Unveiling Intrinsic Dimension of Texts: from Academic Abstract to Creative Story

Vladislav Pedashenko[1*], Laida Kushnareva[2*], Yana Khassan Nibal[2,3], Eduard Tulchinskii[2], Kristian Kuznetsov[2], Vladislav Zharchinskii[1], Yury Maximov[4], Irina Piontkovskaya[2]

[*]Equal Contribution

[1] Moscow State University, [2] Lomonosov Research Institute,
[3] Higher School of Economics [4] Interdata Astana

Abstract

Intrinsic dimension (ID) is an important tool in modern LLM analysis, informing studies of training dynamics, scaling behavior, and dataset structure, yet its textual determinants remain underexplored. We provide the first comprehensive study grounding ID in interpretable text properties through cross-encoder analysis, linguistic features, and sparse autoencoders (SAEs). In this work, we establish three key findings. First, ID is complementary to entropy-based metrics: after controlling for length, the two are uncorrelated, with ID capturing geometric complexity orthogonal to prediction quality. Second, ID exhibits robust genre stratification: scientific prose shows low ID (∼8), encyclopedic content medium ID (∼9), and creative/opinion writing high ID (∼10.5) across all models tested. This reveals that contemporary LLMs find scientific text "representationally simple" while fiction requires additional degrees of freedom. Third, using SAEs, we identify causal features: scientific signals (formal tone, report templates, statistics) reduce ID; humanized signals (personalization, emotion, narrative) increase it. Steering experiments confirm these effects are causal. Thus, for contemporary models, scientific writing appears comparatively "easy," whereas fiction, opinion, and affect add representational degrees of freedom. Our multi-faceted analysis provides practical guidance for the proper use of ID and the sound interpretation of ID-based results.

1 Introduction

Quantifying data complexity is one of the key pillars supporting modern ML and LLM theory. Early information-theoretic views (e.g., the Information Bottleneck) tie learning efficiency to compressibility (Tishby and Zaslavsky, 2015), while the manifold perspective motivates a geometric view of representations. In NLP, complexity assessments have evolved from data–intrinsic measures (e.g., gzip ratio, lexical or syntactic indicators) to model–based ones that depend on how a model organizes and predicts text. The latter divide into *predictive* metrics (entropy, cross-entropy, perplexity) and *geometric* metrics (anisotropy, intrinsic dimension). While entropy-like measures are ubiquitous in training and evaluation, geometric complexity remains comparatively underexplored.

We focus on **intrinsic dimension (ID)** as a geometric measure of degrees of freedom in embedding space. We establish a conceptual separation between ID and prediction-based entropy: ID depends solely on the *geometry* of hidden representations, whereas entropy arises only after *unembedding* and reflects alignment with the vocabulary. Hence, the two quantities are not substitutes. We begin by showing empirically that, after controlling for text length, ID and (cross-)entropy are essentially uncorrelated; we then substantiate this separation with a formal argument.

In this paper, we provide a comprehensive, semantics-grounded account of ID for text. First, we characterize how ID varies across domains and styles, showing that scientific/informational prose is consistently low-ID, while opinionated and fictional writing is high-ID, indicating increased representational degrees of freedom from the model's perspective. Second, we link ID to observable linguistic regularities: higher lexical diversity and lower cross-sentence repetition associate with higher ID, whereas syntactic indicators are comparatively uninformative for well-formed text. Third, using sparse autoencoders and feature steering, we connect ID to interpretable semantic axes (e.g., genre, narrative, personalization, emotional tone) and probe their causal effects on generation. Together, these results position ID as a complementary, geometry-based lens on textual complexity with practical implications for evaluation and training: specifically, the need to balance low-ID (encyclopedic/scientific) and high-ID (personal/narrative) data to faithfully assess model

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7916–7944

capabilities.

The paper is organized as follows: Section 2 formalizes ID and its relation to predictive metrics; Section 3 details our SAE-based analysis methodology; Section 5 presents experimental observations. Finally, Section 6 summarizes our findings and discusses implications.

1.1 Related Work

Intrinsic Dimension in Neural Networks. Intrinsic dimension (ID) analysis provides a geometric framework for understanding and improving large language models (LLMs). Li et al. (2018) established that good solutions lie in low-dimensional subspaces of the parameter space. Extending this idea to NLP, Aghajanyan et al. (2021) demonstrated that task-specific variation in pretrained models is confined to a low-dimensional subspace, inspiring low-rank adaptation methods such as LoRA (Hu et al., 2022), which assumes that effective fine-tuning occurs within compact intrinsic subspaces.

Recent studies have shifted from parameter to representation space, using nonlinear estimators to measure ID in embeddings. While Havrilla and Liao (2024) link ID to model scaling, Razzhigaev et al. (2024) demonstrate that it may reflect training saturation and dynamics. Viswanathan et al. (2025) confirm previous observations and show that higher ID correlates with higher entropy and loss. ID has also been used for architecture search (He et al., 2023), providing an efficient geometric signal for model evaluation. Further, Arnold (2025) relate ID to memorization capacity, while Lee et al. (2025) show that nonlinear ID captures semantic compositionality, unlike linear measures such as PCA.

Persistent Homology Dimension (PHD) in Neural Networks. Intrinsic dimension can be estimated via Persistent Homology Dimension (PHD), which combines local and global properties of point cloud (Schweinhart, 2021). Birdal et al. (2021) link it to generalization properties of neural networks, while Tulchinskii et al. (2023a) show that texts generated by early GPT models exhibit lower PHD than human-written texts, enabling simple AI text detection. However, this distinction diminishes for newer models (Kuznetsov et al., 2024). Building on this, Kushnareva et al. (2024) demonstrate that local changes in intrinsic dimensionality can detect boundaries between human and machine-generated text segments.

Data-Centric View of Intrinsic Dimension. Several other studies also investigate the intrinsic dimension inherent to datasets and downstream tasks themselves. Early work by Narayanan and Mitter (2010) showed the theoretical connection between the intrinsic dimension of a dataset and the number of samples a model requires to learn from it. Pope et al. (2021) empirically confirmed this relationship on image datasets. These results demonstrate that the ID of a dataset can serve as a measure of learning difficulty—essentially quantifying the task complexity for the model.

Tsukagoshi and Sasano (2025) show that substantial dimensionality reduction of task-specific embeddings is possible with minimal performance loss across classification, clustering, and retrieval tasks, using estimators such as TwoNN and metrics like Isoscore. Yin et al. (2024) apply Local Intrinsic Dimension (LID) to hallucination detection, showing that high and highly variable LID signals fabricated or unsupported content, whereas truthful generations exhibit smoother, low-variance LID profiles. Ruppik et al. (2025) further argue that intrinsic dimensionality varies locally across embedding space: common tokens occupy simple, low-LID neighborhoods, while rare or domain-specific terms require higher-dimensional representations. Layer-wise analysis reveals that lower layers exhibit uniformly low LID, whereas higher layers show increased LID and variance, reflecting growing specialization and semantic disambiguation.

This rapidly expanding body of research demonstrates that intrinsic dimension in neural networks is not merely a mathematical curiosity, but rather a practical tool that offers pathways toward more interpretable and capable models and opens new applications in downstream tasks.

2 Intrinsic Dimension

Intrinsic dimension (ID) measures a dataset's degrees of freedom: for linear spaces and smooth manifolds it matches the usual dimension, and extensions to metric spaces aim to remain robust to noise and recoverable from finite samples.

A rich line of work has produced estimators of intrinsic dimension with complementary strengths. *TwoNN* estimates ID from ratios of first/second neighbor distances via a simple log–log fit (Facco et al., 2017); *MLE* fits a local Poisson model to k-NN distances; *TLE* leverages extreme-value theory in tight neighborhoods to reduce variance (Amsa-

leg et al., 2022). Beyond distance methods, PHD uses persistent homology to capture multi-scale connectivity/holes and is comparatively stable under text domain shift or added noise to embeddings. The formal definitions are given in App. A.1.

2.1 Relation between ID and model-prediction entropy

Intrinsic dimension (ID) is computed from the geometry of hidden representations and thus captures text complexity *from the model's embedding space*, complementing prediction-based metrics such as (cross-)entropy (Viswanathan et al., 2025). We show that ID and entropy are *complementary*: ID encodes structural information not reflected in entropy.

Formally, let $h_{1:T}$ be last-layer embeddings (the point cloud used for ID). Entropy arises only after *unembedding*: logits $z_t = W^\top h_t + b$ and softmax. Hence, entropy depends on the *alignment* of h_t with the vocabulary matrix W, whereas ID depends solely on the geometry of $\{h_t\}$. Figure 1 (left) illustrates this consideration: identical geometric structure (same ID) can yield different entropies if the local density of unembedding vectors differs.

Consequently, ID cannot be replaced by entropy: synonym-rich neighborhoods raise entropy without changing geometry; conversely, structurally complex but predictable text can have low entropy yet high ID (see App. A.2). Empirically, after controlling for length, ID and entropy are essentially uncorrelated, indicating that they capture distinct facets of complexity (Fig. 1, right); a residual dependence remains only in the very low-ID regime.

3 Intrinsic Dimension through the Lens of Text Semantics

We aim to characterize which semantic and stylistic properties of text are associated with higher or lower intrinsic dimension (ID). Our methodology combines sparse, interpretable representations learned from model activations with causal interventions and independent linguistic diagnostics.

3.1 Sparse Autoencoders (SAEs) and feature steering

SAEs learn sparse, approximately monosemantic features from LLM activations, addressing polysemantic neurons (Olah et al., 2020) under the superposition view (Elhage et al., 2022). Given a layer activation $\mathbf{x} \in \mathbb{R}^d$, the encoder–decoder (Sharkey et al., 2023)

$$\mathbf{f} = \sigma(\mathbf{W}_{\text{enc}}\mathbf{x} + \mathbf{b}_{\text{enc}}), \quad \hat{\mathbf{x}} = \mathbf{W}_{\text{dec}}\mathbf{f} + \mathbf{b}_{\text{dec}} \quad (1)$$

produces a sparse, nonnegative code $\mathbf{f}$ over decoder columns $\mathbf{d}_j$ (feature directions), with $\mathbf{x} \approx \mathbf{b}_{\text{dec}} + \sum_j \mathbf{f}_j \mathbf{d}_j$. For a text, we aggregate a feature's tokenwise activations (e.g., sum) to obtain its sequence-level value.

To probe causality, we modify the model's hidden state during generation using a chosen SAE feature (Kuznetsov et al., 2025):

$$\mathbf{x}' = \mathbf{x} + \lambda\, A_i\, \mathbf{d}_i, \quad (2)$$

where A_i is the feature's reference-scale (max activation on a held-out set) and λ controls intervention strength.

SAE features give interpretable axes of variation; steering tests their *causal* effects on text style and content. We use this to examine features correlated with intrinsic dimension (ID) and to validate whether they systematically increase or decrease representational complexity in generated text.

SAE features and selection. Building on evidence that sparse autoencoders (SAEs) separate semantic differences between synthetic and natural texts (Kuznetsov et al., 2025), we utilize SAEs trained on LLM activations to obtain sparse latent features. For each feature, we aggregate token-level activations to the sequence level and measure its correlation with ID. Features with the strongest positive and negative correlations are retained for analysis.

Interpreting and probing features. To ascribe semantic meaning to the selected features, we: (i) perform *extremal-activation* analysis by contrasting texts across activation quantiles; (ii) assess *domain specialization* via average activations over diverse human-written corpora; and (iii) conduct *steering* experiments, applying small interventions along a feature direction during generation to observe controlled changes while preserving fluency. The three perspectives jointly provide convergent evidence about how each feature relates to ID.

External linguistic validation. We relate ID to independent text properties using TAACO (Crossley et al., 2019) – the tool for analysing lexical diversity and cohesion, e.g., moving-window type–token ratios and sentence-to-sentence overlap (see Appendix C.3 for details) – and syntactic

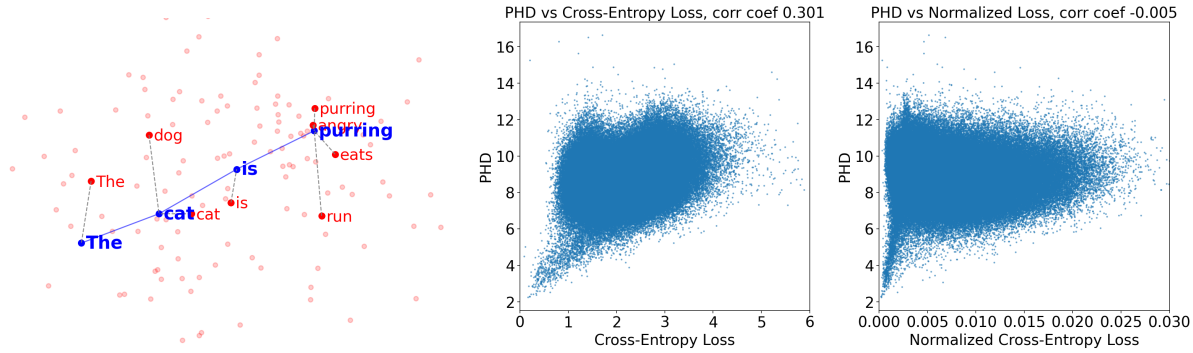

Figure 1: Intrinsic dimension characterizes the geometry of hidden representations (blue points on the leftmost frame), while prediction-based metrics such as entropy and cross-entropy depend on the unembedding dictionary (red points on the leftmost frame). Sequences of embeddings with the same intrinsic dimension may yield very different prediction entropies, depending on how densely the unembedding vectors populate the surrounding space (i.e., on the number of close neighbors, shown by grey connections). Note significant correlation between PHD and Cross-Entropy loss (center frame) and weak correlation between PHD and Cross-Entropy loss, normalized by text length in gemma tokens (rightmost frame).

metrics: syntactic diversity measured as average pairwise distance between dependency-graph representations via a Weisfeiler–Lehman kernel (Guo et al., 2025), and the POS compression ratio based on gzip over POS-tag sequences (Shaib et al., 2024) – see Appendix C.2. These diagnostics corroborate SAE-based findings and ground ID variation in observable lexical and syntactic structure.

4 Experimental details

Data. We use the GenAI Content Detection Task 1 dataset of Wang et al. (2025) (henceforth, COLING). Unless stated otherwise, experiments run on a cleaned split of the development set: from 261,758 texts (human and multiple LLM families), we retain 172,750 after filtering out (i) samples with < 150 GEMMA tokens and (ii) cases where any PHD estimator on GEMMA-2-2B, QWEN-2.5-1.5B, or ROBERTA-BASE was unstable or outside $[2, 18]$. COLING spans diverse domains (news, Wikipedia articles, essays, QA, summaries, reviews, scientific abstracts, forums, technical documentation) and generations from T5/T0, GPT, LLaMA, OPT, Mixtral, and others. Most semantic interpretation experiments are performed on the subset of Human-written texts.

