Linguistic Profiling of Transformer Embedding Geometry

Lucia Domenichelli[1,2], Dominique Brunato[1], Felice Dell'Orletta[1]
[1]Istituto di Linguistica Computazionale "Antonio Zampolli" (CNR-ILC), ItaliaNLP Lab, Pisa
[2]University of Pisa
{name.surname}@ilc.cnr.it

Abstract

Transformer language models embed tokens in high-dimensional spaces, but whether geometry reflects linguistic structure remains unclear. We analyse token representations in BERT and GPT-2, selected as canonical encoder-only and decoder-only Transformer architectures, through a linguistically-grounded geometric lens. We partition tokens from the Universal Dependencies English Web treebank by surface and syntactic features (position, length, POS, head distance and arity) and examine how their representational geometry evolves across layers. We employ complementary diagnostic metrics, including isotropy, linear and nonlinear intrinsic dimensionality, to capture distinct aspects of embedding structure. Our findings reveal that BERT maintains more isotropic and higher-dimensional subspaces, whereas GPT-2 exhibits stronger anisotropy driven by a compact cluster of sentence-initial tokens. Across models, open-class words, longer tokens, and predicates with several dependents occupy more isotropic, higher-dimensional manifolds than short function words and pre-head modifiers, indicating that semantic richness and syntactic centrality play a key role in structuring embedding space. Our analysis provides a reusable framework for profiling how linguistic abstractions organize the geometry of Transformer embeddings.[1]

1 Introduction

Transformer-based language models have driven major advances in natural language processing, with encoder, decoder, and hybrid architectures enabling strong generalization across a wide range of tasks (Vaswani et al., 2017). Despite their empirical success, a central open question remains: *what linguistic information do these models represent internally, and how is it structured across layers and architectures?*

[1]Code to reproduce our experiments is available on GitHub.

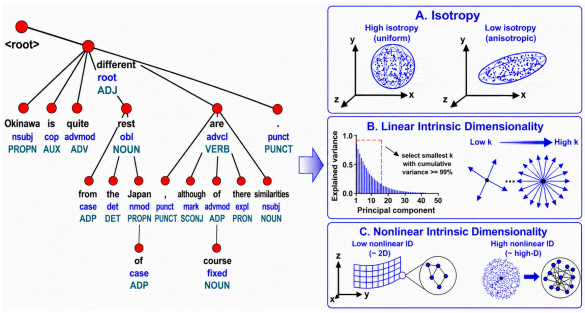


Figure 1: Dependency-parsed sentences are annotated with linguistic features, and token embeddings are grouped by feature class across Transformer layers. Each class-conditioned embedding cloud is characterized using isotropy, linear intrinsic dimensionality, and nonlinear intrinsic dimensionality.

A variety of interpretability paradigms have been proposed to address this question. *Probing* approaches train auxiliary classifiers on hidden representations to assess whether specific linguistic properties are decodable, revealing layerwise progressions from surface to syntactic and semantic information (Tenney et al., 2019). At the same time, concerns about probe expressivity and causal validity have motivated controls based on complexity and information theory (Hewitt and Manning, 2019; Pimentel and Cotterell, 2021). In parallel, *mechanistic interpretability* seeks causal explanations by tracing computations at the level of attention heads, neurons, and circuits (Olsson et al., 2022). However, these paradigms do not directly characterize the global organization of representation spaces themselves.

In this work, we adopt a *geometric* perspective on interpretability that is model- and task-agnostic. This perspective is motivated by the manifold hypothesis (Fefferman et al., 2016), which posits that high-dimensional representations concentrate near lower-dimensional structures whose geometry reflects underlying regularities. Rather than asking whether linguistic information is decodable or

which components implement specific behaviours, we study how token representations are organized in embedding space. From this viewpoint, if linguistic properties meaningfully organize model representations, they should give rise to identifiable geometric signatures in the embedding space.

Recent work has begun to connect geometric properties of representation spaces with linguistic structure. Mamou et al. (2020) show that linguistic categories form progressively more separable and lower-dimensional manifolds across Transformer layers. Hernandez and Andreas (2021) proposed a geometrically informed probing framework showing that several linguistic properties can be recovered from low-dimensional linear subspaces of contextualized representations. More recently, Lee et al. (2025) linked intrinsic geometric properties of language-model representations to compositional structure, showing that linear and nonlinear intrinsic dimensionality capture complementary aspects of linguistic organization. Complementary work has also investigated whether grammatical and structural information can be localized within sentence embeddings through latent bottlenecks and targeted sparsification (Nastase and Merlo, 2024b,a).

Our work builds on this emerging line of research, but differs from probing-based and separability-oriented approaches in an important respect: rather than asking whether linguistic information can be decoded from representations, we directly characterize the geometry of linguistically defined token manifolds. Specifically, we partition contextualized token representations according to surface and morpho-syntactic properties and examine how their geometric organization evolves across layers and architectures.

Methodologically, we adopt a multi-metric perspective on representation geometry. We combine isotropy measures with both linear and nonlinear intrinsic dimensionality estimators, allowing us to capture complementary aspects of embedding organization.

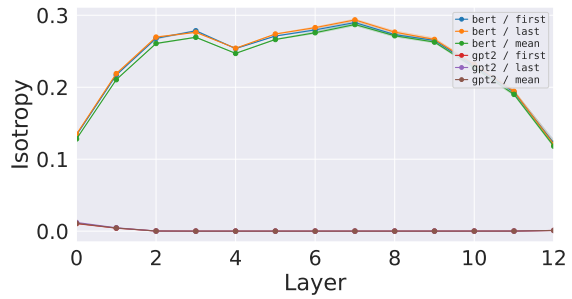
Beyond representation analysis, this framework may also be relevant from a linguistic perspective. Linguistic categories such as part of speech and dependency relations exhibit gradient and cross-linguistically variable boundaries (Haspelmath, 2007). Investigating whether such categories correspond to distinct geometric organization in neural representations provides a useful diagnostic for understanding how distributional learning structures linguistic information in neural language models, complementing ongoing debate on the nature of linguistic knowledge encoded by these models (Piantadosi, 2023; Katzir, 2023).

