim and Scale-Dependent ID. The connection between MST scaling and dimension estimation predates PHD. van de Weygaert et al. (1992) estimate generalized dimensions by analyzing the scaling of MST edge-length statistics across random subsamples. For a subsample of $m+1$ points, whose MST has edge lengths $\{\ell_i(m)\}_{i=1}^{m}$, they define the partition function $S(\tau, m) = \sum_{i=1}^{m} \ell_i(m)^{-\tau}$ and infer generalized dimensions from its scaling with m, averaging over repeated random subsamples to reduce sampling fluctuations. This construction connects directly to PH_0-based estimation: under the edge-length filtration convention, finite PH_0 lifetimes coincide with MST edge lengths, making $S(\tau, m)$ a persistence-energy sum with exponent $\alpha = -\tau$. PHDim recovers the dimension from how fast the MST energy grows with sample size n (Costa & Hero, 2004; Steele, 1988); since this energy sums over all MST edges, the estimate mixes all scales into a single number. Contrary, scale-dependent ID has been studied via the local slope of the correlation integral (Granata & Carnevale, 2016) and via neighbour rank in GRIDE (Denti et al., 2022); using the latter, Osman et al. (2026) show that ID of LLM representations switches between a local, within-word regime and a global, across-word regime. Scale-aware PH-based estimators additionally face noisy subsample statistics (Chazal et al., 2014) and high computational cost beyond PH_0 (Graf et al., 2026).

3 Scale-resolved persistent homology dimension

3.1 PHDim combines information across scales

PHDim estimates dimension from how the connectivity of a point cloud changes as more points are sampled. In zero-dimensional persistent homology (PH_0), points begin as separate components and merge as the distance threshold increases. With edge length as the filtration parameter, the finite death times are exactly the minimum spanning tree (MST) edge lengths. Thus the MST records both short connections within dense groups and longer connections needed to join them.

For n points, let $\ell_{(1)} \le \cdots \le \ell_{(n-1)}$ be these lengths. PHDim uses their α-energy and its growth exponent:

$$S_\alpha(X_n) = \sum_{i=1}^{n-1} \ell_{(i)}^{\alpha}, \qquad S_\alpha(X_n) \asymp n^{\beta}, \qquad d = \frac{\alpha}{1-\beta}. \tag{1}$$

For regular d-dimensional sampling, $\beta = 1-\alpha/d$; in practice it is fitted across subsample sizes (Steele, 1988; Schweinhart, 2021; Tulchinskii et al., 2023).

Unlike estimators based on a few nearest neighbours, PHDim uses the entire MST edge-length spectrum. This gives it access to several geometric scales, but combines them in a prescribed way: every edge contributes its α-power length. The resulting slope does not identify which part of the spectrum caused a change. Our aim is to make this choice explicit while retaining a dimension estimator on regular data.

3.2 Selecting an edge-length band

We select a scale by choosing a quantile interval $I = [a, b]$, with $0 \le a < b \le 1$. Since the $n-1$ MST edges are ordered by nondecreasing length, the corresponding partial energy is simply

$$S_{\alpha,I}(X_n) = \sum_{i=\lfloor a(n-1)\rfloor+1}^{\lfloor b(n-1)\rfloor} \ell_{(i)}^{\alpha}. \tag{2}$$

Thus $I = [0, 1]$ recovers the full PHDim energy, whereas a smaller interval restricts the sum to the selected edge ranks.

To estimate dimension, we draw repeated random subsamples at several sizes n. For each subsample, we build its MST, sort its edges, and compute the partial energy using the *same interval* I. We then average these energies at each size to obtain $\overline{S}_{\alpha,I}(n)$ and fit a straight line to $\log \overline{S}_{\alpha,I}(n)$ against $\log n$. Its slope $\widehat{\beta}_I$ gives $\widehat{d}_I = \alpha/(1-\widehat{\beta}_I)$ by equation 1. Boundary rounding and the precise fit are specified in Appendix B.1.

Varying I produces a dimension profile. The short-edge profile uses $I(q) = [0, 1-q]$ (discarding the longest fraction q of edges); the long-edge profile uses $I(q) = [q, 1]$ (discarding the shortest fraction q). Thus q is the fraction removed in either case. A third option moves an interval of fixed width h:

Profile	Retained interval $I(q)$	Effect of increasing q
Short edges	$[0, 1-q]$	Removes more of the longest edges
Long edges	$[q, 1]$	Removes more of the shortest edges
Interval	$[q, q+h]$	Moves a fixed-width window to longer edges

For the first two profiles $0 \le q < 1$; for the third $0 < h < 1$ and $0 \le q \le 1-h$. The text experiments use $h = 0.5$. Each profile value is $\widehat{d}_{I(q)}$ from a separate fit with q and α fixed across subsample sizes. Thus q selects a relative rank range, whose absolute edge lengths may shrink as n grows.

3.3 Dimension recovery and finite-sample profiles

The key requirement is that selecting an edge-length band should not change the dimension recovered from a regular distribution.

Theorem 1 (Dimension recovery under fixed quantile trimming)**.** *Let X_n consist of n independent samples from a density f on a compact convex polytope $K \subset \mathbb{R}^d$ with nonempty interior, where $0 < c \le f \le C < \infty$ on K. For any fixed $\alpha > 0$ and $0 \le a < b \le 1$, there is a constant $C_{\alpha,I}(f) \in (0, \infty)$ such that*

$$n^{\alpha/d-1} S_{\alpha,I}(X_n) \xrightarrow{\mathbb{P}} C_{\alpha,I}(f), \qquad I = [a, b]. \tag{3}$$

Consequently, $\log S_{\alpha,I}(X_n)/\log n \xrightarrow{\mathbb{P}} 1 - \alpha/d$. The fitted estimator $\widehat{d}_I$ is consistent when its sample-size grid grows proportionally, as specified in Appendix B.4.

The proof uses the MST law of large numbers of Penrose & Yukich (2001), followed by a quantile-integral argument that also covers bands reaching either endpoint (Appendix B). An isometric embedding into a higher-dimensional space leaves the result unchanged.

The theorem calibrates every fixed band to the *same* dimension on regular data. At finite sample sizes, however, different bands can have different effective slopes when the cloud contains structures at different scales. Figure 1 illustrates this contrast: the homogeneous control has an approximately flat profile, whereas structured clouds have band-dependent responses. Appendix C gives the full diagnostics. Such profiles describe finite-sample geometry; the theorem does not assert several distinct asymptotic dimensions, or guarantee accurate fits for arbitrarily narrow bands. We next ask which structures these bands resolve in contextual token embeddings.

4 Token cells, type skeletons, and lexical dependence

We now test whether contextual token embeddings exhibit a two-scale geometry: local cells of repeated token types connected by a type-level skeleton. This provides a geometric interpretation of the scale-dependent estimates introduced above.

4.1 How close are embeddings to a cell structure?

The cell model groups contextual occurrences by token type: occurrences of one type should remain close and connected, with longer connections joining different types. We test how closely the observed clouds follow this model using three diagnostics: same-type nearest neighbours, connectivity of each type in the MST, and the ordering of within-type versus cross-type edge lengths. These test local clustering, whole-cell coherence, and scale separation, respectively.

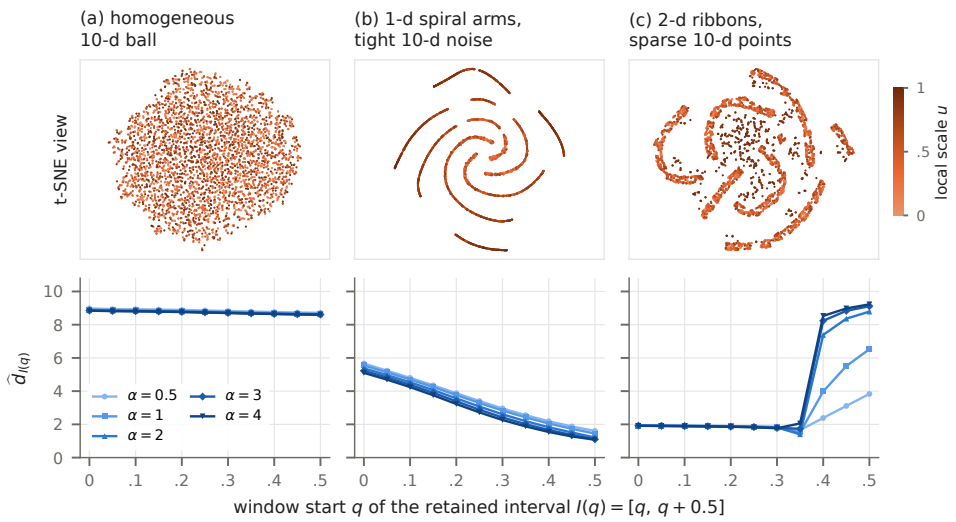


Figure 1: **A dimension profile distinguishes geometric regimes.** (a) A homogeneous 10-dimensional ball; (b) one-dimensional spiral arms with tight isotropic noise in $\mathbb{R}^{10}$; (c) a sparse multiscale cloud: two-dimensional ribbons (85% of points) and sparse points in a 10-dimensional ball (15%). Top: t-SNE views coloured by local scale u, the rank quantile of each point's nearest-neighbour distance among the MST edge lengths. Bottom: interval profiles $\widehat{d}_{I(q)}$, $I(q) = [q, q + 0.5]$, computed in the original space for several energy exponents α. The homogeneous control is flat; the structured clouds respond differently across bands.

ModernBERT-base embeddings (Warner et al., 2025) concentrate near a thin spherical shell, with a separate low-norm group that we call the *attention sink*. We remove this group using a norm threshold of 30. Appendix O documents the radial distribution, the threshold, and a sensitivity check retaining the sink.

In the last layer of ModernBERT, **98.9% of occurrences of repeated token types have a nearest neighbour of the same type**, after this filtering and special-token removal. Thus cells are supported by distances in the original embedding space, not just by the t-SNE illustration in Figure 2. They are unevenly populated: frequent types form large groups, while 48.5% of occurrences are singletons in samples of 201 embeddings.

Is this organization specific to ModernBERT? Layer-wise tests show strong same-type clustering throughout all five encoders examined: BERT, RoBERTa, DeBERTa, ModernBERT, and NeoBERT. In Qwen and Gemma, clustering is strongest in early layers and weakens deeper in the network. Whole-cell MST connectivity is more restrictive: ModernBERT remains nearly unfragmented throughout, whereas other models show layer-dependent departures from connected cells. Appendix D.1 presents the diagnostics and distinguishes local clustering from whole-cell connectivity.

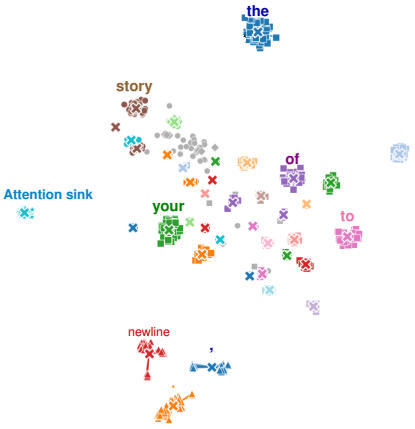


Figure 2: **Token cells.** t-SNE of token embeddings from a single text. Colours mark repeated types and grey points are singletons. (Appendix O).

Finally, edge composition tests whether the cells occupy a distinct range of scales. Within-type edges account for 100% of the shortest MST decile, but only 52.1% at quantiles 0.30–0.40 and 15.6% at 0.40–0.50. Conversely, 89.2% of the longest fifth of edges touch a singleton (Appendix L). Short edges therefore mainly connect occurrences inside cells, while longer edges increasingly connect the type-level skeleton and its singletons.

Together, the three diagnostics support an approximate cell–skeleton structure for ModernBERT: same-type occurrences are locally close and almost always

connected, while edge lengths separate the two levels only approximately. We now quantify what follows from connectivity alone, and what additionally requires separation.

4.2 THE CELL–SKELETON BOUNDARY AND SAMPLING DEPENDENCE

The cell structure connects edge scales to token counts. Let a sample contain n occurrences of N token types. A simple counting argument shows that, when each type forms a connected MST subtree, there are $n - N$ within-type edges and $N - 1$ cross-type edges (Lemma 1, Appendix E). If within-type edges are also shorter, the spectrum splits at

$$q^* = \frac{n - N}{n - 1}. \tag{4}$$

Edges below this quantile lie within cells, while those above it connect the type-level skeleton. With overlapping edge lengths, q^* is an approximate transition rather than a strict boundary. On 14,400 ModernBERT windows, it correlates with the empirical separating quantile at $r = 0.95$, compared with $r = 0.03$ after shuffling type labels (Appendix D.4).

This boundary depends on lexical composition and sample size. As a subsample grows, its within-cell edge count grows as $n - N(n)$ and its skeleton edge count as $N(n) - 1$. Since the number of represented types $N(n)$ generally grows nonlinearly with n, the proportion of short edges changes across sample sizes. A fixed quantile band can therefore mix the two scales in different proportions. Full PHDim is also affected: although it retains every edge, their changing composition alters the relative contributions of cells and skeleton to the energy. Consequently, changing the sampling range can change the fitted dimension, not merely its precision.

To compare texts under a common measurement protocol, we fix the reference sample at $N_{\mathrm{ref}} = 201$ occurrences and use the same regression grid, $n \in \{41, 61, 81, 101, 121, 161, 201\}$, regardless of document length. This controls the sampling budget while preserving lexical differences between texts. We use $\alpha = 1$, 32 subsamples per size, and three external seeds; Appendix J specifies the complete algorithm. The separate 128-token diagnostics use their own fixed protocol.

5 CONTROLLED VALIDATION OF THE GEOMETRIC INTERPRETATION

The cell–skeleton structure suggests that short-edge dimensions track contextual variation within types, intermediate bands respond to lexical repetition, and long-edge dimensions reflect skeleton organization. We test these links through controlled text transformations and geometric measurements; complementary synthetic experiments are reported in Appendix M.

5.1 PAIRED INTERVENTIONS AND FIXED DIMENSION BANDS

We apply 26 transformations to 50 English source texts: thirteen mechanical operations and thirteen LLM-generated rewrites. They vary repetition, vocabulary, topical diversity, and contextual organization. Each transformed text is compared with its source (*Original*), using the fixed ModernBERT (Warner et al., 2025) protocol of Section 4.2, including $N_{\mathrm{ref}} = 201$. Appendix I specifies the transformations.

We summarize each profile by three partly overlapping bands. Each is the arithmetic mean of separately fitted dimensions $\widehat{d}_{I(q)}$ on the following grid, with $I(q)$ fixed throughout every fit:

Band	Retained edge quantiles $I(q)$	Grid nodes q
Fine	$[0, 1 - q]$	$0.60, 0.65, \ldots, 0.90$
Middle	$[q, q + 0.50]$	$0.30, 0.35, \ldots, 0.50$
Coarse	$[q, 1]$	$0.50, 0.55, \ldots, 0.90$

These ranges emphasize cells, the transition, and the upper spectrum, respectively, without assuming perfect scale separation. We denote the median paired percentage change across source texts by Δ. Original means are 4.00, 15.96, and 21.09 (profiles and fit diagnostics in Appendix N).

To relate these responses to geometry, we measure four descriptors: normalized radius of repeated content-token cells r_c; the type-count boundary q^* (Eq. 4); and the normalized mean length μ_p

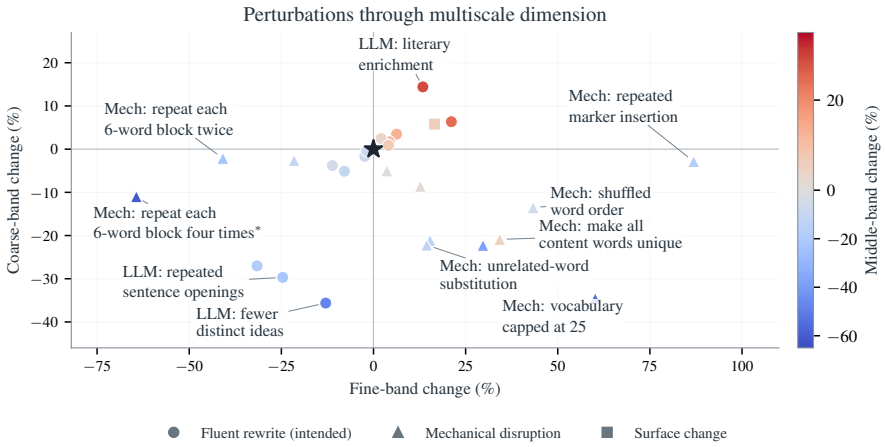


Figure 3: **Linguistic interventions produce distinct geometric responses.** Coordinates are median paired fine- and coarse-band percentage changes; colour gives the middle-band change. Circles: LLM rewrites intended to preserve fluency; triangles: disruptive mechanical operations; square: a surface operation. The star marks Original. The asterisk flags the unreliable fine estimate under strong repetition.

and coefficient of variation c_p of long edges joining singletons to repeated types (definitions in Appendix L).

Across 24 interventions included in the joint analysis, Δd_{F} correlates with Δr_c at $r = 0.898$, and Δd_{M} with Δq^* at $r = -0.926$. These correlations compare intervention-level median paired changes. They support links between fine dimension and cell spread, and between middle dimension and lexical richness: at fixed sample size, more types means a smaller q^*. Coarse changes depend on several descriptors; a linear model using all four gives in-sample $R^2 = 0.927$. Appendix M details the two exclusions, held-out validation, and synthetic experiments testing this geometric explanation.

5.2 From linguistic changes to geometric responses

Figure 3 connects these geometric associations to linguistic changes: position shows fine and coarse responses, while colour shows the middle response.

Fine scale: contextual variability. Cell radius measures the spread of a token's contextual representations. Its association with fine dimension helps explain why recurring patterns and phrase repetition lower this band. Conversely, word shuffling raises it by 43.3%: contextual variability can increase through disruption as well as linguistic enrichment.

Middle scale: lexical distribution. The association with q^* links middle dimension to the balance between token types and repeated occurrences. Phrase repetition approximately preserves relative unigram frequencies, whereas selective word repetition concentrates them. Their different middle-band responses therefore distinguish two forms of repetition through their effects on lexical occupancy.

Coarse scale: diversity and fluency. The linguistic interpretation is less direct, but none of the tested fluency-disrupting mechanical perturbations increases coarse dimension. Literary rewriting raises it by 14.4%, whereas unrelated-word substitution lowers it by 22.2%, despite similar fine increases. Fluent simplification can also lower coarse dimension. The band thus appears sensitive to diversity organized within coherent text, rather than fluency alone.

Together, the bands distinguish changes that a single dimension conflates. We next test whether separating these effects explains reversals in human–AI dimension differences.

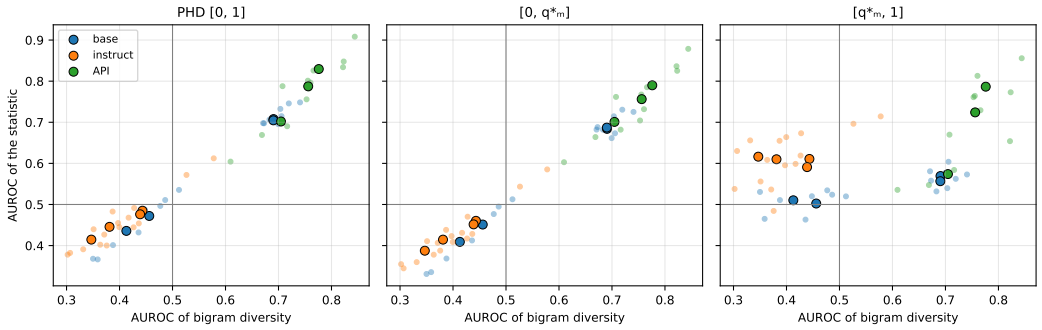


Figure 4: **Scale separation reduces lexical dependence.** Dimension AUROC versus bigram-diversity AUROC for classical PHD (left), lower band (centre), and upper band (right). Large markers summarize each generator across genres; small markers show individual genres. Colours distinguish base, instruction-tuned, and API models. Above 0.5, AI scores are higher; below 0.5, they are lower.

6 What drives dimension-based AI-text detection?

We compare dimension-based and lexical-diversity-based human–AI separation to understand what drives the detection signal and whether separating scales makes its direction more consistent. We use HAP-E-2 (Reinhart et al., 2024): 1,200 human and 13,200 generated continuations from 11 generators across four genres, measured with ModernBERT on fixed 128-token windows (Appendix Q).

AUROC uses AI as the positive class and the measured quantity as the score, without reversing its sign across generators. Values near one therefore mean consistently higher scores for AI texts, values near zero mean consistently lower scores, and 0.5 indicates little directional separation. Distance from 0.5 captures separation strength in either direction.

Classical PHD closely follows bigram diversity: their AUROCs fall on the same side of 0.5 for all 11 generators, with rank correlation 0.94 (Figure 4, left). API texts have higher PHD and diversity than human texts, while open-weight instruction-tuned models have lower values; API separation is also stronger. Base models vary, with Gemma-2 following the API direction. Thus a fixed classical-PHD decision direction does not transfer across generators.

We separate lower and upper bands at $q^*_{\mathrm{m}} = (1 + q^*)/2$, where q^* is the type-count boundary (Eq. 4). The cutoff lies above the cell–skeleton transition and remains fixed throughout subsampling. These diagnostic bands differ from the three summaries in Section 5.1.

The lower band retains the lexical association, whereas upper-band AUROC is at least 0.5 for all generators, with appreciable separation for instruction-tuned and API models. Lexical features explain 61% of classical-PHD variance on human texts but only 7% of upper-band variance. In 630 lexically matched human–AI pairs, the probability of higher AI dimension is 0.52 for classical PHD; for the upper band it remains 0.66 for instruction-tuned and 0.64 for API models, but 0.48 for base models. The upper band thus retains information beyond the matched lexical features for the first two groups, without establishing universal detection (protocol and robustness checks in Appendix Q).

The COLING analysis of earlier generators (Wang et al., 2025) further links dimension profiles to repetition, formulaic patterns, and semantic incoherence despite surface fluency. It also places GPT-3.5/GPT-4-family outputs near human scientific writing (`arxiv`) in the fine–coarse plane (Appendix V). Scale separation therefore helps distinguish lexical, defect-related, and genre-related contributions to the detection signal.

7 Proficiency and naturally occurring text errors

We examine writing proficiency using CEFR-labelled texts (Common European Framework of Reference for Languages, A1–C2) and ELLIPSE, a corpus of learner essays assessed on six aspects of English writing. Scale separation makes proficiency differences more pronounced (Figure 18, Appendix R). In ModernBERT, full PHDim ($q = 0$) only weakly distinguishes EL-

LIPSE Overall-score groups: their median dimensions range from approximately 6.8 to 7.3. Retaining the longest 20% of edges ($q = 0.8$) increases the gap between the lowest- and highest-score groups from about 0.5 to 2.2 dimension units (14.6 versus 16.8). For CEFR, the A1–C2 gap similarly grows from approximately 1.7 to 6.5. Conversely, restricting the profiles to the shortest edges makes higher-proficiency groups converge. This is consistent with our interpretation of coarse-scale geometry as sensitive to richer semantic organization: it reveals both subtle differences among more advanced essays and the pronounced gap between the simplest and more complex texts. The ordering is not strict: adjacent levels can overlap or reverse, and C1/C2 remain close. These are descriptive differences between group medians, not individual prediction results; full profiles across models and criteria appear in Appendices S and T.

We also compare paired reference and damaged texts from six corpora covering OCR, speech recognition, web extraction, learner errors, and vandalism, using a fixed sampling budget and matched formatting or source pages where appropriate (Appendix U). Since errors can introduce unusual word fragments, we track hapax-token share alongside dimension (Figure 5). Recognition errors modestly increase hapax share and primarily raise fine and middle dimensions, with smaller coarse changes. Repetitive web boilerplate instead lowers middle and coarse dimensions; learner errors have weaker effects, and vandalism changes little across bands. These distinct profiles can inform collection-level assessment of processing quality, although aggregate shifts do not imply reliable diagnosis of individual texts.

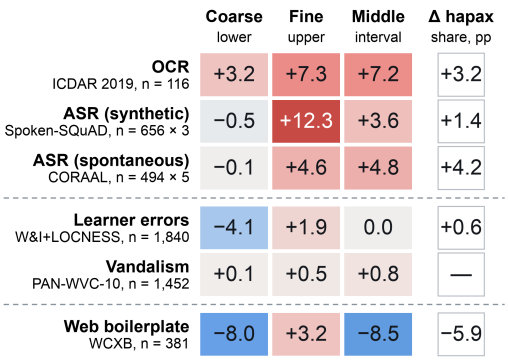


Figure 5: **Responses to text errors.** Median paired dimension changes (%; red: increase, blue: decrease) and hapax-share changes Δh (percentage points).

8 Conclusion and Limitations

Quantile-trimmed PHDim turns a single geometric summary into a scale-resolved profile while preserving dimension recovery on regular data. In contextual token embeddings, these profiles reveal an approximate cell–skeleton organization and distinguish changes in contextual variability, lexical occupancy, and the organization of longer connections. Accounting for sampling dependence is essential to making these comparisons meaningful.

This interpretation clarifies why classical dimension-based AI detection can change direction across generators and why different forms of proficiency, repetition, and text damage produce different responses. The profiles are therefore useful as comparative descriptors of text collections, rather than as a universal ranking of individual texts. Extending the analysis across representations—especially those that do not exhibit a cell–skeleton structure—is a promising next step towards geometric diagnostics of datasets and model outputs.

Our findings have several limitations. Most detailed experiments use ModernBERT; diagnostics across other models provide broader evidence but do not establish a universal linguistic interpretation, especially where the cell–skeleton structure fails. The consistency theorem concerns regular sampling distributions, while finite-sample profiles remain sensitive to sampling budget, band choice, and fit quality. Synthetic experiments support a partial geometric explanation, and the coarse band's connection to fluency and semantic diversity remains empirical. Finally, our conclusions concern aggregate and paired comparisons: lexical matching controls only selected features, and proficiency differences are descriptive. Reliable individual-text prediction and transfer across representations, languages, and unseen generators require further validation. The extended limitations of our analysis are discussed in Appendix A.

AI USE STATEMENT

In this work, we used generative AI tools to assist with proving mathematical claims. Specifically, GPT-6 Astra was used to compose latex code and organize graphs. Claude 5 Opus was used to create software code for text embedding structure analysis, and cvheck hypothese for stable PHDim calculation algorithm development. We did not use generative AI tools to propose hypotheses, design experiments, or interpret results.

Additionally, we used generative AI tools to edit the manuscript for grammar, clarity, and readability.

All AI-assisted proofs were checked line by line by authors, and errors found in the AI-generated drafts were corrected by the authors. No text was included without author review. We take full responsibility for the final content of this work, including text, claims, and proofs produced with the aid of generative AI.