For proficiency analysis we use ONESTOPENGLISH (Vajjala and Lučić, 2018): 189 human-authored texts, each in Elementary/Intermediate/Advanced versions (567 total). For controlled generation we use RAID (Dugan et al., 2024): (i) 1,000 prompts with temperatures 0.2–2.0 for temperature–ID studies; (ii) 20 prompts from each of five domains (abstracts, books, news, Reddit, Wikipedia) for fine-grained steering.

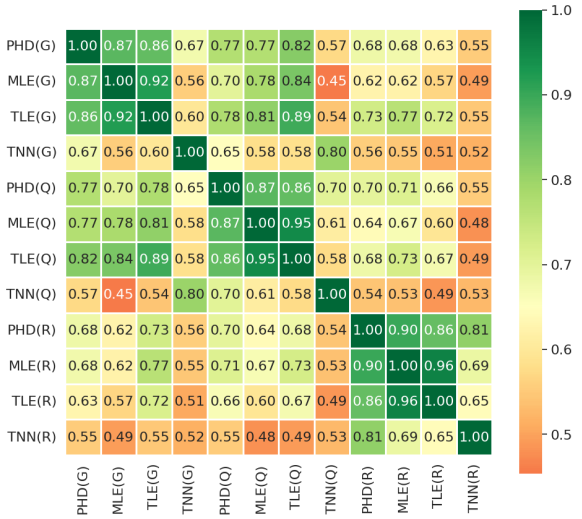

Figure 2: Correlations among various ID estimators. (G) denotes ID estimators upon Gemma, (R) - RoBERTa, (Q) - Qwen. Note that PHD estimators upon all three models have correlation more than 0.5 with all other estimators, making it a solid compromise. See Appendix B for scatterplots and further discussion.

SAE setup. We employ GEMMA-SCOPE (Lieberum et al., 2024) SAEs for GEMMA-2 (Team, 2024), using the canonical residual SAE (width 16k) on GEMMA-2-2B (base and instruct). For steering, we apply $\lambda \in \{0.25, 0.5, 1, 5, 10, 15\}$ and report the largest λ per feature that preserves coherence (very large λ can induce repetitive or broken outputs).

5 Results

5.1 Comparison of ID estimators

We compare PHD, MLE, TLE, and TwoNN on Gemma, Qwen, and RoBERTa embeddings (Fig. 2). All estimators are substantially correlated (pairwise $r > 0.45$). TwoNN deviates most, whereas PHD tracks the others best ($r > 0.67$ with each on the same embeddings), so we use PHD as the primary estimator.

Across embeddings, Gemma and Qwen agree closely for a fixed estimator, while RoBERTa diverges more; we revisit this in Appendix.

Within a model family, ID grows slightly with model size (Fig. 11). Nevertheless, across methods, sizes, and families, well-formed texts concentrate in a narrow band (ID $\approx$ 5–15), with outliers due to very short texts (Fig. 14) or broken syntax. These IDs remain far below ambient embedding dimensions (hundreds to thousands).

5.2 PHD and data properties

Text length. We find PHD to be stable for sufficiently long samples, with high variance only for short texts (App. C.1). Hence, we exclude texts shorter than 150 tokens.

Intrinsic data complexity. PHD is positively correlated with gzip compression ratio (CR); for texts longer than 150 tokens the correlation exceeds 0.3 (Fig. 3). The scatter exhibits a wedge-shaped support: $10\,\mathrm{CR} \lesssim \mathrm{ID} \lesssim 25\,\mathrm{CR}$. Thus, CR bounds but does not pointwise predict ID.

Embedding geometry. Appendix B.3 shows only weak correlations between ID and isotropy measures, indicating that radial/directional uniformity is not a driver of PHD. By contrast, the correlation between PHD and explained variance at rank k (EV-k) is sharply peaked at $k \approx 60$ (Fig. 12). This suggests a *global linear embedding dimension* of ~ 60 despite a local ID of ~ 10, highlighting strong manifold nonlinearity and cautioning against purely linear characterizations.

Lexical properties. PHD correlates positively with lexical diversity (type–token ratio and n-gram diversity) and negatively with sentence-level overlap/repetition; syntax/discourse metrics show weak associations (Fig. 4). Appendix experiments further show: (i) robustness to perturbations that preserve repetition patterns while destroying semantics; (ii) extreme out-of-range values on broken generations from weaker LLMs (very low for repetitive, very high for disconnected text); and (iii) limited informativeness of syntactic metrics for well-formed texts.

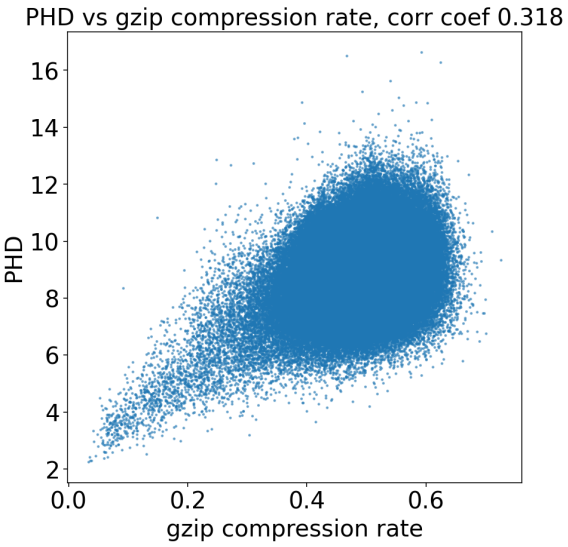

Figure 3: PHD and gzip

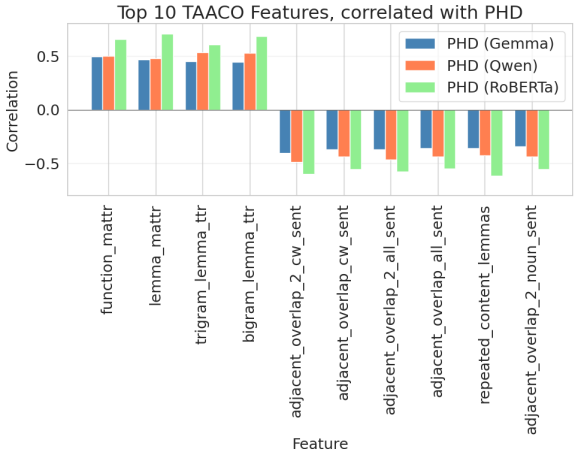

Figure 4: Top-10 features from TAACO with the strongest correlation with PHD(Gemma). See Appendix C.3 for similar barplots with MLE, TLE, TwoNN.

5.3 ID divergence by text source and domain

Contrary to Tulchinskii et al. (2023b), we observe a clear shift of ID values between domains (Fig. 5). The domains can be divided into three groups with distinct semantic and stylistic characteristics:

1. **Low-dimensional group** (*mean* PHD *7.5–8*): includes scientific and technical sources such as `arxiv`, `medicine`, `pubmed`, `wiki_csai`, and `sci_gen`.

2. **Medium-dimensional group** (*mean* PHD *8–9*): consists of factual and reporting texts with

encyclopedic or formal news style, such as `cnn`, `xsum`, and `wikipedia`.

3. **High-dimensional group** (*mean* PHD *9–10.5*): includes essays, opinionated writing, and user-generated content with more informal or creative style, such as `cmv`, `eli5`, `wp`.

The lowest ID values appear in `pubmed` papers, while the highest (among public-domain corpora) occur in fictional stories from `wp`. Notably, `peerread` (OpenReview discussions), which blends scientific and opinionated styles, lies in the mid-PHD range. The ordering of domains by PHD is nearly identical across encoders (Fig. 6).

We next estimate PHD for student-written stories at three proficiency levels—elementary, intermediate, and advanced (Fig. 25). PHD rises with linguistic complexity: elementary texts have mean PHD 9.5 (near the lower edge of the *fictional* group), whereas advanced texts approach or slightly exceed `wp` fiction (mean 12.5). Although ID correlates with proficiency, genre is the stronger factor: even elementary *fiction* surpasses *scientific* and *informational* genres.

These results imply that, for modern LLMs, scientific texts are the "simplest," while informal, opinionated, and creative writing drives higher intrinsic dimensionality, peaking in fiction. This pattern aligns with our linguistic analysis: scientific prose shows strong topical/terminological coherence (high overlap/repetition, negatively correlated with PHD), whereas fiction demands greater lexical diversity. Simplified syntax at the elementary level has little effect on embedding dimensionality.

5.4 SAE-based interpretation

We selected a set of middle-layer features with the strongest absolute correlation with intrinsic dimension (ID), including both positively and negatively correlated cases. First, we computed mean activations across domains. Figure 7 shows a clear bifurcation: a *scientific* cluster (typified by `arXiv`) and a *humanized* cluster with strongest activations in `wp` (writing prompts/short fiction). The former correlates negatively with ID, the latter positively, confirming a close link between ID and content (Table 1). We then interpreted features via (i) extremal-activation examples and (ii) targeted steering. The evidence is consistent: features *positively* correlated with ID yield more emotional, personalized, rhetorically complex text, whereas *negative* features push generations toward academic/reporting

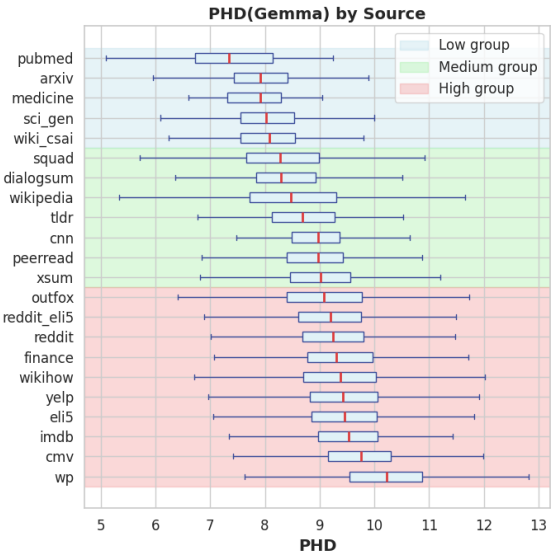

Figure 5: PHD(Gemma) by source with group differentiation.

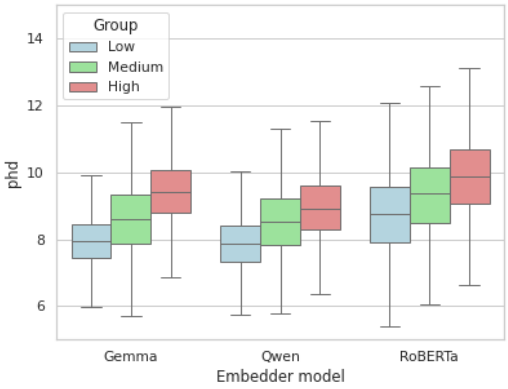

Figure 6: PHD by groups with different embedders models.

styles with formal tone and rigid structure (Table 1). See Appendix D.2 for additional details and results.

Layer-wise, late layers (24–25) set global discourse templates, e.g., movie-plot, social-media article, BBC-style report, analytical template (cf. `14085`, `4610`, `15879`, `2409`), whereas layer 16 modulates local stylistic details (e.g., scientific framing `5159`, population statistics `8104`, claim foregrounding `2433`). Among informal signals, `5228` induces an uncertain or subdued tone, `9868` injects engaging elements into abstracts, and `6978` steers toward a "writing-assistant" voice. Table 2 shows concise steering transformations; additional examples appear in App. D.3.

To further explore causality, we conduct an additional small-scale experiment. We select 40 texts

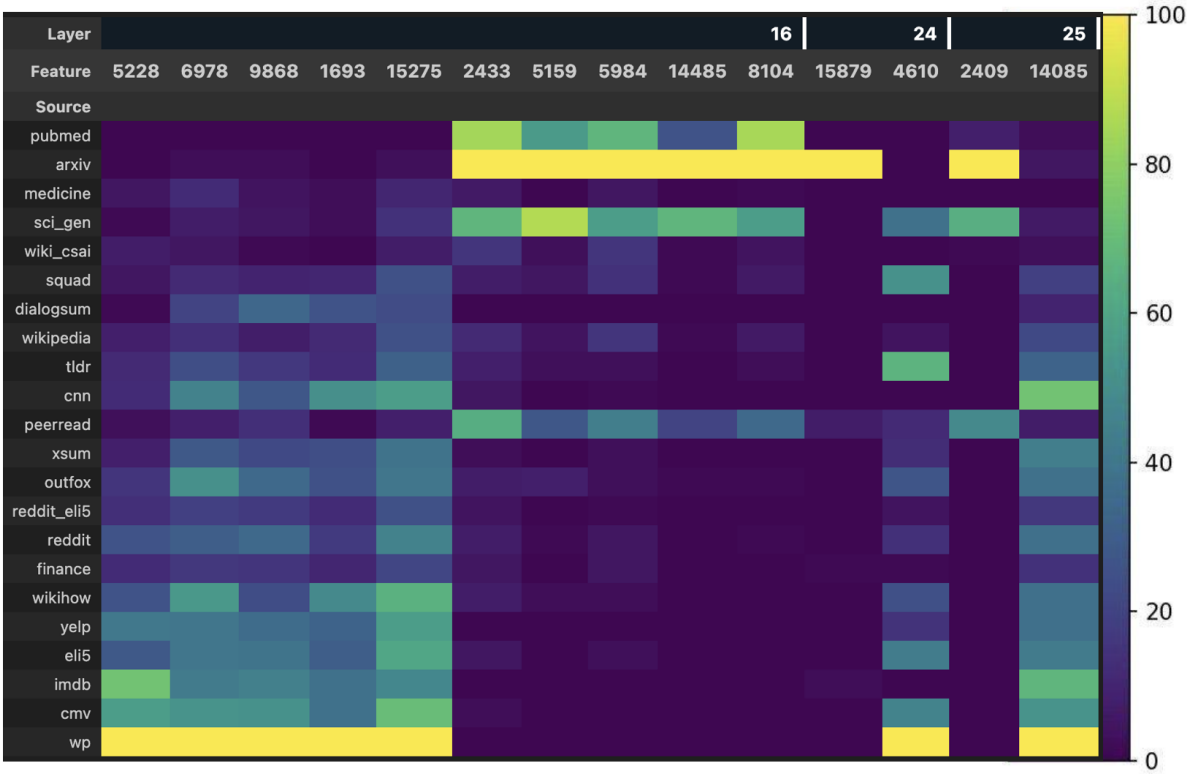

Figure 7: Colormap of SAE feature activations table by source of the text (domain)