Contributions This paper makes three contributions. We present a *linguistically grounded geometric analysis* of Transformer representations, showing how token-embedding geometry evolves across layers when conditioned on surface, lexical, and syntactic features. We compare Encoder and Decoder architectures, revealing systematic differences tied to their training objectives and inductive biases. Finally, we show that a *multi-metric* view is required for reliable geometric characterization: combining isotropy measures with linear and nonlinear intrinsic-dimensionality estimators, we find that different metric families capture complementary structure, yielding a reusable diagnostic toolkit for comparing how linguistic abstractions organize embedding space across models and layers.

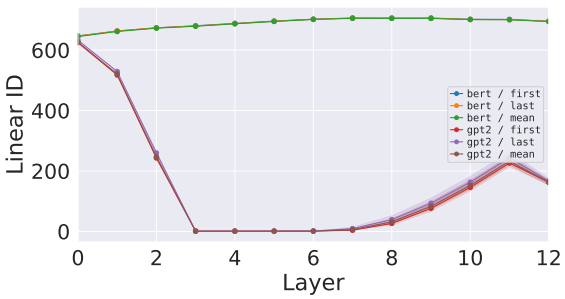
2 Related Work

A recurring finding in the study of Transformer representations is that their geometry is far from uniform. Early analyses showed that contextual vectors are typically *anisotropic*: random tokens have spuriously high cosine similarity and cluster into a "narrow cone" of directions (Ethayarajh, 2019). At the same time, these representations appear to respect the *manifold hypothesis*, i.e. natural data concentrate on low-dimensional structures [2] (Bengio et al., 2013; Fefferman et al., 2016). In this view, *isotropy* captures how evenly representations spread around the manifold. A range of fixes aim to restore uniformity: contrastive objectives such as SimCSE promote a balanced hypersphere (Gao et al., 2021), normalizing flows (BERT-flow) map embeddings toward an isotropic Gaussian (Li et al., 2020), post-hoc whitening improves directional balance (Su et al., 2021). At the lexical level, mean-centering and "all-but-the-top" remove dominant components in static embeddings, motivating analogous procedures for contextual ones (Mu and Viswanath, 2018). Yet uniformity is not an unconditional goal, since over-flattening can wash out meaningful cluster structure (Mickus et al., 2024). A similar view develops a geometric account via *intrinsic dimensionality* (ID), the effective degrees of freedom used by a model or its

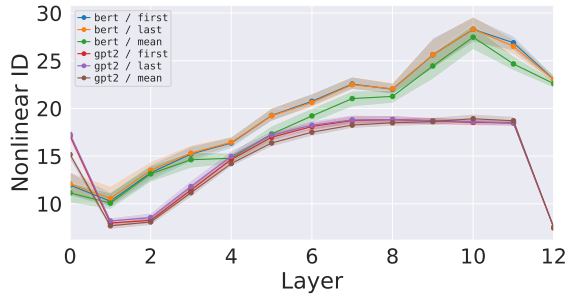
[2] We use *manifold* in a relaxed sense, assuming for simplicity that points lie on measurable geometric objects.



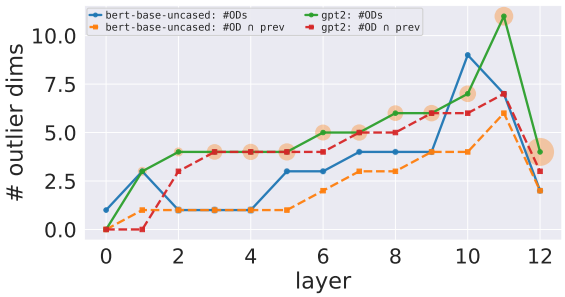
(a) Isotropy.



(b) Linear intrinsic dimensionality.



(c) Nonlinear intrinsic dimensionality.



(d) Outlier dimensions.

Figure 2: Isotropy (a), linear (b) and nonlinear (c) intrinsic dimensionality on Encoder and Decoder hidden states across layers. Subfigure (d) shows outlier dimensions for BERT (blue) and GPT-2 (green) (FIRST and LAST pooling methods respectively), with marker size proportional to the average exceedance $(|v_i|(\ell) - (\mu + 3\sigma))_+$. Dashed line shows the number of overlapping outliers with the previous layer.

representations. In parameter space, early work showed that high-quality solutions often reside in surprisingly low-dimensional subspaces. In NLP, this helps explain why fine-tuning moves within compact intrinsic subspaces and motivates low-rank adaptation (Li et al., 2018; Aghajanyan et al., 2021; Hu et al., 2022). Shifting to representation space, ID traces how information is organized across depth. A recurrent pattern is expansion in early layers followed by compression and stabilization, with Transformers exhibiting a distinct *high-dimensional abstraction* band whose earlier onset correlates with better language-modeling performance (Cheng et al., 2025). Moreover, theory links scaling-law exponents to the intrinsic dimensionality of the data manifold, and empirical analyses show ID rising and then contracting as training saturates (Havrilla and Liao, 2024; Razzhigaev et al., 2024). Beyond description, ID serves as a practical signal. Token-level ID tracks next-token uncertainty, and sequence-level ID modulates memorization in over-parameterized models (higher-ID sequences are less likely to be memorized) (He et al., 2023; Viswanathan et al., 2025; Arnold, 2025). Finally, not all "dimensions" are equal: nonlinear ID captures semantic compositionality that linear proxies can miss, highlighting the value of manifold-aware estimators (Lee et al., 2025).

3 Our Approach

In this study, we focus on *token-level manifolds* defined by linguistically coherent classes that capture surface level, lexical, and syntactic phenomena. In this sense, when we refer to *linguistic manifolds*, we mean the subsets of token representations induced by an *a priori* linguistic partition.

Our analysis has two main goals: first, to trace how the geometry of token representations evolves across layers as a function of linguistic features, and second, to examine how these patterns differ between Encoder and Decoder architectures. We concentrate on BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019) because they are canonical encoder-only and decoder-only architectures. Both models share comparable depth and dimensionality, allowing us to control for scale while isolating architectural and objective-driven differences in representation geometry. Although not state-of-the-art in performance, these models remain foundational to many contemporary language models and provide a clean testbed for studying how architectural and objective differences shape the geometry of token representations. To our knowledge, while prior

work has studied grammatical and structural information in Transformer embeddings, systematic comparisons of word-level contextual representation geometry across encoder-only and decoder-only models using multiple geometric metrics remain limited.