REFERENCES

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 7319–7328, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.568. URL https://aclanthology.org/2021.acl-long.568/.

Stefan Arnold. Memorization in language models through the lens of intrinsic dimension. In Robin Jia, Eric Wallace, Yangsibo Huang, Tiago Pimentel, Pratyush Maini, Verna Dankers, Johnny Wei, and Pietro Lesci (eds.), *Proceedings of the First Workshop on Large Language Model Memorization (L2M2)*, pp. 23–28, Vienna, Austria, August 2025. Association for Computational Linguistics. ISBN 979-8-89176-278-7. doi: 10.18653/v1/2025.l2m2-1.2. URL https://aclanthology.org/2025.l2m2-1.2/.

Adrien Barbaresi. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In Heng Ji, Jong C. Park, and Rui Xia (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pp. 122–131, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-demo.15. URL https://aclanthology.org/2021.acl-demo.15/.

Tolga Birdal, Aaron Lou, Leonidas J Guibas, and Umut Simsekli. Intrinsic dimension, persistent homology and generalization in neural networks. *Advances in Neural Information Processing Systems*, 34:6776–6789, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/35a12c43227f217207d4e06ffefe39d3-Paper.pdf.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. The BEA-2019 shared task on grammatical error correction. In Helen Yannakoudakis, Ekaterina Kochmar, Claudia Leacock, Nitin Madnani, Ildikó Pilán, and Torsten Zesch (eds.), *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 52–75, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4406. URL https://aclanthology.org/W19-4406/.

Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, and Larry Wasserman. Stochastic convergence of persistence landscapes and silhouettes. In *Proceedings of the 30th Annual Symposium on Computational Geometry (SoCG 2014)*, pp. 474–483. ACM, 2014. doi: 10.1145/2582112.2582128.

José A. Costa and Alfred O. Hero. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Transactions on Signal Processing*, 52(8):2210–2221, 2004. doi: 10.1109/TSP.2004.831130.

Francesco Denti, Diego Doimo, Alessandro Laio, and Antonietta Mira. The generalized ratios intrinsic dimension estimator. *Scientific Reports*, 12(1):20005, 2022. doi: 10.1038/s41598-022-20991-1.

Murrough Foley. WCXB: A multi-type web content extraction benchmark, 2026. URL https://arxiv.org/abs/2605.21097.

Yuri Gardinazzi, Karthik Viswanathan, Giada Panerai, Alessio Ansuini, Alberto Cazzaniga, and Matteo Biagetti. Persistent topological features in large language models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 18811–18830. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/gardinazzi25a.html.

Florian Graf, Paolo Pellizzoni, Martin Uray, Stefan Huber, and Roland Kwitt. The flood complex: Large-scale persistent homology on millions of points. In *Advances in Neural Information Processing Systems*, volume 38, pp. 118356–118389, 2026. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/aba03e6f25decb32bda9c5bf81c58305-Paper-Conference.pdf.

Daniele Granata and Vincenzo Carnevale. Accurate estimation of the intrinsic dimension using graph distances: Unraveling the geometric complexity of datasets. *Scientific Reports*, 6:31377, 2016. doi: 10.1038/srep31377.

Alex Havrilla and Wenjing Liao. Understanding scaling laws with statistical and approximation theory for transformer neural networks on intrinsically low-dimensional data. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 42162–42210, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4a68bb0e075a3215d4ff37382ea598dc-Paper-Conference.pdf.

Xin He, Jiangchao Yao, Yuxin Wang, Zhenheng Tang, Ka Chun Cheung, Simon See, Bo Han, and Xiaowen Chu. NAS-LID: efficient neural architecture search with local intrinsic dimension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7839–7847, 2023. URL https://ojs.aaai.org/index.php/AAAI/article/download/25949/25721.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Saahith Janapati and Yangfeng Ji. A comparative study of learning paradigms in large language models via intrinsic dimension. In Vaibhav Adlakha, Alexandra Chronopoulou, Xiang Lorraine Li, Bodhisattwa Prasad Majumder, Freda Shi, and Giorgos Vernikos (eds.), *Proceedings of the 10th Workshop on Representation Learning for NLP (RepL4NLP-2025)*, pp. 59–86, Albuquerque, NM, May 2025. Association for Computational Linguistics. ISBN 979-8-89176-245-9. doi: 10.18653/v1/2025.repl4nlp-1.5. URL https://aclanthology.org/2025.repl4nlp-1.5/.

Tyler Kendall and Charlie Farrington. The corpus of regional african american language. Eugene, OR: The Online Resources for African American Language Project, 2023. Version 2023.06.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020. doi: 10.1073/pnas.1915768117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1915768117.

Laida Kushnareva, Tatiana Gaintseva, German Magai, Serguei Barannikov, Dmitry Abulkhanov, Kristian Kuznetsov, Eduard Tulchinskii, Irina Piontkovskaya, and Sergey Nikolenko. Ai-generated text boundary detection with RoFT. In *1st Conference on Language Modeling (COLM)*, volume 2024, 2024. URL https://hal.science/hal-04734023/.

Kristian Kuznetsov, Eduard Tulchinskii, Laida Kushnareva, German Magai, Serguei Barannikov, Sergey Nikolenko, and Irina Piontkovskaya. Robust AI-generated text detection by restricted embeddings. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 17036–17055, Miami, Florida,

USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.992. URL https://aclanthology.org/2024.findings-emnlp.992/.

Chia-Hsuan Lee, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. Spoken SQuAD: A study of mitigating the impact of speech recognition errors on listening comprehension. In *Interspeech 2018*, pp. 3459–3463. ISCA, 2018. doi: 10.21437/Interspeech.2018-1714.

Jin Hwa Lee, Thomas Jiralerspong, Lei Yu, Yoshua Bengio, and Emily Cheng. Geometric signatures of compositionality across a language model's lifetime. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5292–5320, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.265. URL https://aclanthology.org/2025.acl-long.265/.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://yosinski.com/media/papers/Li__2018__Measuring_the_Intrinsic_Dimension_of_Objective_Landscapes__.pdf.

Hariharan Narayanan and Sanjoy Mitter. Sample complexity of testing the manifold hypothesis. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (eds.), *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/8a1e808b55fde9455cb3d8857ed88389-Paper.pdf.

Arwa Osman, Marco Baroni, and Iuri Macocco. Local and global regimes of geometric complexity in language model representations, 2026. URL https://arxiv.org/abs/2608.14361.

Vladislav Pedashenko, Laida Kushnareva, Yana Khassan Nibal, Eduard Tulchinskii, Kristian Kuznetsov, Vladislav Zharchinskii, Yury Maximov, and Irina Piontkovskaya. Unveiling intrinsic dimension of texts: from academic abstract to creative story, 2025. URL https://arxiv.org/abs/2511.15210.

Mathew D Penrose and JE Yukich. Weak laws in geometric probability. *arXiv preprint math/0107149*, 2001. URL https://arxiv.org/abs/math/0107149.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *9th International Conference on Learning Representations, ICLR 2021, Austria, May 3-7, 2021*, 2021. URL https://openreview.net/forum?id=XJk19XzGq2J.

Martin Potthast. Crowdsourcing a wikipedia vandalism corpus. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pp. 789–790, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450301534. doi: 10.1145/1835449.1835617. URL https://doi.org/10.1145/1835449.1835617.

Alexandre Quemy. A hub of short rows inflates intrinsic dimension estimation of token embeddings, 2026. URL https://arxiv.org/abs/2608.29702.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264/.

Anton Razzhigaev, Matvey Mikhalchuk, Elizaveta Goncharova, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov. The shape of learning: Anisotropy and intrinsic dimensions in transformer-based models. In Yvette Graham and Matthew Purver (eds.), *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 868–874, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-eacl.58/.

Alex Reinhart, David West Brown, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, and Gordon Weinberg. Do llms write like humans? variation in grammatical and rhetorical styles, 2024. URL https://arxiv.org/abs/2410.16107.

Christophe Rigaud, Antoine Doucet, Mickaël Coustaty, and Jean-Philippe Moreux. ICDAR 2019 Competition on Post-OCR Text Correction. In *15th International Conference on Document Analysis and Recognition*, pp. 1588–1593, Sydney, Australia, September 2019. URL https://hal.science/hal-02304334.

Benjamin Matthias Ruppik, Julius von Rohrscheidt, Carel van Niekerk, Michael Heck, Renato Vukovic, Shutong Feng, Hsien-chin Lin, Nurul Lubis, Bastian Rieck, Marcus Zibrowius, and Milica Gasic. Less is more: Local intrinsic dimensions of contextual language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 67795–67828. Curran Associates, Inc., 2025. doi: 10.52202/085713-2276. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/61c2c6338033da68885e0226881cbe71-Paper-Conference.pdf.

Benjamin Schweinhart. Persistent homology and the upper box dimension. *Discrete & Computational Geometry*, 65(2):331–364, 2021. URL https://link.springer.com/article/10.1007/s00454-019-00145-3.

Yusheng Song, Lirong Qiu, Xi Zhang, and Zhihao Tang. Hallucination detection via internal states and structured reasoning consistency in large language models, 2026. URL https://arxiv.org/abs/2510.11529.

J. Michael Steele. Growth rates of euclidean minimal spanning trees with power weighted edges. *The Annals of Probability*, 16(4):1767–1787, 1988. doi: 10.1214/aop/1176991596.

Hayato Tsukagoshi and Ryohei Sasano. Redundancy, isotropy, and intrinsic dimensionality of prompt-based text embeddings. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25915–25930, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1330. URL https://aclanthology.org/2025.findings-acl.1330/.

Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Sergey Nikolenko, Evgeny Burnaev, Serguei Barannikov, and Irina Piontkovskaya. Intrinsic dimension estimation for robust detection of ai-generated texts. *Advances in Neural Information Processing Systems*, 36: 39257–39276, 2023.

Rien van de Weygaert, Bernard JT Jones, and Vincent J Martinez. The minimal spanning tree as an estimator for generalized dimensions. *Physics Letters A*, 169(3):145–150, 1992.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Karthik Viswanathan, Yuri Gardinazzi, Giada Panerai, Alberto Cazzaniga, and Matteo Biagetti. The geometry of tokens in internal representations of large language models, 2025. URL https://arxiv.org/abs/2501.10573v1.

Pedashenko Vladislav, Laida Kushnareva, Yana Khassan Nibal, Eduard Tulchinskii, Kristian Kuznetsov, Vladislav Zharchinskii, Yury Maximov, and Irina Piontkovskaya. Unveiling intrinsic dimension of texts: from academic abstract to creative story. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7916–7944, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.370. URL https://aclanthology.org/2026.eacl-long.370/.

Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Vladislav Mikhailov, Rui Xing, Zhuohan Xie, Jiahui Geng, Giovanni Puccetti, Ekaterina Artemova, Jinyan Su, Minh Ngoc Ta, Mervat Abassy, Kareem Ashraf Elozeiri, Saad El Dine Ahmed El Etter, Maiya Goloburda, Tarek

Mahmoud, Raj Vardhan Tomar, Nurkhan Laiyk, Osama Mohammed Afzal, Ryuto Koike, Masahiro Kaneko, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. GenAI content detection task 1: English and multilingual machine-generated text detection: AI vs. human. In Firoj Alam, Preslav Nakov, Nizar Habash, Iryna Gurevych, Shammur Chowdhury, Artem Shelmanov, Yuxia Wang, Ekaterina Artemova, Mucahid Kutlu, and George Mikros (eds.), *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pp. 244–261, Abu Dhabi, UAE, January 2025. International Conference on Computational Linguistics. URL https://aclanthology.org/2025.genaidetect-1.27/.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2526–2547, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.127. URL https://aclanthology.org/2025.acl-long.127/.

Helen Yannakoudakis, Øistein E. Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, 31(3):251–267, 2018. doi: 10.1080/08957347.2018.1464447.

Fan Yin, Jayanth Srinivasa, and Kai-Wei Chang. Characterizing truthfulness in large language model generations with local intrinsic dimension. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 57069–57084, 2024. URL https://dl.acm.org/doi/abs/10.5555/3692070.3694426.

Appendix

A LIMITATIONS

Dependence on the representation. Most of the detailed geometric analysis and controlled experiments use ModernBERT. Layer-wise diagnostics across seven models and proficiency profiles across four models provide broader evidence, but do not establish that the same interpretation holds for every encoder or layer. The estimator itself applies to point clouds independently of how they are constructed; its linguistic interpretation nevertheless depends on the embedding model. Our results establish a theoretically grounded method and demonstrate its practical interpretability in one extensively studied representation. Replication across models, languages, and embedding geometries remains necessary, particularly where the cell–skeleton approximation fails.

Finite-sample and interpretive scope. The consistency theorem concerns regular sampling distributions and fixed quantile intervals. It does not make every finite-sample band estimate an intrinsic property of a text, or establish distinct asymptotic dimensions for different bands. Sampling budget, band choice, and scaling-fit quality remain relevant. Synthetic experiments support a partial geometric explanation but do not fully reproduce response magnitudes. Likewise, the connection between coarse dimension, fluency, and semantic diversity is an empirical interpretation of the tested interventions, rather than a calibrated measure of these properties.

Collection-level analysis. Our primary aim is to explain aggregate differences and paired responses, rather than optimize decisions on individual texts. Genre, style, topic, authorship, and context can obscure these effects at the document level. We therefore do not provide a comprehensive comparison with dedicated text-level detectors or proficiency predictors. Lexical matching controls selected measured features, not every lexical or stylistic difference; proficiency-group comparisons are descriptive. Reliable individual prediction and transfer to unseen generators require separate validation.

B QUANTILE-TRIMMED MST ENERGIES: DEFINITIONS AND PROOF

This appendix justifies the estimator of Section 3. We first give an exact finite-sample definition, then prove the energy limit in Theorem 1, and finally connect that limit to the regression used in practice. The assumptions describe a regular sampling model; dependent contextual embeddings are investigated empirically in the subsequent sections.

B.1 EXACT TRIMMING AND ESTIMATION

Let $m = n - 1$. The main definition sums whole edges at ranks $\lfloor am \rfloor + 1, \dots, \lfloor bm \rfloor$. In the text experiments, the two boundary edges are instead weighted fractionally to avoid discrete jumps as n varies. This implementation detail does not affect the asymptotic limit, although the two finite-sample energies need not be identical. We denote the fractional version by T to distinguish it from the rank sum S:

$$T_{\alpha,[a,b]}(X_n) = \sum_{i=1}^{m} w_{i,n}([a,b])\ell_{(i)}^{\alpha}, \qquad w_{i,n}([a,b]) = m\,\big|[(i-1)/m, i/m] \cap [a,b]\big|. \tag{5}$$

Here the bars denote interval length. Interior weights equal one, exterior weights equal zero, and only boundary weights can be fractional. Set $Q_n(u) = \ell_{(\lceil mu \rceil)}$ for $0 < u < 1$; endpoint values do not affect integrals. Then

$$T_{\alpha,[a,b]}(X_n) = m \int_a^b Q_n(u)^{\alpha}\, du, \qquad |S_{\alpha,[a,b]} - T_{\alpha,[a,b]}| \le 2 \max_i \ell_{(i)}^{\alpha}. \tag{6}$$

Both definitions select edge ranks after constructing the MST of the full subsample. Neither requires recomputing the MST on selected points. For very small samples a whole-edge band can be empty; only positive energies can enter the log–log fit.

For subsample sizes n_j and R repeats, the practical fit is

$$\overline{S}_j = \frac{1}{R}\sum_{r=1}^{R} S_{\alpha,I}(X_{n_j}^{(r)}), \qquad \widehat{\beta}_I = \frac{\sum_j (\log n_j - \overline{\log n}) \log \overline{S}_j}{\sum_j (\log n_j - \overline{\log n})^2}, \qquad \widehat{d}_I = \frac{\alpha}{1-\widehat{\beta}_I}. \tag{7}$$

For fractional trimming, replace S by T throughout equation 7. We average energies before taking logs. All experiments using the three primary bands use $\alpha = 1$; their sampling grids and repetitions are specified in Appendix J. No subsample is independently rescaled. A common dilation of every cloud only changes the fitted intercept, since $S_{\alpha,I}(cX) = c^{\alpha} S_{\alpha,I}(X)$.

B.2 The limiting distribution of rescaled edge lengths

Work under the assumptions of Theorem 1. For $d \geq 2$, Theorem 2.3(ii) of Penrose & Yukich (2001) gives the following MST law of large numbers. For bounded continuous g, and for $g(t) = t^p$ with any fixed $p > 0$,

$$\frac{1}{n}\sum_{e \in \mathrm{MST}(X_n)} g(n^{1/d}|e|) \xrightarrow{L^2} \frac{1}{2}\int_K f(x)\, \mathbb{E}\left[\sum_{e \in E_0} g\Big(f(x)^{-1/d}|e|\Big)\right] dx. \tag{8}$$

Here E_0 consists of edges incident to the origin in the Euclidean minimal spanning forest of $\mathcal{P}_1 \cup \{0\}$, where $\mathcal{P}_1$ is a unit-intensity Poisson process in $\mathbb{R}^d$. The factor $1/2$ counts each undirected edge once. The support and density assumptions ensure the moment bounds for polynomial weights. For $d = 1$, the same formula follows from their general stabilizing-graph theorem: the two incident edges join the closest points on the left and right; these distances stabilize and have uniformly bounded rescaled moments under our density bounds.

Define a deterministic measure ν by

$$\int g\, d\nu = \frac{1}{2}\int_K f(x)\, \mathbb{E}\left[\sum_{e \in E_0} g\Big(f(x)^{-1/d}|e|\Big)\right] dx. \tag{9}$$

Taking $g = 1$ in equation 8 shows that ν has mass one because the finite MST has $n-1$ edges. It is supported on $(0, \infty)$: Poisson points are distinct, and all incident edge lengths are positive and finite. Its positive moments are finite. Thus, with

$$\nu_n = \frac{1}{n-1}\sum_{i=1}^{n-1} \delta_{n^{1/d}\ell_{(i)}}, \qquad Q(u) = \inf\{t : \nu((0,t]) \geq u\}, \tag{10}$$

we have weak convergence $\nu_n \Rightarrow \nu$ in probability and, for every $p > 0$, $\int t^p\, d\nu_n \to \int t^p\, d\nu$ in probability. Weak convergence follows by applying equation 8 to a countable convergence-determining class of bounded continuous functions. This step handles the dependence between MST edges; they are not treated as independent observations.

B.3 Proof of Theorem 1

Proof. The empirical quantile of ν_n is $\widetilde{Q}_n(u) = n^{1/d} Q_n(u)$. We show

$$\int_a^b \widetilde{Q}_n(u)^{\alpha}\, du \xrightarrow{\mathbb{P}} \int_a^b Q(u)^{\alpha}\, du. \tag{11}$$

First truncate lengths at $M < \infty$. Weak convergence implies convergence of quantiles at almost every $u \in (0,1)$, and bounded convergence therefore gives convergence of the integrals of $(\widetilde{Q}_n \wedge M)^{\alpha}$. For convergence in probability, apply this deterministic argument along an almost surely weakly convergent further subsequence of any subsequence.

To remove truncation, choose $p > \alpha$. Uniformly in $[a,b]$,

$$0 \leq \int_a^b \Big[\widetilde{Q}_n^{\alpha} - (\widetilde{Q}_n \wedge M)^{\alpha}\Big]\, du \leq \int_{(M,\infty)} t^{\alpha}\, d\nu_n(t) \leq M^{\alpha-p}\int t^p\, d\nu_n(t). \tag{12}$$

The last moment converges to a finite constant. The same bound holds for ν, so letting $M \to \infty$ proves equation 11, including $a = 0$ and $b = 1$. This avoids assuming that the quantile endpoints themselves converge or that the limiting distribution has a strictly positive density.

By equation 6,

$$n^{\alpha/d-1} T_{\alpha,[a,b]}(X_n) = \frac{n-1}{n} \int_a^b \widetilde{Q}_n(u)^\alpha \, du \xrightarrow{\mathbb{P}} C_{\alpha,[a,b]}(f) := \int_a^b Q(u)^\alpha \, du. \tag{13}$$

The limit is finite by the moment bound and strictly positive because $b > a$ and $Q(u) > 0$ for almost every $u > 0$.

It remains to pass from fractional weights to the whole-edge rank sum of the theorem. By equation 6, their normalized difference is at most $2 \max_i (n^{1/d} \ell_{(i)})^\alpha / n$. For $p > \alpha$,

$$\frac{\max_i (n^{1/d} \ell_{(i)})^\alpha}{n} \le n^{\alpha/p - 1} \left(\frac{1}{n} \sum_i (n^{1/d} \ell_{(i)})^p \right)^{\alpha/p} = o_{\mathbb{P}}(1).$$

Thus the same positive limit holds for $S_{\alpha,I}$. Taking logs gives

$$\log S_{\alpha,I}(X_n) = (1 - \alpha/d) \log n + \log C_{\alpha,I}(f) + o_{\mathbb{P}}(1), \tag{14}$$

which proves the claimed exponent for both boundary conventions. □

B.4 Consistency of the sample-size regression

The exponent limit alone does not justify an arbitrary regression grid. Here is a sufficient condition matching the usual proportional-subset construction.

Corollary 1 (Regression on a growing grid)**.** *Fix $J \ge 2$, distinct constants $0 < t_1 < \cdots < t_J \le 1$, and a fixed number $R \ge 1$ of repeats. From an i.i.d. pool of size N, draw uniform subsets of sizes $n_j(N) = \lfloor t_j N \rfloor$, using the same α and I at every size. Apply equation 7. As $N \to \infty$, $\widehat{\beta}_I \to 1 - \alpha/d$ and $\widehat{d}_I \to d$ in probability. The conclusion also holds for the fractional energy T. Subsets may overlap and need not be independent of each other.*

Proof. Each uniform subset has the marginal distribution of an i.i.d. sample of its size. For the finitely many size–repeat pairs, Theorem 1 and a union bound give simultaneous convergence of $n_j^{\alpha/d-1} S_{\alpha,I}$ to the same positive constant. Averaging over repeats preserves this limit, so

$$\log \overline{S}_j = (1 - \alpha/d) \log n_j + \log C_{\alpha,I}(f) + o_{\mathbb{P}}(1)$$

uniformly over j. Centered predictors converge to $\log t_j - J^{-1} \sum_k \log t_k$, with strictly positive sum of squares. Substitution into equation 7 proves slope consistency. Since $1 - (1 - \alpha/d) = \alpha/d > 0$, the continuous mapping theorem gives dimension consistency. □

Scope. This is an increasing-data guarantee, not a guarantee obtained by repeating subsampling of one fixed text indefinitely. Narrow bands, near-unit fitted slopes, or non-power-law energy curves can make finite-sample dimension estimates unstable. Our empirical fit diagnostics address this separately. The theorem covers full-dimensional densities on regular supports and their isometric embeddings in higher dimensions; it does not establish the corresponding result for arbitrary curved manifolds, fractals, mixtures of dimensions, or dependent token sequences.

B.5 Connection to fixed-rank and threshold descriptions

The function Q_n is also the ascending PH_0 barcode envelope: all components are born at zero, and the finite lifetimes equal MST lengths under our filtration convention. Equation 6 therefore integrates a chosen part of that envelope. At any continuity point u of the limiting quantile with $0 < Q(u) < \infty$, weak convergence gives $n^{1/d} Q_n(u) \to Q(u)$ in probability. A fixed-rank log-length slope then tends to $-1/d$. This explains the fixed-rank diagnostics in Appendix C; the primary estimator remains the band energy.

For a deterministic rescaled threshold t, the same edge law yields

$$n^{\alpha/d-1} \sum_{e \in \mathrm{MST}(X_n)} |e|^{\alpha} \mathbf{1}\{n^{1/d}|e| > t\} \xrightarrow{\mathbb{P}} \int_{(t,\infty)} s^{\alpha}\, d\nu(s),$$

provided $\nu(\{t\}) = 0$. Quantile selection avoids requiring the unknown d to set such a threshold. It holds a rank fraction fixed while allowing its physical length to adapt to sample size.

C SYNTHETIC CONTROLS FOR THE SCALE-RESOLVED ESTIMATOR

The controls distinguish two questions: whether band selection preserves a stable dimension on a homogeneous cloud, and whether it reveals different effective regimes on structured finite clouds. These are empirical illustrations, not additional assumptions of Theorem 1.

C.1 HOMOGENEOUS AND STRUCTURED POINT CLOUDS

The existing experiments compare a homogeneous 10-dimensional ball, a low-dimensional structure with local high-dimensional perturbations, and a sparse multiscale cloud. Their t-SNE views and interval-energy profiles appear in Figure 1; all distances and MSTs are computed before projection.

C.2 BARCODE AND FIXED-RANK DIAGNOSTICS

Barcode curves display ordered merge lengths at different sample sizes. Fixed-rank estimates compare the same relative rank across sizes, using the log-length scaling described in Appendix B.5. Their agreement on the homogeneous control and differences on the structured clouds motivate comparing several bands rather than relying on one endpoint of the spectrum.

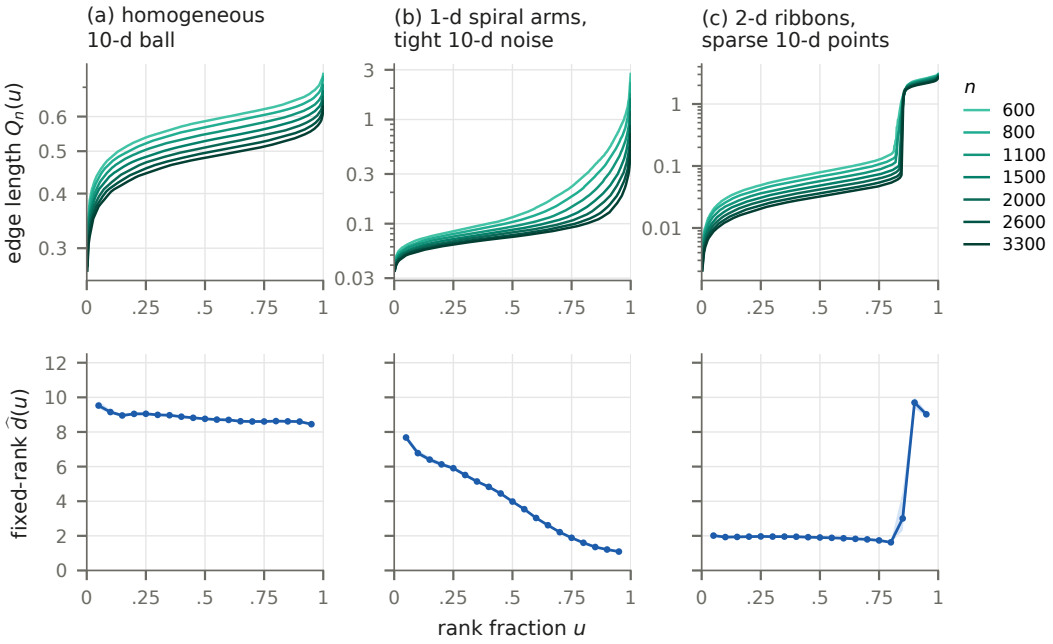


Figure 6: **Barcode curves and fixed-rank dimension estimates.** Columns follow Figure 1. Top: barcode envelopes $Q_n(u) = \ell_{(\lceil mu \rceil)}$, $m = n - 1$, averaged over subsamples of each size n (log scale); under $Q_n(u) \propto n^{-1/d}$ the curves are vertically shifted copies, spaced by d^{-1} times the log size ratio. Bottom: fixed-rank estimates $\widehat{d}(u) = -1/\widehat{\beta}(u)$, where $\widehat{\beta}(u)$ is the slope of $\log Q_n(u)$ against $\log n$; bands are 16–84% bootstrap intervals over subsamples.