Feature	Domain	r	Extreme values	Steering
Typical features with negative correlation with ID:				
16-5984	arxiv, sci-gen	−0.3	Polished, formal, logically structured documents (e.g., academic, reports)	Academic reframing; nominalizations; enumerations
16-14485	arxiv, sci-gen	−0.36	Complex arguments, structured academic or expository texts	Formal academic tone; expository structure; sectioning
16-2433	arxiv, pubmed, sci-gen, peerread	−0.35	Unified compositions with clear rhetorical purpose	Outline/placeholder scaffold; bulletization; meta-headers
16-8104	arxiv, pubmed, sci-gen	−0.33	Continuous formal prose (abstracts, summaries)	Social-science framing; survey stats; institutional voice
16-5159	arxiv, sci-gen	−0.34	Academic/scientific style, technical vocabulary, objective tone	Analytical reportese; self-reference; repetitive passive
24-15879	arxiv	−0.31	Highly formal, technical, objective academic prose	BBC/news article template; anecdotal lede; call-to-action
25-2409	arxiv, sci-gen, peerread	−0.43	Formal, self-contained, expository prose	Stronger macro-structure; cohesive sections; “finally” cadence
Typical features with positive correlation with ID:				
16-15275	wp, cmv, wikihow, eli5, yelp, cnn	0.36	Coherent, human-readable, publication-quality texts	Personalization; concrete characters; richer plot detail
16-5228	wp, imdb	0.33	Conversational, emphatic, lengthened texts with repetition	Uncertainty/hedging; introspection; open-endedness
16-6978	wp, wikihow, outfox, cnn	0.36	Multi-source, composite, sometimes incoherent synthesis	Informal/meta voice; bracketed notes; assistant aside
16-9868	wp	0.37	Informal, noisy, user-generated style; robust to corrupted input	Media/teaser tone; punchy fragments; promo cadence
16-1693	wp, cnn, wikihow	0.37	Long, elaborated, multi-threaded arguments/narratives	Forum/discussion vibe; colloquial; community framing
25-14085	wp, cnn, imdb	0.4	Personal voice; 1st/2nd person; informal, conversational, opinionated; narratives and reviews.	Movie/outline format; headings; character bios
24-4610	wp, tldr	0.36	Paragraphs or list items start with ‘,‘, ‘„‘, sometimes exaggerated ‘„„‘	Social-media article; bold subheads; quotes + how-to steps

Table 1: Short interpretations of selected features and their effect on document style depending on activation.

from the COLING dataset that were written by humans and exceeded 150 Gemma tokens in length (these criteria ensure both text quality and robustness of PHD estimator).

For each feature listed in Table 1, we prompted Gemini 2.5 Pro (Comanici et al., 2025) to reformulate the corresponding texts so that they matched the descriptions in the Extreme Values and Steering columns of the Table 1. We then compare ID of the reformulated texts and the original texts. To assess whether the differences were statistically significant, we applied an independent-samples t-test for differences in means, assuming unequal variances. The results are presented in Table 14.

Among the features that have a negative impact on PHD, we observe heterogeneous behavior: for some signs, the PHD scores increase; for others, they decrease or show no difference relative to the original texts. However, both the mean and the median PHD values remain within the 7.6–8.8 range (low to medium).

For the features that have a positive impact on PHD, the average PHD scores of the reformulated texts increase relative to the originals, placing them within the "high" range: 8.87–9.6 (by mean and median). These findings provide additional grounds to discuss potential causal effects, at least for the positively correlated features. The inconsistent behavior observed for negatively correlated features warrants further investigation.

6 Discussion

Intrinsic dimension (ID) remains an underexplored measure of textual complexity. Our extensive experiments (with details in the Appendix) elucidate its semantic and structural correlates and support the following conclusions.

Complementarity to prediction-based metrics. ID is complementary to (cross-)entropy and related prediction-based measures: after controlling for length, the two are largely uncorrelated, indicating that ID captures geometric/structural information not reflected in likelihood.

Narrow operating band, far below ambient dimension. For well-formed texts, ID concentrates in a narrow range (≈5–15) and is remarkably stable across domains, encoder families/sizes, and ID estimators. Yet ID *cannot* be substituted by linear dimensionality reduction: the manifold is highly curved, with an effective *global linear embedding dimension* of ∼60–100 (e.g., peak PHD–EV_k correlation near $k \approx 60$).

Failure detection. ID reliably flags broken generations: extremely low values for repetitive/looped text and extremely high values for disconnected fragments or incoherent mixtures.

Compression is bounding, not predictive. ID correlates moderately with gzip compression ratio, but the relationship is not pointwise predictive; compression provides bounds on plausible ID rather than accurate estimates.

Linguistic correlates. ID increases with lexical diversity and decreases with cross-sentence repetition/overlap; it is comparatively insensitive to syntactic indicators for well-formed text.

Style and genre effects. ID is strongly conditioned by style/genre. Scientific and purely informational prose exhibits lower ID, whereas personalized, opinionated, and fictional writing attains the highest ID. Adding "semantic dimensions", such as personality, emotion, narrative/plot, stance, informality, raises ID.

Implications and cautions. Prior work has used ID for training dynamics evaluation, synthetic text detection, scaling-law validation, architecture search, and memorization studies. However, treating ID as a monolithic proxy for "difficulty" risks misinterpretation: domain shifts, spurious correlations, and evaluations confined to low-ID corpora (e.g., Wikipedia-like text) can yield misleading conclusions. Our findings indicate that complexity-based LLM analyses should explicitly target higher-ID domains (forums, fiction, opinionated writing) and report local (ID) linear structure, rather than relying solely on linear projections or prediction-based metrics.

7 Conclusion

In this work, we have analyzed intrinsic dimensionality as a lens through which to interrogate the geometric structure of text representations in large language models. Our findings reveal that ID is not a monolithic indicator of complexity, but rather a domain- and style-sensitive property: scientific and informational texts occupy low-ID regions, while opinionated, narrative, and personalized writing exhibit substantially higher ID, reflecting greater semantic degrees of freedom. Critically, ID is largely orthogonal to prediction-based measures like cross-entropy, underscoring its value in exposing structural regularities that likelihood alone cannot capture.

Feature	Sign	Shift	Example change (steered)
Default generation (excerpt): *The classic Ford Falcon, a symbol of American muscle in the 1960s, is a popular choice for car enthusiasts. … The problem can stem from a number of factors, including a clogged fuel line, a faulty fuel pump, or a blocked fuel filter.*			
16-14485	–	Research-report structure	"The article identifies three primary causes for the fuel supply deficiency; firstly..., secondly..., thirdly...."
16-5159	–	Analytical, self-referential	"This paper examines potential causes and provides a diagnostic framework; this paper is intended to be a guide."
16-15275	+	Storytelling, humanized	"John, a 30-year restoration veteran, replaced the pump, lines, and filter — yet the Falcon still refuses to start."
16-1693	+	Forum-like conversational	"This can be a real headache — we're gonna get that carb working again!"

Table 2: Steering effect of several representative features. "Sign" reflects if the feature positively or negatively correlated with intrinsic dimension

8 Limitations

Our PHD analysis uses embeddings from only three models (Gemma, RoBERTa, Qwen); results may vary with other encoders, as intrinsic dimension estimates are sensitive to embedding choice. Moreover, PHD is computed on random subsamples of the data, and due to this stochastic sampling procedure, PHD estimates may exhibit some variability across runs. Finally, common estimators—TLE, MLE, TwoNN, and PHD—rest on different assumptions about the geometry and distribution of token embeddings. Consequently, these estimators capture complementary but non-equivalent aspects of intrinsic dimensionality and are not directly comparable.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328, Online. Association for Computational Linguistics.

Laurent Amsaleg, Oussama Chelly, Michael E. Houle, Ken ichi Kawarabayashi, Miloš Radovanović, and Weeris Treeratanajaru. 2022. Intrinsic dimensionality estimation within tight localities: A theoretical and experimental analysis. *arXiv preprint arXiv:2209.14475*.

Stefan Arnold. 2025. Memorization in language models through the lens of intrinsic dimension. In *Proceedings of the First Workshop on Large Language Model Memorization (L2M2)*, pages 23–28, Vienna, Austria. Association for Computational Linguistics.

Rajendra Bhatia. 1997. *Matrix Analysis*, volume 1. Springer.

Tolga Birdal, Aaron Lou, Leonidas J Guibas, and Umut Simsekli. 2021. Intrinsic dimension, persistent homology and generalization in neural networks. *Advances in neural information processing systems*, 34:6776–6789.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

S. A. Crossley, K. Kyle, and M. Dascalu. 2019. The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavioral Research Methods 51(1), pp. 14-27*.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, Bangkok, Thailand. Association for Computational Linguistics.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Preprint*, arXiv:2209.10652.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. 2017. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2025. Benchmarking linguistic diversity of large language models. *Preprint*, arXiv:2412.10271.

Alex Havrilla and Wenjing Liao. 2024. Understanding scaling laws with statistical and approximation theory for transformer neural networks on intrinsically low-dimensional data. In *Advances in Neural Information Processing Systems*, volume 37, pages 42162–42210.

Xin He, Jiangchao Yao, Yuxin Wang, Zhenheng Tang, Ka Chun Cheung, Simon See, Bo Han, and Xiaowen Chu. 2023. Nas-lid: Efficient neural architecture search with local intrinsic dimension. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7839–7847.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net.

Laida Kushnareva, Tatiana Gaintseva, German Magai, Serguei Barannikov, Dmitry Abulkhanov, Kristian Kuznetsov, Eduard Tulchinskii, Irina Piontkovskaya, and Sergey Nikolenko. 2024. Ai-generated text boundary detection with roft. *Preprint*, arXiv:2311.08349.

Kristian Kuznetsov, Laida Kushnareva, Anton Razzhigaev, Polina Druzhinina, Anastasia Voznyuk, Irina Piontkovskaya, Evgeny Burnaev, and Serguei Barannikov. 2025. Feature-level insights into artificial text detection with sparse autoencoders. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25727–25748, Vienna, Austria. Association for Computational Linguistics.

Kristian Kuznetsov, Eduard Tulchinskii, Laida Kushnareva, German Magai, Serguei Barannikov, Sergey Nikolenko, and Irina Piontkovskaya. 2024. Robust AI-generated text detection by restricted embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17036–17055, Miami, Florida, USA. Association for Computational Linguistics.

Jin Hwa Lee, Thomas Jiralerspong, Lei Yu, Yoshua Bengio, and Emily Cheng. 2025. Geometric signatures of compositionality across a language model's lifetime. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5292–5320, Vienna, Austria. Association for Computational Linguistics.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. 2018. Measuring the intrinsic dimension of objective landscapes. *Preprint*, arXiv:1804.08838.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300, Miami, Florida, US. Association for Computational Linguistics.

Philip M. McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Hariharan Narayanan and Sanjoy Mitter. 2010. Sample complexity of testing the manifold hypothesis. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024.001.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. 2021. The intrinsic dimension of images and its impact on learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Anton Razzhigaev, Matvey Mikhalchuk, Elizaveta Goncharova, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov. 2024. The shape of learning: Anisotropy and intrinsic dimensions in transformer-based models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 868–874, St. Julian's, Malta. Association for Computational Linguistics.

Olivier Roy and Martin Vetterli. 2007. The effective rank: A measure of effective dimensionality.

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 Gašić. 2025. Less is more: Local intrinsic dimensions of contextual language models. *Preprint*, arXiv:2506.01034.

Benjamin Schweinhart. 2021. Persistent homology and the upper box dimension. *Discrete & Computational Geometry*, 65(2):331–364.

Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C Wallace. 2024. Detection and measurement of syntactic templates in generated text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6416–6431, Miami, Florida, USA. Association for Computational Linguistics.

Lee Sharkey, Hoagy Cunningham, Aidan Ewart, Logan Riggs, and Robert Huben. 2023. Sparse autoencoders find highly interpretable features in language models. *Preprint*, arXiv:2309.08600.

Lucas Shen. 2022. Lexicalrichness: A small module to compute textual lexical richness.

Hermann H Somers. 1966. Statistical methods in literary analysis. *The computer and literary style*, 128:140.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. Ieee.

Hayato Tsukagoshi and Ryohei Sasano. 2025. Redundancy, isotropy, and intrinsic dimensionality of prompt-based text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25915–25930, Vienna, Austria. Association for Computational Linguistics.

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

Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Sergey Nikolenko, Evgeny Burnaev, Serguei Barannikov, and Irina Piontkovskaya. 2023b. Intrinsic dimension estimation for robust detection of AI-generated texts. In *Advances in Neural Information Processing Systems*. 37th NeurIPS (poster).

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Karthik Viswanathan, Yuri Gardinazzi, Giada Panerai, Alberto Cazzaniga, and Matteo Biagetti. 2025. The geometry of tokens in internal representations of large language models. *Preprint*, arXiv:2501.10573.

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, and 7 others. 2025. GenAI content detection task 1: English and multilingual machine-generated text detection: AI vs. human. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 244–261, Abu Dhabi, UAE. International Conference on Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Fan Yin, Jayanth Srinivasa, and Kai-Wei Chang. 2024. Characterizing truthfulness in large language model generations with local intrinsic dimension. In *ICML*.

Viacheslav Yusupov, Danil Maksimov, Ameliia Alaeva, Anna Vasileva, Anna Antipina, Tatyana Zaitseva, Alina Ermilova, Evgeny Burnaev, and Egor Shvetsov. 2025. From internal representations to text quality: A geometric approach to llm evaluation. *Preprint*, arXiv:2509.25359.

A Mathematics behind Intrinsic Dimension

A.1 Intrinsic dimension definitions

We outline several common notions of intrinsic dimension used in data analysis, and illustrate them with uniformly distributed points in a d-dimensional Euclidean space. Throughout, $\sim$ denotes asymptotic proportionality.

MLE (Maximum Likelihood Estimation dimension). For a point x consider the probability that a random point x' falls in its ε-neighborhood. For a uniform distribution, this probability scales as:

$$\Pr(\|x - x'\|_2 < \varepsilon) \;\sim\; C(x)\,\varepsilon^d. \tag{1}$$

Hence, the exponent in Eq. (1) defines the *MLE dimension*. Formally, we compute

$$\dim_{\mathrm{MLE}} = \mathbb{E}_x \left[\lim_{\varepsilon \to 0} \frac{\log N_\varepsilon(x)}{\log 1/\varepsilon} \right], \tag{1a}$$

where $N_\varepsilon(x)$ is the number of points in the ε-neighborhood of x.