As a testbed, we leverage the Universal Dependencies English Web treebank (UD-EWT) described in Section 3.1. Methodologically, for each sentence of this dataset, we first apply the model's native tokenizer and pass the entire sentence through the model. We then extract, for each token position, its layerwise hidden state, yielding a contextualized representation for every token in context.[3] For each word w at layer ℓ, we construct a word-level vector $\mathbf{v}^{(w)}(\ell) \in \mathbb{R}^D$ by aggregating the hidden states of its constituent tokens, where D denotes the model's hidden-state dimensionality. We consider three pooling strategies: FIRST, which takes the hidden state of the first sub-token, LAST, which takes the hidden state of the last sub-token, and MEAN, which averages the hidden states of all sub-tokens in the word. Unless otherwise specified, we use FIRST pooling for BERT and LAST pooling for GPT-2. Appendix B reports correlations between the MEAN-pooling results and the corresponding FIRST- and LAST-pooling results. Full MEAN-pooling results are available in our repository.

As regards the diagnostic methods, we characterize embedding space using **three geometric metrics**: isotropy, linear intrinsic dimensionality, and nonlinear intrinsic dimensionality. Together, these measures capture distinct aspects of representation structure, namely global directional uniformity and the effective degrees of freedom of token-level manifolds under both linear and nonlinear assumptions. Specifically, on the isotropy side, we adopt *IsoScore*, a global covariance–spectrum–based measure that provides a compact summary of how evenly representations occupy the ambient space (Rudman et al., 2022). For intrinsic dimensionality, we choose a linear estimator, *Global PCA* (PCA_{99}) (Cangelosi and Goriely, 2007), and a nonlinear, neighbourhood-based estimator, *GRIDE* (Denti et al., 2022). These metrics are standard in the representation geometry literature, and in our experiments they are correlated with a broad suite of alternative isotropy and intrinsic dimensionality estimators (Figure 13 in Appendix C.7).

[3] In Hugging Face, `hidden_states` is returned when `output_hidden_states=True` and contains the embedding output plus one tensor for each layer.

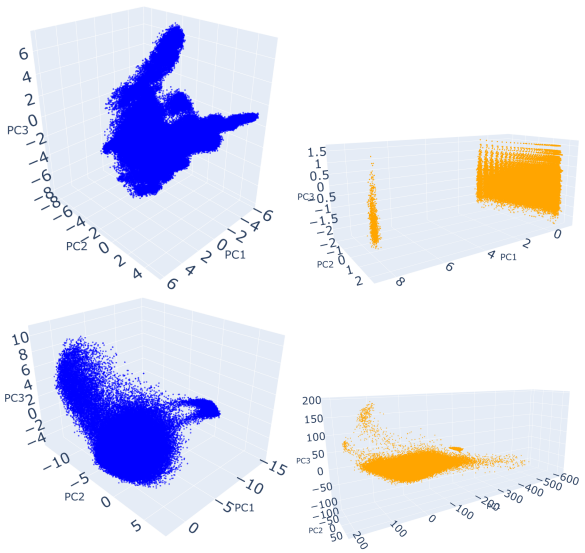


Figure 3: 3D PCA of token representations of layers 1 (top) and 12 (bottom) of BERT (left) and GPT-2 (right). In GPT-2, a cluster is already visible at layer 1 and separates further in the subsequent layers, consistent with the drop in isotropy and linear ID in Figures 2a and 2b. The cluster then largely stabilizes (primarily along PC1) and contracts again by the final layer, similarly to the encoder.

In Section 3.1, we introduce the linguistic features used to partition the data. For each feature f, we divide the word occurrences into a finite set of classes C_f, corresponding either to annotation labels (e.g., POS tags) or to discretized bins (e.g., token position or head distance). We then compute each geometric metric separately for every class $c \in C_f$ and every model layer ℓ, allowing us to trace how representation geometry varies with depth across the classes of each feature. Our linguistic partitions are naturally imbalanced. Rather than artificially balancing the corpus, we preserve these empirical distributions and quantify the resulting finite-sample uncertainty using a within-class bootstrap: for each class, we repeatedly resample tokens with replacement and compute each geometric statistic on the resample, reporting bootstrap means with percentile confidence intervals. Rare classes therefore appear with wider uncertainty intervals, and we interpret such estimates more cautiously than those of high-support classes. More details on bootstrap hyperparameters are in Appendix C.1, and sample-size sensitivity is discussed in C.6.

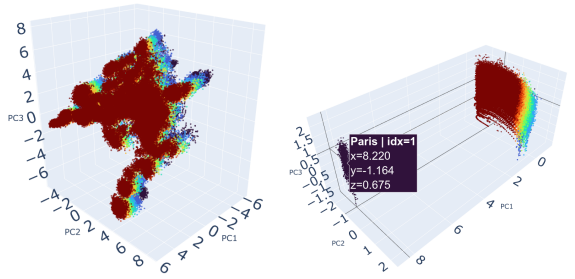


Figure 4: Index clusters in BERT (left) and GPT-2 (right). Because GPT-2 sentence-initial tokens form a compact anisotropic subcluster, we exclude GPT-2 `index=1` tokens from the subsequent feature-conditioned analyses, except for the positional analysis in Figure 5 where the index feature itself is examined.

3.1 Dataset and Linguistic Features

We use the **UD-EWT Treebank** (Silveira et al., 2014), which contains 254,820 words across 16,622 sentences drawn from weblogs, newsgroups, emails, reviews, and Yahoo! Answers. This corpus provides gold standard morphological and syntactic annotations, thus allowing precise alignment between linguistic categories and embedding geometry. We derive five token-level features spanning surface form and syntax. **INDEX** is the token's left-to-right position in the sentence; we restrict this feature to the first ten positions (1-10). **LENGTH** is the number of characters in a word, capped at 10. **POS** is the part-of-speech tag from the Universal POS tagset: the classes used in our experiments are provided in Table 3[4].

HEAD DIST is the signed linear distance (in tokens) from a dependent to its syntactic head in surface order: negative values indicate that the head precedes the dependent, and positive values indicate that it follows. We discretize distances into the range $[-6, 6]$. We additionally report a focused analysis of **nominal head distance**.

Finally, **VERBAL ARITY** is the number of syntactic dependents attached to a verbal head, capped at 6[5]. Results for arity across all POS heads are reported in Appendix D.

Together, these features span complementary levels of linguistic description, from surface to morpho-syntactic and syntactic (both linear and hierarchical) structure. This means we can ask whether a token's geometry is driven by *where* it appears, *what type* of word it is, or *how* it participates in syntax.