C.3 COMPARING BAND CHOICES

Long-edge, short-edge, and interval energies select different regions of the same MST spectrum. The diagnostics show that a particular trimming rule can miss a transition visible to another: removing short edges alone does not resolve all the local high-dimensional perturbations of the low-dimensional structure. A profile is therefore more informative than a single selected estimate, but the observed values should not be equated with exact dimensions of individual components.

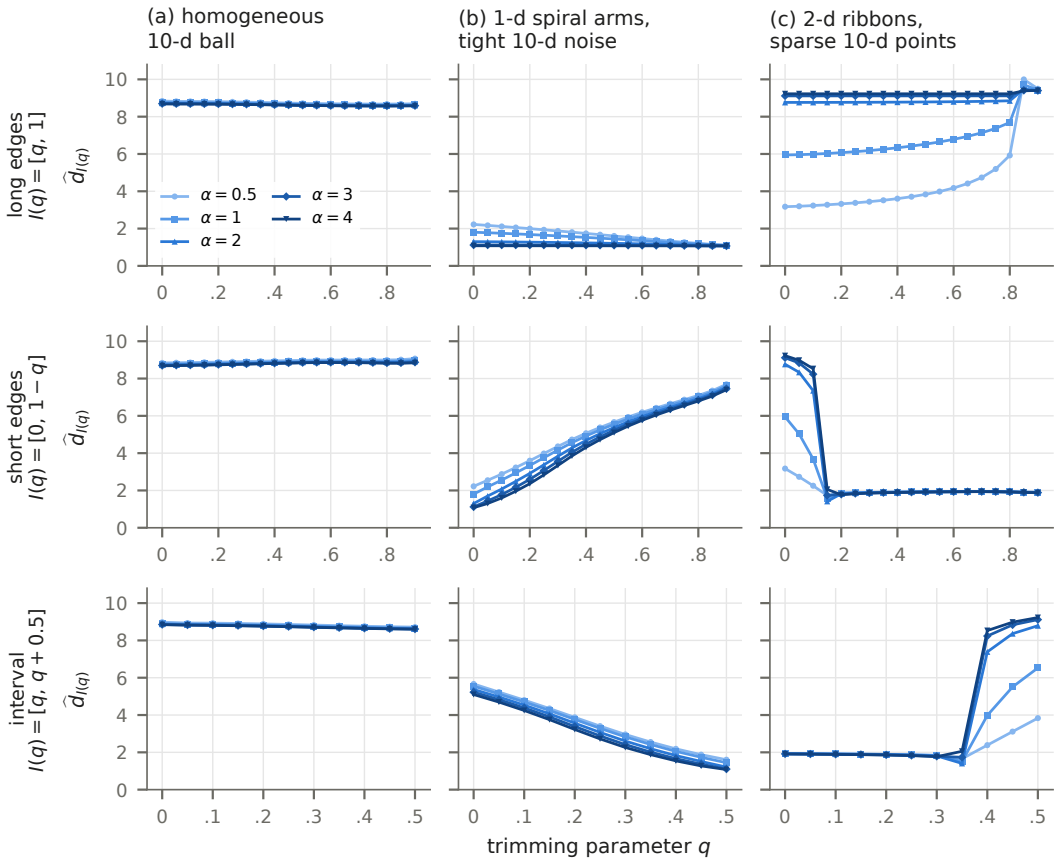


Figure 7: **Dimension profiles for three band choices.** Columns follow Figure 1. Rows: long-edge energies $I(q) = [q, 1]$, short-edge energies $I(q) = [0, 1-q]$ and interval energies $I(q) = [q, q+0.5]$; in every row q is the trimming parameter of Section 3.2. Colours and markers give the energy exponent α. These finite-sample profiles illustrate sensitivity to relative scale; consistency under regular sampling is established separately in Theorem 1.

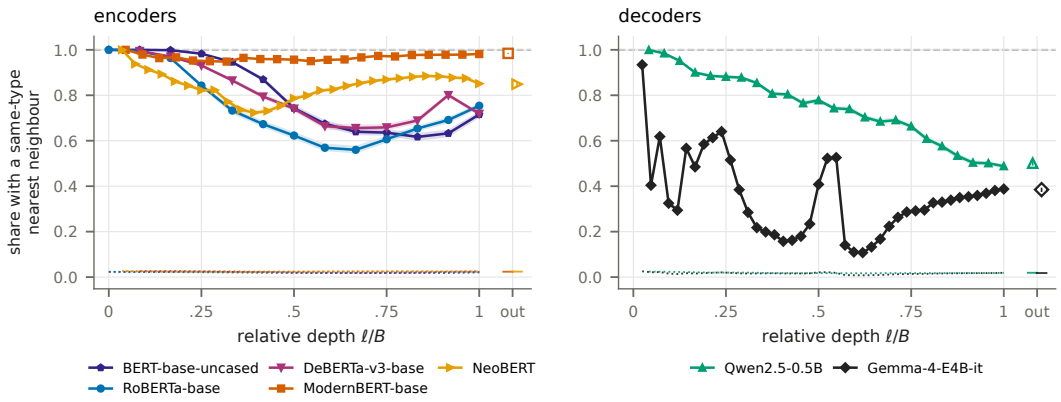


Figure 8: **Same-type clustering across models and layers.** Share of repeated-type occurrences whose nearest neighbour has the same type. Left: five encoders; right: two decoders. Dotted curves show the control obtained by shifting type labels along each text. Relative depth is the layer index divided by the number of blocks; `out` denotes the representation after final normalization. All models use 400 windows of 128 tokens.

D Cell diagnostics and separation conditions

D.1 Cell structure across models and layers

We examine whether the cell structure observed in ModernBERT extends to other representations using two complementary diagnostics: same-type nearest-neighbour agreement and within-type MST connectivity. Both are evaluated across layers on 400 windows of 128 tokens from seven models: BERT-base, RoBERTa-base, DeBERTa-v3-base, ModernBERT-base, NeoBERT, Qwen2.5-0.5B, and Gemma-4-E4B-it.

Local clustering. For occurrences belonging to repeated token types, we measure the proportion whose nearest neighbour has the same type. Figure 8 compares this proportion with a control that shifts type labels along each text. All five encoders retain a strong excess over the control throughout the network. ModernBERT stays close to one, while the other encoders show more pronounced layer dependence. Among the decoders, agreement declines gradually in Qwen and drops sharply after the first layer in Gemma, with substantial fluctuations in later layers.

Whole-cell connectivity. Nearest-neighbour agreement alone does not establish that all occurrences of a type form a connected cell: several disconnected groups can each contain same-type neighbours. We therefore also count the connected components of each type in the MST. The fragmentation count is zero when every type forms one connected subtree (Section 4.2). ModernBERT has almost no fragmentation at any layer, whereas the other models exhibit varying degrees of fragmentation (Figure 9).

These diagnostics establish different levels of support for the cell model. Local same-type clustering is widespread, particularly among encoders, but whole-cell connectivity has a more restricted model- and layer-dependent scope. Consequently, strong nearest-neighbour agreement alone is insufficient to justify the full cell–skeleton approximation.

D.2 Connectivity and separation

Let the n tokens of a window, with pairwise distinct distances, be partitioned by type into N groups X_t, so that $q^* = (n - N)/(n - 1)$. An MST edge is *within-type* if both endpoints have the same type and *cross-type* otherwise.

Lemma 1 (Counting identity)**.** *For any point cloud and type labels, the numbers of within-type and cross-type MST edges are*

$$E_{\text{within}} = n - N - F, \qquad E_{\text{cross}} = N - 1 + F. \tag{15}$$

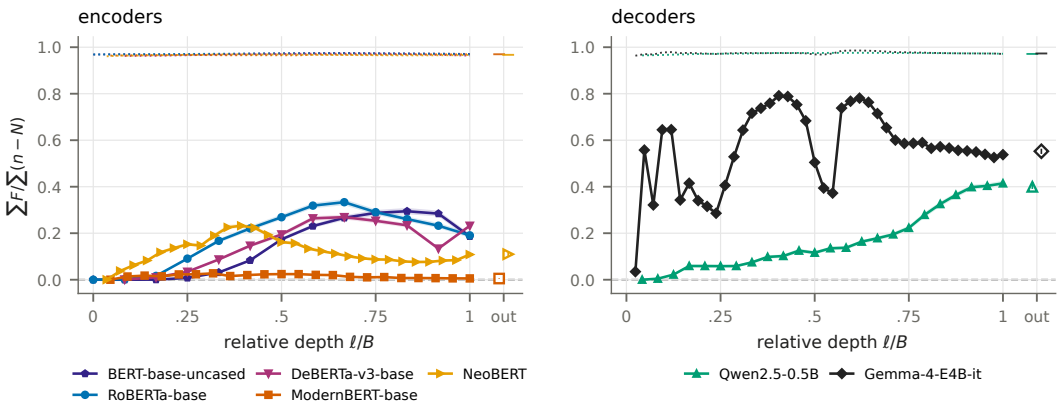


Figure 9: Fragmentation of types, $\sum F / \sum (n - N)$ over windows, by relative depth. Zero means that the occurrences of every type form a subtree of the MST (Lemma 1). Encoders left, decoders right; dotted: labels moved along the text.

Lemma 1 holds for any cloud and any labelling. Geometry enters only through the *fragmentation* F, which is zero exactly when the occurrences of every type form a subtree of the MST. The next statement gives a condition for this, and for the within-type edges to also take the short end of the spectrum.

Proposition 1 (Splitting)**.** *Let m_t be the longest edge of* $\mathrm{MST}(X_t)$ *($m_t = 0$ if X_t is a single token) and $d_t = \mathrm{dist}(X_t, X \setminus X_t)$. (i) If $m_t < d_t$ for every t, then $F = 0$, and the $N - 1$ cross-type edges (the* frame*) form the MST of the types with single-linkage distances. (ii) If $\max_t m_t < \min_t d_t$, every within-type edge is also shorter than every cross-type edge, and the spectrum splits exactly at the quantile q^*.*

For a type with two occurrences, (i) says that they are mutual nearest neighbours. The separated-cell model of Appendix K satisfies (ii) (Appendix E). Condition (i) is local, whereas (ii) is global and fails as soon as one type, typically punctuation or a function word, is spread wider than the gaps around it. A cloud can therefore have $F \approx 0$, so that the count $n - N$ is approximately correct, and still show the overlapping transition of Section 4.1: in ModernBERT windows of 128 tokens a median 8% of the $n - N$ shortest edges are cross-type (Appendix Q).

D.3 Nearest neighbours and fragmentation across layers

We embed 400 windows of 128 tokens with seven models at every layer: five encoders (BERT-base, RoBERTa-base, DeBERTa-v3-base, ModernBERT-base, NeoBERT) and two decoders (Qwen2.5-0.5B and Gemma-4-E4B-it). Fig. 8 repeats the nearest-neighbour test of Section 4.1, a local diagnostic of type clustering. This test alone does not imply condition (i). In all five encoders the share of repeated tokens whose nearest neighbour has the same type stays far above chance at every layer; Qwen2.5 shows a gradual decline, while Gemma-4 drops sharply after its first block and fluctuates thereafter. Fig. 9 measures F, which checks the whole tree. ModernBERT keeps it near zero at every depth. Other encoders show layer-dependent fragmentation: BERT, RoBERTa and DeBERTa become more fragmented in intermediate and upper layers, whereas NeoBERT peaks earlier and partially recovers. Qwen fragmentation increases with depth; Gemma-4 fragments sharply after its first block and remains strongly nonmonotonic.

D.4 Testing the type-count boundary

The independent HAP-E-2 diagnostic uses 14,400 ModernBERT windows of 128 tokens (1,200 human and 13,200 generated). The empirical boundary is the quantile that best separates within-type and cross-type edges. Regressing it on the count-based q^* gives slope 0.99, correlation 0.95, and mean offset -0.012. After shuffling type labels within each window, the slope is 0.00 and correlation 0.03. However, a median 8% of the first $n - N$ edges are cross-type; the transition spans

roughly quantiles 0.2–0.45. This supports a count-predicted, overlapping transition rather than exact separation. Appendix Q describes this corpus and its separate preprocessing protocol.

E TOKEN-CELL GEOMETRY AND BAND FORMULAS

E.1 MST FACTS AND PROOFS OF THE CELL STATEMENTS

Throughout, pairwise distances in X are distinct, so the MST is unique. We use three standard facts about the MST of a complete graph with distinct weights. *Cycle property:* an edge $e = \{x, y\}$ is not in the MST if and only if some x–y path uses only edges shorter than e. *Minimax property:* the MST path between x and y minimises, over all x–y paths, the length of the longest edge on the path. *Threshold property:* for every t, the number of MST edges of length at most t equals $|V| - \kappa(G_t)$, where $\kappa(G_t)$ is the number of connected components of the graph G_t of all edges of length at most t. Indeed, after Kruskal's algorithm has processed all edges of length at most t, its forest spans exactly the components of G_t.

Lemma 2 (Ordered MST edges depend monotonically on the weights)**.** *Let w and w' be weights on the complete graph on a vertex set V, and let $\ell_{(k)}$ and $\ell'_{(k)}$ be the k-th shortest edges of the corresponding MSTs. If $w'(e) \le \phi(w(e))$ for every edge e and a nondecreasing ϕ, then $\ell'_{(k)} \le \phi(\ell_{(k)})$ for every k. In particular, $|w' - w| \le \varepsilon$ on all edges implies $|\ell'_{(k)} - \ell_{(k)}| \le \varepsilon$, and $\lambda_1 w \le w' \le \lambda_2 w$ implies $\lambda_1 \ell_{(k)} \le \ell'_{(k)} \le \lambda_2 \ell_{(k)}$.*

Proof. By the threshold property, $\ell_{(k)} \le t$ if and only if $\kappa(G_t) \le |V| - k$. Every edge with $w(e) \le t$ has $w'(e) \le \phi(t)$, so $G'_{\phi(t)} \supseteq G_t$ and $\kappa(G'_{\phi(t)}) \le \kappa(G_t)$. Taking $t = \ell_{(k)}$ gives $\ell'_{(k)} \le \phi(\ell_{(k)})$. The two special cases follow with $\phi(s) = s + \varepsilon$ and $\phi(s) = \lambda_2 s$, and by exchanging the roles of w and w'. □

Proof of Lemma 1. The MST edges with both endpoints in X_t form a subgraph of a tree, hence a forest; it has $k_t = |X_t|$ vertices and c_t components, hence $k_t - c_t$ edges. Every MST edge is either within one type or cross-type. Summing over types, the number of within-type edges is $\sum_t (k_t - c_t) = n - \sum_t c_t = n - N - F$, and the remaining $(n-1) - (n - N - F) = N - 1 + F$ edges are cross-type. Dividing by $n - 1$ gives $(n - N - F)/(n - 1) = q^* - F/(n-1)$. □

Proof of Proposition 1. (i) Fix t with $k_t \ge 2$ and $x, y \in X_t$. The path from x to y inside $\mathrm{MST}(X_t)$ has its longest edge at most m_t. Any x–y path that leaves X_t contains an edge from X_t to its complement, of length at least $d_t > m_t$. By the minimax property the MST path from x to y does not leave X_t. Hence the MST edges inside X_t connect it, $c_t = 1$, and $F = 0$.

This subtree T_t is $\mathrm{MST}(X_t)$. Otherwise some edge $e \in \mathrm{MST}(X_t) \setminus T_t$ closes a cycle in T_t containing an edge longer than e; exchanging the two in the MST of X gives a lighter spanning tree, a contradiction.

Contract each X_t to a vertex. The $N - 1$ cross-type edges become a spanning tree T' of the complete graph on types. A simple path in a tree cannot leave a subtree and come back to it, so the MST path between points of X_t and X_u projects onto the t–u path of T'.

Each edge of T' between t and u is the shortest edge between X_t and X_u. Otherwise that shortest edge would close an MST cycle whose only cross-type edge is longer than it, contradicting the cycle property. So T' carries the weights $w_{tu} = \mathrm{dist}(X_t, X_u)$.

T' is the MST of the type graph with these weights. For $\{t, u\} \notin T'$, the edge realising w_{tu} is not in the MST. By the cycle property, the MST path between its endpoints uses only shorter edges, and its cross-type edges are the edges of the t–u path in T'. This is the cycle property in the type graph.

(ii) The condition implies (i), so the within-type edges are the edges of the trees $\mathrm{MST}(X_t)$ and have length at most $\max_t m_t$. A cross-type edge joins some X_t to its complement and has length at least $d_t \ge \min_t d_t$. Hence all within-type edges are shorter than all cross-type edges. By Lemma 1 with $F = 0$ there are $n - N$ of them, so they occupy ranks $1, \dots, n - N$, i.e. the quantiles up to q^*. □

The separated-cell model. In the model of Appendix K every X_t lies in a ball of radius r around its centre C_t and the centres are more than $4r$ apart. Then $m_t \le 2r$ and, by the triangle inequality, $d_t > 2r$, so condition (ii) of Proposition 1 holds. The converse fails: (ii) needs no common radius, no centres and no uniform gap. In this model the frame is also close to the MST of the centres. With

$D_{tu} = \|C_t - C_u\|$, the triangle inequality gives $D_{tu} - 2r \le w_{tu} \le D_{tu} + 2r$, and $2r < D_{tu}/2$, so $\frac{1}{2}D_{tu} \le w_{tu} \le \frac{3}{2}D_{tu}$. By Lemma 2 every ordered frame edge lies between $\frac{1}{2}$ and $\frac{3}{2}$ of the corresponding ordered edge of $\mathrm{MST}(C_1, \dots, C_N)$, so the energy of any fixed-rank band of the frame differs from that of the centres by a factor in $[2^{-\alpha}, (3/2)^{\alpha}]$, which does not change a growth exponent.

E.2 How type growth enters the fitted slope

The estimators of Section 3.2 fit the log energy of a band against $\log n'$ over subsamples of one window (Appendix J); write β for this slope and $\hat{d} = \alpha/(1-\beta)$. A subsample with N' types has its own boundary $q^*(n') = (n' - N')/(n' - 1)$, and $\gamma = d\log N'/d\log n'$ is the Heaps exponent measured over subsample sizes. For a band of $m(n')$ edges let $\mu = d\log m/d\log n'$. Throughout we assume condition (i) of Proposition 1 on every subsample, so that the frame of $N' - 1$ cross-type edges is defined. Let β_N be the growth exponent of the longest half of the frame against $\log N'$, and $\hat{d}_N = \alpha/(1-\beta_N)$.

Bands inside the frame. The first channel of Section 6 is absent when every edge of the band is a frame edge; the following lemma quantifies the second.

Lemma 3 (Frame bands, general form)**.** *Assume that on every subsample the $m \le N' - 1$ longest MST edges are cross-type, so that they are the m longest frame edges. Assume also that the energy of the m longest frame edges is proportional to $f^{\,1-\eta}$ in $f = m/N'$, with a shape parameter η constant over subsample sizes. Then $\beta = \gamma\,\beta_N + (1-\eta)(\mu - \gamma)$, and $0 \le \eta < 1$.*

Proof. By the assumptions the band is the top fraction f of the frame and has energy $S = S_{1/2}\,(2f)^{1-\eta}$, where $S_{1/2}$ is the energy of the longest half of the frame. So

$$\log S = \log S_{1/2} + (1-\eta)\big(\log f - \log \tfrac{1}{2}\big), \qquad \log f = \log m - \log N'.$$

Differentiating in $\log n'$:

$$\frac{d\log S_{1/2}}{d\log n'} = \frac{d\log S_{1/2}}{d\log N'} \cdot \frac{d\log N'}{d\log n'} = \beta_N\gamma, \qquad \frac{d\log f}{d\log n'} = \mu - \gamma.$$

Hence $\beta = \gamma\beta_N + (1-\eta)(\mu-\gamma)$. The energy of the top fraction f of the frame increases with f, so $1 - \eta > 0$. Further, $S/m \propto f^{-\eta}$ is the mean α-power length of the m longest frame edges; it does not increase when shorter edges are added, so $\eta \ge 0$. □

The condition of Lemma 3 concerns only the m edges of the band. It is weaker than condition (ii) of Proposition 1, which fails in real clouds near q^* but not in their upper tail (Appendix D.4). A band's lower boundary is either recomputed on every subsample ($m = N' - 1$ or $m \propto N'$) or kept at a fixed quantile of all edges ($m \propto n'$).

Proposition 2 (Frame bands)**.** *Under the assumptions of Lemma 3: (a) a recomputed boundary with the slope against $\log n'$ gives $\hat{d} = \alpha/(1-\gamma\beta_N)$, an increasing function of the Heaps exponent; (b) the same band with the slope against $\log N'$ gives $\hat{d}_N$; (c) a fixed-quantile boundary gives $1/\hat{d} = \gamma/\hat{d}_N + \eta(1-\gamma)/\alpha$. With $\chi = \eta\hat{d}_N/\alpha$ this equals $1/\hat{d}_N$ when $\chi = 1$ and otherwise moves with γ at the rate $(1-\chi)/\hat{d}_N$.*

Case (a) is the second channel of Section 6: isolating the frame is not enough, because the slope against $\log n'$ turns a clean frame band into a statistic of the Heaps exponent. Case (b) removes the lexical part exactly. Case (c) is the upper band $[q^*_{\mathrm{m}}, 1]$: it removes most of the lexical part, because a fixed-quantile band gains edges like n' rather than like N', and the residual vanishes at $\chi = 1$. If the frame edges grow like the MST of N' points of a d_{fr}-dimensional set, then $\hat{d}_N = d_{\mathrm{fr}}$; the scope remarks below list what this reading requires.

Proof of Proposition 2. (a) A recomputed boundary keeps $m = N' - 1$ or $m \propto N'$, so $\mu = \gamma$ and the second term of Lemma 3 vanishes: $\beta = \gamma\beta_N$ and $\hat{d} = \alpha/(1-\gamma\beta_N)$. For $0 < \beta_N < 1$ this increases with γ and equals $\alpha/(1-\beta_N) = \hat{d}_N$ only at $\gamma = 1$.

(b) By the chain rule, $d \log S / d \log N' = \beta/\gamma = \beta_N$, hence $\hat{d} = \hat{d}_N$.

(c) A band $[q, 1]$ with fixed q contains $(1-q)(n'-1)$ edges, so $\mu = n'/(n'-1) \approx 1$. With $\mu = 1$,

$$1 - \beta = 1 - \gamma\beta_N - (1-\eta)(1-\gamma) = \gamma(1-\beta_N) + \eta(1-\gamma).$$

Dividing by α gives $1/\hat{d} = \gamma/\hat{d}_N + \eta(1-\gamma)/\alpha$. Its derivative in γ is $1/\hat{d}_N - \eta/\alpha = (1-\chi)/\hat{d}_N$ with $\chi = \eta\hat{d}_N/\alpha$. At $\chi = 1$ the two terms combine to $1/\hat{d}_N$ for every γ; at $\gamma = 1$ only the first term remains. □

E.3 Mixtures of within-type and skeleton energies

This is the first channel of Section 6. A band with fixed quantiles $[a, b]$ that reaches below $q^*(n')$ contains edges of both kinds. Split its energy as $S = S_W + S_C$ into within-type and frame parts, and let $w = S_W/S$ and β_W, β_C be the slopes of S_W and S_C against $\log n'$.

Proposition 3 (Mixed bands)**.** *Assume condition (ii) of Proposition 1 on every subsample. (i) The band contains $(n'-1)\,\big|[a,b] \cap [0, q^*(n')]\big|$ within-type edges, and its frame edges lie at the relative frame positions $u = (x - q^*(n'))/(1 - q^*(n'))$ for $x \in [a, b]$, $x > q^*(n')$. Both depend on counts only. (ii) $\beta = w\,\beta_W + (1-w)\,\beta_C$. (iii) Rescaling every cell about its centroid by a factor c, while (ii) of Proposition 1 still holds, multiplies S_W by c^α and leaves β_W unchanged. If the frame part is held fixed, $\partial\beta/\partial \log c = \alpha\, w(1-w)(\beta_W - \beta_C)$.*

Proof of Proposition 3. (i) With fractional trimming (Appendix J) the band $[a, b]$ occupies the rank interval $[a(n'-1),\ b(n'-1)]$. By Proposition 1(ii) and Lemma 1 the within-type edges occupy the ranks up to $n' - N' = q^*(n')(n'-1)$. The overlap of the two intervals has length $(n'-1)\,|[a,b] \cap [0, q^*(n')]|$. The frame occupies the quantiles above $q^*(n')$, and $u = (x - q^*)/(1 - q^*)$ is the position of quantile x within it. Both quantities are functions of n', N', a and b alone.

(ii) With $'$ denoting $d/d \log n'$, $\beta = (S_W + S_C)'/S = (S_W/S)(\log S_W)' + (S_C/S)(\log S_C)' = w\,\beta_W + (1-w)\,\beta_C$.

(iii) Let $x \mapsto C_t + c\,(x - C_t)$ on each X_t, where C_t is the centroid. All distances within X_t are multiplied by c, so $\mathrm{MST}(X_t)$ keeps its edges and every edge length is multiplied by c; since the factor is common to all types, the order of within-type edges is also preserved. If condition (ii) of Proposition 1 still holds, the within-type edges of the MST are the edges of the trees $\mathrm{MST}(X_t)$ (proof of Proposition 1), they occupy the same ranks by part (i), and the within-type part of the band consists of the same edges scaled by c. Hence $S_W \mapsto c^\alpha S_W$ at every n' and β_W is unchanged. With S_C fixed, $w = c^\alpha S_W/(c^\alpha S_W + S_C)$ and $\partial w/\partial \log c = \alpha\, w(1-w)$; by part (ii) with β_W and β_C fixed, $\partial\beta/\partial \log c = \alpha\, w(1-w)(\beta_W - \beta_C)$.

The frame is not exactly fixed. Every point of X_t moves by $|c-1|\,\|x - C_t\| \le |c-1|R_t$ with $R_t = \max_{x \in X_t} \|x - C_t\|$, so every single-linkage weight w_{tu} changes by at most $|c-1|(R_t + R_u)$. By Proposition 1(i) the frame is the MST of the type graph with these weights, and by Lemma 2 every ordered frame edge moves by at most $2|c-1| \max_t R_t$. □

E.4 Interpretation of the band summaries

The grids are defined in Section 5.1.