TwoNN (Two Nearest Neighbors dimension). For each point x, let r_1 and r_2 be the distances to its nearest and second-nearest neighbors, and define the ratio

$$\mu = r_2/r_1.$$

For uniform distribution in $\mathbb{R}^d$, the following holds:

$$\Pr(\mu \le t) = 1 - t^{-d}, \quad t \ge 1. \tag{2}$$

Thus, the distribution of μ depends only on d. We can estimate d by fitting this model CDF to the

empirical CDF $\hat{F}(\mu)$, for example by minimizing a divergence:

$$\dim_{\mathrm{TwoNN}} = \underset{d}{\operatorname{argmin}} \; D_{\mathrm{KL}}\Big(\hat{F}(\mu) \,\|\, (1-\mu^{-d})\Big), \quad \text{(2a)}$$

where $\hat{F}(\mu)$ is the empirical cumulative distribution function of the ratios.

PHD (Persistent Homology Dimension). Let N_ε be the number of nontrivial homological features (e.g., connected components or cycles) that appear in a simplicial complex built from the data at scale ε. For a set in $\mathbb{R}^d$ we get

$$N_\varepsilon \;\sim\; C\,\varepsilon^{-\dim_{\mathrm{PH}}}. \quad (3)$$

Thus, the *PH dimension* is defined as

$$\dim_{\mathrm{PH}} = \limsup_{\varepsilon\to 0} \frac{\log N_\varepsilon}{\log 1/\varepsilon}. \quad \text{(3a)}$$

An equivalent characterization uses the total length of the minimum spanning tree (MST). Let L_n be the MST length on n points sampled uniformly from the set. Asymptotically, one has

$$L_n \;\sim\; n^{\frac{d-1}{d}}, \quad \text{(3b)}$$

which yields

$$\dim_{\mathrm{PH}} = \lim_{n\to\infty} \frac{1}{1-\frac{\log L_n}{\log n}}. \quad \text{(3c)}$$

A.2 Three Arguments Against Functional Dependence Between Entropy and Intrinsic Dimension

In prior work (Viswanathan et al., 2025), the authors present several mathematical arguments suggesting a possible dependency between intrinsic dimension and entropy, although no formal proof is provided. In contrast, we offer three complementary arguments in the opposite direction. Each of them serves as a constructive counterexample, demonstrating that no functional dependency between entropy and intrinsic dimensionality can exist in the general case.

Of course, neural networks do not represent the fully general case: their embeddings and parameter spaces are restricted by training dynamics and architectural constraints. However, our empirical results support these theoretical considerations, indicating that even in practical settings the observed correlations arise empirically rather than by necessity.

(1) Dependence Argument. Let $U = \{u_i\}_{i=1}^V \subset \mathbb{R}^m$ be the unembedding matrix and $p(h) = \mathrm{softmax}(Uh)$ the model's prediction for a hidden state h. The prediction entropy is

$$H(h) = -\sum_i p_i(h)\log p_i(h),$$

which depends explicitly on U.

In contrast, the intrinsic dimension (ID) of a set $H \subset \mathbb{R}^m$ depends only on the *geometry of embeddings*—it is a property of their relative arrangement in $\mathbb{R}^m$ and does not depend on the unembedding matrix U.

Even within a single model, one can select disjoint subsets of tokens whose projections under U differ strongly, leading to distinct entropy distributions while the geometry of their embeddings (and thus ID) remains unchanged.

Example. Assume that U consists of two parts: a dense cluster of "positive" tokens u_i with nonnegative components, and a single "negative" token $u_0 = (-1,-1,\ldots,-1)$. Consider a set H_+ of hidden vectors h with positive coordinates. For such h, all scalar products $u_i^\top h$ are positive, so $\mathrm{softmax}(Uh)$ is approximately uniform, and the average entropy $H(h)$ is close to $\log V$. If we flip the signs of all these vectors, obtaining $H_- = \{-h : h \in H_+\}$, the geometric structure and intrinsic dimension are identical, but now all scalar products $u_i^\top(-h)$ are negative, so $\mathrm{softmax}(Uh)$ becomes nearly one-hot and $H(h) \approx 0$. Thus, entropy can vary drastically while intrinsic dimension remains the same.

(2) Scale Invariance Argument. Consider a scaling map $T_\alpha(h) = \alpha h$. Since affine transformations preserve local dimensionality,

$$\mathrm{ID}(T_\alpha(H)) = \mathrm{ID}(H).$$

However, the corresponding prediction distribution transforms as

$$p(T_\alpha(h)) = \mathrm{softmax}(U(\alpha h)) = \mathrm{softmax}(\alpha U h),$$

which is equivalent to inverse-temperature rescaling:

$$\alpha \to 0 \Rightarrow p \to \tfrac{1}{V}\mathbf{1}, \qquad \alpha \to \infty \Rightarrow p \to e_k,$$

and therefore

$$H(T_\alpha(h)) \in [0, \log V].$$

Thus, entropy can be continuously adjusted from its maximum to minimum while intrinsic dimension remains invariant. Entropy and ID are therefore *affinely independent quantities*.

(3) Continuity Argument. With a fixed unembedding matrix, (token-level) entropy induces a continuous function $H : \mathbb{R}^m \to \mathbb{R}$ on the embedding space: it depends only on a point's location in the ambient space (hence on distances to the unembedding vectors), not on its neighbors along the data manifold.

For a measurable set $A \subset \mathbb{R}^m$, write $\bar{H}(A) := |A|^{-1} \int_A H(h)\, dh$. Let $\mathcal{M}_d$ be a compact d-dimensional submanifold and $N_\varepsilon(\mathcal{M}_d) = \{x : \mathrm{dist}(x, \mathcal{M}_d) \le \varepsilon\}$ its tubular ε-neighborhood, then we have:

$$\bar{H}\big(N_\varepsilon(\mathcal{M}_d)\big) \xrightarrow[\varepsilon \to 0]{} \bar{H}(\mathcal{M}_d).$$

Assume further that for any compact d-manifold $\mathcal{M}_d$ the mean $\bar{H}(\mathcal{M}_d)$ depends only on d (denote this value by $\mu(d)$). A compact D-manifold $\mathcal{M}_D$ admit a finite cover by such tubular sets. Then

$$\bar{H}(\mathcal{M}_D) \;=\; \sum_{i=1}^{N} \alpha_i(\varepsilon)\, \bar{H}\big(N_{\varepsilon_i}(\mathcal{M}_d^{(i)})\big) \;=\; \mu(d),$$

where $\alpha_i(\varepsilon) = |N_{\varepsilon_i}(\mathcal{M}_d^{(i)})| / \big|\bigcup_j N_{\varepsilon_j}(\mathcal{M}_d^{(j)})\big|$ are convex weights.

Conclusion. The mean entropy is continuous under infinitesimal "thickenings" of dimension: it cannot exhibit jumps across intrinsic dimensions and therefore cannot be a functional of intrinsic dimension alone.

B The connection between various ID Estimators

B.1 MLE, TLE and TwoNN upon embeddings from main models

Figs. 8, 9, and 10 illustrate pairwise relationships among intrinsic-dimension estimators computed on embeddings produced by the same model. MLE and TLE exibit very similar tracks but differ substantially from TwoNN. Notable, that PHD remains strongly correlated with all other estimators.

Most ID estimators are monotonically related; for Qwen and Gemma embeddings, the relationships are nearly linear. RoBERTa stands out as an exception having substantially non-linear relationships. The TwoNN estimator on RoBERTa is non-monotonic with respect to the others. See Fig. 10 for details.

B.2 PHD dependence on model size

We wondered how the number of parameters in the embedder model affects the internal dimension of the text. To ensure consistency, we used only Qwen 3 base models with sizes 0.6B, 1.7B, 4B, 8B, 14B, 32B. We calculated PHD on human texts with a length of more than 150 tokens and obtained the following results, see Tab. 3.

We see a general trend towards an increase in PHD as the model size increases. Qwen3-4B stands out from this picture. This may be due to the fact that only this model has a 128k context window and uses the Embeddings Tie. For smaller models, the context window is 32k, and Embedding Tie is used. Embedding Tie is a technique for reducing the number of parameters for small LLMs. Embedding Tie is not used for Qwen3 models with more than 4B parameters, but the context window increases to 128k for these models..(Yang et al., 2025).

Model	PHD (median)	Emb. size	Num. layers
Qwen3-0.6B	9.429	1024	28
Qwen3-1.7B	9.534	2048	28
Qwen3-4B	10.063	2560	36
Qwen3-8B	9.601	4096	36
Qwen3-14B	10.046	5120	40
Qwen3-32B	10.449	5120	64

Table 3: Median PHD dependence on Qwen3 size

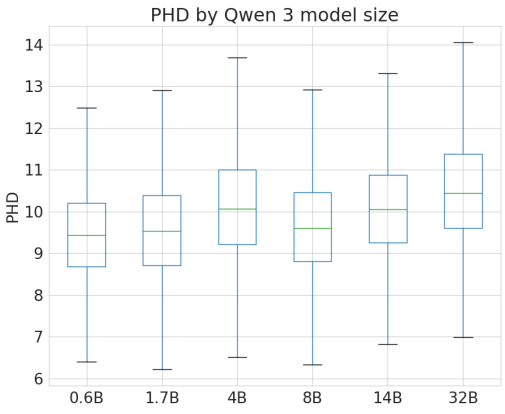

Figure 11: Boxplots of PHD across Qwen3 model sizes

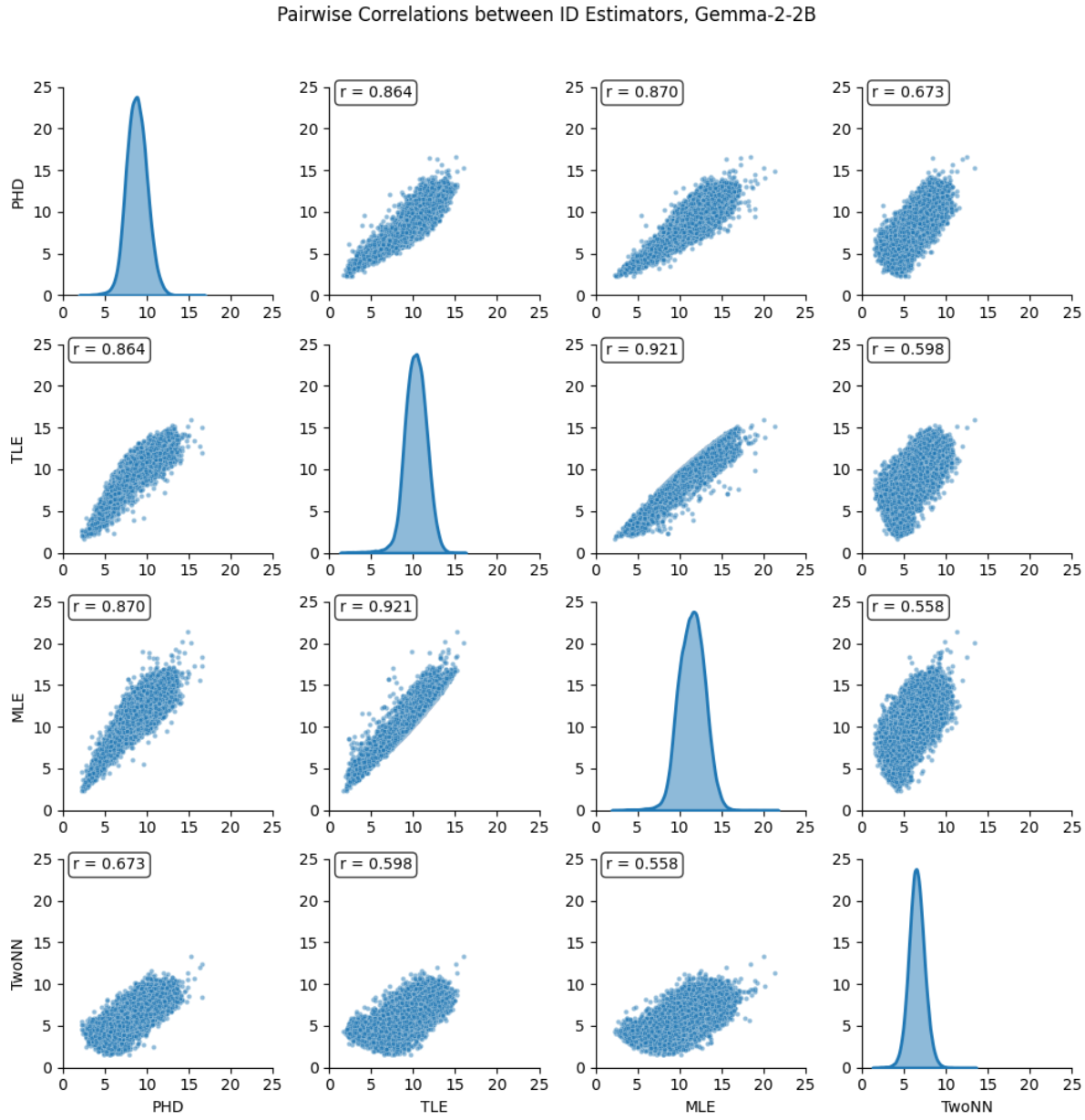

Figure 8: ID Estimators Gemma

B.3 Intrinsic Dimensionality and data geometry

In this section, we study the connection between ID and other geometric properties of data in the embedding space, related to dataset complexity.

There are several methods to measure the "diversity" (anisotropy) of a given dataset and here we will study their correlation with the intrinsic dimensionality. Suppose we have a text and the set of its tokens embeddings $X = (x_1, \ldots, x_N), x_i \in \mathrm{R}^d$. As $\sigma_1, \sigma_2, \ldots, \sigma_r$ ($r = \min(N, d)$) we denote the singular values of the matrix of stacked embeddings X.