4 Results

4.1 Feature-Agnostic Analysis

We first report results of our metrics on the whole dataset without considering the linguistically defined manifolds. Consistent with prior findings on contextual embeddings (Razzhigaev et al., 2024; Skean et al., 2025), in Figure 2 we observe that GPT-2 is slightly more anisotropic than BERT. We hypothesize that GPT-2's causal, unidirectional objective aligns token representations along a few dominant directions predictive of next-token generation, while BERT's masked-LM objective and bidirectional attention mix information more uniformly across directions, yielding greater isotropy. Nevertheless, anisotropy is high in both models, and is linked to strong outlier dimensions in the representation distribution (Ethayarajh, 2019; Hämmerl et al., 2023). We compute the corpus-level mean activation $v_i^{(w)}(\ell)$ and declare a dimension i an outlier at layer ℓ when $v_i^{(w)}(\ell) \geq \mu + 3\sigma$, where μ and σ are the global mean and standard deviation over all layers and dimensions, following (Kovaleva et al., 2021). Figure 2d shows that GPT-2 has more and larger outliers than BERT, and that these persist between adjacent layers. In PCA space this appears as a compact cluster drifting away from the main cloud (Figure 3). A fine-grained analysis (Figure 4) reveals that this cluster contains most of the tokens with `index=1`. In the decoder, some initial tokens form a consistent boundary case, so the model can maintain a dedicated offset direction that encodes sentence initial state information across layers. The cluster then reconnects in the final layer possibly because the network must map all positions into a shared space suitable for the common next token prediction interface, which encourages late alignment of position specific states with the global representation geometry.

Regarding intrinsic dimensionality, Figures 2b and 2c show that the two architectures follow distinct trajectories. In BERT, linear ID starts high, rises slightly toward the early–middle layers, then tapers off. In GPT-2, it drops rapidly in the earliest layers (because of the separate `index=1` subcluster), rebounds at intermediate depths, and then de-

[4]Additional results by typed dependency relation are reported in Figure 14 in Appendix D. These effects largely overlap with POS-based patterns.

[5]Arity is intended as a coarse proxy for the linguistic notion of *valency*, but it simply measures the number of dependents attached to a predicate in the dependency structure and does not distinguish between arguments, adjuncts, or complement types.

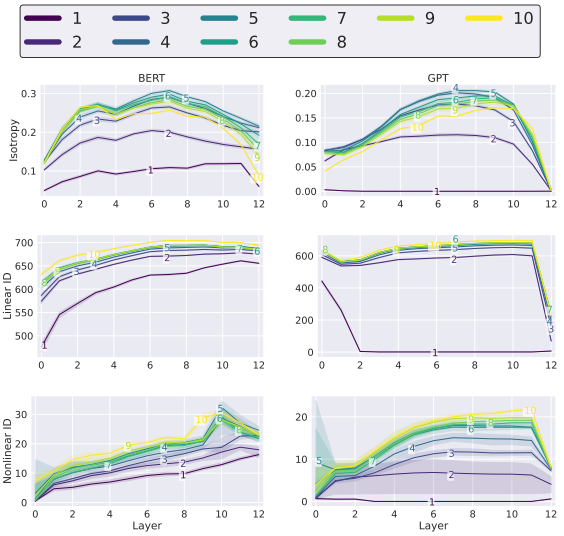


Figure 5: Encoder and Decoder layerwise geometry by token index.

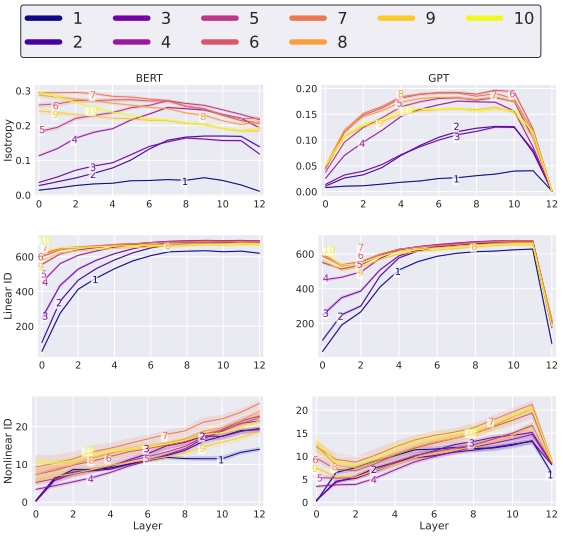


Figure 6: Encoder and Decoder layerwise geometry by word length.

clines mildly again toward the top. Across most depths, especially in the middle layers, encoder layers exhibit higher linear intrinsic dimensionality than decoder layers, indicating that BERT's representations distribute variance across a larger set of orthogonal components. This pattern aligns with extensive findings in the "BERTology" and probing literature, which show that BERT's middle layers encode a rich mixture of lexical, semantic, and morpho-syntactic information, while upper layers increasingly specialize toward the masked language modeling objective (Tenney et al., 2019; Rogers et al., 2020; Liu et al., 2021). Under a neighbourhood-based view, Encoder and Decoder profiles become more similar. GRIDE shows that both BERT and GPT-2 have relatively high local intrinsic dimensionality in the middle layers and a slight decrease toward the top, with BERT holding only a small edge over GPT-2. Nonlinear ID thus increases from lower to middle layers and then declines, replicating the pattern of Ansuini et al. (2019): at local scales, both models' representation spaces have similar intrinsic dimensionality, even if GPT-2's global geometry is skewed by a few principal axes. Overall, all scores decline in the upper layers and especially in the decoder, consistent with increasing specialization toward the training objective.

4.2 Feature-Conditioned Analysis

In what follows, we analyse the representational geometry of the manifolds induced by each linguistic feature. Each figure compares encoder scores (left) and decoder scores (right), with isotropy, linear intrinsic dimensionality, and nonlinear intrinsic dimensionality shown from top to bottom. As discussed in Section 4, the `index=1` class in the decoder contains a *subset* of tokens that forms a compact, strongly anisotropic cluster clearly separated from the main cloud. Importantly, many other `index=1` tokens remain intermingled with the bulk distribution. Nevertheless, this separated subgroup is sufficient to disproportionately bias global isotropy and linear intrinsic dimensionality estimates, so we exclude `index=1` tokens from all analyses except the positional study (Figure 5). More statistics about how each feature is distributed can be found in Appendix A.