- *Coarse* ($[q, 1]$, $q \ge 0.5$). The boundary lies above the Original value $q^* = 0.36$ and above $q^*(n')$ of smaller subsamples, so the coarse band is a frame band of kind (c). Its top is the singleton tail of Section 4.1: a singleton is a type of its own, so every edge attaching it is a frame edge. Interventions that raise q^* above 0.5 push the lowest grid nodes across the split.
- *Fine* ($[0, u]$, $u \le 0.4$). The band is within-type at the reference size, but $q^*(n')$ falls below u on smaller subsamples, where the band crosses the split. As n' grows it gains within-type edges and loses frame edges, so typically $\beta_W > \beta_C$, and by Proposition 3(iii) wider cells raise its slope through the weight w, although the slope of the within-type part does not change. This is a mechanism for the association of the fine band with cell radius (Section 5.1).

- *Middle* ($[q, q + 0.5]$, $q \in [0.3, 0.5]$). The band starts near q^*; its within-type share and the position of its frame part are both set by q^* (Proposition 3(i)), and its lowest nodes sit in the overlapping transition. Its changes follow q^*.
- *Classical PHD* ($[0, 1]$). The band contains all within-type edges and the whole frame, a recomputed band of kind (a); this decomposition needs only condition (i). Hence $\beta = w\,\beta_W + (1 - w)\,\gamma\beta_N$: the frame part is diluted by the Heaps exponent, and the weights move with q^* as the subsample grows, so classical PHD depends on counts even at fixed geometry.

E.5 Scope and finite-grid approximations

(1) *Finite grids.* Both statements are local in n'. On a finite grid of subsample sizes the least-squares slopes satisfy them exactly when γ, μ, η and w are constant over the grid; otherwise the slopes are weighted averages of the local values.

(2) *Which bands are which.* A band $[q, 1]$ with q fixed across subsample sizes, such as the coarse band or the upper band $[q^*_{\mathrm{m}}, 1]$ of Section 6, is of kind (c) as long as q lies above $q^*(n')$ on every subsample. Subsampling removes repeats, so $q^*(n')$ is typically below the window's q^*, and the condition typically holds for $q \geq q^*_{\mathrm{m}}$ and, in Original texts, for $q \geq 0.5$. The recomputed band $[q^*(n'), 1]$ is the whole frame, $m = N' - 1$ (kinds a and b). Classical PHD contains the whole frame and all within-type edges; its decomposition above uses Proposition 3(ii) with the frame part given by kind (a), and needs only condition (i) of Proposition 1.

(3) *Reading $\hat{d}_N$ as a dimension.* The shape assumption says only that the distribution of the longest frame edges keeps its form as the subsample grows; β_N is whatever growth exponent the frame has. Reading $\hat{d}_N$ as the dimension d_{fr} of the frame requires three further things. First, the frame must grow like the MST of N' points of a d_{fr}-dimensional set; Theorem 1 is stated for densities in $\mathbb{R}^d$, and a d_{fr}-dimensional subset of the embedding space needs its extension to d_{fr}-regular measures, which we do not prove here. Second, the frame uses single-linkage distances; the paragraph on the separated-cell model bounds its energy by that of the centroids, but under condition (i) alone no such bound is available. Third, the centroids present in a subsample must be an unbiased sample of all centroids. Subsampling keeps frequent types preferentially, and frequency correlates with position in the cloud (function words, punctuation), so none of the three is automatic.

F GEOMETRIC SIDE RESULT: A SHARED UNCERTAINTY DIRECTION

A text perturbation can change the mean location of embeddings as well as their relative geometry. This supplementary analysis identifies a shared direction of displacement under several corruptions, then tests whether that displacement explains changes in dimension. It is a geometric side result, separate from the band-validation argument.

F.1 A SHARED DISPLACEMENT COMPONENT ACROSS CORRUPTION OPERATIONS

Shuffling changes both the arrangement of embeddings and their mean displacement. To distinguish these effects, we evaluate a direction constructed in an earlier shuffling experiment and held fixed here. The construction averages type-centred displacements among occurrences in the lowest third of pointwise-mutual-information association scores. Type reference vectors ("homes") average at least five occurrences from the first 200 human texts. We project content-word displacements onto the resulting unit direction, normalize by the cloud's mean pairwise distance, and average within a text.

Fourteen sources in the present cohort also occur in the axis or reference-vector construction set. We therefore report the full cohort and the 36-source non-overlapping subset separately. Positive paired changes indicate movement in the direction defined by shuffling, which we call the *Universal Uncertainty Direction* (the Semantic Uncertainty Direction in the earlier draft). Here the name refers to the shared component tested across token classes and perturbations; its empirical scope is specified below.

Table 1: Mean projection changes relative to each text's Original. "Positive" counts sources with a positive paired change. Available sample sizes differ for coherent insertion.

	All sources			No construction overlap		
Condition	n	Shift	Positive	n	Shift	Positive
Original	50	+0.000	0/50	36	+0.000	0/36
Word shuffling	50	+0.336	50/50	36	+0.336	36/36
Unrelated hapax swap	50	+0.146	50/50	36	+0.144	36/36
Limited hapax swap	50	+0.035	50/50	36	+0.033	36/36
Coherent word insertion	42	-0.001	23/42	30	+0.000	17/30

The non-overlapping sources provide evidence of transfer: shuffling, unrelated hapax substitution, and limited substitution give positive changes in all 36 sources, with means $+0.336$, $+0.144$, and $+0.033$. Coherent insertion gives approximately $+0.0005$ in its 30 available sources. These interventions are not matched by word list or dose, so the contrast does not isolate semantic coherence. Nor is positive displacement universal under semantic impoverishment: fewer ideas and strong repetition need not move positively, while removal of punctuation or function words can do so. The random-direction control has condition-mean projections within $|0.0312|$.

An earlier class-level diagnostic found positive shuffling shifts for punctuation ($+0.187$), subwords ($+0.218$), rare content words ($+0.308$), frequent content words ($+0.338$), and function words ($+0.389$). Thus the shared component extends across classes and is strongest for function words in that diagnostic. The earlier shuffling and hapax-corruption directions have cosine similarity 0.743; the shuffling and sink directions have cosine 0.022. These preliminary class measurements are separate from the current cohort estimates in Table 1.

F.2 WHY DIRECTIONAL DISPLACEMENT DOES NOT EXPLAIN THE BAND RESPONSE

A common translation preserves all distances, since $\|(x_i + t) - (x_j + t)\| = \|x_i - x_j\|$, and consequently leaves every quantile-trimmed MST energy unchanged. Class-dependent shifts do not satisfy this exact invariance. However, the synthetic class-shift control changes fine and coarse bands by only -0.8% and $+0.2\%$. Inflating cells by 17% changes them by $+29.7\%$ and -7.2%, and combining both operations gives $+28.2\%$ and -7.0%.

The real shuffling response likewise includes a change in relative edge geometry: petiole mean falls by 1.64%, standard deviation by 1.23%, and CV rises by 1.07% (medians of the respective paired changes). Dispersion therefore contracts less than the mean. The directional shift is a

reproducible correlate of corruption, while cell spread and the long-edge distribution provide the stronger explanation of the dimension response. A shared component in mean displacement does not imply that individual words all move in exactly parallel directions.

Scale	Typical cluster content	Examples
Low q	punctuation / function words / common discourse markers	`. . .,the the,to to,he he he,said said,( ( (,cent cent,You you`
Middle q	repeated lexical items and domain-level semantic phrases	`code code code, shot shot shot, genomic genomic genetic, Language Language, proportion number,associated contributing, picture left right,initial The`
High q	rare fragments / names / domain-specific terminology	`narr narr, curv curv, Se Se, X X, Duch Duch, Manit Manit, proprio proprio, EW EW, No significant differences`

Table 2: Representative homogeneous clusters across MST scales. Low quantiles are dominated by frequent syntactic tokens and punctuation, middle quantiles contain repeated lexical or semantic patterns, while high quantiles capture rare domain-specific fragments, named entities, and specialized terminology.

G Embeddings geometry

G.1 Linguistic feature correlation vs text length

See Fig. 10 for details.

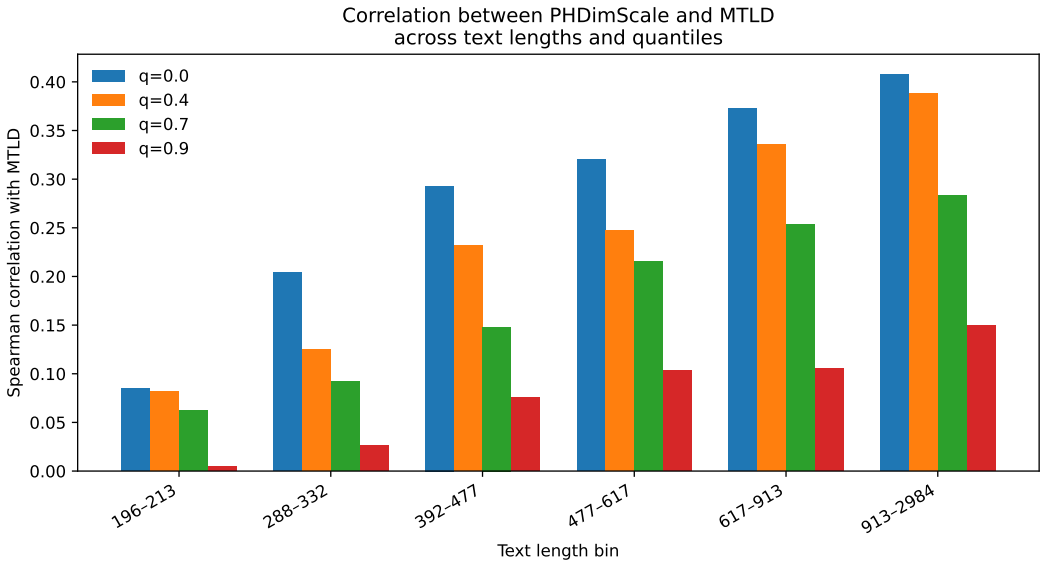


Figure 10: Dependence of the correlation between PHDimScale and lexical richness (MTLD) on text length for different MST quantiles. The relationship strengthens substantially for longer texts, while larger MST quantiles systematically reduce the effect.

G.2 Examples of homogeneous MST clusters

Table 2 shows representative homogeneous clusters sampled at different MST quantiles. Low-quantile clusters are dominated by punctuation and frequent function words. Intermediate quantiles contain repeated content words and short lexical units. High quantiles increasingly contain rare subword fragments, names, symbols, and domain-specific pieces.

H TECHNICAL DETAILS

H.1 COMPUTATION RESOURCES

For our experiments we used servers with the fol- lowing computational resources:

- 2 x NVIDIA GeForce RTX 4090 GPU + Intel(R) Core(TM) i7-14700K CPU
- 4 x NVIDIA A100 80 Gb + Intel(R) Xeon(R) Gold 6348 CPU @ 2.60GHz

I STUDY DESIGN AND INTERVENTION CATALOGUE

The study uses 50 English human-written texts from ten domains, with 282–688 words per source. Each source is assigned 13 mechanical interventions and 13 LLM rewrites, and every transformed text is compared with its own source. All embedding measurements use the same source cohort.

The catalogue varies lexical recurrence, contextual compatibility, topical diversity, and rhetorical structure. Intervention names specify operations or generation requests rather than independently validated semantic categories. We therefore interpret responses to these transformations without assuming that they isolate single linguistic variables.

I.1 MECHANICAL INTERVENTIONS

Mechanical operators act before encoder tokenization, using fixed lexical and function-word inventories; eligible content words generally exceed three characters. Word-level uniqueness therefore need not imply subword uniqueness. Randomized operations use recorded source-specific seeds. Table 3 specifies each operation and its side effects.

Table 3: Mechanical interventions. Operator identifiers link the descriptions to the analysis code.

Intervention / operator	Construction
Word shuffling `shuffle_words`	Randomly permute all whitespace-separated words, carrying attached punctuation with them.
Local repetition $\times 2$ `loop_local_1`	Repeat each six-word chunk twice, concatenate, then truncate to the original word count. Later original content is consequently lost.
Local repetition $\times 4$ `loop_local_3`	Repeat each six-word chunk four times and truncate to the original word count. This is stronger local repetition and content truncation, not a fourfold increase in output length.
Partial repetition `echo_p25`	Independently duplicate a six-word chunk once with probability 0.25, then truncate the concatenated output to the original word count.
Vocabulary cap 25 `collapse_vocab_25`	Retain the 25 most frequent eligible content words; map other eligible words to a retained word of similar length. Function words are retained.
Vocabulary cap 60 `collapse_vocab_60`	Apply the same vocabulary-capping procedure with 60 retained eligible content words.
All-unique expansion `expand_vocab_100`	Replace eligible content-word occurrences with fresh vocabulary matched approximately by length and frequency rank. Function words remain. Repeated content cells can disappear.
Unrelated hapax swap `hapax_swap_wide`	Replace eligible once-used words with length-matched unused words drawn across unrelated domains.

Intervention / operator	Construction
Single-domain hapax swap `hapax_swap_local`	Replace eligible once-used words from one foreign domain vocabulary; the replacements share a domain but need not fit the surrounding text.
Limited hapax swap `hapax_swap_wide_20w`	Apply cross-domain hapax substitution to a 0.2 share of eligible positions (approximately twenty words in the original design, not a fixed count).
Periodic marker `marker_p3`	Insert *item* after every three words, then truncate to the original word count. The suffix of the source is therefore removed.
Remove punctuation `strip_punctuation`	Replace non-word, non-whitespace characters by spaces and collapse whitespace. This also splits contractions and hyphenated words and erases punctuation boundaries.
Remove function words `drop_function_words`	Delete words in the implementation's fixed function-word inventory; no replacement is inserted. Output length and the available embedding pool decrease.

I.2 LLM-GENERATED REWRITES

The shared generation instruction requests English output, the source genre and register, a specified word count, and no explanatory commentary. The individual prompts in Table 4 specify the intended transformation. These outputs were generated and stored before the common-cohort embedding measurements; they were not regenerated to improve length compliance or geometric results. Measurement coverage is documented with the accompanying code.

Table 4: LLM-generated interventions. Descriptions state the requested change; they do not assume perfect generator compliance.

Intervention / stored key	Requested transformation
Fewer ideas `ideas_down_v2`	Express one central idea repeatedly, varying vocabulary and sentence shape without introducing a second distinct fact.
Single topic `topics_down_v2`	Keep every sentence on one narrow subject; remove excursions, outside examples, and references to other domains.
Lexically poorer `lexical_diversity_down_s`	Reuse a small content-word vocabulary while retaining the ideas and meaning.
No synonyms `syn_none`	Name each recurring notion with exactly the same word; avoid synonymous, pronominal, or paraphrastic substitutions.
Locked term context `context_locked`	Whenever a term recurs, repeat the surrounding phrase verbatim rather than changing or shortening its wording.
Multiple topics `topics_up_hard`	Make each sentence concern a different subject area while preserving continuous prose; the whole text should not have one assignable topic.
Lexically richer `lexical_diversity_up_s`	Avoid repeating content words and vary the words used for recurring ideas without intentionally adding ideas.
No repeated content word `syn_max`	Use each content word at most once, choosing context-appropriate synonyms or short paraphrases while keeping natural prose and the original content.
Rare synonyms `words_rare_s`	Replace common words with rarer, specialized, or literary synonyms; retain proper names, ideas, and sentence structure.
Literary rewrite `style_literary_T1.3`	Use literary or essayistic prose, varied rhythm, imagery, and an authorial voice, retaining subject matter and requested length. The stored condition uses temperature 1.3.
Coherent word insertion `inject_topic`	Integrate twenty sampled words from one foreign domain into a natural, factually coherent rewrite on the original subject. The word list is not matched to the mechanical hapax-swap condition.
Anaphora `style_anaphora`	Start every sentence with the same three or four words, retaining deliberate, well-formed prose and the original content.
Repeated record fields `style_record_fields`	Rewrite as structured records with identical field labels in identical order, one Label: value field per line, retaining the content.

J MEASUREMENT PROTOCOL

We extract the last hidden layer of `answerdotai/ModernBERT-base`. Special tokens `[CLS]`, `[SEP]`, and `[PAD]` are removed. The principal analysis also removes embeddings with Euclidean norm below 30 before sampling; Appendix O motivates this attention-sink criterion and reports the corresponding analysis with the sink retained.

The reference sample size is $N_{\text{ref}} = 201$ (denoted L in the experimental summaries). Each available pool contains $\min(402, T)$ surviving token occurrences, where T is the number available in the text. We sample without replacement using external seeds 1000, 1001, and 1002, with 32 replicates at each sample size

$$n \in \{41, 61, 81, 101, 121, 161, 201\}. \tag{16}$$

For a fixed edge-quantile interval I, let $S_I(n)$ denote its MST energy. We use $\alpha = 1$, so this energy is the sum of retained Euclidean edge lengths. Fractional weights on boundary edges implement exact quantile trimming. Energies are averaged over the 32 replicates at each n, after which ordinary least squares fits

$$\log \mathbb{E}_{\text{rep}}[S_I(n)] = a_I + b_I \log n, \qquad \hat{d}_I = \frac{1}{1 - b_I}. \tag{17}$$

Here $\mathbb{E}_{\text{rep}}$ denotes the empirical replicate average. The interval I is fixed throughout each regression. Energies use raw distances; individual subsamples are not rescaled by their own mean pairwise distance, which could otherwise change the fitted dependence on n.

Table 5: Universal bands used in the primary analysis. Each band averages dimension estimates over the stated q grid, with spacing 0.05.

Band	Retained edge quantiles	Averaged values of q	Code mode
Fine	$[0, 1 - q]$	$0.60, \ldots, 0.90$	`q_large`
Middle	$[q, q + 0.5]$	$0.30, \ldots, 0.50$	`q0.5_range`
Coarse	$[q, 1]$	$0.50, \ldots, 0.90$	`q_small`

For each intervention, we report the median percentage change across available texts, comparing each perturbed text with its own Original.

K From two-level geometry to quantile summaries

K.1 Why separated cells produce an MST boundary

The geometric model associates each token type with a centre C_j and places its contextual occurrences in a cell X_j. Singletons are cells containing one point. Suppose

$$X = \bigcup_{j=1}^{N} X_j, \qquad X_j \subset B(C_j, r), \qquad \delta = \min_{j \neq k} \|C_j - C_k\| > 4r. \tag{18}$$

Every within-cell edge has length at most $2r$. By the triangle inequality, every between-cell edge has length at least $\delta - 2r > 2r$. Kruskal's algorithm therefore connects each cell internally before selecting an edge between cells. Its within-cell forest contains

$$\sum_{j=1}^{N} (|X_j| - 1) = n - N \tag{19}$$

edges, and the remaining $N - 1$ edges connect the cells. This establishes the boundary $q^* = (n - N)/(n - 1)$. Between-cell distances differ from their corresponding centre distances by at most $2r$, so this second part of the tree approximates a skeleton on type centres.

This result supplies an exact interpretation under strong separation. The observed nearest-neighbour purity supports local cell structure, but does not verify the theorem's minimum-separation and maximum-radius conditions. We therefore use q^* as a geometric reference in real clouds, where edge populations may overlap.

K.2 Grouping dimensions without changing the estimator

Define the short- and long-edge profiles by $D_-(u) = \hat{d}_{[0,u]}$ and $D_+(u) = \hat{d}_{[u,1]}$. Each value is estimated using a fixed u in Equation equation 17. The short-edge convention corresponds to $u = 1 - q$ in `q_large`; the long-edge convention uses $u = q$ in `q_small`.

For each source–condition pair, let Q^* be its mean measured boundary at the reference size $L = 201$. The geometry-based summaries average the existing estimates over

$$\bar{d}_{<Q^*} = \frac{1}{|\mathcal{U}_-|} \sum_{u \in \mathcal{U}_-} D_-(u), \qquad \mathcal{U}_- = \{u \in \{.10, .15, \ldots, 1\} : u < Q^*\}, \tag{20}$$

$$\bar{d}_{>Q^*} = \frac{1}{|\mathcal{U}_g|} \sum_{u \in \mathcal{U}_g} D_+(u), \qquad \mathcal{U}_g = \{u \in \{0, .05, \ldots, .90\} : u > Q^*\}, \tag{21}$$

$$\bar{d}_{>q^{**}} = \frac{1}{|\mathcal{U}_+|} \sum_{u \in \mathcal{U}_+} D_+(u), \qquad \mathcal{U}_+ = \{.85, .90\}, \quad q^{**} = .80. \tag{22}$$

The profile values are also averaged over external seeds. These operations select which already estimated dimensions to average; they never introduce a sample-size-dependent trimming boundary. The primary geometric triplet is Q^*, $\bar{d}_{<Q^*}$, and $\bar{d}_{>q^{**}}$. In Original, their means are 0.3596, 3.88, and 23.36, respectively; the additional whole-skeleton summary $\bar{d}_{>Q^*}$ is 20.42.

K.3 Why universal bands remain useful

The fixed fine window averages $D_-(u)$ for $u = .10, \ldots, .40$, close to the typical cell boundary. The middle window samples the transition, and the coarse window averages a broader upper region than the two-point tail above q^{**}. They provide common measurement ranges across interventions. In contrast, a Q^*-based response can reflect both a changed profile and a changed set of quantiles included in its average. Appendix M reports the corresponding associations. Neither summary presumes that the empirical skeleton is perfectly homogeneous or that q^{**} is a second sharp structural boundary.

Uniform dilation preserves these dimensions: $S_I(cX) = cS_I(X)$ for $\alpha = 1$, so scaling every point changes the log-energy intercept but preserves its slope. The synthetic responses in Appendix M therefore arise from changes in relative structure.

L GEOMETRIC MEASUREMENTS AND EDGE POPULATIONS

L.1 FOUR DESCRIPTORS AND THEIR EMPIRICAL SUPPORT

The four descriptors distinguish changes within cells from changes in the type-level skeleton. For each sample, a repeated type's centre is its mean embedding and its radius is the mean distance of its occurrences from that centre. The descriptor r_c averages these radii over repeated content types, then divides by the sample's mean pairwise distance. At least three repeated content cells are required. This average radius characterizes typical cell spread; it differs from the maximum radius in the separation theorem.

The boundary $q^* = (n - N)/(n - 1)$ measures the balance of repeated occurrences and distinct types across all token classes. A singleton is defined within the sampled cloud. Petioles are between-type MST edges joining a singleton to a repeated type and lying at or above the 80th edge-length percentile. Their normalized mean is μ_p, and their standard deviation divided by mean is c_p. The definition does not require a singleton endpoint to have degree one. Pooled averages of these measurements are converted into percentage changes using the source-wise comparison described in Appendix J.

In Original, mean $N = 129.07$, singleton occurrence share is 0.4852, $r_c = 0.2285$, $\mu_p = 0.8363$, and $c_p = 0.0539$. Samples contain 8.82 repeated content cells and 12.29 petioles on average. Mean normalized radius across all repeated types is 0.2598; these are mean centroid distances, not cell diameters. The nearest-neighbour purity statistic excludes singletons, whereas the singleton share counts occurrences, not types.

The complete intervention-level dimensional comparison is summarized in Figure 12. It brings observations, geometric regressions, and the synthetic experiment onto the same rows and colour scale.

L.2 EDGE COMPOSITION ACROSS GEOMETRIC SCALES

We classify MST edges by whether they connect occurrences of the same type, distinct repeated types, a singleton to a repeated type, or two singletons. Figure 11 shows how their proportions change with edge-length quantile. Within-type edges account for 100% of the lowest decile, 91.3% at quantiles 0.20–0.30, 52.1% at 0.30–0.40, and 15.6% at 0.40–0.50. Their gradual replacement by between-type edges supports the cell–skeleton interpretation, with an overlapping transition near q^*.

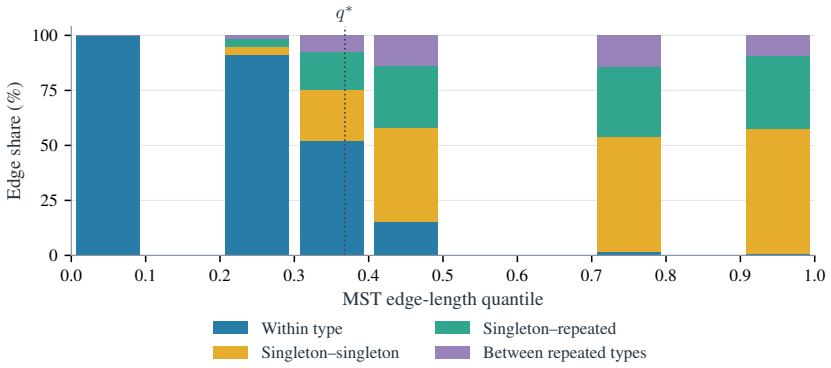


Figure 11: **MST edge composition across quantiles.** Short edges predominantly connect occurrences of the same type; longer edges increasingly connect distinct types, most often involving singletons. Stacked bars show percentages in the reported quantile bins; gaps indicate unreported bins. The dotted line marks the type-count boundary q^*.

The longest edges are predominantly incident to singletons. In the upper 20%, 58.51% join two singletons and 30.73% join a singleton to a repeated type; only 0.99% remain within-type edges. Thus 89.24% of this tail touches a singleton. This enrichment motivates distinguishing a singleton-sensitive coarse scale within the broader type-level skeleton.

M GEOMETRIC ASSOCIATIONS AND SYNTHETIC MECHANISMS

Four geometric measurements strongly track the dimensional responses of real text embeddings. Figure 12 compares these observations with fitted geometric regressions and an exploratory synthetic reconstruction. The shared layout makes both agreement and failures visible; the two model columns answer different questions.