Following (Yusupov et al., 2025), we explore the following metrics:

- **Maximal Explained Variance (MEV)** (Razzhigaev et al., 2024) – the proportion of variance explained by the first principal component of the data (i.e., $\frac{\sigma_1^2}{\sum_{i=1}^{r}\sigma_i^2}$. Additionally, we calculate the proportion of variance, explored by the first 20 components ($\frac{\sigma_1^2+\ldots\sigma_{20}^2}{\sum_{i=1}^{r}\sigma_i^2}$, we denote it as **20-EV**); we properly explore the effect of the number of components further down in this section.
- **Resultant Length** (Ethayarajh, 2019). Let x_i' be the unit-normalized token embeddings (i.e., $x_i' = \frac{x_i}{||x_i||_2}$, $i = 1..n$). Then, the Resultant

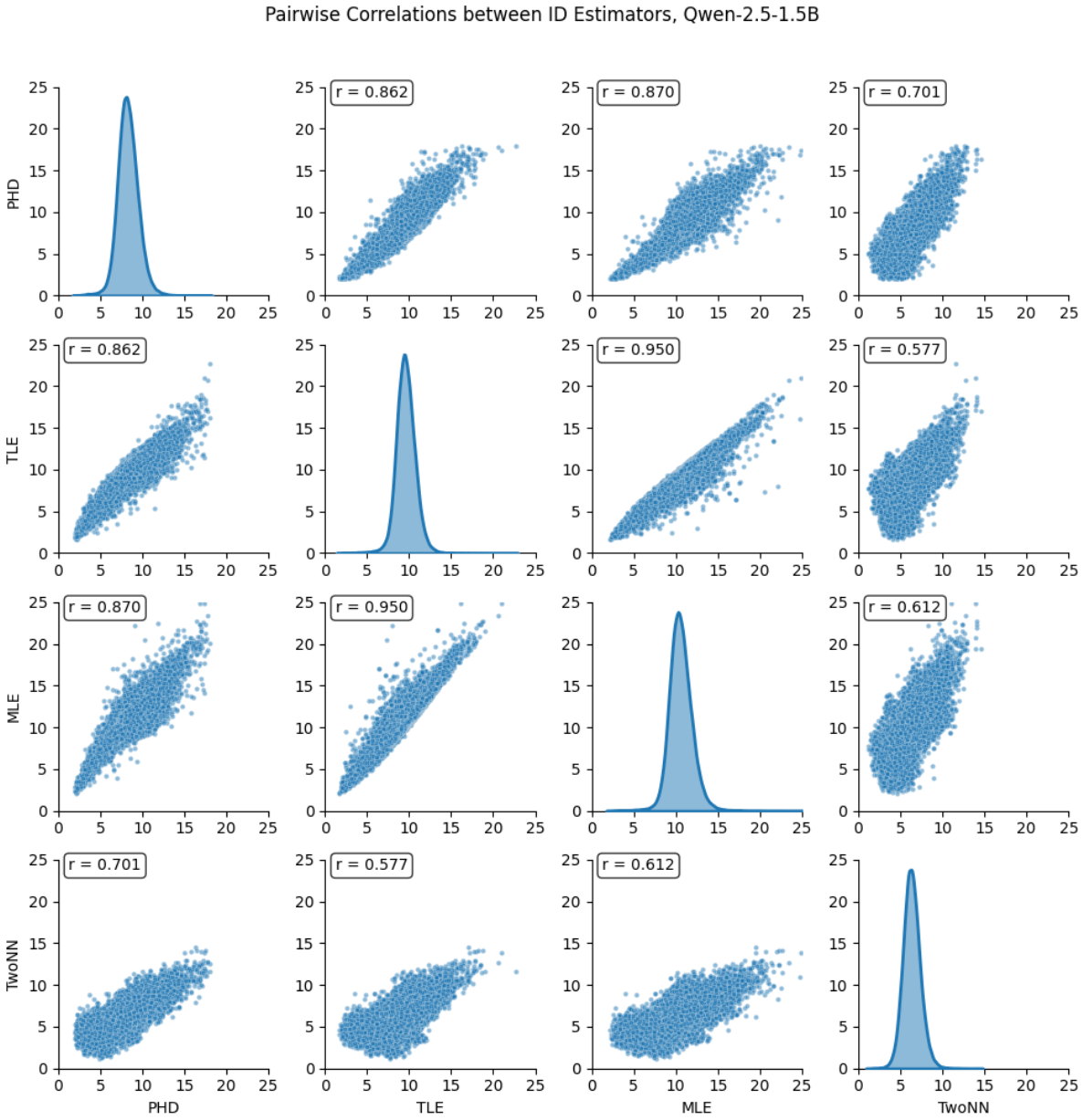

Figure 9: ID Estimators Qwen

Length of X is the length of the mean directional vector $R(X) = ||\frac{1}{N}\sum_{i=1}^{N} x_i'||_2$. For any text, $0 \leq R(X) \leq 1$, where $R = 0$ implies perfect isotropy and $R = 1$ implies that all tokens are perfectly aligned and lay on the same line.

- **Schatten-*p* Norm** (Bhatia, 1997) quantifies the global spectral energy in the matrix X. It is defined as $||X||_{S_p} = (\sum_{i=1}^{r} \sigma_i^p)^{1/p}$. In this section we study only the case of $p = 2$.

- **Effective Rank** (Roy and Vetterli, 2007) estimates the *effective dimensionality* of the embedding space. It is proposed as an entropy-based continuous approximation of matrix rank that is robust to minor perturbations in the data. Let $p_k = \frac{\sigma_k}{\sum_{i=1}^{r} \sigma_i}$ then Effective Rank of X is calculated as $ERank(X) = \exp\left(-\sum_{i=1}^{r} p_k \log p_k\right)$

We calculate the Pearson correlation coefficient between these metrics and 4 intrinsic dimensionality estimators on human-written text data from COLING dataset. Figure 12 presents the results for embeddings obtained from RoBERTa-base, Gemma2-2B and Qwen2.5-1.5B models. We can see that different estimators of intrinsic dimensionality (PHD, TwoNN, MLE, and TLE) are highly correlated with each other; for Gemma2 and

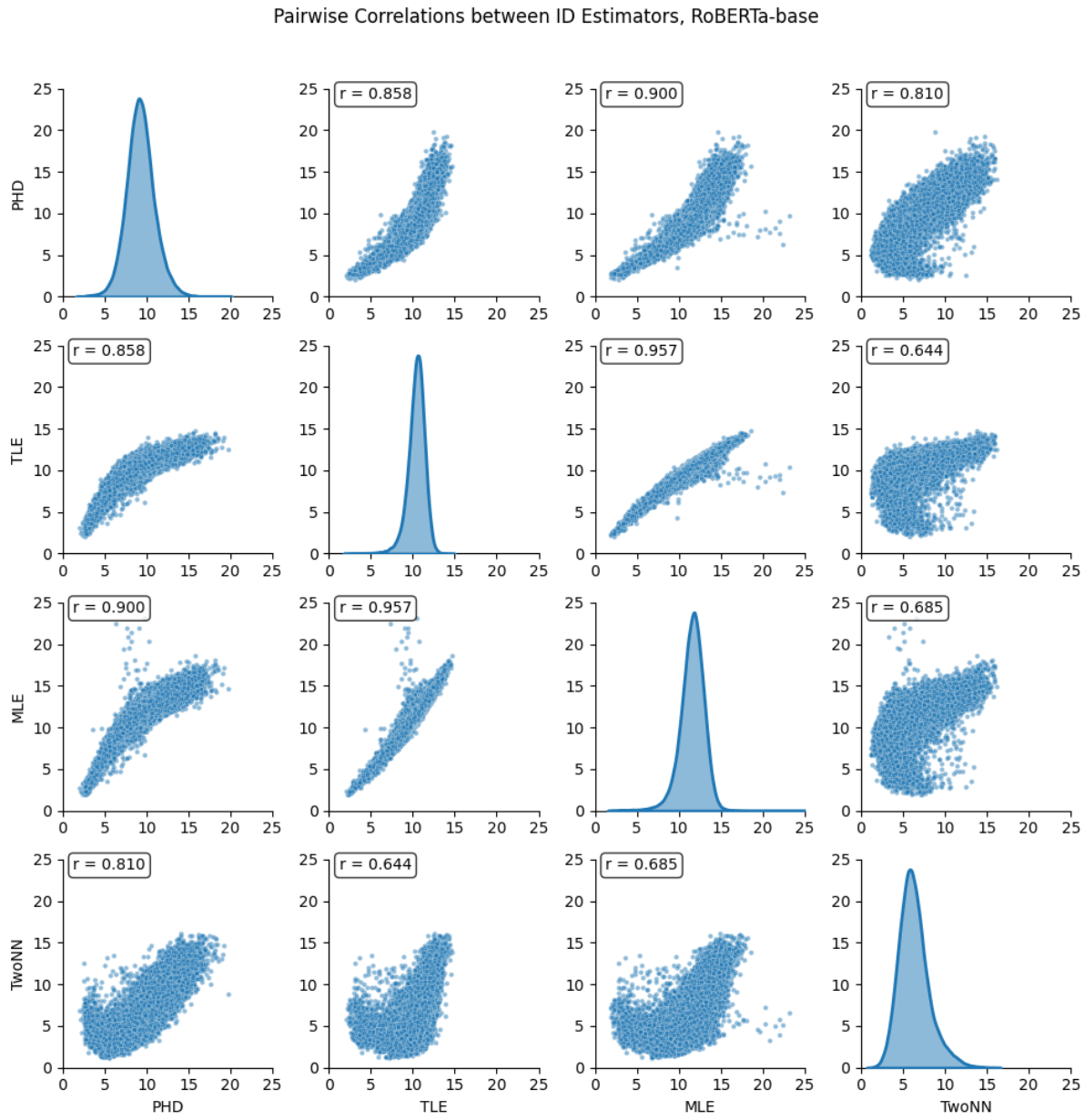

Figure 10: ID Estimators RoBERTa

Qwen2.5 models, there is also strong negative correlation between them and maximal explained variance. The overall picture for RoBERTa is slightly different from other two models, likely because it is an encoder model while Gemma and Qwen are decoders.

From plots on the right half of the Figure 12 we can see that the negative correlation between intrinsic dimensionalities and cumulative explained variance grows with the number of components up to a certain point and then start to decrease. In all cases, the weakest (i.e., closest to zero) correlation is for the TwoNN estimator, which is to be expected because it relies almost entirely on the local structure of the data while principal components are computed for the whole dataset.

C Linguistic, syntax and other text properties

C.1 PHD dependence on text length (tokens)

Fig. 13 shows the standard deviation of PHD as a function of human text length. The variance is high for short texts and stabilizes beyond about 150 tokens. In addition, Fig. 14 shows a large spread of PHD(Gemma) values for short texts.

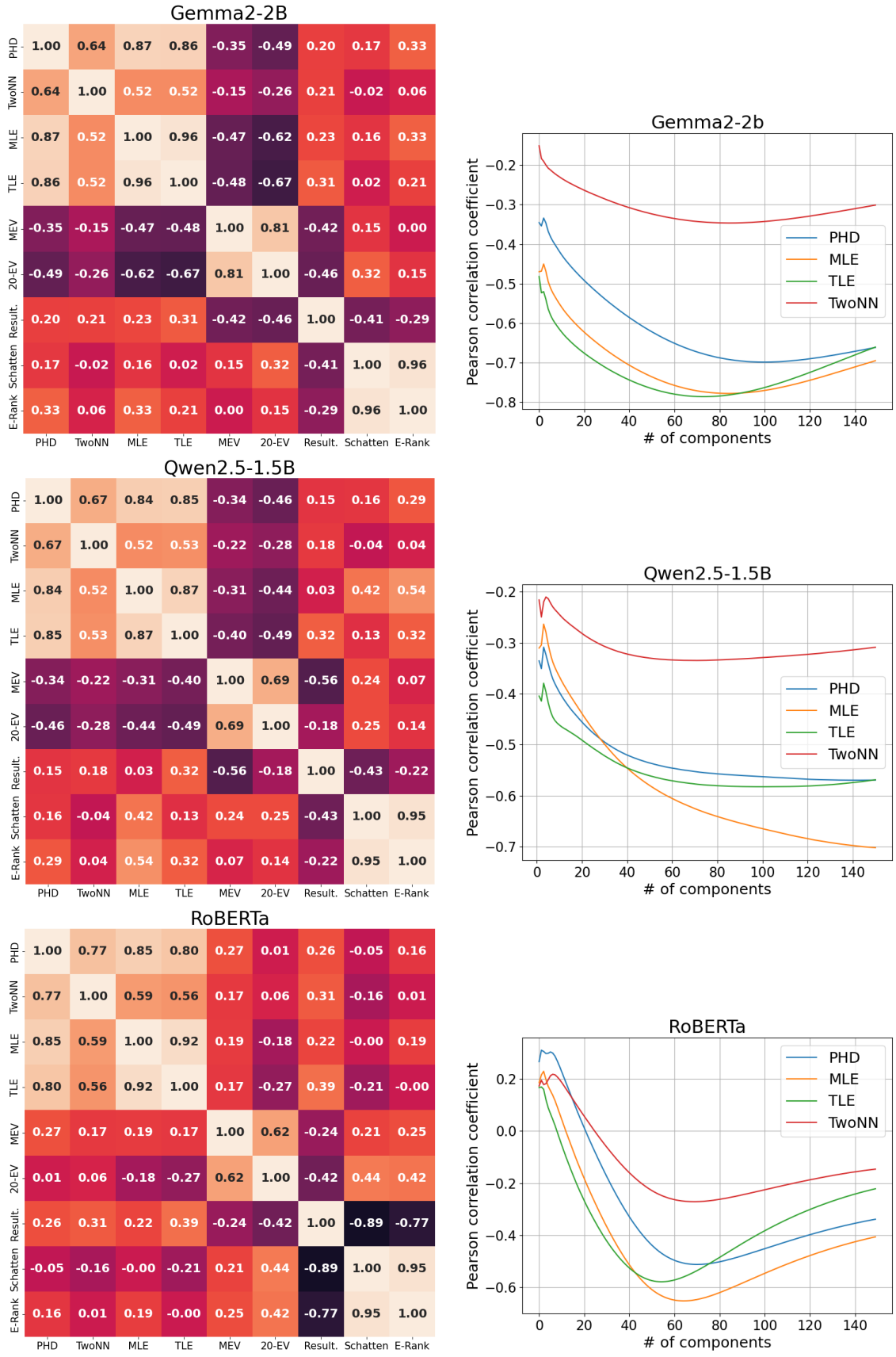

Figure 12: Correlation between Intrinsic Dimensionality (ID) and distribution metrics for human text embeddings (left), correlation between ID and cumulative explained variance (right).

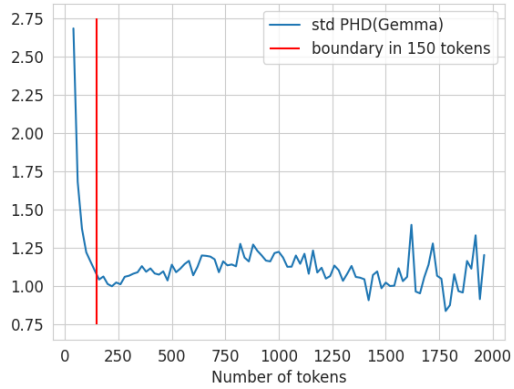

Figure 13: Standard deviation of PHD (Gemma) in 20-token bins.