Index Figure 5 analyses the feature *index*, the token's position from the start of the sentence (indices 1-10). **In both models, all three metrics increase with depth and peak in the upper–middle layers**, then drop in the final layers. In the decoder, the near degenerate trajectory for `index=1` echoes the outlier behaviour in Figure 2d, and is consistent with the causal setting, suggesting that with minimal left context for the first token, the model may rely more on positional priors, yielding a more compressed space. Across layers, index classes maintain a stable ordering. Early tokens exhibit the lowest isotropy and intrinsic dimensionality, while later tokens (indices 5-10) consistently rank highest. Corpus statistics in Table 1 provide informative cues: **early positions are dominated by closed class words with fewer types and stronger**

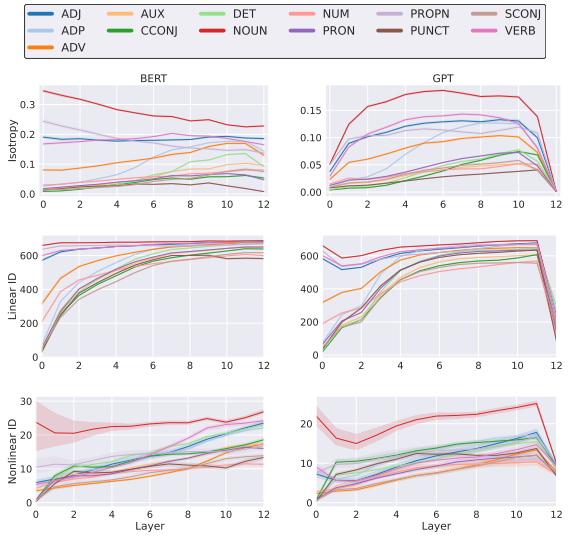


Figure 7: Encoder and Decoder layerwise geometry by tokens grouped by POS.

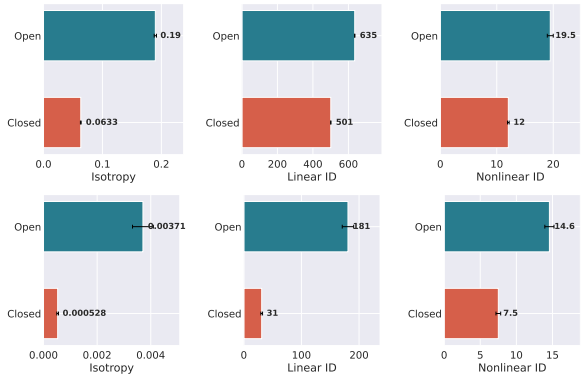


Figure 8: Mean geometric scores of selected open- and closed-class POS groups in BERT (top) and GPT-2 (bottom). For each model and metric, bars report the mean across layers for open-class categories (ADJ, ADV, INTJ, NOUN, PROPN, VERB) and closed-class categories (ADP, AUX, CCONJ, DET, NUM, PART, PRON, SCONJ). Error bars show percentile bootstrap confidence intervals.

syntactic constraints, producing narrower, more anisotropic representations, whereas mid-to-late positions contain mostly nouns, whose greater lexical diversity and looser constraints increase all metrics.

Word length Figure 6 groups words into character length classes. *In both models*, the **linear and nonlinear intrinsic dimensionality** exhibit similar trends. They **increase with depth across all length bins and tend to saturate in the upper-middle layers**. The **deepest layers then compress these subspaces** along task-relevant semantic axes. In the *encoder*, the scores decline slightly toward the final layers, whereas in the *decoder* the drop is more pronounced. *In the encoder*, very short tokens (1-3 characters) start out highly anisotropic but become progressively more isotropic with depth. In contrast, longer tokens (7-9 characters) begin relatively isotropic and gradually lose isotropy toward the top layers. *In the decoder*, isotropy follows a characteristic bell shaped trajectory, with shorter words tracing a smaller curve from the first to the last layer. A robust ordering holds across all metrics and layers; indeed, **a clear gradient is visible: longer tokens consistently show higher linear and nonlinear intrinsic dimensionality, while shorter tokens remain at the bottom of all curves**. This ranking mirrors their distributional profiles summarised in Table 2: short tokens are mostly functional words, tend to have shorter head distances, and are influenced by strong positional priors.

Part-of-speech Figure 7 examines tokens according to their POS classes. Across both models and all metrics, the trends closely resemble those observed for the word-length feature in Section 4.2. Open-class POS categories tend to pattern with longer content words, whereas closed-class categories with shorter function words. Figure 8 summarizes this contrast by collapsing POS tags into open- and closed-class groups: **in both architectures, content-word categories occupy more isotropic and higher-dimensional subspaces than function-word categories, especially under the linear ID metric**. This is consistent with their broader lexical and syntactic variability: NOUNs and VERBs have high type-token diversity and participate in a wider range of dependency relations, whereas ADP and DET typically occur in more constrained local configurations, with short dependency distances and high rates of left-of-head attachment. For nonlinear ID, the effect is weaker and more layer-dependent: both models show a small early-layer decrease, but only GPT-2 shows a second drop in the final layer, while BERT continues to increase toward the top.

Dependency head distance Figure 9 groups tokens by their signed dependency distance to the syntactic head, with $d < 0$ indicating that the head occurs to the left of the token and $d > 0$ indicating that the head occurs to its right. **In BERT, words that follow their head show the highest isotropy and the largest linear and nonlinear**

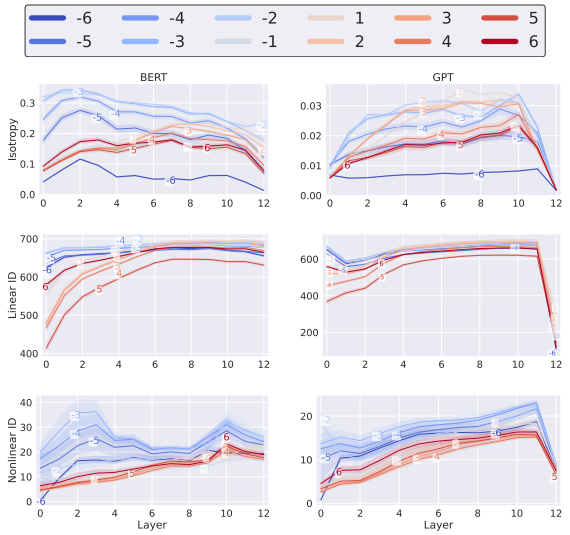


Figure 9: Encoder and Decoder layerwise geometry by token distance from syntactic head.