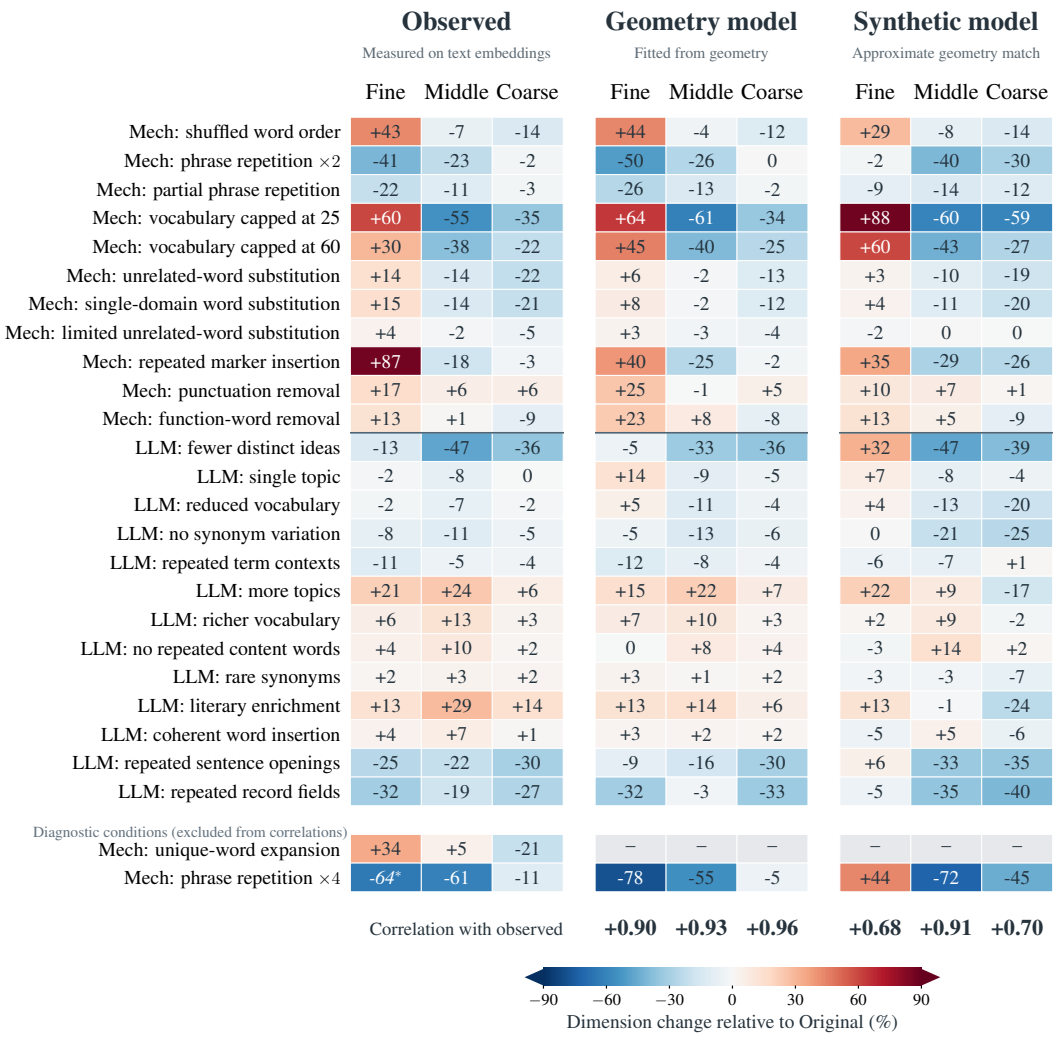


Figure 12: **Observed, geometric-model, and synthetic band changes.** All panels share intervention rows and a colour scale. The geometric panel shows in-sample fitted values. The synthetic panel reports median paired changes on twelve new clouds per condition, calibrated using geometry alone. Correlations use the 24 main conditions. Geometry matching is approximate; the synthetic column is a partial mechanistic reconstruction. The star marks unreliable observed fine scaling; grey cells have no comparable radius measurement.

M.1 STRONG EMPIRICAL ASSOCIATIONS

For each intervention, Δ denotes the median paired percentage change defined in Appendix J. Fine changes correlate with cell-radius changes ($r = 0.898$), while middle changes correlate with the type-count boundary ($r = -0.926$). Coarse changes correlate positively with mean petiole length ($r = 0.920$) and negatively with petiole CV ($r = -0.747$). The regressions exclude all-unique expansion, which leaves too few repeated content cells, and phrase repetition $\times 4$, whose fine-scale fit is unreliable; both remain in the descriptive results. These associations motivate ordinary least-squares summaries on the same 24 retained interventions:

$$\widehat{\Delta d}_{\mathrm{F}} = 1.765 + 2.262\Delta r_c, \tag{23}$$

$$\widehat{\Delta d}_{\mathrm{M}} = -3.727 - 0.744\Delta q^*, \tag{24}$$

$$\widehat{\Delta d}_{\mathrm{C}} = -0.646 - 0.113\Delta r_c - 0.120\Delta q^* + 5.396\Delta \mu_p - 0.331\Delta c_p. \tag{25}$$

Original supplies the paired denominators and is not added as a regression observation at zero. Correlations between fitted and observed responses are 0.898, 0.926, and 0.963 for fine, middle, and coarse bands; the corresponding R^2_{model} values are 0.806, 0.858, and 0.927. Thus a small set of geometric measurements accounts for much of the variation between interventions, although individual errors remain: marker insertion raises fine dimension by 86.9%, compared with a fitted 40.0%.

Leaving out one intervention at a time and refitting the coefficients gives pooled held-out R^2 values of 0.758, 0.826, and 0.818, with RMSEs of 13.59, 8.14, and 5.81 percentage points. This tests transfer across intervention types on the same source cohort, not to new texts; feature selection and diagnostic exclusions are not repeated within each fold. Predictor correlations also prevent interpreting individual regression coefficients as identified causal effects.

Sensitivity of the association. Restoring repetition $\times 4$ gives in-sample/held-out R^2 of $.842/.806$, $.888/.865$, and $.925/.812$; including all 26 conditions gives $.832/.798$, $.882/.859$, and $.885/.709$, using the selectively available radius for all-unique expansion. The q-based summaries show similar associations: the short-edge average below q^* correlates with radius at 0.875; the long-edge averages above q^* and q^{**} correlate with mean petiole length at 0.925 and 0.916, and with CV at -0.777 and -0.750. The short-edge radius model has $R^2 = 0.766$ (held-out 0.704), and the four-predictor upper-tail model has 0.877 (0.634). Changing the petiole cutoff from .80 to .75 or .85 gives coarse-model R^2 of .931 or .936 (held-out .820 or .866). The association is therefore not confined to a single band definition or tail cutoff.

M.2 SYNTHETIC INTERVENTIONS AND PARTIAL CAUSAL SUPPORT

Direct manipulation of a point cloud complements the observational regressions. In the earlier controlled sweeps, increasing the cell-radius parameter from .232 to .271 raises fine dimension by 32.4%, with middle and coarse changes of $+4.3\%$ and -6.1%. Increasing singletons from 94 to 110 raises middle dimension by 10.0%; moving all singletons 10% farther from their nearest repeated-cell centres raises coarse dimension by 14.0%. These are changes induced by geometric interventions in the synthetic system. They support a partial causal mechanism within that system, while each intervention may change several geometric properties simultaneously.

The right panel of Figure 12 makes a stronger, intervention-by-intervention comparison. It uses a new geometry-targeted reconstruction rather than treating the earlier sweeps as synthetic counterparts of particular texts. Calibration uses the four geometric descriptors only; no dimensional response is used to choose parameters. Dimension changes are evaluated afterwards on new clouds.

M.3 CONSTRUCTION AND SCOPE OF THE SYNTHETIC RECONSTRUCTION

The generator contains 201 points in 24 dimensions and 31 repeated-type cells. Gaussian centres and centred Gaussian within-cell offsets follow the earlier exploratory construction. Given q^*, the nearest feasible integer type count is $N = \mathrm{round}(201 - 200q^*)$; the other $N - 31$ types are singletons. Repeated-cell multiplicities are drawn from $[2, 2, 2, 3, 3, 3, 4, 4, 5, 6]$ and adjusted to reach 201 points, retaining at least two occurrences per cell. Offsets are scaled to match normalized mean cell radius. Each singleton's displacement from its nearest repeated-cell centre is multiplied by $\exp(p + hz - h^2/2)$, where $z \sim \mathcal{N}(0, 1)$; p controls common displacement and h its heterogeneity.

The Original configuration is approximately calibrated to empirical geometry on six fixed clouds, then frozen. Because the simple family does not match all absolute moments accurately, intervention targets apply the measured relative geometric changes to this achieved synthetic baseline. Differential evolution fits $p \in [-.7, .7]$ and $h \in [0, .45]$ by minimizing squared log errors in mean petiole length and CV over the same six clouds (population multiplier 5, 24 iterations, fixed search seed). Petioles use the current upper-tail singleton–repeated-type definition from Appendix L.

Each condition is evaluated on twelve new cloud seeds. Sampling uses 32 draws without replacement at the same seven sample sizes, the same fractional fixed-quantile trimming, and the same bands as the real-text protocol. The synthetic pool contains 201 points, whereas real pools contain up to 402. The plotted response is the median of seed-paired percentage changes relative to synthetic Original. The experiment comprises 312 clouds: Original, 24 main conditions, and repetition $\times 4$, each with twelve realizations. All-unique expansion is omitted because its radius is not comparably measured.

What transfers, and what does not. Synthetic responses correlate with observed responses at 0.683, 0.915, and 0.697 across the three bands. These correlations show that the reconstruction captures a substantial part of the variation across interventions, with the strongest agreement in the middle band. Absolute response magnitudes are less accurately reproduced: synthetic R^2 values are 0.421, 0.749, and -0.182. For literary enrichment, observed coarse dimension rises by 14.4%, while the reconstruction predicts a 24.3% decline. The negative coarse R^2 means that these predictions have larger squared error than the constant observed-mean benchmark.

Matching accuracy limits the interpretation of these discrepancies. Across the 24 main conditions, median absolute errors in reproducing the requested percentage changes are approximately 0.00, 0.27, 1.30, and 7.21 percentage points for radius, q^*, mean petiole length, and CV; corresponding maxima are 0.00, 0.69, 9.10, and 20.21. Residual geometry mismatches, finite-sample variation, and the different pool sizes can therefore contribute to the dimensional errors.

An informative but incomplete geometric description. The empirical associations and synthetic interventions jointly support a partial geometric explanation of band changes. The four scalars summarize cell spread, occupancy, and the location and dispersion of one long-edge population; they leave cell anisotropy, the full occupancy distribution, and the arrangement and dependence of cell centres and singletons unspecified. These are plausible additional geometric determinants whose separate causal contributions remain unmeasured. The incomplete transfer is consistent with omitted geometric structure, but does not isolate its contribution from matching and sampling errors. We therefore regard the four measurements as useful explanatory descriptors with partial mechanistic support, rather than a complete characterization of the geometry determining multiscale dimension.

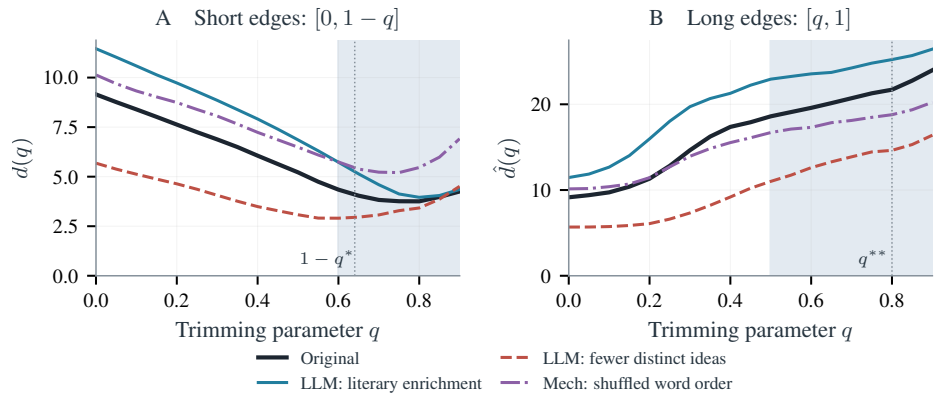


Figure 13: **Different edge ranges give different dimension profiles.** Short- and long-edge profiles for Original and selected interventions on 49 shared sources; shading marks the fine and coarse averaging ranges. Both panels use the text convention: q is the fraction of longest (left) or shortest (right) edges removed. The middle-band summary additionally uses interval profiles.

N RELIABILITY OF THE FITTED DIMENSIONS

Before interpreting band changes, we check the underlying power-law fit. This tests the dimension estimate itself, separately from its geometric explanation.

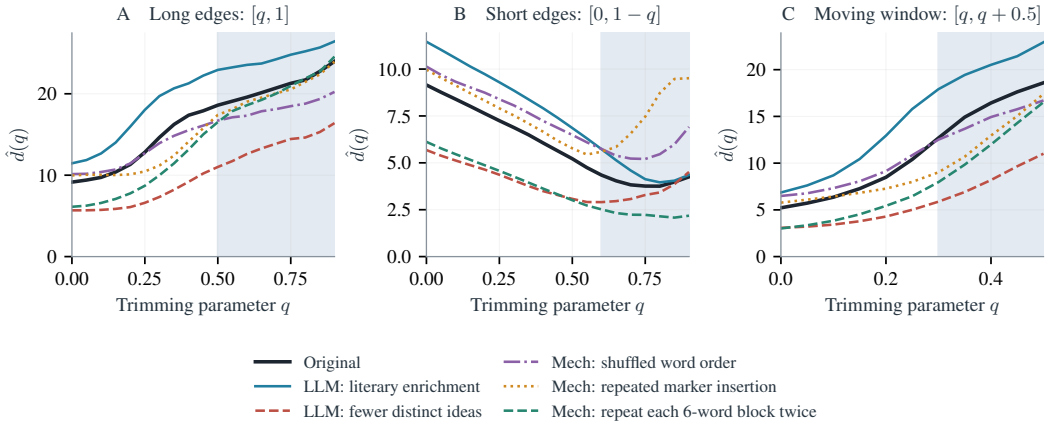


Figure 14: Original and five contrasting interventions on 49 shared sources. Shaded regions define the coarse, fine, and middle bands, respectively. Each panel uses a different retained edge interval, so increasing q has a different meaning across panels.

Table 6: Fine-band diagnostics, pooling source–seed–q fits. The italic asterisk flags the unreliable dimension summary.

Condition	Median R_{scale}	Minimum	Fraction < 0.9	Mean $\mathcal{F}$
Original	.9985	.977	0.0%	4.00
Local repetition $\times 2$	.9934	.932	0.0%	2.29
Local repetition $\times 4$	.8632	.159	61.5%	*1.44**
Partial repetition	.9975	.965	0.0%	3.04

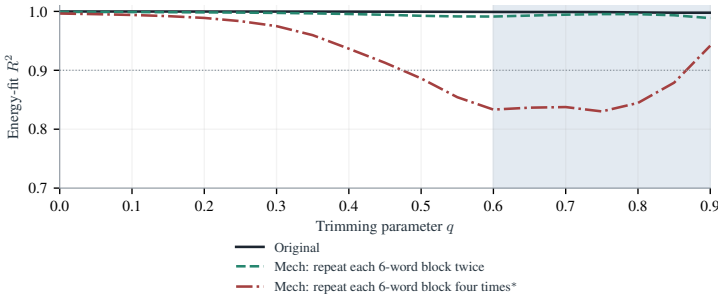


Figure 15: Median energy-fit quality as longer edges are removed. Strong local repetition departs from the approximate power law within the fine-band window (shaded), whereas Original and repetition $\times 2$ remain well fitted.

The median R_{scale} across sink-removed profiles is 0.9998. Of 1,266 fits below 0.9, 1,265 belong to repetition $\times 4$ and one to repeated record fields. No fit in the middle or coarse windows falls below that threshold. The problem is therefore concentrated in the fine-scale response to strong repetition.

Under repetition $\times 4$, mean edge length falls from 12.29 at $n = 41$ to 8.90 at $n = 201$ (Original: 14.71 to 13.37), consistent with short-edge reorganization. The fine response *-64.3%** remains a flagged diagnostic: averaging below Q^* cannot repair poor constituent fits.

O SENSITIVITY TO ATTENTION-SINK REMOVAL

O.1 RADIAL CONCENTRATION AND STABILITY OF THE DESCRIPTOR

To establish the radial geometry before interpreting type cells, we compare the norm distribution of the main population with that of the low-norm group. The sink criterion separates a compact group of low-norm embeddings from the main shell. A preliminary diagnostic of 15,452 non-sink occurrences gives norms 37.9 ± 1.9 (mean $\pm$ SD), with 5th and 95th percentiles 35.1 and 41.1. The low-norm group has mean 26.3 ± 0.4, motivating the threshold of 30. In the present Original cohort, the sink share is 2.40% and mean norm before removal is 37.63. We use "attention sink" as the operational name of this group. The standard deviation of the main population is about 5% of its mean, supporting a thin-shell approximation. Similar norms do not imply a uniform angular distribution or, by themselves, a cell structure. The latter is tested separately in Section 4.1. The threshold is specific to this model and representation; it is not a universal norm cutoff for other encoders or layers.

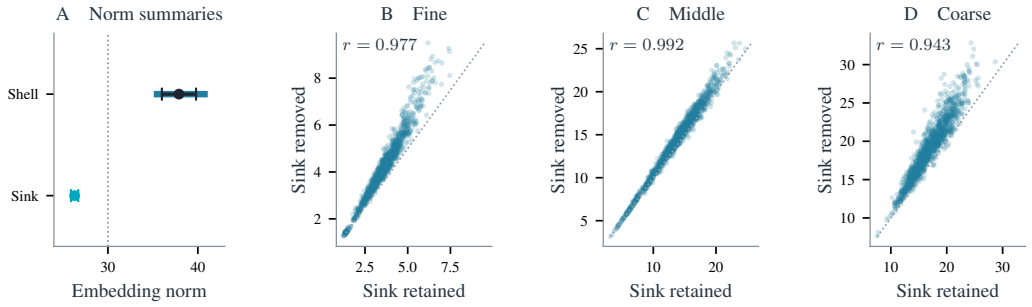


Figure 16: **Norm concentration and sensitivity to sink removal.** A: measured norm summaries from the preliminary diagnostic. Points and capped bars show means and SDs; the thick shell interval spans the 5th–95th percentiles. The vertical line marks the removal threshold. B–D: dimension levels for 1,299 matched source–condition pairs in the current run, with Pearson correlations. Dotted diagonals indicate equality. Strong correlation coexists with systematic shifts in dimension level.

Sink removal changes dimension levels, but largely preserves intervention responses. Across the 1,299 matched source–condition pairs, the D–K correlations of levels are 0.977, 0.992, and 0.943 for fine, middle, and coarse bands. Correlations of paired percentage changes at the same level are 0.969, 0.984, and 0.926. For condition-median changes, including Original at zero, they are 0.996, 0.999, and 0.985. These distinct aggregation levels answer different questions: individual-pair stability and preservation of the intervention-level pattern. The principal geometric models consistently use sink-removed measurements.

O.2 DIMENSION LEVELS FOR ALL CONDITIONS

Table 7 reports the complete condition-level comparison. These means summarize levels separately under each policy; they are not ratios used to construct the paired intervention responses. The main models and all principal response plots consistently remove the sink.

Table 7: Mean dimensions with the sink removed (D) and retained (K), using available pairs for each condition. Special tokens are always removed. Starred fine estimates remain flagged fit summaries.

	Fine		Middle		Coarse	
Condition	D	K	D	K	D	K
Original	4.00	3.48	15.96	15.34	21.09	18.82
Word shuffling	5.80	4.74	14.65	14.00	18.33	15.72
Local repetition $\times 2$	2.29	2.12	12.22	11.76	20.32	18.56
Local repetition $\times 4$	*1.44**	*1.41**	6.83	6.77	17.88	16.93
Partial repetition	3.04	2.71	14.14	13.59	20.24	18.29
Vocabulary cap 25	6.45	5.21	7.10	6.92	13.85	12.48
Vocabulary cap 60	5.10	4.22	9.82	9.51	16.49	14.70
All-unique expansion	5.46	4.56	16.67	15.63	17.53	14.83
Unrelated hapax swap	4.53	3.89	13.71	13.08	16.66	14.74
Single-domain hapax swap	4.60	3.94	13.75	13.08	16.73	14.73
Limited hapax swap	4.14	3.57	15.54	14.72	20.02	17.48
Periodic marker	7.55	5.83	13.06	12.37	20.79	18.06
Remove punctuation	4.69	4.05	16.96	16.13	22.64	18.58
Remove function words	4.55	3.91	16.36	15.52	19.09	16.44
Fewer ideas	3.49	3.02	8.77	8.49	14.13	13.00
Single topic	3.92	3.39	14.80	14.18	20.92	18.70
Lexically poorer	3.85	3.34	14.77	14.16	20.80	18.66
No synonyms	3.65	3.15	14.09	13.50	19.64	17.65
Locked term context	3.46	3.05	14.80	14.16	19.89	17.84
Multiple topics	4.81	4.11	19.61	18.62	22.63	19.57
Lexically richer	4.38	3.81	18.67	17.69	22.50	19.42
No repeated content word	4.23	3.70	18.04	17.21	21.94	19.09
Rare synonyms	4.06	3.53	16.47	15.77	21.54	19.19
Literary rewrite	4.62	4.00	20.45	19.27	24.46	20.86
Coherent word insertion	4.14	3.58	17.00	16.19	21.32	18.87
Anaphora	3.03	2.64	12.69	12.02	15.17	14.00
Repeated record fields	2.77	2.49	12.97	12.36	15.07	14.43

P PROJECTION AS A QUALITATIVE ILLUSTRATION

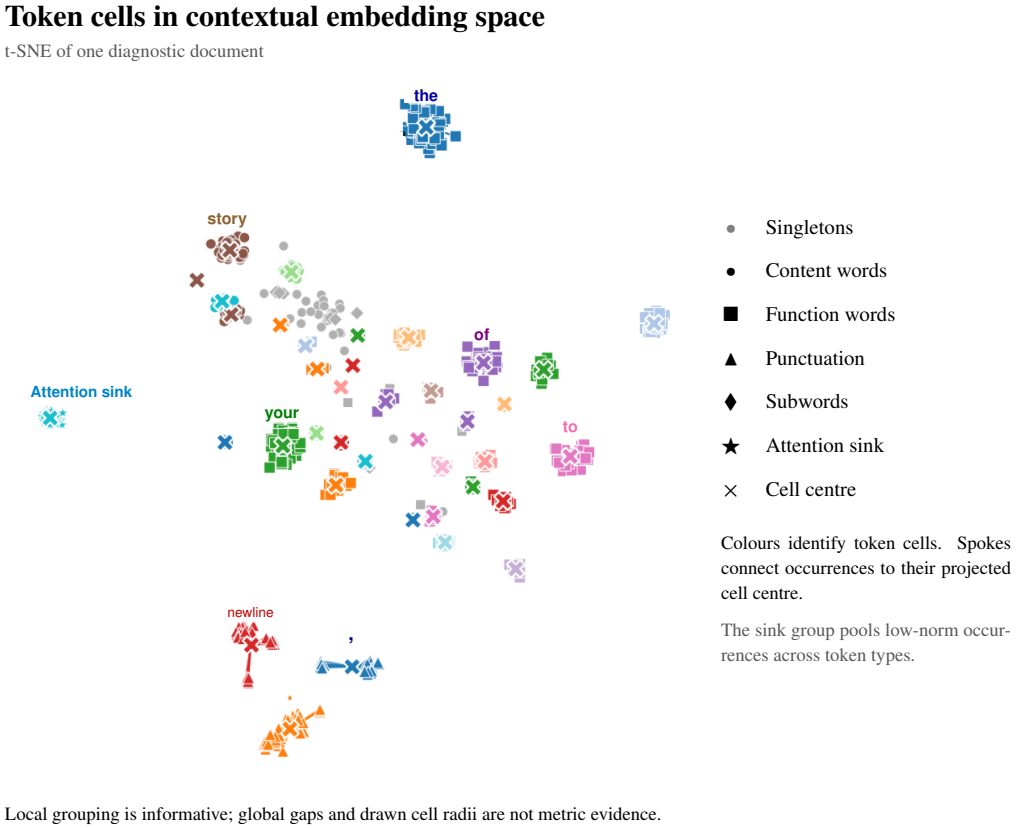


Figure 17: An all-token t-SNE diagnostic from one example document. Colours distinguish repeated token cells, shapes indicate token categories, and grey points are singletons. Crosses and spokes show projected cell centres and membership. The sink combines low-norm occurrences across types. The example contains 40 displayed repeated groups and illustrates local organization rather than an average geometry of the cohort.

The projection uses t-SNE (van der Maaten & Hinton, 2008) on Euclidean embedding distances, with perplexity 30 and a fixed seed. For the illustrated subset, a repeated type's own occurrences occupy 95% of its $k-1$ nearest-neighbour positions in the original space and 98% in the display; nearest-neighbour type agreement is 100% in both. These subset diagnostics support using the image to illustrate local grouping. They differ from the 98.9% nearest-neighbour agreement measured across the main cohort.

Global gaps, relative projected cell radii, and radial positions are not reliable metric summaries of the original space. All quantitative claims about geometry therefore use the original embeddings. The point layer is the retained diagnostic projection with re-typeset labels; the quantitative profile and regression figures are generated from the common-cohort measurements.

Q HUMAN–AI COMPARISON ON HAP-E-2

Q.1 COHORT AND EXCLUSION CHECKS

HAP-E-2 cuts each document into two parts of about 500 words; the generators receive the first part and continue it, and the human continuation is the second part of the same document. We use the registers acad, news, blog and fic; the spoken registers spok and tvm, with transcription and script

markup, are not used. In each register we sample 300 documents for which every author has a text of at least 128 tokens, giving 1200 human and 13200 generated texts. Claude Haiku 4.5 is excluded: the nearest human text by TF-IDF is its own document for 0.2% of its texts, against 51.2% for GPT-4o and 46.5% for Llama-3-8B-Instruct (mean similarity to the own document 0.02 against 0.21 and 0.25).

Q.2 Preprocessing and fixed sampling

All authors are treated identically: typographic quotes and dashes are mapped to ASCII, COCA-style spaces before punctuation (about 1.3 per 100 words in human news, almost none in API output) and markdown markers are removed. The first 128 tokens are embedded as one window with ModernBERT-base (last hidden state). Attention-sink tokens (2.6%: *the*, period and comma with norm ≈ 26) are merged into one pseudo-type. The exact MST is computed on Euclidean distances and the type of a token is its id. Subsamples have $n' \in \{21, 41, 61, 81, 101, 121\}$, 32 per size, drawn without replacement. q^* and q^*_{m} are computed on the whole window and kept for all its subsamples; across texts $q^* \in [0.15, 0.45]$ (1st–99th percentile), hence $q^*_{\mathrm{m}} \in [0.57, 0.72]$.

Q.3 Validation of the type-count boundary

For each text, the observed break of the spectrum is the quantile that best separates within-type from cross-type edges. Against q^*, computed from counts alone, it has slope 0.99, correlation 0.95 and mean offset -0.012 over all 14400 texts; with type labels shuffled within each text the slope is 0.00 ($r = 0.03$). The separation is not exact: a median 8% of the $n - N$ shortest edges join different types, and the transition is spread over quantiles of about 0.2–0.45. Above 0.5 there are practically no within-type edges.