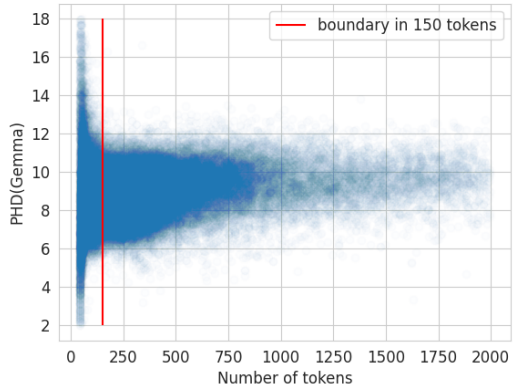

Figure 14: Scatter plot of Text length in tokens and PHD(Gemma)

C.2 Syntatic diversity

We observe that the PHD metric captures little of the syntactic diversity present in texts as in Fig. 15. Correlation between PHD and syntax diversity and correlation between PHD and POS compression ratio is small and differs among generation LLM models as depicted by Table 6.

In addition we see that correlation for different sub sources of text: such as arxiv, imbd is diffenrent in sign, but small as it is below forty percent as shown by Table 4.

This is because PHD is largely insensitive to syntax in high-quality texts. Any apparent correlation with syntactic features is likely driven by noisy or low-quality outputs rather than genuine structural variation. Moreover, domain differences do not significantly affect the relationship between PHD and syntactic diversity or POS compression ratio for good texts. The primary drivers of PHD are, in fact, lexical and thematic diversity, not syntax.

Model	CR_pos	Syntax Diversity
peerread	0.235	0.247
cmv	0.196	0.185
outfox	-0.055	-0.016
eli5	0.187	0.146
wikipedia	-0.086	0.122
wp	0.167	0.240
reddit	0.063	0.129
hswag	0.317	0.267
reddit_eli5	0.044	0.147
finance	0.155	0.197
xsum	0.050	0.011
sci_gen	0.002	-0.054
wikihow	-0.060	0.124
arxiv	0.036	0.089
medicine	-0.037	-0.034
tldr	0.099	0.043
yelp	-0.004	-0.074
roct	0.223	0.038
squad	0.139	0.091
cnn	0.239	0.355
open_qa	0.162	0.025
imdb	-0.200	-0.164
wiki_csai	0.004	-0.012
dialogsum	-0.223	0.291
pubmed	-0.016	-0.055

Table 4: Pearson correlation of CR_pos and Syntax Diversity with phd dimension by source of text

In terms of syntactic diversity, human texts are, on average, the most diverse among current families of LLM-generated texts (Fig. 16), making syntax another factor that should be considered in detecting AI-generated content.

Part-of-speech (POS) tagging compression ratios are relatively high for human-written texts (Fig. 17), though not the highest among all model groups. However, human texts exhibit notably low variance in this metric, indicating consistent syntactic structure across samples.

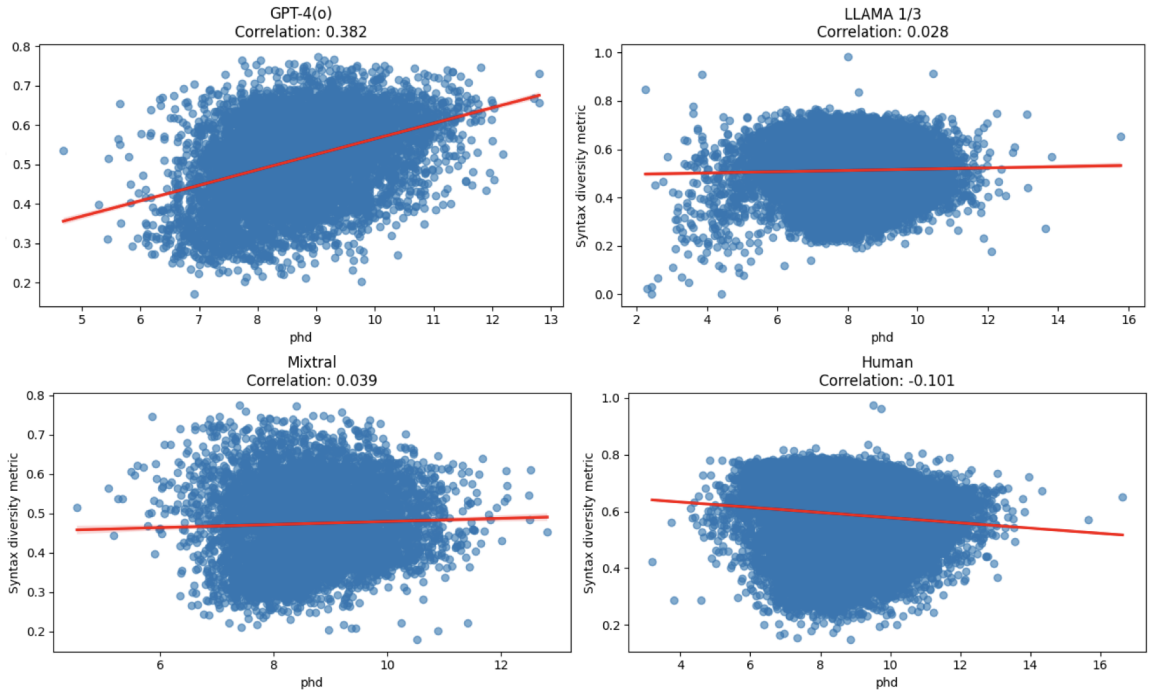

Figure 15: Correlation between syntactic diversity and PHD metric

Table 5: Model Group Performance Metrics

Model Group	CR_pos	Syntax_Diversity
GPT-4(o)	-0.206	0.382
LLAMA 1/3	0.201	0.028
Cohere	-0.042	0.190
GPT-3/3.5	-0.044	0.185
Mixtral	-0.015	0.039
Human	0.030	-0.101
Bloom/BloomZ	0.074	0.061
Gemma	0.016	0.124
OPT	-0.135	0.170
GPT-J/NeoX	-0.320	0.062
Dolly	0.195	-0.030
GLM	0.603	0.414
Flan-T5/T0	0.744	0.297

Table 6: Pearson correlation of OS compression ratio (CR_pos) and Syntax Diversity with phd dimension

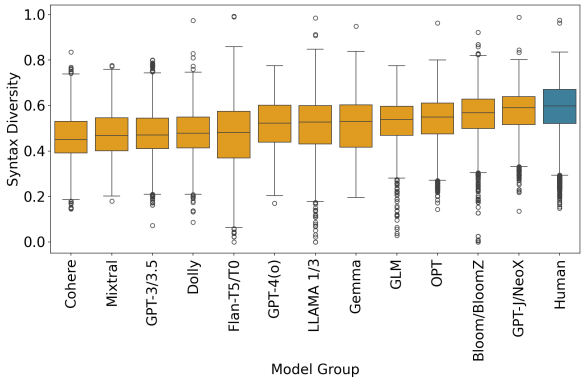

Figure 16: Syntax diversity across text models (human-written and generated by different models)

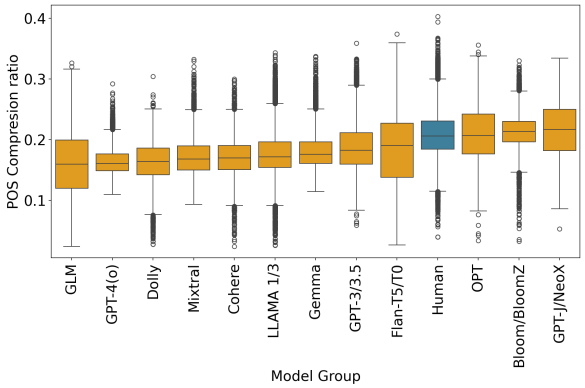

Figure 17: POS compression ratio across text models (human-written and generated by different models)

Moreover,we categorized the texts by source into three distinct groups based on their PHD values: Low-dimensional, Medium-dimensional, and High-dimensional groups.

The Low-dimensional group, which we designate as "Science & Tech", comprises sources such as PubMed, arXiv, Medicine, SciGen, and Wiki -CSAI. This category represents technical and scientific literature from various specialized fields, including medical research and scientific publications.

The Medium-dimensional group, labeled "News & Info", encompasses news and informational resources including SQuAD, DialogSum, Wikipedia, TLDR, CNN, PeerRead, and XSum. These sources consist primarily of factual reporting, encyclopedia entries, and summarized information.

The High-dimensional group, categorized as "Opinion & Forum", contains sources such as Outfox, Reddit-ELI5, Reddit, Finance, WikiHow, Yelp, ELI5, IMDb, CMV, and WP. This group represents opinion-based content, including forum discussions, user reviews, crowd-sourced advice, and personal perspectives across various platforms.

Across these three groups the syntatic diversity metric do not differ a lot, as well as POS Compression ratio Fig. 18.

This can be certainly attributed to generation model itself as distribution of syntax diversity across different domains are similar with insignificantly smaller values for "Science&Tech" domain as shown by Fig. 19.

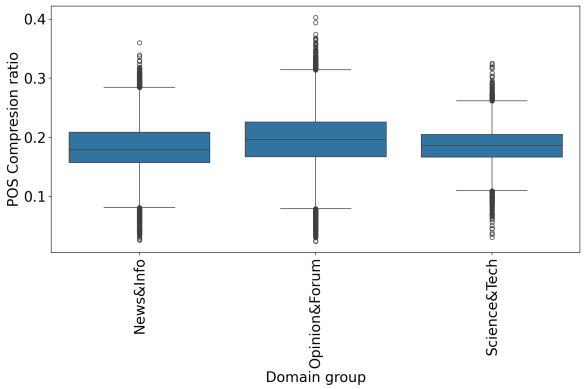

Figure 18: POS Compression ration across text domains

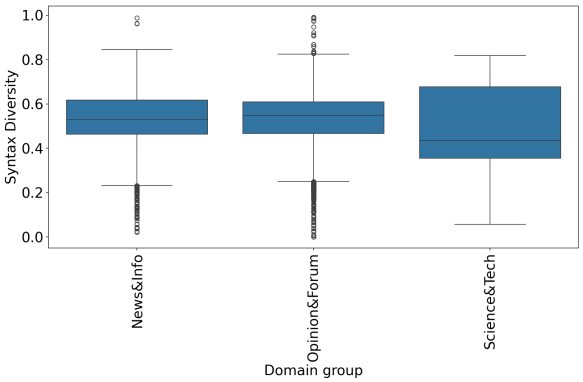

Figure 19: Syntax diversity across text domains

C.3 More details on TAACO features and their relations with ID

These are descriptions of the text metrics from TAACO, used on Figures 4 and 20:

- `function_mattr` (Moving Average Type-Token Ratio): metric calculates the lexical diversity of function words (articles, prepositions, conjunctions, auxiliary verbs, pronouns) using a sliding window approach to reduce text-length sensitivity. Higher values may indicate varied syntactic structures or register shifts, while lower values suggest more formulaic or consistent grammatical patterns.

- `lemma_ttr`, `bigram_lemma_ttr` and `trigram_lemma_ttr`: calculate the type-token ratio of word unigrams, bigrams or trigrams, using their lemmatized forms. Lower values may indicate repetitive words or phrase patterns while higher values suggest more varied lexical combinations.

- `adjacent_overlap_2_cw_sent`: measures the overlap of content words (nouns, verbs, adjectives, adverbs) between adjacent sentences, counting multiple occurrences of the same word. Higher values indicate stronger thematic continuity through repeated emphasis on the same content words, while lower values suggest more varied vocabulary between sentences, potentially indicating topic shifts, diverse expression, or weaker local cohesion. `adjacent_overlap_2_noun_sent` is very similar, except that only nouns are counted.

- `adjacent_overlap_2_all_sent` and `adjacent_overlap_all_sent`: measure overlap of all words between adjacent

sentences. The first metric is counting multiple occurrences of repeated words, while the second one is using binary counting (word presence or absence). High values indicate strong repetition and tight cohesion between consecutive sentences, while low values suggest more varied vocabulary and potentially weaker local connections or topic transitions.

- `repeated_content_lemmas`: measures the proportion of content word lemmas (nouns, verbs, adjectives, and adverbs in their base forms) that appear more than once throughout the text. Higher values indicate greater lexical cohesion through repetition of meaningful words, while lower values indicate greater lexical diversity but potentially weaker thematic unity or cohesion across the text.

- `all_logical`: counts the frequency of logical connectives and discourse markers that express logical relationships in text, including words and phrases like "therefore", "thus", "hence", "consequently", "if-then", "because", and "since". Higher values indicate more explicit logical structuring and argumentative discourse, while lower values may characterize more narrative or descriptive texts with implicit rather than explicit logical connections.

- `sentence_linking`: counts explicit sentence-linking connectives (e.g., "moreover", "furthermore", "however", "in addition"). High values indicate explicit discourse structuring with clear transitions between ideas, typical of formal or academic writing, while low values suggest more implicit connections or informal, narrative styles.

- `lexical_subordinators`: counts subordinating conjunctions that introduce dependent clauses (e.g., "although", "because", "while", "unless"). High values indicate complex syntactic structures with embedded clauses and sophisticated reasoning, while low values suggest simpler, more direct sentence constructions with coordinate rather than subordinate relationships.

- `pronoun_density`: measures the proportion of pronouns relative to total words in the text. High values indicate referential cohesion and informal, interactive discourse or narrative style, while low values suggest more explicit naming, formal register, or introductory text where entities are being established rather than referenced.

Figure 20 and Table 7 show additional correlations between TAACO features and ID estimators upon various models.

C.4 Text transformations and their effect on PHD

We describe three text transformations and examine their effect on the persistent homology dimension across several embedding models.

We consider the following text transformations:

- Type 1: each letter is independently replaced by a homoglyph with probability 0.2;

- Type 2: each letter is replaced by a homoglyph with probability 0.2, but replacements are applied per word type: if two words are identical before the transformation, they remain identical after;

- Type 3: the internal letters of each word are shuffled while keeping the first and last letters fixed; identical words remain identical after the transformation.

Results for the PHD across transformation types and embedding models are given in Tab. 8.

We observe two notable differences between RoBERTa and Gemma/Qwen. First, applying Transformation 3 has little to no effect on the median PHD for RoBERTa, but increases it for Gemma/Qwen. Second, applying Transformations 1 or 2 decreases PHD for RoBERTa, while for Gemma/Qwen these transformations have no effect or slightly increase it.

This experiment shows that there are differences between RoBERTa and Gemma/Qwen, which are confirmed by Fig 2.