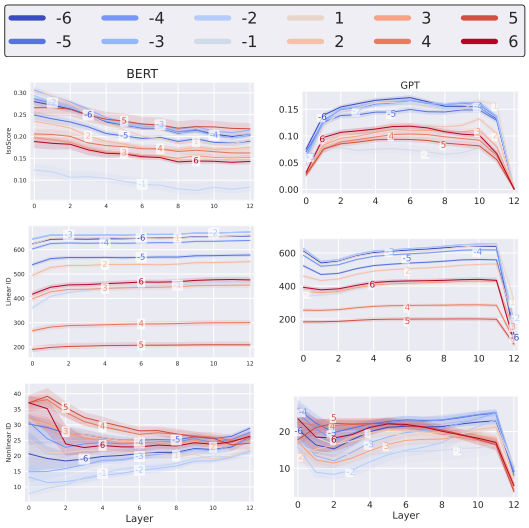


Figure 10: Encoder and Decoder layerwise geometry by nominal head distance (nouns). Token counts per distance bin (n_d): n_{-6}=4 000, n_{-5}=1 882, n_{-4}=3 138, n_{-3}=5 790, n_{-2}=6 275, n_{-1}=1 213, n_1=4 846, n_2=1 650, n_3=1 073, n_4=585, n_5=378, n_6=1 167 (N=31, 997).

intrinsic dimensionality, whereas tokens that precede their head occupy smaller and more anisotropic subspaces. A plausible explanation is that content words appearing after an already available head are both structurally licensed by that head and semantically diverse, allowing their representations to spread across richer subspaces. By contrast, function words that precede a still-unseen head are more syntactically constrained and predictable, yielding lower-dimensional and more anisotropic representations. The dependency-label statistics support this interpretation: large negative distances are mainly associated with nominal dependents such as *obj*, *obl*, and *nmod*, while positive distances are more often associated with configurations involving closed-class or highly constrained elements such as *case*, *det*, and *punct*.

For GPT-2, the pattern is weaker and less monotonic. This may be due to the model's causal objective: tokens at the beginning of a sentence, or tokens whose syntactic head has not yet appeared, must be represented under greater contextual uncertainty. This can increase isotropy and linear intrinsic dimensionality, since these positions are compatible with a broader range of possible continuations. Nonlinear intrinsic dimensionality, however, appears less affected by this left-to-right uncertainty. Overall, GPT-2 seems to attenuate the dependency-distance effects observed in BERT, possibly because causal prediction partially blurs the contrasts induced by syntactic position and type frequency.

Nominal head distance To disentangle the effect of syntactic position from lexical category, we repeat the head distance analysis by restricting attention to *noun* tokens only. Analogous analyses for the other major POS classes are available in the repository. This control allows us to isolate how dependency distance shapes representation geometry *within* a single open class category. As shown in Figure 10, under this constraint, the same geometric ordering largely persists for isotropy and linear intrinsic dimensionality: nouns whose syntactic head lies to the left continue to exhibit higher isotropy and larger linear intrinsic dimensionality than nouns preceding their head. This indicates that the head distance effect is not merely a by-product of POS composition, but reflects a genuine sensitivity to syntactic configuration. By contrast, the behaviour of *nonlinear* intrinsic dimensionality differs from the aggregate pattern. According to this metric, **nouns following their head tend to occupy comparatively more compact local manifolds in many layers, while nouns preceding their head exhibit higher nonlinear dimensionality**. From a linguistic point of view, this reversal suggests that the high nonlinear intrinsic dimensionality values observed in the aggregate analysis were largely driven by the inclusion of semantically heterogeneous open class tokens relative to functional ones: preverbal

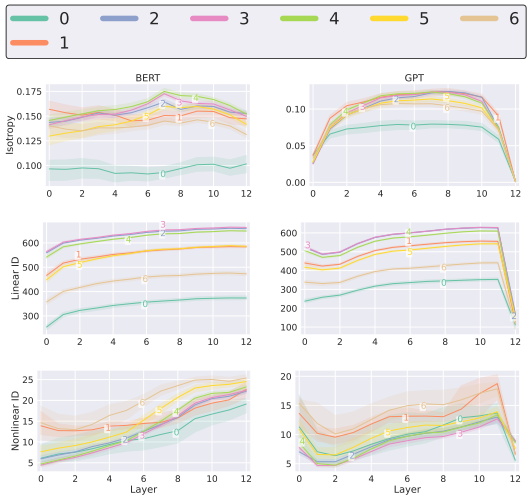


Figure 11: Encoder and Decoder layerwise geometry by verbal arity (verbs). Token counts per arity bin (n_a): n_0=822, n_1=2 011, n_2=5 006, n_3=5 978, n_4=4 281, n_5=2 201, n_6=1 196 (N=21, 495).

nouns aggregate a wider range of discourse and constructional contexts, yielding higher neighbourhood complexity, whereas postverbal dependents occur in more stereotyped, head-governed configurations and therefore form more compact local manifolds.

Verbal arity Figure 11 groups verbs by *arity* class. Across both Encoder and Decoder models, isotropy and linear intrinsic dimensionality show a clear dependence on this feature, with **mid-arity predicates (2–4 dependents) occupying higher dimensional and more isotropic subspaces** than low and high arity verbs. This pattern closely mirrors frequency effects. Highly productive verbal predicates occur in a wider range of syntactic contexts, requiring more linear degrees of freedom to encode their distributional variability. In contrast, nonlinear intrinsic dimensionality scores exhibit substantially more overlap across arity classes. This suggests that **while frequency and syntactic productivity strongly shape the linear organization of verb representations, they have a weaker effect on the underlying nonlinear semantic manifold**, which appears comparatively more stable across arity.

5 Conclusions

We introduced a linguistically grounded framework for analyzing the geometry of neural representations in Transformer language models. By studying class conditioned token manifolds defined by surface and morpho-syntactic features, and by comparing Encoder and Decoder architectures, we showed that representational geometry evolves in systematic, linguistically meaningful ways across layers. Both models exhibit a broad expansion–compression trajectory, but they allocate variance differently: BERT maintains more isotropic, higher dimensional subspaces, whereas GPT-2 exhibits stronger anisotropy and sharper late layer compression, consistent with differences in architecture and training objective.