Q.4 Evaluation and lexical matching

Directed AUROC uses the AI-higher convention (with half credit for ties) for each generator against the human texts of the same registers. 95% intervals come from a bootstrap over documents within registers (1000 replicates). Lexical dependence R^2_{lex} is the share of variance across human texts explained by a linear regression on TTR and hapax share. Lexical measures are TTR, hapax share and bigram diversity (the share of distinct pairs of adjacent tokens, which also registers repeated short phrases). Lexically matched pairs are human and AI continuations of the same document whose TTR, hapax share and bigram diversity differ by at most 0.02 (630 pairs).

Q.5 Results and estimator variants

Table 8 compares band variants, and Table 9 gives AUROC per generator. The sign of classical PHD agrees with the sign of bigram diversity for 11 of 11 generators (with TTR for 10) and in 43 of 44 generator×register cells; the rank correlation is 0.94 across generators and 0.97 across cells. A linear model of PHD on TTR and hapax share, fitted on human texts only, predicts the AI–human difference in the cells with $r = 0.89$; with bigram diversity, $r = 0.94$ at slope 1.02. In the matched pairs, $P(\text{AI higher})$ for the upper band is 0.66 for instruction-tuned and 0.64 for API models (± 0.07) and 0.48 for base models; for classical PHD it is 0.61, 0.55 and 0.39. The upper band is weaker than classical PHD where lexical and geometric effects point the same way (API models, base Gemma-2). Case (b) of Proposition 2 removes the lexical dependence but moves base models below 0.5 as a group. Part of the residual signal of the upper band may be stylistic: fine-tuned models produce fewer names and numbers, which form dense groups in the cloud.

The band $[q^*, 1]$ without margin is "mixed" because q^* is frozen at the window value and the transition around it is spread over 0.2–0.45, so the band takes in within-type edges.

Table 8: Band variants on HAP-E-2. R^2_{lex}: variance across human texts explained by TTR and hapax share; ρ_S: rank correlation between the AUROC of the statistic and the AUROC of bigram diversity over 11 generators; "≥ 0.5": generators with AUROC not below 0.5; last three columns: AUROC by group. $q^{*\prime}$: q^* recomputed on every subsample. Kinds refer to Proposition 2.

statistic	kind	R^2_{lex}	ρ_S	≥ 0.5	base	instruct	API
PHD $[0, 1]$	mixed	0.61	0.94	5/11	0.58	0.46	0.77
$[0, q^*_{\text{m}}]$	mixed	0.61	0.95	5/11	0.56	0.43	0.75
$[q^*_{\text{m}}, 1]$	(c)	0.07	0.20	11/11	0.53	0.61	0.69
$[0, q^*]$ (no margin)	mixed	0.10	0.45	3/11	0.37	0.48	0.53
$[q^*, 1]$ (no margin)	mixed	0.17	0.75	9/11	0.51	0.56	0.73
$[0, 0.7]$ (fixed)	mixed	0.69	0.96	5/11	0.58	0.42	0.76
$[0.7, 1]$ (fixed)	(c)	0.07	0.19	11/11	0.54	0.60	0.69
$[q^{*\prime}, 1]$, slope vs $\log n'$	(a)	0.85	0.90	7/11	0.68	0.44	0.81
$[q^{*\prime}, 1]$, slope vs $\log N'$	(b)	0.05	0.17	8/11	0.48	0.63	0.70

Table 9: AUROC ($P(\text{AI} > \text{human})$) per generator, four registers pooled.

generator	group	TTR	bigrams	PHD $[0, 1]$	$[0, q^*_{\text{m}}]$	$[q^*_{\text{m}}, 1]$	$[0.7, 1]$
Llama-3-8B	base	0.50	0.41	0.44	0.41	0.51	0.51
Llama-3-70B	base	0.55	0.46	0.47	0.45	0.50	0.51
Gemma-2-9B	base	0.85	0.69	0.71	0.68	0.57	0.57
Gemma-2-27B	base	0.84	0.69	0.70	0.69	0.56	0.56
Llama-3-8B-Instruct	instruct	0.29	0.35	0.41	0.39	0.62	0.61
Llama-3-70B-Instruct	instruct	0.34	0.38	0.45	0.41	0.61	0.61
Gemma-2-9B-it	instruct	0.40	0.44	0.48	0.46	0.61	0.61
Gemma-2-27B-it	instruct	0.41	0.44	0.48	0.45	0.59	0.59
GPT-4o	API	0.77	0.76	0.79	0.76	0.72	0.72
GPT-4o-mini	API	0.79	0.78	0.83	0.79	0.79	0.78
GPT-5-mini	API	0.74	0.70	0.70	0.70	0.57	0.57

R OVERALL PROFICIENCY PROFILES

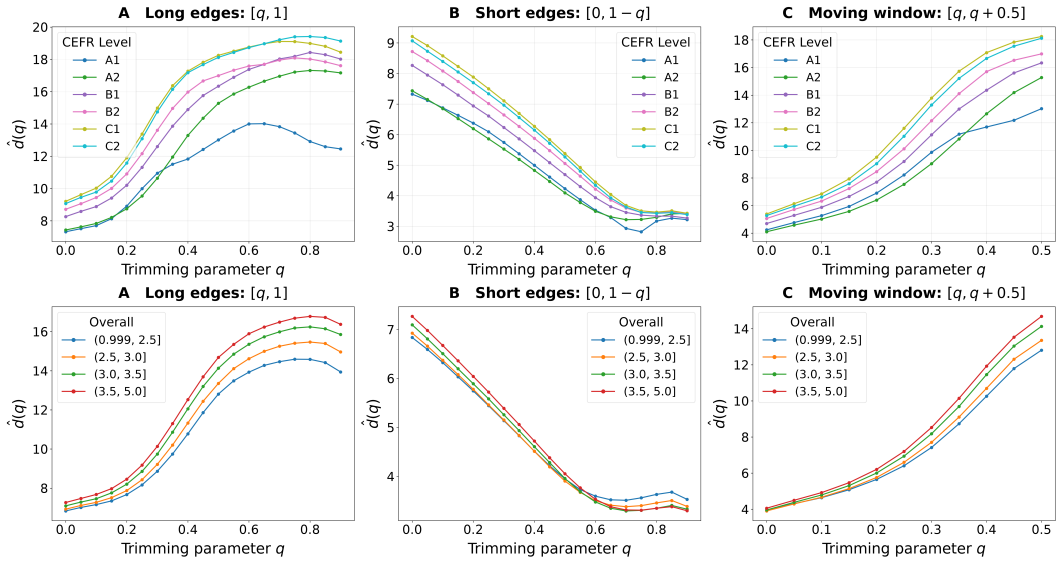


Figure 18: **Proficiency profiles in ModernBERT.** Top: CEFR levels; bottom: ELLIPSE Overall-score groups. Columns retain long edges $[q, 1]$, short edges $[0, 1-q]$, and a moving interval $[q, q+0.5]$. Long-edge profiles preserve group differences that diminish at the shortest-edge scales.

S CEFR PROFILES ACROSS EMBEDDING MODELS

All panels compare long-edge $[q, 1]$, short-edge $[0, 1-q]$, and moving-window $[q, q+0.5]$ profiles. ModernBERT is shown in the main text. The additional models exhibit different degrees of group separation: RoBERTa and Gemma retain a clearer proficiency association, while Qwen distinguishes A1 from other levels more readily than it orders the remaining levels. C1 commonly lies above C2, so the curves do not define a strict proficiency ranking. These are group-level profiles without displayed uncertainty intervals.

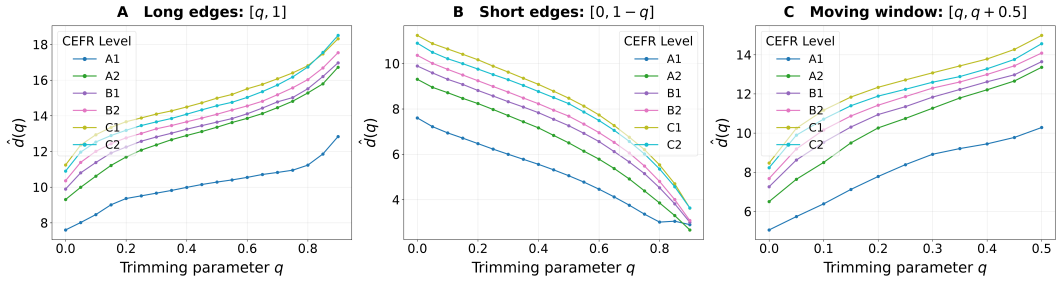


Figure 19: CEFR profiles for RoBERTa-base. Panel order and retained intervals follow Figure 18.

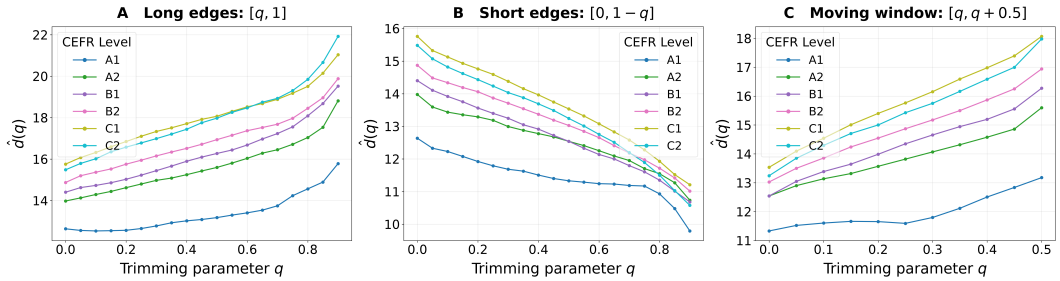


Figure 20: CEFR profiles for Gemma-4-E4B-it. Panel order and retained intervals follow Figure 18.

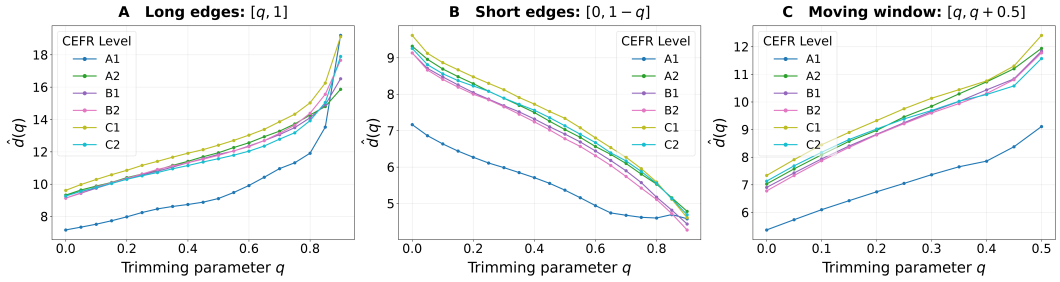


Figure 21: CEFR profiles for Qwen2.5-0.5B. Panel order and retained intervals follow Figure 18.

T ELLIPSE PROFILES BY SCORING CRITERION AND EMBEDDING MODEL

The following profiles compare grammar, cohesion, conventions, phraseology, syntax, vocabulary, and the reported Overall score. Each model uses the score bins displayed in its legend; binning is not assumed identical between criteria. Every figure contains long-edge $[q, 1]$, short-edge $[0, 1 - q]$, and moving-window $[q, q + 0.5]$ profiles in that order.

Across these diagnostics, the encoder models RoBERTa and ModernBERT generally show clearer separation of score groups than the decoder models Gemma and Qwen. This is descriptive evidence of representation dependence, not a test of predictive accuracy. The profiles alone do not identify which linguistic property causes a difference, and overlapping curves do not establish statistical equivalence.

T.1 OVERALL

In contrast, the decoder models, Qwen2.5-0.5B and Gemma-4-E4B-it, show no clear separation of the curves according to score when retaining longer edges, with quantile interval ([q,1])). The moving-window approach, which retains edges within ([q,q+0.5]), likewise shows no consistent ordering of scaled dimension by score as the window shifts towards longer edges. When retaining shorter edges, with interval ([0,1-q])), higher scores tend to correspond to higher scaled dimensions, although the separation is less distinct than for the encoder models. Across the assessed dimensions of English proficiency in the ELLIPSE Corpus, phraseology, cohesion, and syntax show the clearest separation, with higher scores consistently corresponding to higher scaled dimensions. A similar pattern is observed for conventions, grammar and vocabulary, although the differences between score levels are less pronounced.

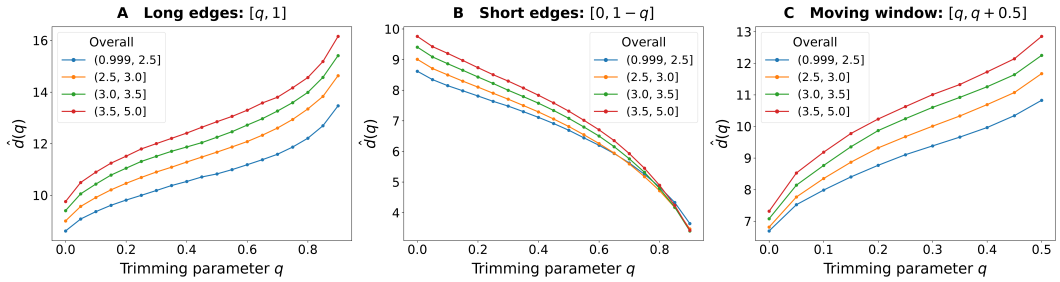


Figure 22: Overall-score profiles for RoBERTa-base. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

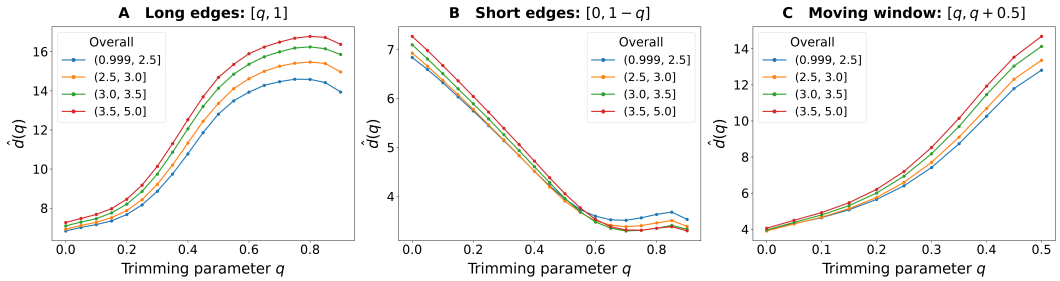


Figure 23: Overall-score profiles for ModernBERT-base. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

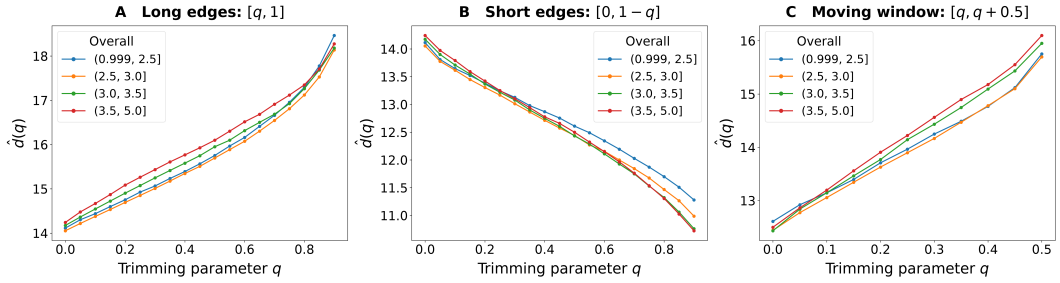


Figure 24: Overall-score profiles for Gemma-4-E4B-it. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

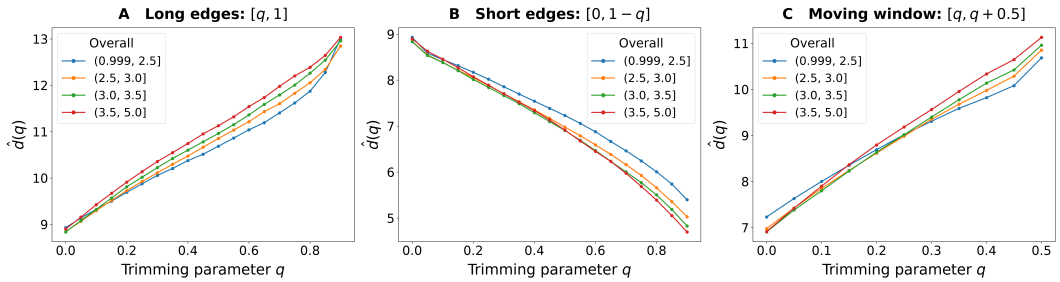


Figure 25: Overall-score profiles for Qwen2.5-0.5B. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

T.2 GRAMMAR

Higher grammar scores are associated with higher scaled dimensions, with clearly separated curves for RoBERTa-base embeddings in Figure 26.

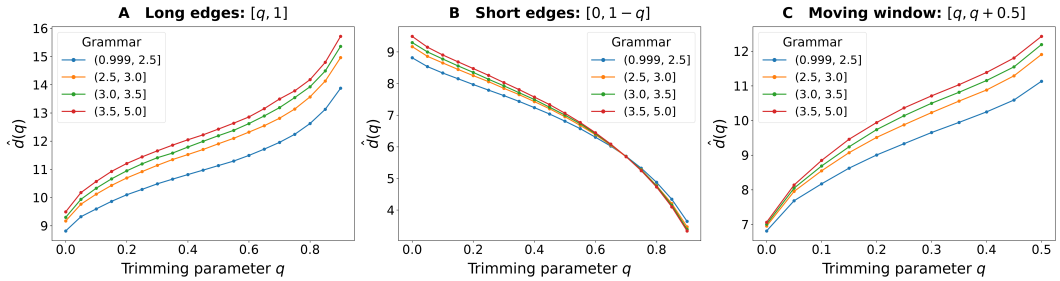


Figure 26: Grammar-score profiles for RoBERTa-base. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

For ModernBERT-base embeddings27, the differences in scaled dimension between grammar scores become more pronounced as the trimming parameter (q) increases.

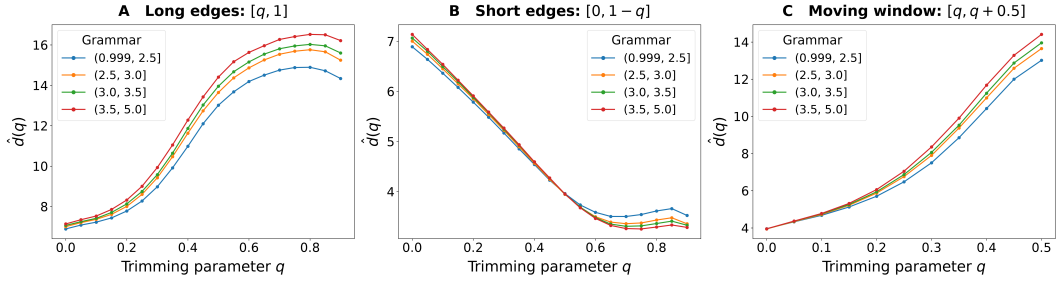


Figure 27: Grammar-score profiles for ModernBERT-base. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

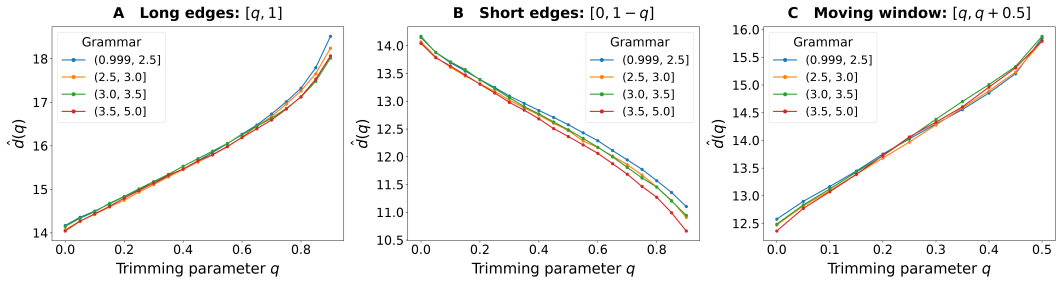


Figure 28: Grammar-score profiles for Gemma-4-E4B-it. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

T.3 COHESION

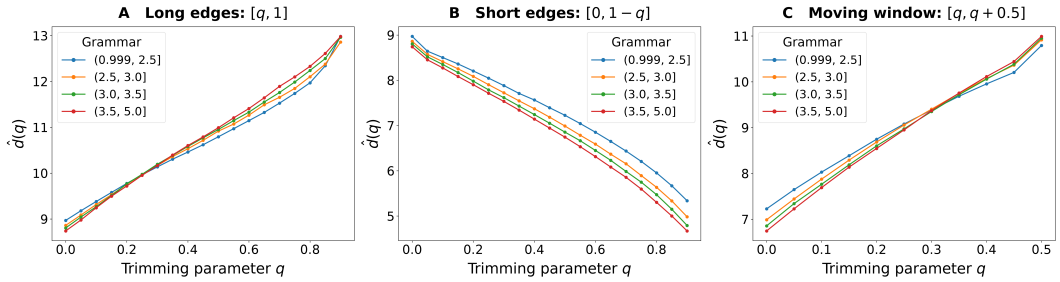


Figure 29: Grammar-score profiles for Qwen2.5-0.5B. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

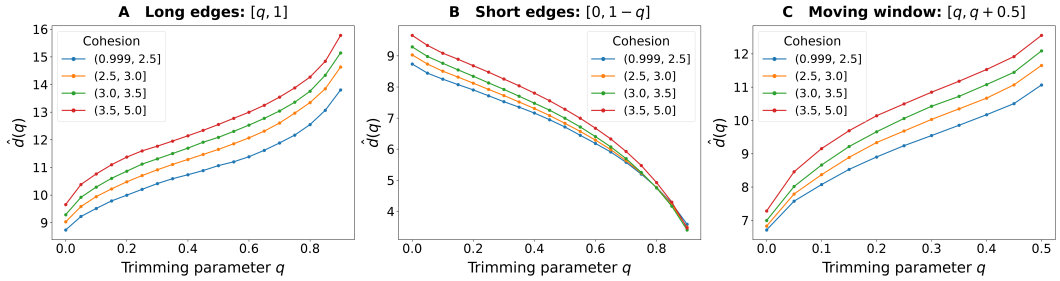


Figure 30: Cohesion-score profiles for RoBERTa-base. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

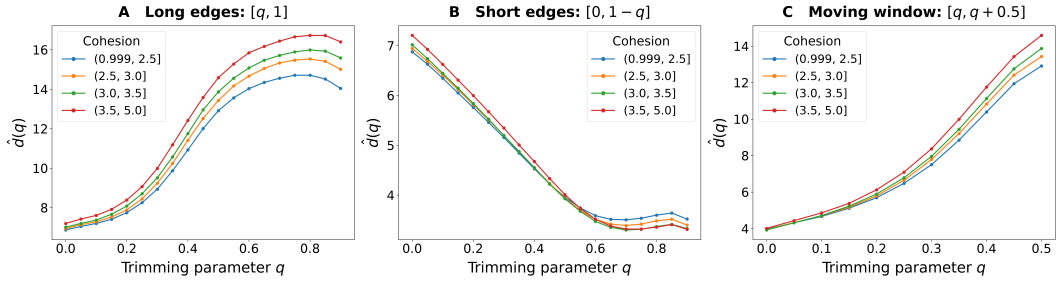


Figure 31: Cohesion-score profiles for ModernBERT-base. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

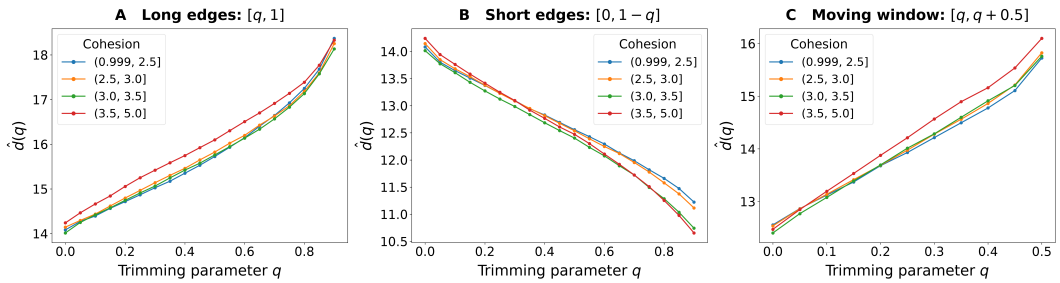


Figure 32: Cohesion-score profiles for Gemma-4-E4B-it. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

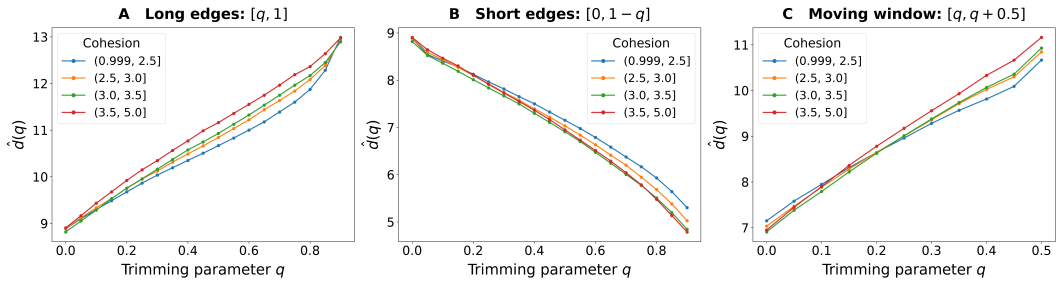


Figure 33: Cohesion-score profiles for Qwen2.5-0.5B. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

T.4 CONVENTIONS

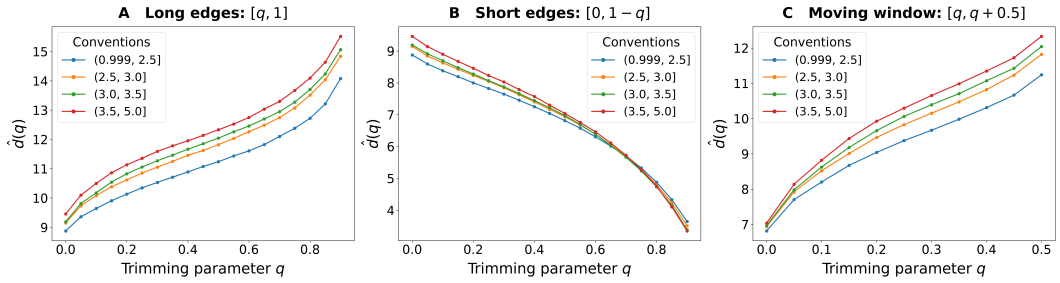


Figure 34: Conventions-score profiles for RoBERTa-base. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