C.5 Distribution of distances to the nearest neighbor for tokens

To understand why the intrinsic dimensionality of a text positively correlates with its intrinsic dimensionality, we decided to examine the distance between nearest neighbors. To do this, we divided tokens into two types: those with the same nearest neighbor as the token itself (Type I) and those without (Type II). It turned out that the nearest

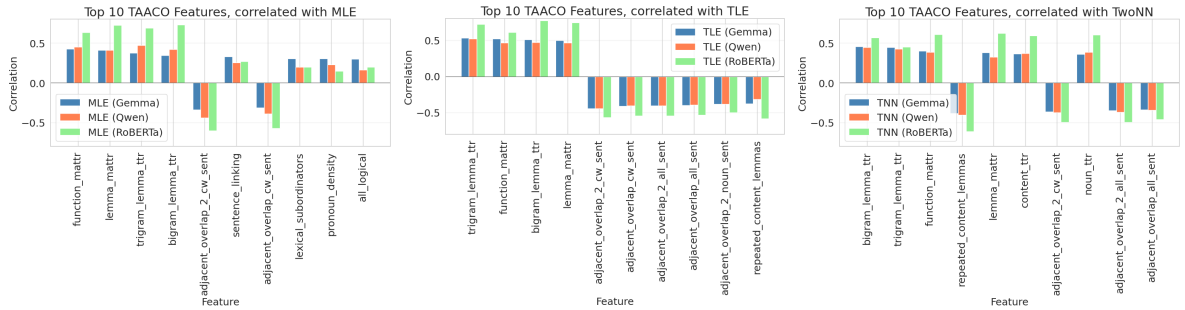

Figure 20: Top-10 features from TAACO with the strongest correlation with MLE, TLE, TwoNN, calculated upon Gemma embeddings, correspondingly. Correlations with ID, calculated upon other embeddings are given as well

TAACO Feature	**PHD** (Gemma)	**PHD** (Qwen)	**PHD** (RoBERTa)	**TAACO Feature**	**PHD** (Gemma)	**PHD** (Qwen)	**PHD** (RoBERTa)
bigram_lemma_ttr	0.446	0.530	0.684	adjacent_overlap_argument_sent	-0.227	-0.306	-0.447
function_mattr	0.493	0.503	0.659	adjacent_overlap_cw_sent_div_seg	-0.270	-0.328	-0.370
lemma_mattr	0.469	0.479	0.706	adjacent_overlap_2_verb_sent	-0.233	-0.297	-0.363
trigram_lemma_ttr	0.452	0.534	0.609	adjacent_overlap_binary_noun_sent	-0.241	-0.306	-0.337
adjacent_overlap_2_cw_sent	-0.404	-0.486	-0.600	adjacent_overlap_binary_2_noun_sent	-0.239	-0.316	-0.327
adjacent_overlap_2_all_sent	-0.368	-0.466	-0.576	adjacent_overlap_adj_sent	-0.247	-0.279	-0.341
noun_ttr	0.339	0.433	0.636	sentence_linking	0.320	0.254	0.285
repeated_content_lemmas	-0.357	-0.426	-0.613	adjacent_overlap_2_argument_sent_div_seg	-0.227	-0.295	-0.336
content_ttr	0.330	0.416	0.629	adjacent_overlap_argument_sent_div_seg	-0.213	-0.272	-0.331
adjacent_overlap_cw_sent	-0.370	-0.434	-0.554	function_ttr	0.153	0.217	0.446
negative_logical	0.279	0.235	0.295	adjacent_overlap_verb_sent	-0.219	-0.262	-0.333
adjacent_overlap_all_sent	-0.358	-0.438	-0.545	all_temporal	0.274	0.253	0.272
adjacent_overlap_2_noun_sent	-0.344	-0.436	-0.552	opposition	0.279	0.232	0.287
repeated_content_and_pronoun_lemmas	-0.274	-0.365	-0.565	determiners	-0.216	-0.238	-0.342
adjacent_overlap_noun_sent	-0.309	-0.387	-0.508	adjacent_overlap_binary_cw_sent	-0.199	-0.277	-0.284
adjacent_overlap_2_fw_sent	-0.310	-0.386	-0.437	adjacent_overlap_all_sent_div_seg	-0.195	-0.274	-0.283
adj_ttr	0.284	0.360	0.481	adjacent_overlap_2_adj_sent_div_seg	-0.223	-0.263	-0.264
adjacent_overlap_fw_sent	-0.316	-0.379	-0.425	adjacent_overlap_binary_2_adj_sent	-0.214	-0.263	-0.258
lemma_ttr	0.226	0.314	0.565	verb_ttr	0.127	0.216	0.375
adjacent_overlap_2_argument_sent	-0.257	-0.354	-0.493	adjacent_overlap_2_all_sent_div_seg	-0.173	-0.264	-0.259
adjacent_overlap_2_noun_sent_div_seg	-0.298	-0.351	-0.382	adjacent_overlap_adj_sent_div_seg	-0.208	-0.238	-0.249
adjacent_overlap_2_cw_sent_div_seg	-0.280	-0.352	-0.376	adjacent_overlap_binary_adj_sent	-0.202	-0.239	-0.247
argument_ttr	0.182	0.284	0.522	adjacent_overlap_binary_2_cw_sent	-0.171	-0.267	-0.249
adjacent_overlap_noun_sent_div_seg	-0.282	-0.328	-0.375	all_negative	0.259	0.191	0.217
adjacent_overlap_2_adj_sent	-0.276	-0.324	-0.382	pronoun_density	0.276	0.202	0.183

Table 7: Top-50 TAACO features: correlations with PHD across embedding models

	Gemma	Qwen	Roberta
PHD	8.69	8.11	9.11
PHD Type 1	9.47	9.83	7.91
PHD Type 2	8.90	8.87	6.54
PHD Type 3	11.17	10.25	9.14

Table 8: Impact of text transformations on PHD

neighbor distances for the Type I tokens are, on average, smaller than those for the Type II, see Fig. 22. This indirectly confirms the positive correlation above. A decrease in the average nearest neighbor distance leads to a decrease in the MST, which in turn may affect its asymptotic behavior.

C.6 PHD and proficiency level of text

Figure 25 shows the distributions of PHD, `lemma_mattr`, and `repeated_content_lemmas` for texts written for readers of different proficiency levels. We observe that the PHD of texts written for beginners is lower than that of texts written for intermediate and advanced readers. This can be explained by the lower type–token ratio and higher lemma repetition across consecutive sentences in beginner-level texts.

C.7 PHD and generation temperature

We generated 10,000 texts with Qwen-3-8B-base and another 10,000 with Qwen-3-8B-instruct, using 1,000 prompts randomly sampled from the RAID dataset. For each prompt, we generated texts at 10 different temperatures (ranging from 0.2 to 2) and measured their PHD using embeddings from Gemma. The results are shown in the figure.

We observe that the PHD of texts generated by Qwen-3-8B-base increases rapidly as the temperature rises from 0.2 to 0.8, then gradually levels off. In contrast, for Qwen-3-8B-instruct, PHD grows more slowly and almost linearly with tem-

perature. We hypothesize that this difference arises because Qwen-3-8B-base tends to produce many repetitions at low temperatures, while Qwen-3-8B-instruct is less prone to such artifacts and maintains consistent generation quality across all temperatures. The lower panels of the figure support this: low-temperature generations from the base model exhibit very low lexical diversity (measured via `lemma_mattr`) and high repetition rates (measured via `repetitive_lemmas`), whereas the instruct model does not show this issue.

Figure 23 shows an example of an anomalous text produced by Qwen-3-8B-base and a well-formed text produced by Qwen-3-8B-instruct using the same temperature and prompt. We observe that the base version of Qwen failed to follow the prompt properly and generated many repetitions.

These observations show that anomalies in PHD distribution may serve as indicators of low-quality text generation.

D SAE-based experiments

D.1 Linear approximation for PHD

We built a simple interpretable model using SAE features, see Tab. 1 for details, and lexical diversity metrics. We trained the model on human texts longer than 150 tokens to predict the PHD(Gemma). We calculated the Pearson and Spearman correlation coefficients for the model's test predictions for various internal dimensionality estimates; see Tab. 9 for details.

Also, to obtain information about the importance of features, we built linear regressions on the same features, each of which was trained to predict its own estimate of the internal dimension, calculated its importance for each feature, and calculated the rank in such a way that the lower the rank, the more important the feature. Then, for each feature, we took the median of its ranks and got the following Tab. 9. Here MTLD (McCarthy, 2005) and Summer (Somers, 1966) are metrics of lexical diversity which are calculated using the Python library *LexicalRichness* (Shen, 2022).

	pearson	spearman
PHD(RoBERTa)	0.597	0.643
MLE(RoBERTa)	0.548	0.587
TwoNN(RoBERTa)	0.512	0.556
TLE(RoBERTa)	0.416	0.440
PHD(Gemma)	0.684	0.702
MLE(Gemma)	0.744	0.771
TwoNN(Gemma)	0.449	0.450
TLE(Gemma)	0.709	0.749
PHD(Qwen)	0.593	0.589
MLE(Qwen)	0.680	0.682
TwoNN(Qwen)	0.430	0.402
TLE(Qwen)	0.706	0.710

Table 9: Pearson and Spearman correlations for linear model predictions on a test for various estimates of the internal dimension.

D.2 Top SAE features, correlated with PHD upon RoBERTa and Qwen embeddings

Table 1 in the main text reports the SAE features most strongly correlated with PHD, computed upon Gemma embeddings. Here, Tables 11 and 12 provide brief overview of the SAE features most strongly correlated with PHD calculated upon RoBERTa and Qwen embeddings, respectively. These tables show substantial overlap with the features in Table 1. The features that appear only in Tables 11 and 12 are also semantically close to those in Table 1. Overall, this suggests that our interpretations are not strongly dependent on the choice of embedding model.

D.3 Steering examples

Table 13 demonstrates the steering effect of the selected features by illustrative examples.

D.4 Causality experiment details

Table 14 shows full results of additional small-scale causality experiment.

Layer	Feature	Median rank
-	Summer	1
-	MTLD	2
16	1693	3
16	9868	3.5
25	14085	6
16	5228	7
16	14885	8
16	8104	9
16	2409	9

Table 10: Median rank of importance of each feature for the linear approximation model PHD

Feature	r	Extreme values
16-941	-0.27	High levels correlate with a formal, authoritative, and objective style. As the feature weakens, the style becomes more informal, personal, and subjective
16-6408	-0.24	At the "strong" end, it enforces strict grammatical rules, formal punctuation, and consistent formatting. As it weakens, it allows for more "naturalistic" text, including common errors, colloquial structures, and the kind of abbreviations and shortcuts seen in unedited digital communication.
16-12844	-0.23	Strong presence results in a formal, objective, and authoritative tone suitable for technical documentation. A moderate or weak presence allows for a more accessible or even subjective and inspirational tone.
24-10164	-0.19	High value produces a formal, encyclopedic, or technical style suitable for manuals, detailed reports, and reference materials. Low value aligns with the concise, direct style of news briefs, social media posts, headlines, and simple answers.
25-9044	-0.18	A high presence of formal data creates a technical, academic, or encyclopedic style. The absence of this feature is characteristic of informal, conversational, literary, or persuasive styles.
16-9868	+0.29	–
16-15275	+0.28	–
16-12453	+0.26	A high value corresponds to texts about disputes, controversies, violence, or problems. A low value corresponds to descriptive encyclopedia-style entries, neutral instructions, or simple factual statements.
24-10369	+0.30	High values yield a conversational, personal, and elaborative style, suitable for forums, tutorials, or opinion pieces. Low values generate a formal, objective, and dense style appropriate for academic abstracts, news snippets, and reference materials.
25-14085	+0.29	–

Table 11: SAE features, most correlated with PHD calculated upon RoBERTa embeddings. "–" marks features that already appeared in Table 1.

E Speeding up the running time of the PHD calculation.

The most expensive part of the PHD calculation is a calculation of pairwise distances. For long texts with a length of more than 1024 tokens, this may take nearly a week on Intel(R) Core(TM) i7-14700K processor for full COLING-dev dataset. Therefore, we rewrote the code for calculating the matrix of pairwise distances with CPU to GPU, which significantly accelerated the work with long texts. For shorter texts, this gives a slight speedup, which is associated with a long transfer of data between the CPU and GPU. The code is located here [1].

F Computational resources

For our experiments we used servers with the following 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

G Artifact licenses

- COLING dataset - The MIT License,
- RAID dataset - The MIT license,
- TAACO - Creative Commons Attribution-NonCommercial-Sharealike license (4.0),
- OneStopEnglish - CC by-SA 4.0

[1] https://github.com/candelabrum/TopAnDat

Feature	r	Extreme values
16-2433	-0.21	–
16-2257	-0.21	High values: text demonstrates the highest level of formality: an objective, impersonal tone, complex sentence structures, and highly specialized, technical vocabulary. Low values: text is almost entirely informal, using subjective, personal, and often emotional language, simple sentence structures, and colloquialisms.
16-8193	-0.20	High values: the texts are exclusively academic or scientific abstracts. Low values: the texts are simple, often short, and conversational. They consist of anecdotes, brief news blurbs, simple definitions, personal reviews, and basic instructions.
24-978	-0.20	This feature controls the nature and density of information. At its strongest level, it generates text focused on novel, complex research findings presented with specialized terminology. As the feature weakens, the content shifts from reporting new knowledge to explaining existing concepts, then to providing general descriptions, and finally to conveying personal opinions or simple anecdotes.
25-2409	-0.20	High values: a formal, objective, and impersonal style, mandating the use of a specialized vocabulary and complex syntax typical of scientific papers. Low values: allows for a transition to semi-formal or informal styles, including encyclopedic, didactic, or evaluative tones.
16-1693	+0.38	–
16-10480	+0.35	High values: a formal, authoritative, and analytical style. Low values: the style becomes more descriptive, then more direct and conversational, and finally, very simplistic and elementary.
16-15275	+0.35	–
24-10369	+0.36	High values yield a conversational, personal, and elaborative style, suitable for forums, tutorials, or opinion pieces. Low values generate a formal, objective, and dense style appropriate for academic abstracts, news snippets, and reference materials.
25-8591	+0.34	A strong presence creates instructional texts (how-to guides), biographies, and narrative reports. A weak or absent presence is characteristic of reference texts, factual summaries, reviews focused on static qualities, and descriptive lists.

Table 12: SAE features, most correlated with PHD calculated upon Qwen embeddings. "–" marks features that already appeared in Table 1.