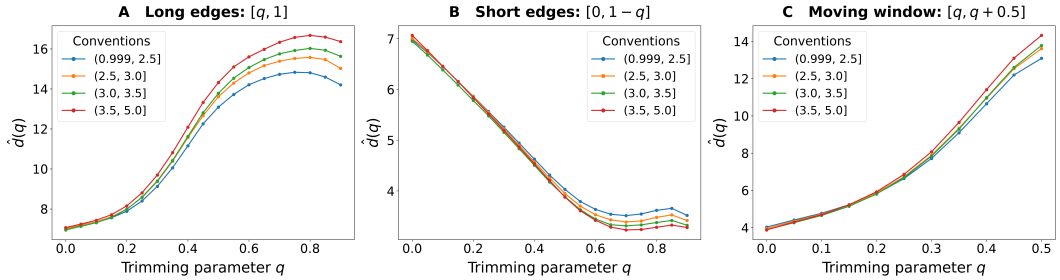


Figure 35: Conventions-score profiles for ModernBERT-base. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

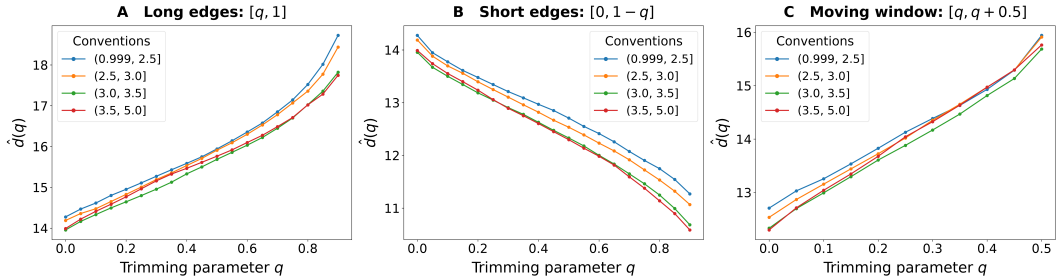


Figure 36: Conventions-score profiles for Gemma-4-E4B-it. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

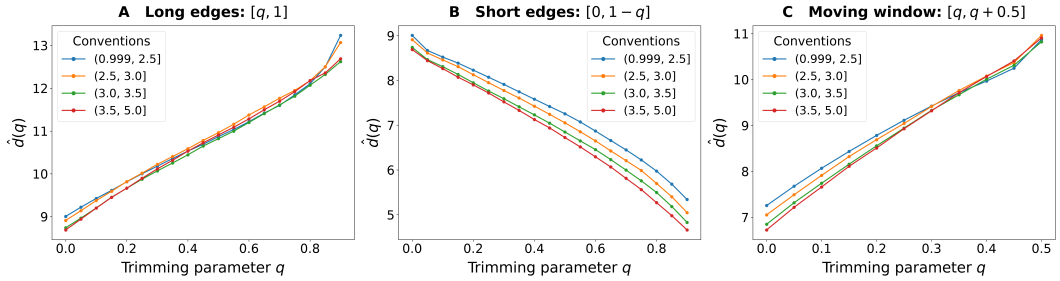


Figure 37: Conventions-score profiles for Qwen2.5-0.5B. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

T.5 PHRASEOLOGY

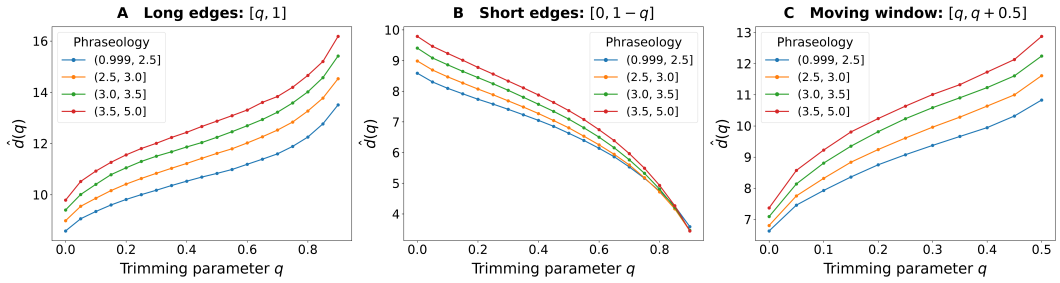


Figure 38: Phraseology-score profiles for RoBERTa-base. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

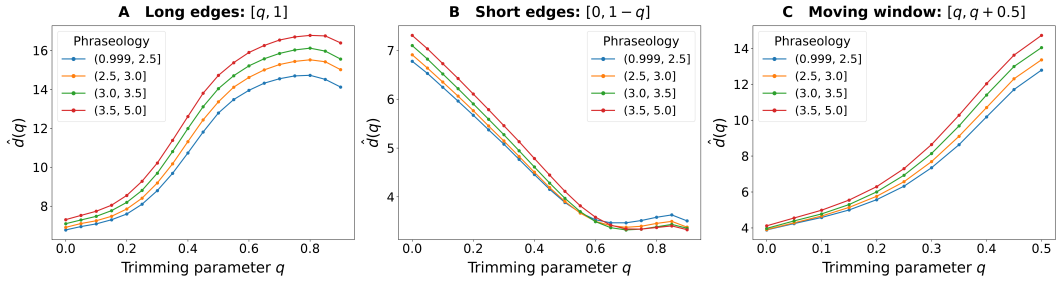


Figure 39: Phraseology-score profiles for ModernBERT-base. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

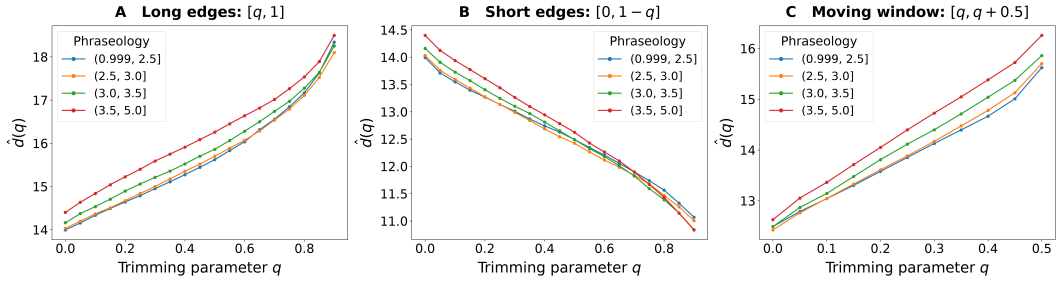


Figure 40: Phraseology-score profiles for Gemma-4-E4B-it. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

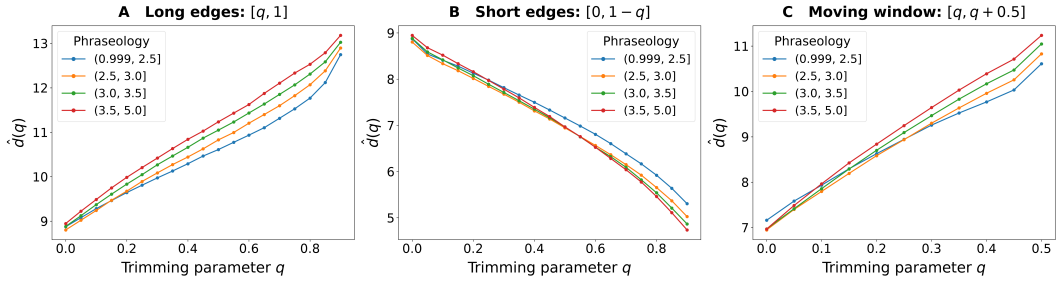


Figure 41: Phraseology-score profiles for Qwen2.5-0.5B. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

T.6 SYNTAX

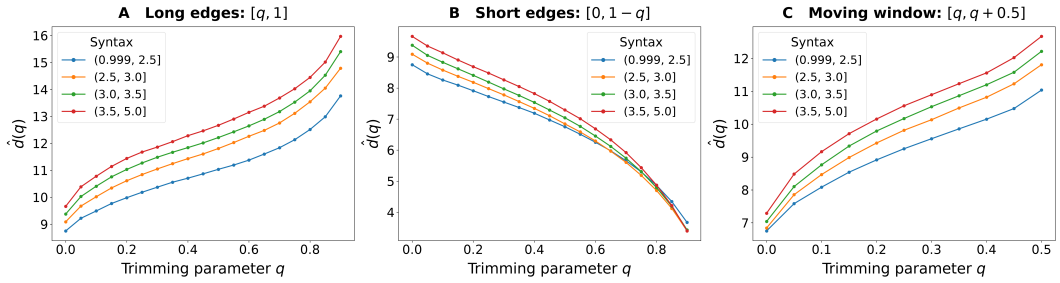


Figure 42: Syntax-score profiles for RoBERTa-base. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

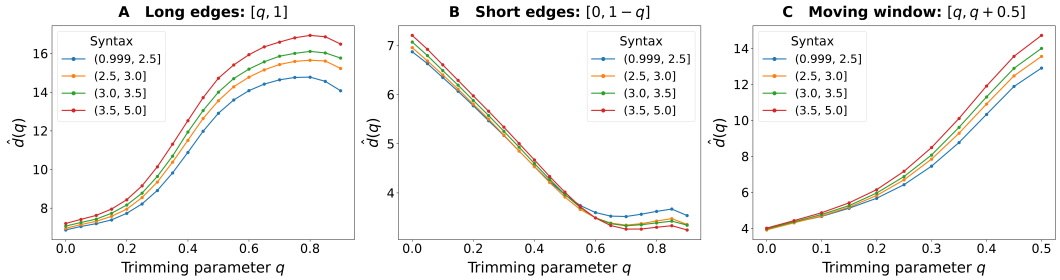


Figure 43: Syntax-score profiles for ModernBERT-base. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

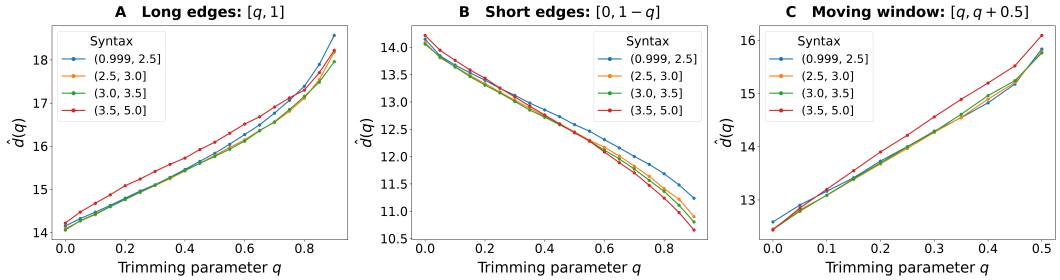


Figure 44: Syntax-score profiles for Gemma-4-E4B-it. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

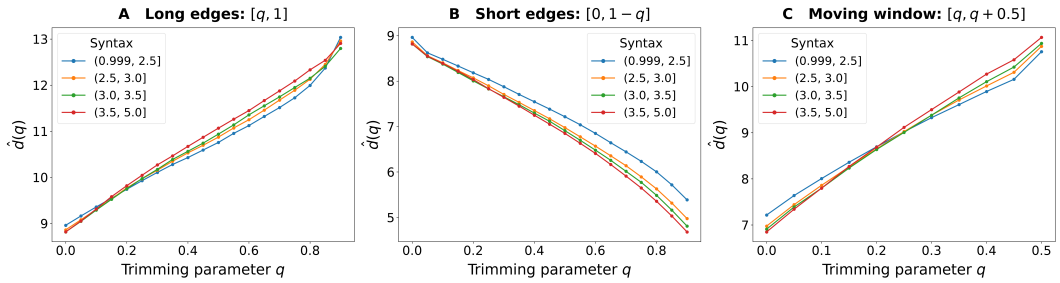


Figure 45: Syntax-score profiles for Qwen2.5-0.5B. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

T.7 VOCABULARY

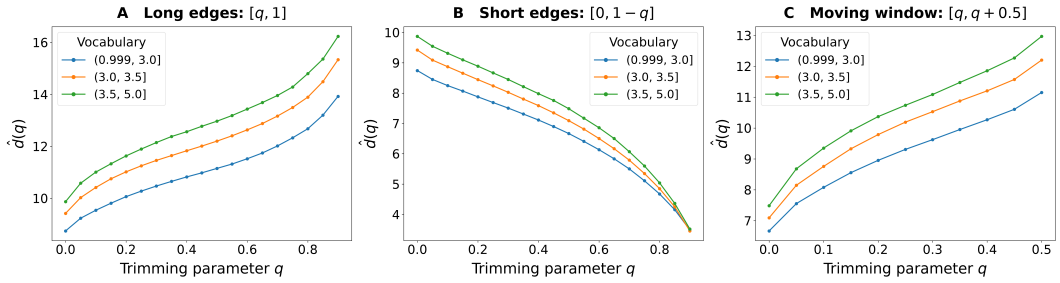


Figure 46: Vocabulary-score profiles for RoBERTa-base. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

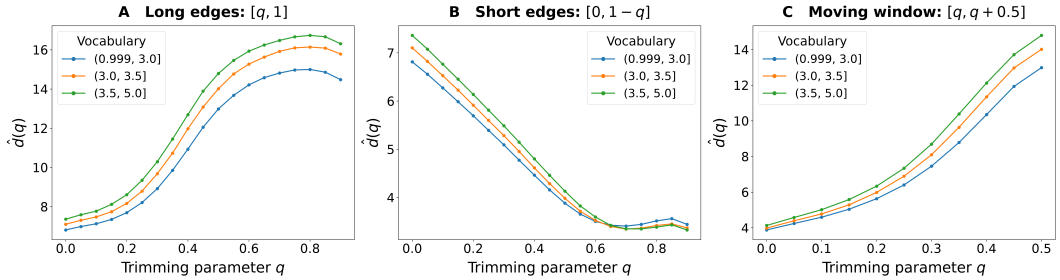


Figure 47: Vocabulary-score profiles for ModernBERT-base. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

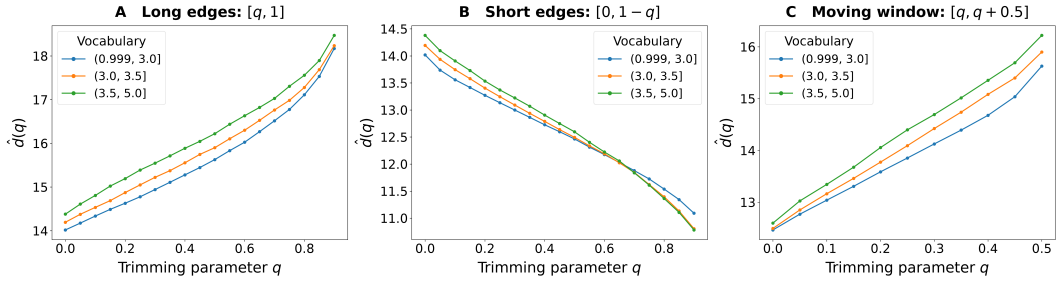


Figure 48: Vocabulary-score profiles for Gemma-4-E4B-it. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

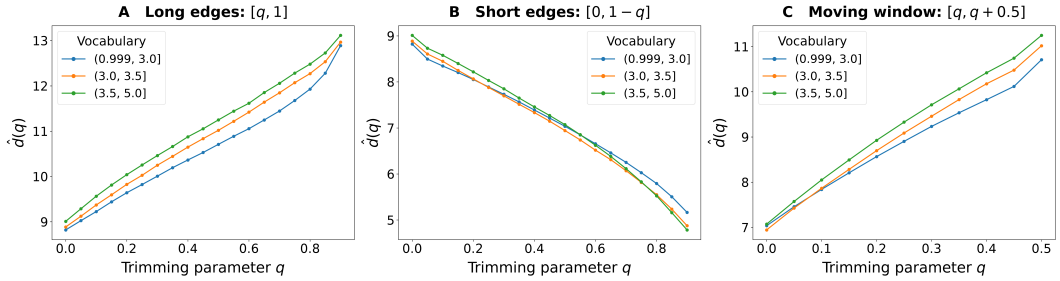


Figure 49: Vocabulary-score profiles for Qwen2.5-0.5B. Colours identify the displayed score bins; panels retain long edges, short edges, and a moving interval.

U PAIRED TEXT-DAMAGE EXPERIMENTS

U.1 CORPORA AND PAIRED BASELINES

We use the fine, middle and coarse grids in Section 5.1. Each summary averages separately fitted dimensions at fixed retained intervals.

To test whether the bands react to the naturally occuring types of damage of the human texts, we use six English corpora in which every text exists in two versions, a damaged one and a clean one. Each document serves as its own control, reducing differences in topic and genre; the common sampling budget controls the number of measured occurrences. The corpora cover five kinds of damage:

- **Optical character recognition (OCR) errors.** ICDAR 2019 (Rigaud et al., 2019): OCR output of historical English books and newspapers, paired with a hand-corrected transcription (116 pairs).
- **Automatic speech recognition (ASR) errors.** Spoken-SQuAD (Lee et al., 2018): Wikipedia paragraphs from SQuAD (Rajpurkar et al., 2016), converted to speech and transcribed by an ASR system at three noise levels, with a reported word error rate (WER) of 23–55% (656 documents). CORAAL (Kendall & Farrington, 2023): recorded sociolinguistic interviews, each with a manual transcript and the transcripts of five commercial ASR systems collected by Koenecke et al. (2020) (494 documents).
- **Web boilerplate.** WCXB (Foley, 2026):web pages with a verified main text. We extract each page twice: with TRAFILATURA (Barbaresi, 2021) (a widely used open-source library that extracts the main text of a web page and discards navigation, advertisements and other boilerplate), and by naively stripping HTML tags, which keeps menus, buttons and footers (381 pages).
- **Human writing errors.** W&I+LOCNESS (Bryant et al., 2019; Yannakoudakis et al., 2018): essays by learners of English and by native students, each with a version corrected by an annotator (1,840 essays).[1]
- **Semantic damage.** PAN-WVC-10 (Potthast, 2010): Wikipedia articles before and after an edit labelled as vandalism (1,452 edits).

Both versions of each text (with and without damage) are embedded with ModernBERT-base (Warner et al., 2025), and the bands are computed at $L = 201$ tokens using a common sampling budget for both versions, averaged over three random seeds. We report the median relative change of each band from the clean to the damaged version. Two comparisons need a matched baseline. ASR output is lowercase and almost unpunctuated, and this formatting alone changes the bands by 3–12%, so we compare it with the clean transcript converted to the same format. The same caution applies to WCXB. Even TRAFILATURA, whose output nearly matches the verified text (word F1 0.95), shifts the coarse band by 3.7% relative to it, so part of any difference from the verified text comes from extraction itself rather than from boilerplate. We therefore compare naive extraction with TRAFILATURA: both are extracted from the same HTML, and they differ mainly in whether the boilerplate is kept.

U.2 BAND RESPONSES AND DOCUMENT-LEVEL DETECTION

The direction of the shift depends on the mechanism of damage (Fig. 5). Damage that breaks text into rare fragments raises the fine and middle bands: OCR by $+7.3\%$ and $+7.2\%$, ASR by $+12.3\%$ on Spoken-SQuAD and $+4.6\%$ on CORAAL in the fine band. Repetitive boilerplate lowers the coarse and middle bands by 8.0% and 8.5% (all $p < 10^{-7}$). Vandalism, which changes meaning rather than form, moves no band by more than 1%. One associated change is the share of hapaxes, tokens that occur once and attach to the MST through long edges: it grows under OCR and ASR, falls under boilerplate, and its change tracks the middle-band shift with ρ between $+0.42$ and $+0.54$ across corpora. The effect grows with the dose: on Spoken-SQuAD the fine band rises by $+7.3\%$, $+13.7\%$ and $+16.9\%$ at WER 0.25, 0.44 and 0.57.

[1]The LOCNESS essays are used courtesy of the Centre for English Corpus Linguistics (CECL), Université catholique de Louvain.

Learner errors behave differently: the coarse band falls by 4.1% and the middle band does not move, consistent with a different response from recognition errors that fragment tokens. A rule that flags documents whose band exceeds the corpus mean by 20% catches ASR output at 57–82% precision (base rate 50%) but not learner texts (48%), so it separates pipeline damage from human errors. Paired effects do not always carry over to single documents: on WCXB the paired fall is significant at $p = 10^{-10}$, but a document-level rule reaches only AUC 0.60, because the shift varies widely between pages rather than because of measurement noise.

kind	n	what it is, and one quoted example
Repetition (121 texts)		
`opening`	56	the same sentence opening, again and again *"Does the animal have horns? Does the animal have claws? Does the animal"*
`skeleton`	24	a template skeleton left in the text *"[title] Go to twitter.com in the second tab. [step]"*
`insistent`	17	one phrase or construction repeated insistently *"then you'll wait until the 2, and then you'll play the 2"*
`list`	14	list or recipe scaffolding repeated verbatim *"1 tbsp (15 ml) reduced-fat mayonnaise 1 tbsp (15 ml)"*
`connective`	5	the same connective opening every paragraph *"Furthermore, Venus is an extreme environment"*
`closing`	3	a formulaic summing-up closing *"So, distance learning is an invaluable resource that can provide students with"*
`hedge`	2	the same hedging formula throughout *"a complex and multifaceted one."*
Semantic failure (117 texts)		
`jump`	71	a jump to an unrelated subject between adjacent sentences *"stop the car at certain speeds. Amazon and Walmart have started a war."*
`wordsalad`	36	locally grammatical text that states nothing coherent *"But here is the problem, your stomach is in the living room and there"*
`referentless`	7	a reference with nothing to refer to *"The proof below gives a similar bound when M 1."*
`unlinked`	2	adjacent clauses with no link between them *"management of the development and deployment of the models. The authors need to provide"*
`empty`	1	words that carry no content at all *"content creation tools such as content creator apps like Instagram Stories created that use"*

Table 10: The two defect classes of Fig. 50b, as the annotation reports them over all 1048 texts. Every example is a verbatim span the judge was required to quote and which was checked against the document. Grammatical smoothness kinds (agreement, sentence fragments, dangling modifiers, register breaks; 10 texts in all) are excluded from the semantic class, so that class contains only locally well-formed text.

V SUPPLEMENTARY GENERATOR AND GENRE COMPARISON

We apply the method to the COLING corpus (Wang et al., 2025), which is convenient for two reasons: it holds a wide range of human genres alongside texts from many different generators, and it has been analysed before with the *full* PH dimension (Pedashenko et al., 2025), so the two readings can be compared directly. The cohort is every "model" with at least 25 texts of 280 words or more, 26–30 texts each, and the 13 genres of the human half on the same footing, 23–30 texts each: 23 generators and 681 generated texts against 13 genres and 378 human ones. All of them are measured under one protocol — the three bands are averages of $\hat{d}(q)$ over fixed quantile-trimming windows, *coarse* over $q \in [0.5, 0.9]$, *fine* over $q \in [0.6, 0.9]$, *middle* over $q \in [0.3, 0.5]$ — and annotated by an LLM judge that is required to quote a verbatim span for every finding, checked back against the document; the prompt is in App. Y, and it returned a verdict for 1048 of the 1059 texts.

Fig. 50a places the median of each generator (circles and triangles) and of each human genre (squares) in the plane of the coarse and fine bands. Panel b is the same plane, one point per text, with the two defect classes the annotation finds marked by colour; Table 10 gives their taxonomy and quotes an example of every kind.

The zone of human texts is clearly separated from the artificial one, and the artificial texts split into two clusters. One has both bands high and coincides roughly with the zone of semantic failure: syntactically adequate text whose meaning does not hold together, median fine band 5.11 against 3.89 for the undamaged texts. The other has a low fine band and overlaps substantially with the repetition defects, median fine band 2.80. The human genres lie between them: 3.49–4.65 on the fine band against 2.75–3.57 for the low cluster and 4.65–6.39 for the high one, three intervals that abut almost exactly.

Note that the full dimension, while it does isolate the high cluster, separates the low one substantially worse, and it discards information that matters. On the full dimension the seven high-cluster generators (11.26–14.57) fall entirely outside the genre range (7.93–9.94), but 7 of the 16 low-cluster generators land inside it, against 3 of 16 on the fine band alone. And in the plane we can see that of all the human genres only `arxiv` interferes with the generated ones, sitting next to the GPT-3.5/GPT-4 family (0.13 from `davinci`, 0.38 from `gpt-3.5-turbo`, 0.41 from `gpt4o`, in units of the standard deviation of each axis over all 36 medians; the next-closest genre, `wiki_csai`, is 0.40 from any generator at all); whereas `eli5`, whose full dimension is indistinguishable from that of `wikipedia` (9.53 against 9.65), is set apart by the coarse band (21.1 against 18.0) — coherence and semantic variety, which the single number averages away.