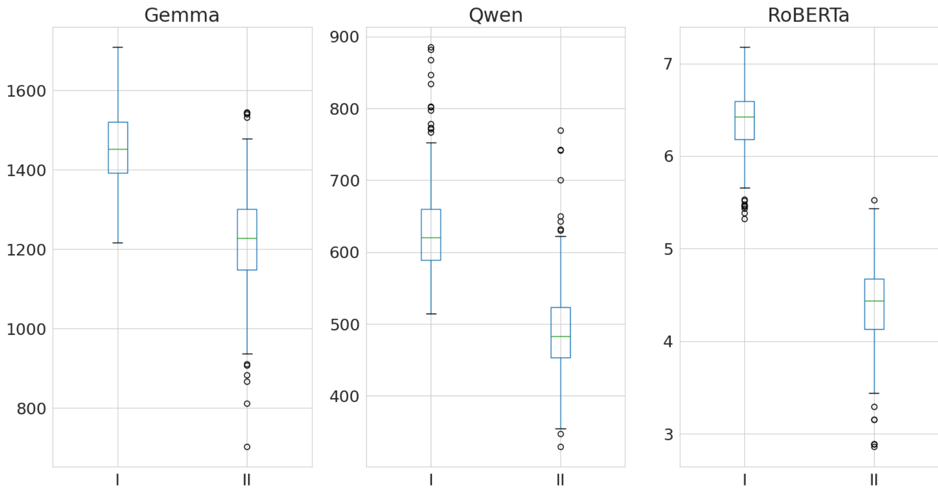

Figure 22: Boxplots of nearest neighbor distances for different types of token

Prompt: `Write the body of a plot summary for a novel titled "The Drawing of the Dark". Do not give it a title.`

Qwen-3-8B-base

```
Do not write in the first person. Do not use contractions. Do not use dialogue. Do not use italics. Do not use bold. Do not use hyperlinks. Do not use lists. Do not use tables. Do not use footnotes. Do not use endnotes. Do not use references. Do not use citations. Do not use bibliographies. Do not use indexes. Do not use glossaries. Do not use appendices. Do not use prefaces. Do not use forewords. Do not use introductions. Do not use conclusions. Do not use summaries. Do not use abstracts. Do not use acknowledgments. Do not use dedications. Do not use epigraphs. Do not use prologues. Do not use epilogues. Do not use interludes. Do not use intermissions. Do not use transitions. Do not use flashbacks. Do not use flashforwards. Do not use dream sequences. Do not use fantasy sequences. Do not use hallucinations. Do not use illusions. Do not use visions. Do not use prophecies. Do not use omens. Do not use portents. Do not use auguries. Do not use auguries. Do not use auguries. Do not use auguries. Do not use auguries. Do not use auguries. Do not use auguries. Do not use auguries. Do not use auguries. Do not use auguries. Do not use auguries. Do not use auguries. Do not use auguries. Do not use auguries. Do not use auguries. Do not use auguries. Do not use auguries. Do not use auguries. Do not use auguries. Do not use auguries. Do not use auguries. Do not use auguries. Do not use auguries. Do not use auguries.

...
```

Qwen-3-8B-instruct

```
In a world where the line between reality and illusion is increasingly blurred, a young art student named Elara discovers a mysterious sketchbook hidden in the attic of her late grandmother's house. The pages are filled with intricate drawings that seem to shift and change when viewed from different angles, revealing hidden symbols and cryptic messages. As Elara begins to study the drawings, she notices that they begin to manifest in the real world—objects appearing in her room, shadows moving on their own, and whispers echoing through the halls of her home.
Drawn deeper into the mystery, Elara learns that the sketchbook is a relic from an ancient order of artists known as the Keepers of the Veil, who once protected the world from a force known as the Dark. This force, a sentient void that feeds on fear and uncertainty, was sealed away centuries ago by the Keepers, who used their art to bind it in a place called the Hollow. But the seal is weakening, and the drawings are not just a record of the past—they are a warning.
As Elara delves further into the history of the Keepers, she uncovers a prophecy that speaks of a chosen artist who will either restore the balance or unleash the Dark upon the world. With the help of a reclusive scholar named Kael, who has spent his life studying the Keepers' lost works, Elara begins to understand the true nature of the drawings and the power they hold. She learns that each drawing is a fragment of a larger, unfinished masterpiece that can either reinforce the seal or break it entirely.

...
```

Figure 23: Comparison of texts generated by Qwen-3-8B-base and Qwen-3-8B-instruct with the same prompt and temperature (0.2).

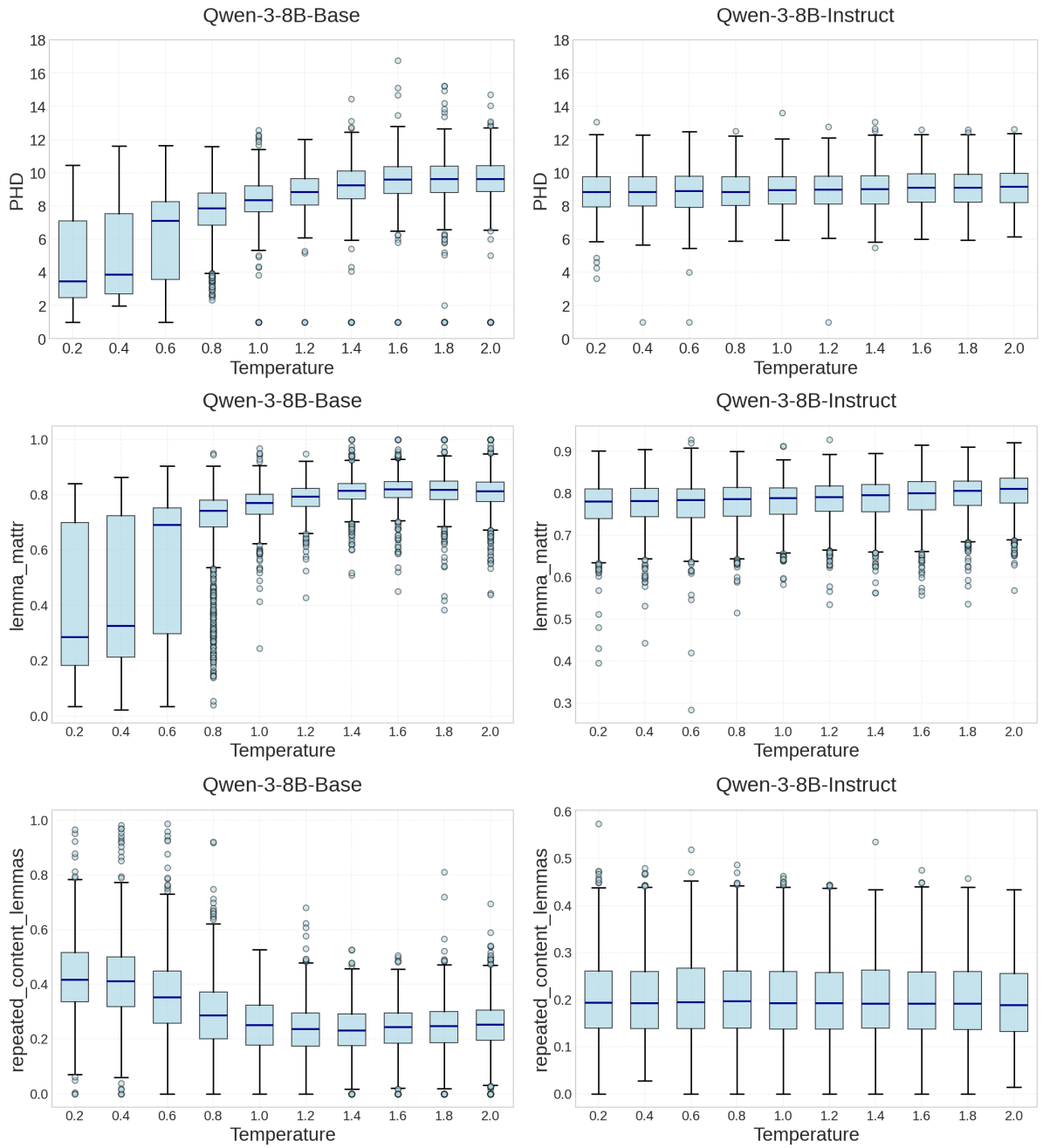

Figure 24: PHD, `lemma_mattr` and `repeated_content_lemmas` as a functions of generation temperature

Feature	Sign	Shift	Example change (steered)
Default generation (excerpt): *The internet is awash with advice for anyone struggling to return to education after a long break. … The first step is to acknowledge the gap in your education and accept that it's okay to feel nervous. Local colleges often offer adult courses to help you catch up.*			
24-15879	–	BBC-style templated explainer	"'I'm not sure what I'm doing with my life,' she said — the anecdotal lede introducing a structured explainer: paragraph 1 (anecdote), paragraph 2 (context and stats), paragraph 3 (options for learners), paragraph 4 (call to action)."
16-8104	–	Social-science report	"A UK survey found that 12% of adults had never attended a formal school or college."
24-4610	+	Social-media explainer	"**Feeling lost?** You're not alone. **1. Assess your goals 2. Research options 3. Talk to someone**."
16-6978	+	Informal helper tone	"It's okay — let's take it step by step and outline a plan that fits your life."
Default generation (excerpt): *The awkward silence hung heavy in the air as Sarah and I sat across from each other at the cafe. We were supposed to be discussing her art, but all I could think about was the way her eyes lingered on me.*			
16-2433	–	Outline / topic-sentence scaffolding	"The story focuses on: • context of the situation • comparison with other cases."
16-5159	–	Analytical / detached commentary	"The conversation is analyzed through a framework of social and cultural factors influencing interaction."
16-5228	+	Emotional uncertainty; introspective tone	"I'm trying to move on, but it's hard — I still don't know if she likes me as a friend or more."
15275	+	Character-focused vividness; narrative expansion	"Emily's fingers trembled as she met Sarah's gaze, unsure if the warmth in her smile meant love or fear."
Default generation (excerpt): *Elias Thorne, a brilliant but troubled young programmer, is caught in the middle of this clash of worlds. He's a man of contradictions, haunted by the ghosts of his own making.*			
16-5984	–	Academic analytical tone	"The novel's narrative structure is based on competing architectural firms' design histories within the 1980s urban landscape."
16-14485	–	Expository / research-like narrative	"The narrative emphasizes the city's socio-economic contrasts and their impact on collective memory."
15275	+	Richer plot and personal emotion	"Our protagonist, Alex, a struggling journalist, fights to reclaim his place in a city that forgot his name."
16-5228	+	Psychological depth / uncertainty	"Anya must race against time, yet her own doubts and past failures threaten to destroy her resolve."
16-6978	+	Informal helper tone	"It's okay — let's take it step by step and outline a plan that fits your life."
Default generation (excerpt): *This paper explores the theoretical framework for unconventional superconductivity in strongly correlated systems, focusing on the interplay between real-space pairing and mean-field theory.*			
16-5984	–	Analytical / formalized research tone	"The study reveals the critical role of the fuel supply system, including the pump, lines, and pressure regulator, in performance metrics."
16-5159	–	Repetitive academic phrasing; self-referential	"This paper examines the causes of this problem and provides a diagnostic framework; this paper is intended to be a guide."
16-6978	+	Informal / conversational rewrite	"Here's the abstract — I've kept it short and jargon-free so it's easier to follow."
16-9868	+	Media / popular-science tone	"The emergence of superconductivity in correlated systems has fascinated scientists for decades — and the story is still unfolding."

Table 13: Steering effect of selected features for four different prompts. "Sign" denotes positive or negative correlation of the feature with ID.

Feature	r	Median PHD (Gemma)	Mean PHD (Gemma)	Std PHD (Gemma)	p-value	The level of significance
16-5984	-0.30	8.04	8.00	0.58	8.41×10^{-1}	–
16-14485	-0.36	8.68	8.80	0.57	4.45×10^{-5}	***
16-2433	-0.35	7.89	7.82	0.48	1.88×10^{-1}	–
16-8104	-0.33	7.76	7.77	0.57	1.34×10^{-1}	–
16-5159	-0.34	7.56	7.64	0.57	2.68×10^{-2}	*
24-15879	-0.31	8.78	8.66	1.33	1.77×10^{-2}	*
25-2409	-0.43	8.40	8.30	0.60	1.50×10^{-1}	–
16-15275	0.36	9.49	9.57	0.61	1.44×10^{-12}	***
16-5228	0.33	8.99	8.96	0.49	9.28×10^{-7}	***
16-6978	0.36	8.98	8.91	0.52	3.01×10^{-6}	***
16-9868	0.37	8.87	8.94	0.76	1.05×10^{-5}	***
Original texts	–	8.10	8.04	0.93	–	–

Table 14: Causal relationship between SAE features and text ID

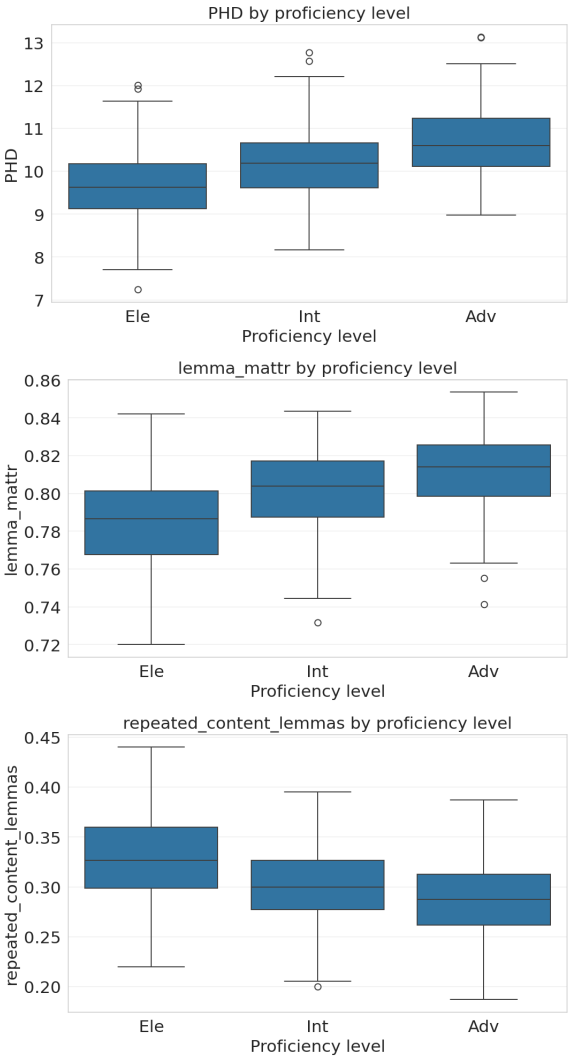

Figure 25: PHD, `lemma_mattr` and `repeated_content_lemmas` for texts of various proficiency levels