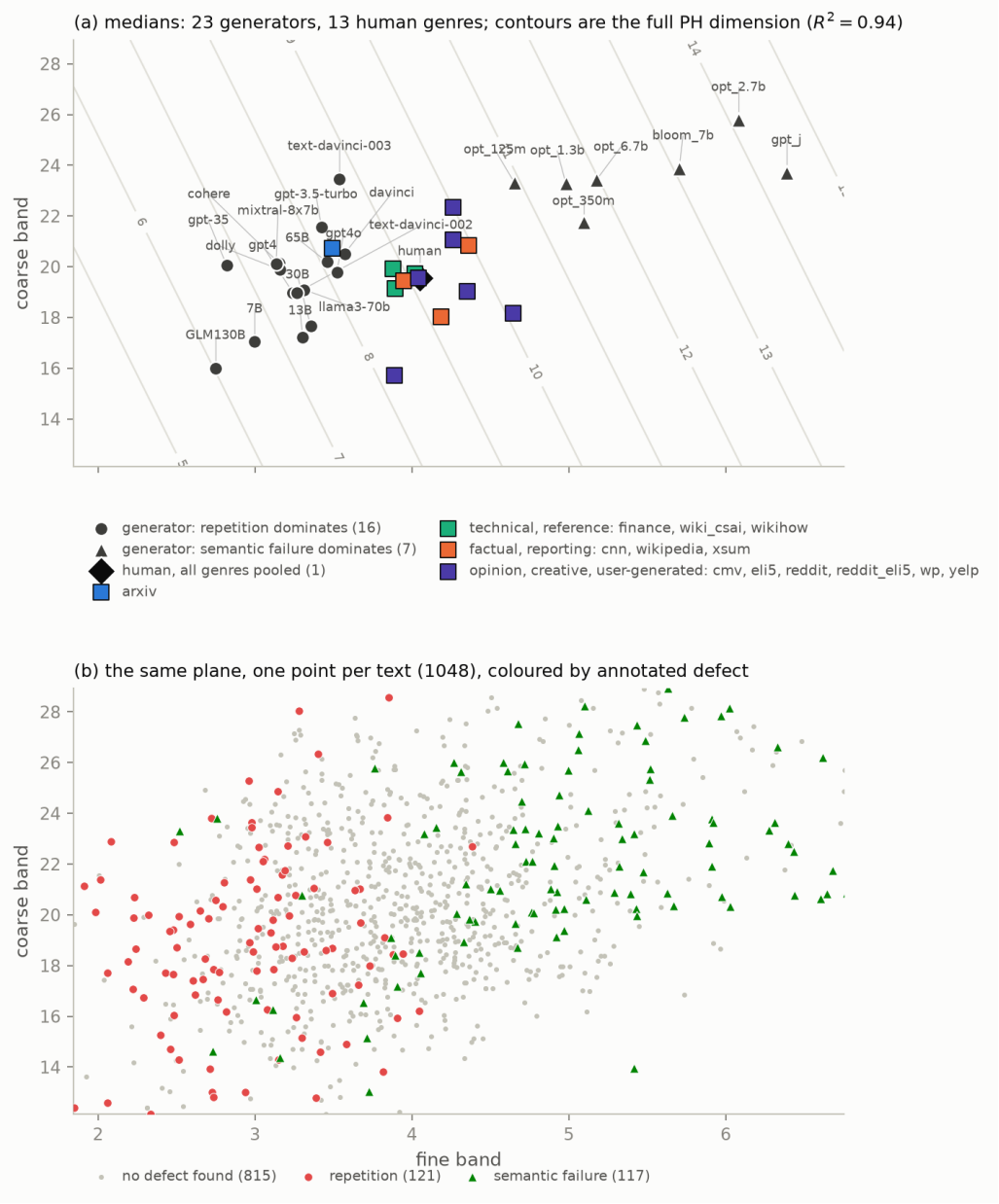


Figure 50: Generators, human genres and annotated defects in one plane, both panels drawn on the same scale so the zones can be compared by eye. **(a)** medians over the 1059 measured texts: 23 generators, labelled individually and drawn in one colour, with marker shape carrying the dominant defect, and the human half pooled into a single point (diamond). The 13 genres are squares, coloured by the intrinsic-dimension strata of Pedashenko et al. (2025) (their Fig. 5), with `arxiv` separate as the lowest-dimension source both there and here; that work names `arxiv`, `wiki_csai`, `cnn`, `xsum`, `wikipedia`, `cmv`, `eli5` and `wp` explicitly, and the remaining five genres are assigned by its description of each stratum. Its encoder is not ours (Gemma/Qwen/RoBERTa against ModernBERT-base), so absolute dimensions and parts of the ordering are not expected to agree. Contours are the full, untrimmed PH dimension reconstructed from the plane ($R^2 = 0.94$); the contour value differs from the directly measured one by 0.38 on average and 0.92 at worst, so they are a ± 0.4 guide rather than an exact scale. **(b)** one point per text, coloured by annotated defect, shape as in panel a; the 1048 texts are those the judge returned a parseable verdict for, 11 short of the measured set. The shared window is the 2nd–98th percentile of the texts: it contains every median and leaves 74 texts (7.1%) outside the frame.

Table 11: Mean MST dimension across embedders (generalized D_q, middle band $[q, q{+}0.2]$). $\langle D_q \rangle$ is the corpus mean; the τ-spread and band range are defined in the text above.

Embedder	$\langle D_q \rangle$	τ-spread	band range
RoBERTa-large	13.6	2.31	16.4
ModernBERT-base	11.9	2.07	14.0
Qwen2.5-1.5B	9.2	1.10	5.4
Qwen2.5-0.5B	9.0	1.08	6.2

W MEAN MST DIMENSION ACROSS EMBEDDERS

The band summaries of the main text follow from trimmed MST energies. A closely related object is the generalized dimension of the MST partition function,

$$S(\tau, m) = \frac{1}{m} \sum_{i=1}^{m} \lambda_i^{-\tau} \;\sim\; m^{\,q(\tau)-1}, \qquad D_q = \frac{\tau}{q(\tau) - 1}, \tag{26}$$

where $\{\lambda_i\}$ are the normalized edge lengths of the MST of a random subsample of m points and $q(\tau)$ is the log–log slope (van de Weygaert et al., 1992). Restricting the sum to a middle band of edges (the $[q, q{+}0.2]$ quantile window) makes the partition function scale-resolved, and sweeping the moment exponent τ gives the generalized-dimension surface $D_q(\tau, q)$. From this surface we form two summaries: the *band range*, $\max_q \operatorname{med}_\tau D_q(\tau, q) - \min_q \operatorname{med}_\tau D_q(\tau, q)$, which measures dependence on the edge-length scale, and the *τ-spread*, $\operatorname{med}_q(\max_\tau D_q(\tau, q) - \min_\tau D_q(\tau, q))$, which measures dependence on the moment order at fixed scale. Figure 51 shows the text-averaged surface for four embedders under the main protocol ($L = 101$, pool $2L$, eight replicates; 200 texts selected, ≈ 550 retained per embedder). Within each panel the generalized-dimension estimate is shown next to the algebraically equivalent estimator of $\hat{d}$ used elsewhere in the paper.

Table 11 summarizes the level and the two spreads. The encoder-style models, ModernBERT-base and RoBERTa-large, give the largest dimensions ($\langle D_q \rangle = 11.9$ and 13.6) and the widest scale-band range (14.0 and 16.4); the decoder models, Qwen2.5-1.5B and Qwen2.5-0.5B, are lower (9.2 and 9.0) and much flatter across bands (5.4 and 6.2). In every case the spread over the moment exponent τ is small relative to the band range (τ-spread 1.1–2.3 against band ranges 5.4–16.4), so once a band is fixed the point cloud is approximately monofractal in the moment order. This is the scale-dominates-moment pattern: the dimension varies across edge-length bands but is approximately monofractal in the moment order once a band is fixed.

X SCALE-DEPENDENT SYNTHETIC CONTROLS

A standard stress test asks what a scale-free estimator does when the effective dimension itself changes with scale. The construction is a modified random β-model dust that has dimension $D = 1$ below a homogeneity scale $L_h = L/4$ and $D = 3$ (homogeneous) above it. Because a single MST partition function mixes all scales, the local dimension on such a dust never plateaus: it sweeps monotonically between 1 and 3, and limited statistics only blur the transition. This is precisely the failure mode that motivates the scale-resolved summaries used throughout the paper.

We verify the estimator on three synthetic controls whose true dimensions are known: a monofractal dust (`mono`, $D = 1$), a homogeneous dust (`homog`, $D = 3$), and the scale-dependent dust (`scaledep`, $D = 1 \to 3$). On the two single-regime controls the generalized dimension is recovered and stays flat in the moment exponent (mean $\langle D_q \rangle = 1.22$ with τ-spread 0.29 for `mono`; $\langle D_q \rangle = 2.91$ with τ-spread 0.12 for `homog`); on the scale-dependent dust the profile sweeps between the two regimes rather than settling on a plateau. The corresponding generalized $D_q(\tau, q)$ heatmaps for these controls are shown alongside the clouds in Fig. 52.

Table 12: Closed list of kinds for the PATTERN axis: which specific thing the document repeats.

kind	meaning
`opening`	the same sentence opening recurs
`closing`	the same wrap-up formula recurs
`skeleton`	one sentence skeleton refilled with different words
`insistent`	the same phrase or construction returns insistently, far more often than the subject requires
`connective`	the same transition opens every paragraph
`hedge`	the same hedging or filler phrase recurs
`list`	the same list-item shape repeats
`disclaimer`	assistant-style disclaimers or meta-commentary about itself
`other`	something else, which you must name in three words

Table 13: Closed list of kinds for the SMOOTH axis: how the surface of the text stops being continuous.

kind	meaning
`jump`	adjacent sentences are about unrelated things
`dangling`	a pronoun or definite reference with no antecedent
`agreement`	tense, person or number drifts inside the document
`unlinked`	sentences abut with no connective where one is needed
`register`	register or voice switches back and forth
`fragment`	sentences are systematically incomplete
`other`	something else, which you must name in three words

Y THE ANNOTATION PROMPT

Every text of the generator cohort was annotated by a single LLM call with the prompt below, followed by the document itself under a `DOCUMENT:` header. Model `gemini-3.6-flash` at temperature 0.0; requests are cached on disk, so the annotation is reproducible without the network.

The prompt reports four separate things: what the collection pipeline did to the document (five axes, carried over verbatim from an earlier judge so that this block stays comparable with it), and three properties of the text itself — the pattern it repeats, where its surface stops being smooth, and where it stops meaning anything. Each of the latter three is scored by *extent*, so that level 2 means "recurs through the document" rather than "is unpleasant"; the analysis therefore thresholds at 2.

Three design choices are inherited from the earlier judge because they are what made it work where a judge of prose quality had failed. First, every score above zero must carry a verbatim quotation of at most fifteen words, and the quotation is checked against the document afterwards: a score justified by text that is not there counts as a failure of the judge, not as evidence. Second, each defect is charged to one axis only. Third, the descriptive axes — the form of the document and how carefully the person wrote — are declared not to move any score, so that "a person wrote this carelessly" cannot be confused with "a pipeline damaged this".

Of the 702 annotated texts, 23 (3.3%) carried at least one quotation that could not be found in the document and 2 named a kind outside the closed list. Those rows are kept and flagged rather than dropped, so that the rate is visible and any statistic can be recomputed without them.

PROMPT, VERBATIM

```
You are auditing documents collected into a machine-learning training
corpus. You will report four different things about one document, and
↪  keeping
them apart is the whole task.

  A. what the collection pipeline did to the document
  B. what specific pattern the document repeats
  C. where its surface stops being smooth, systematically
```

Table 14: Closed list of kinds for the SENSE axis, judged from internal evidence only.

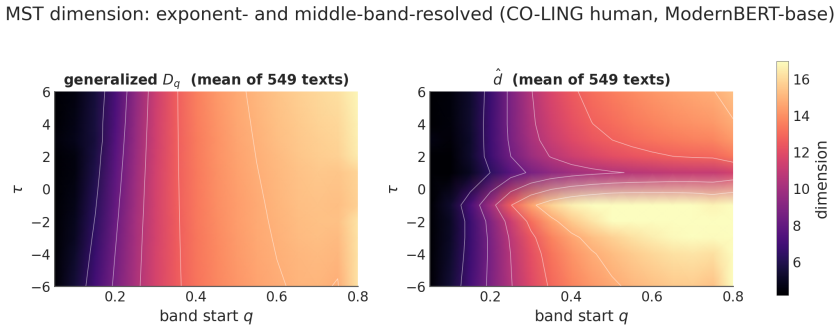
kind	meaning
`wordsalad`	words that look grammatical but combine into nothing a reader could restate
`referentless`	impossible to tell what or whom the passage is about
`empty`	sentences parse and assert nothing restatable
`circular`	explains or defines a thing by itself
`other`	something else, which you must name in three words

```
  D. where it stops meaning anything, systematically

A rambling forum comment full of typos and slang is undamaged prose
↪  written by a
person: good data. A polished article whose formulas were replaced by
↪  rendering
junk is damaged however well its sentences read. Do not confuse the two.

Score every axis from 0 to 3. Every score above 0 must be justified by
↪  quoting
the offending text verbatim, up to fifteen words, copied exactly from the
document. If you cannot quote it, the score is 0.

Charge each piece of damage to one axis only -- the one that names what
happened. A stripped link that leaves a stray space before a full stop is
↪  WEB,
not also CHARS.

For blocks B, C and D score by EXTENT, not by how unpleasant the text is:
  1 = isolated: one or two instances
  2 = systematic: recurs through the document, a reader would notice it
  ↪  as a
      habit of the text
  3 = dominant: the document is largely made of it

For block D judge only from INTERNAL evidence: contradictions inside the
document, statements that carry no proposition, text that never arrives
↪  at a
subject. You cannot check facts against the world and must not try; an
unfamiliar but coherent claim is not a defect.

WEB -- damage from turning a web page into a document: delimiters
lost, service blocks kept, separate pieces of the page run together,
↪  markup left
behind.
  1 = small: list or paragraph separators gone so entries run together,
  ↪  an
      isolated leftover tag or bracket where a link was, a signature or a
      caption kept
  2 = a whole service block kept: navigation, a moderator footer, a
  ↪  cookie
      notice, a related-articles list
  3 = the document is mostly page furniture, or several unrelated pieces
  ↪  of the
      page are concatenated without any separator

LOSS -- part of what the document said is gone. Not formatting and
not markup, which belong to WEB: meaning the document promises and does
↪  not
deliver.
  1 = a pointer to material that is not there ("see the table below", and
  ↪  no
      table follows)
  2 = a heading, a colon or an enumeration opening onto nothing
```

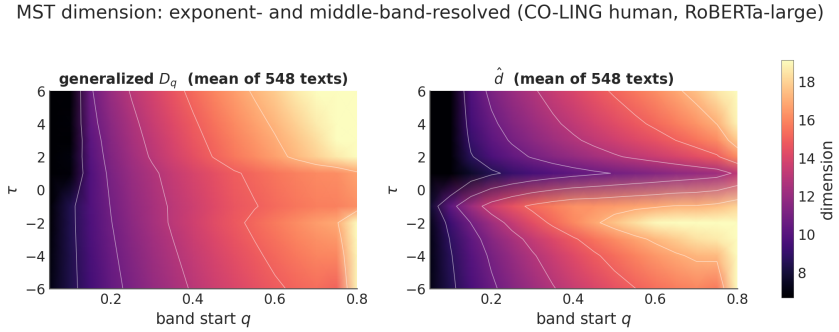
```
  3 = whole sections gone: a thought begins, breaks off, and the next
  ↪ sentence
      is about an unrelated subject

DUP -- the same material appears more than once because the
pipeline copied it, not because the author repeated himself.
  1 = a sentence or a heading duplicated once
  2 = a paragraph or a block repeated verbatim
  3 = the document is largely a repetition of its own parts

CHARS -- the writing itself was corrupted: characters or spacing
were replaced by rendering junk, a broken encoding turning text to
↪ mojibake, a
tokeniser applied and not undone.
  1 = isolated: fewer than about five instances in the document
  2 = systematic: most occurrences of the affected element are damaged --
  ↪ every
      apostrophe, every sentence boundary, every line end
  3 = readability is destroyed: the words can no longer be reliably made
  ↪ out

CUT -- the document's boundaries were lost.
  1 = the ending is unfinished but the document is substantially whole
  2 = it breaks off mid-sentence, or begins mid-sentence
  3 = neither beginning nor end is present: a fragment cut from the
  ↪ middle

PATTERN -- a specific thing the document repeats. Not repetition
of content, which is DUP, but a recurring shape: the same opening, the
↪ same
skeleton, the same filler. Name which one.

NOT this, and this was the error to avoid: ordinary pronoun reference. A
↪ text
that says "he", "it", "this", "they" to point back at something it named
↪ earlier
is using the language as the language works, and that is never a pattern.
↪ Nor is
a text that simply stays on its subject, nor a list whose items are
↪ naturally
alike because the things listed are alike. A pattern is a habit of the
↪ writing,
visible as the same wording coming back when nothing required it to.

PATTERN kinds, pick exactly one:
  opening = the same sentence opening recurs
  closing = the same wrap-up formula recurs
  skeleton = one sentence skeleton refilled with different words
  insistent = the same phrase or construction returns insistently, far
  ↪ more often than the subject requires
  connective = the same transition opens every paragraph
  hedge = the same hedging or filler phrase recurs
  list = the same list-item shape repeats
  disclaimer = assistant-style disclaimers or meta-commentary about
  ↪ itself
  other = something else, which you must name in three words

SMOOTH -- systematic breaks in the surface of the text: the places
where it stops reading as one continuous piece. A single awkward sentence
↪ is not
this; a habit of awkwardness is.

NOT this: informal writing, typos, missing commas, slang, or long loose
sentences -- those are how a person writes and belong to HUMAN. Nor a
↪ change of
subject that the text announces and carries out.
```

```
SMOOTH kinds, pick exactly one:
  jump = adjacent sentences are about unrelated things
  dangling = a pronoun or definite reference with no antecedent
  agreement = tense, person or number drifts inside the document
  unlinked = sentences abut with no connective where one is needed
  register = register or voice switches back and forth
  fragment = sentences are systematically incomplete
  other = something else, which you must name in three words

SENSE -- the passage cannot be understood. Not "is wrong", not "is
doubtful", not "is badly written": cannot be understood.

**The test, and apply it literally.** Could you restate the passage in
↪  your own
words, so that someone who had not read it would know what it said? If
↪  yes, the
score is 0 -- however false, strange, clumsy or unfamiliar the passage is.
↪  A lie
restates perfectly. A mistake restates perfectly. Only nonsense does not.

You do not know whether what the document says about the world is true,
↪  and you
are not being asked. **Fact-checking is not this axis and is not any
↪  axis.**

What earns a score, with the quotation showing it:

  wordsalad     the words are there and nothing assembles: "thjoidering
                intellectual partner. For o nn. And scorers will and
                ↪  object"
  referentless  the passage cannot be tied to any subject at all -- not
  ↪  an
                unnamed subject you could infer, but none available
  empty         it parses and asserts nothing restatable
  circular      "it's the case because it's the case"
Every one of the following was an observed error of this judge. None of
↪  them
scores above 0:

  * a self-contradiction. This one is worth stating plainly, because it
  ↪  is the
    error the judge kept making: a document that asserts X and also not-X
    ↪  is
    perfectly understandable -- you can restate it as "it says X and also
    not-X". Contradiction belongs to no axis here. "Because you cannot
    ↪  create
    consciousness || I don't say it is impossible" is a writer qualifying
    himself, and even a flat contradiction would score 0
  * a claim about the world you believe false -- "Apple introduces
  ↪  Android 10",
    "The methane gas contains carbon dioxide", "Wilhelm Scheele published
    ↪  X".
    Restate them and you will see they restate fine
  * an ordinary or true statement that looks odd out of context -- "The
  ↪  company
    has been active for over 50 years", "Most species are hermaphrodites"
  * a speaker in dialogue who denies, disagrees or changes his mind
  * a distinction, a hedge, a concession, an unfamiliar subject, an open
  ↪  question
  * garbled characters from a broken pipeline -- that is CHARS, not this

Score by how much of the document you cannot understand:
  1 = one passage
  2 = several, through the document
  3 = the document as a whole cannot be understood
```

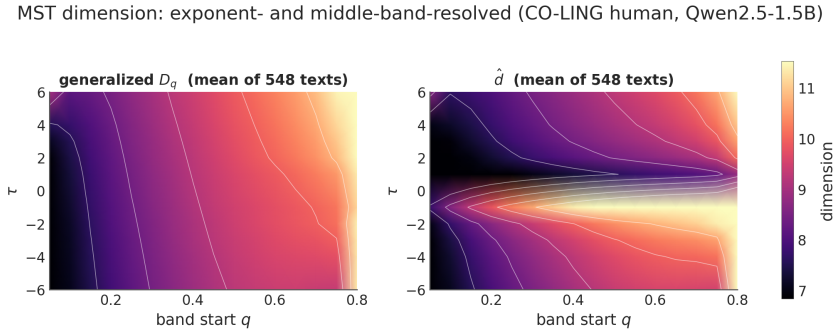
```
This axis is meant to be rare. If in doubt, score 0.

SENSE kinds, pick exactly one:
  wordsalad = words that look grammatical but combine into nothing a
  ↪  reader could restate
  referentless = impossible to tell what or whom the passage is about
  empty = sentences parse and assert nothing restatable
  circular = explains or defines a thing by itself
  other = something else, which you must name in three words

Record the form of the document. This is description only and moves
no score.
  1 = continuous prose
  2 = prose with formal apparatus: headings, citations, quoted blocks
  3 = predominantly an enumeration: list, table, catalogue, running
  ↪  commentary
  4 = a template: a form, a questionnaire, an automatically generated
  ↪  page
  5 = a mixture: a prose section together with an enumerated one

How the person wrote. This is NOT a defect and enters no total.
  0 = careful standard writing
  1 = informal but controlled
  2 = conversational: slang, missing commas, typos, loose sentences
  3 = barely organised: stream of thought, heavy misspelling, no
  ↪  structure

Answer in exactly these lines, each beginning with its tag. Use "-"
where no quotation is required. Quote verbatim, never paraphrase.

WEB: <0-3> | <quote or ->
LOSS: <0-3> | <quote or ->
DUP: <0-3> | <quote or ->
CHARS: <0-3> | <quote or ->
CUT: <0-3> | <quote or ->
PATTERN: <0-3> | <kind> | <quote or ->
SMOOTH: <0-3> | <kind> | <quote or ->
SENSE: <0-3> | <kind> | <quote or ->
NATURE: <1-5>
HUMAN: <0-3>
DAMAGE: <0-5, overall severity of PIPELINE damage, from
↪  WEB/LOSS/DUP/CHARS/CUT
only and never from how well the document is written:
  0 = nothing found; the document arrived intact
  1 = isolated blemishes a reader would pass over
  2 = clearly present but local; the document is fully usable
  3 = systematic: the defect runs through the whole document, or an
  ↪  intruding
      block takes up a real share of it
  4 = severe: the damage competes with the content
  5 = useless as data>
TEXTUAL: <0-5, overall severity of the TEXT's own failings, from
↪  PATTERN/SMOOTH/
SENSE only and never from pipeline damage. Same scale: 0 nothing, 3
↪  systematic,
5 the document is unreadable as a piece of writing>
VERDICT: keep | keep-with-caveats | drop
WHY: <one line>
```



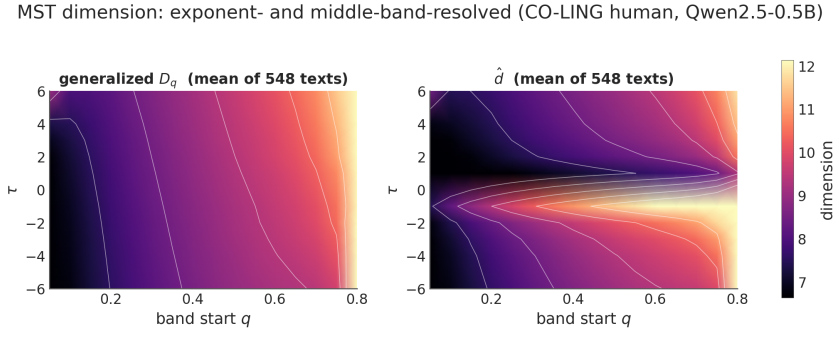
(a) ModernBERT-base



(b) RoBERTa-large

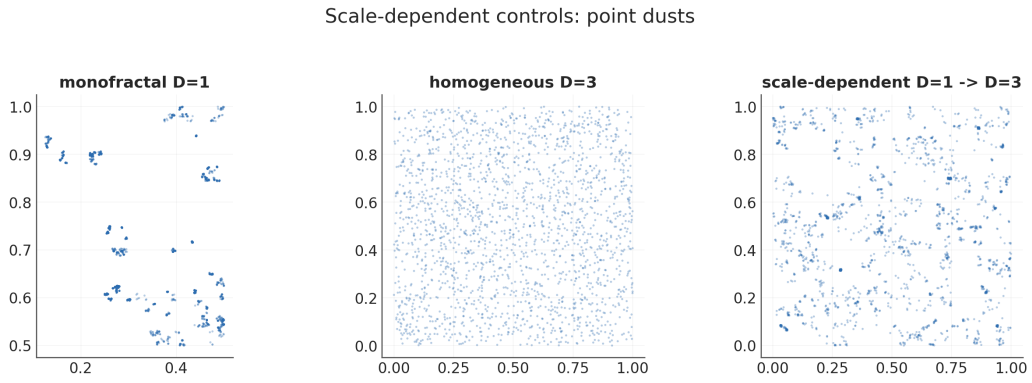


(c) Qwen2.5-1.5B

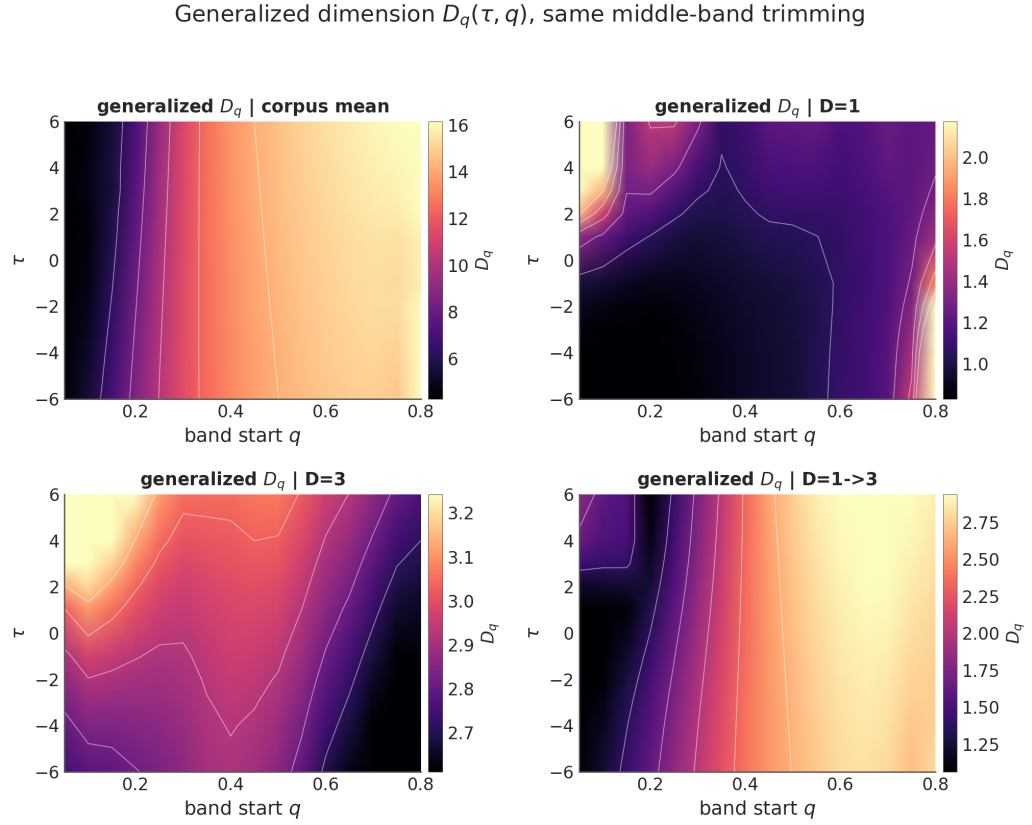


(d) Qwen2.5-0.5B

Figure 51: Mean MST dimension. Each panel is a heatmap of the text-averaged generalized dimension D_q over the moment exponent τ (rows) and the edge-length band q (columns), shown for the generalized-dimension estimator and the equivalent estimator $\hat{d}$, for four embedders. The dimension varies strongly across edge-length bands and weakly across the moment exponent, as quantified in Table 11.



(a) Synthetic control point clouds



(b) Generalized $D_q(\tau, q)$ heatmaps

Figure 52: Scale-dependent synthetic controls. The point clouds show the `mono`, `homog` and `scaledep` constructions; the $D_q(\tau, q)$ heatmaps show that the estimator recovers the true dimension on the single-regime controls and sweeps between the two regimes on the scale-dependent dust, where a single scale-free number cannot summarize the geometry.