Our results position geometric analysis as a principled diagnostic for interpretability, disentangling changes driven by linguistic structure from those driven by distributional skew. We show that different geometric metrics capture complementary aspects of representation space: isotropy and linear intrinsic dimensionality are strongly shaped by token frequency and distributional variance (highlighting how function words, early sentence positions, and low arity predicates concentrate into compact, low-rank regions), while nonlinear intrinsic dimensionality reveals finer manifold structures that are less sensitive to corpus skew and lexical heterogeneity. Together, these findings highlight the need for multi-metric geometric profiling to reliably characterize linguistic organization in embedding spaces and compare models across architectures and layers.

Looking forward, this framework enables practical applications beyond analysis: it can be used to monitor and diagnose representation pathologies during pretraining and fine-tuning, to compare objectives and architectural choices via measurable geometric signatures, and to guide representation post-processing. More broadly, feature-conditioned geometric profiling can support model auditing under domain shift and inform targeted interventions that improve robustness and interpretability without relying on task-specific probes.

6 Limitations

Although UD-EWT provides high-quality morphological and dependency annotations, our analysis relies on a single English treebank; therefore, the resulting geometric/linguistic patterns may not generalize to other languages, domains, or annotation conventions, especially with richer morphology or different word order properties. We focus on BERT-base and GPT-2 as canonical encoder-only / decoder-only Transformers with comparable depth

and hidden size, but these models do not cover the diversity of modern architectures, tokenizers, training recipes, or scaling regimes, so the reported architectural differences should be viewed as evidence from these specific instances rather than universal claims. Several linguistic features are discretized or capped (e.g., position, distance, length), which stabilizes estimates but can hide long tail phenomena. Moreover, many factors (position, POS, frequency, dependency roles) are correlated, so our analysis is primarily descriptive and does not fully disentangle causal drivers. Isotropy and intrinsic dimensionality estimators depend on sampling and hyperparameters (e.g., kNN neighbourhood sizes) and may be sensitive to noise and high dimensional effects, so we emphasize relative trends across layers and classes rather than absolute values. Finally, the geometric signatures we report do not identify the mechanisms that produce them, nor establish that these properties are necessary for downstream behaviour.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of ACL-IJCNLP 2021 (Long Papers)*, pages 7319–7328, Online. Association for Computational Linguistics.

Luca Albergante, Jonathan Bac, and Andrei Zinovyev. 2019. Estimating the effective dimension of large biological datasets using fisher separability analysis. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Laurent Amsaleg, Oussama Chelly, Teddy Furon, Stéphane Girard, Michael E Houle, Ken-ichi Kawarabayashi, and Michael Nett. 2018. Extreme-value-theoretic estimation of local intrinsic dimensionality. *Data Mining and Knowledge Discovery*, 32(6):1768–1805.

Laurent Amsaleg, Oussama Chelly, Michael E Houle, Ken-Ichi Kawarabayashi, Miloš Radovanović, and Weeris Treeratanajaru. 2019. Intrinsic dimensionality estimation within tight localities. In *Proceedings of the 2019 SIAM international conference on data mining*, pages 181–189. SIAM.

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. 2019. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32.

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.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Paola Campadelli, Elena Casiraghi, Claudio Ceruti, and Alessandro Rozza. 2015. Intrinsic dimension estimation: Relevant techniques and a benchmark framework. *Mathematical Problems in Engineering*, 2015(1):759567.

Richard Cangelosi and Alain Goriely. 2007. Component retention in principal component analysis with application to cDNA microarray data. *Biology direct*, 2(1):2.

Kevin M Carter, Raviv Raich, and Alfred O Hero III. 2009. On local intrinsic dimension estimation and its applications. *IEEE Transactions on Signal Processing*, 58(2):650–663.

Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Lei Yu, Alessandro Laio, and Marco Baroni. 2025. Emergence of a high-dimensional abstraction phase in language transformers. In *The Thirteenth International Conference on Learning Representations*.

Francesco Denti, Diego Doimo, Alessandro Laio, and Antonietta Mira. 2022. The generalized ratios intrinsic dimension estimator. *Scientific Reports*, 12(1):20005.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Vittorio Erba, Marco Gherardi, and Pietro Rotondo. 2019. Intrinsic dimension estimation for locally undersampled data. *Scientific reports*, 9(1):17133.

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(1):12140.

Amir Massoud Farahmand, Csaba Szepesvári, and Jean-Yves Audibert. 2007. Manifold-adaptive dimension estimation. In *Proceedings of the 24th international conference on Machine learning*, pages 265–272.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. 2016. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049.

Keinosuke Fukunaga and David R Olsen. 1971. An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on computers*, 100(2):176–183.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of EMNLP 2021*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peter Grassberger and Itamar Procaccia. 1983. Measuring the strangeness of strange attractors. *Physica D: nonlinear phenomena*, 9(1-2):189–208.

Katharina Hämmerl, Alina Fastowski, Jindřich Libovický, and Alexander Fraser. 2023. Exploring anisotropy and outliers in multilingual language models for cross-lingual semantic sentence similarity. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7023–7037, Toronto, Canada. Association for Computational Linguistics.

Martin Haspelmath. 2007. Pre-established categories don't exist: Consequences for language description and typology.

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

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7839–7847.

Evan Hernandez and Jacob Andreas. 2021. The low-dimensional linear geometry of contextualized word representations. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 82–93, Online. Association for Computational Linguistics.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Kurt Hornik and Bettina Grün. 2014. On maximum likelihood estimation of the concentration parameter of von Mises-Fisher distributions. *Computational statistics*, 29(5):945–957.

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 *International Conference on Learning Representations*.

Kerstin Johnsson, Charlotte Soneson, and Magnus Fontes. 2014. Low bias local intrinsic dimension estimation from expected simplex skewness. *IEEE transactions on pattern analysis and machine intelligence*, 37(1):196–202.

James D. Johnston. 1988. Transform coding of audio signals using perceptual noise criteria. *IEEE Journal on Selected Areas in Communications*, 6(2):314–323.

Iain M Johnstone and Debashis Paul. 2018. PCA in high dimensions: An orientation. *Proceedings of the IEEE*, 106(8):1277–1292.

Roni Katzir. 2023. Why large language models are poor theories of human linguistic cognition: A reply to Piantadosi. *Biolinguistics*, 17:1–12.

Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. 2021. BERT busters: Outlier dimensions that disrupt transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3392–3405, Online. 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.

E. Levina and P. J. Bickel. 2005. Maximum likelihood estimation of intrinsic dimension. *NIPS*.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. 2018. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*.

Zhuang Liu, Wayne Lin, Ya Shi, and Jun Zhao. 2021. A robustly optimized BERT pre-training approach with post-training. In *China National Conference on Chinese Computational Linguistics*, pages 471–484.

N. Madhu. 2009. Note on measures for spectral flatness. *Electronics Letters*, 45(23):1195–1196.

Jonathan Mamou, Hang Le, Miguel Del Rio, Cory Stephenson, Hanlin Tang, Yoon Kim, and SueYeon Chung. 2020. Emergence of separable manifolds in deep language representations. *CoRR*, abs/2006.01095.

Timothee Mickus, Stig-Arne Grönroos, and Joseph Attieh. 2024. Isotropy, clusters, and classifiers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 75–84.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective post-processing for word representations. In *6th International Conference on Learning Representations, ICLR 2018*.

Vivi Nastase and Paola Merlo. 2024a. Are there identifiable structural parts in the sentence embedding whole? In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 23–42, Miami, Florida, US. Association for Computational Linguistics.

Vivi Nastase and Paola Merlo. 2024b. Tracking linguistic information in transformer-based sentence embeddings through targeted sparsification. In *Proceedings of the 9th Workshop on Representation Learning for NLP (RepL4NLP-2024)*, pages 203–214, Bangkok, Thailand. Association for Computational Linguistics.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Steven T Piantadosi. 2023. Modern language models refute Chomsky's approach to language. *From fieldwork to linguistic theory: A tribute to Dan Everett*, 15:353–414.

Tiago Pimentel and Ryan Cotterell. 2021. A Bayesian framework for information-theoretic probing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2869–2887, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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.

Stefano Recanatesi, Matthew Todd Farrell, Madhu S. Advani, Timothy Moore, Guillaume Lajoie, and Eric Shea-Brown. 2019. Dimensionality compression and expansion in deep neural networks. *ArXiv*, abs/1906.00443.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how BERT works. *Transactions of the association for computational linguistics*, 8:842–866.

Olivier Roy and Martin Vetterli. 2007. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pages 606–610. IEEE.

Alessandro Rozza, Gabriele Lombardi, Claudio Ceruti, Elena Casiraghi, and Paola Campadelli. 2012. Novel high intrinsic dimensionality estimators. *Machine learning*, 89(1):37–65.

William Rudman, Nate Gillman, Taylor Rayne, and Carsten Eickhoff. 2022. IsoScore: Measuring the uniformity of embedding space utilization. In *Findings of ACL 2022*, pages 3316–3330.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. 2025. Layer by layer: Uncovering hidden representations in language models. In *Forty-second International Conference on Machine Learning*.

Suvrit Sra. 2012. A short note on parameter approximation for Von Mises-Fisher distributions: and a fast implementation of $i_s(x)$. *Computational Statistics*, $27(1) : 177--190$.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *ArXiv*, abs/2103.15316.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Joel A. Tropp. 2015. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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

Appendix A. Dataset statistics

This section summarizes the distribution of our feature classes in the UD-EWT and their overlap with other features. Tables 1-6 report, for each class, token/type counts and the most frequent co-occurring values of the remaining features.

	Counts		POS		Length		Head dist.		Relation		Arity	
Index	n_{tok}	n_{types}	top	%	top	%	top	%	top	%	top	%
1	10067	1600	PRON	33.24	3	21.10	1	31.79	nsubj	32.95	0	83.68
2	9647	2309	AUX	17.76	3	20.74	1	33.02	root	18.01	0	65.37
3	9800	2483	VERB	18.20	4	20.05	1	32.01	root	17.38	0	62.15
4	9932	2701	VERB	15.88	4	19.84	1	32.69	root	14.03	0	62.90
5	9948	2876	NOUN	18.69	3	19.14	1	31.39	root	9.63	0	62.78
6	9972	2891	NOUN	18.66	3	18.61	1	30.14	case	10.00	0	64.37
7	9559	2820	NOUN	18.87	3	18.84	1	29.58	punct	9.93	0	64.16
8	9142	2709	NOUN	18.22	3	18.99	1	29.74	punct	10.75	0	65.38
9	8704	2699	NOUN	18.64	3	19.05	1	28.37	punct	11.35	0	64.84
10	8252	2566	NOUN	18.82	3	18.90	1	29.50	punct	12.02	0	66.02

Table 1: Per-index distribution of words in the UD-EWT training set.

	Counts		POS		Index		Head dist.		Relation		Arity	
Length	n_{tok}	n_{types}	top	%	top	%	top	%	top	%	top	%
1	29626	74	PUNCT	71.80	1	5.98	1	21.08	punct	71.80	0	97.21
2	29918	334	ADP	37.84	1	6.61	1	41.01	case	35.28	0	92.30
3	36097	911	DET	26.98	1	5.88	1	41.34	det	26.02	0	81.98
4	33008	1660	NOUN	18.66	2	6.03	1	26.63	advmod	10.33	0	55.09
5	18892	2012	NOUN	30.75	2	5.92	1	26.32	amod	9.01	0	47.07
6	14274	2373	NOUN	37.89	4	5.73	1	21.52	amod	8.81	0	36.53
7	12188	2426	NOUN	36.82	4	5.29	1	19.49	amod	9.56	0	33.13
8	7895	2044	NOUN	39.61	5	5.78	1	20.39	amod	12.89	0	31.91
9	5526	1557	NOUN	39.32	4	5.52	1	21.08	amod	12.76	0	31.94
10	3526	1150	NOUN	45.86	3	6.66	1	21.67	amod	10.78	0	30.20

Table 2: Per-length distribution of words in the UD-EWT training set.

	Counts		Index		Length		Head dist.		Relation		Arity	
POS	n_{tok}	n_{types}	top	%	top	%	top	%	top	%	top	%
NOUN	33607	6903	6	5.54	4	18.33	-2	18.67	obj	20.22	2	27.81
PUNCT	22123	94	10	4.48	1	96.14	-1	15.46	punct	100.0	0	100.0
VERB	22095	3538	3	8.07	4	27.70	0	30.91	root	30.91	3	27.54
PRON	18255	138	1	18.33	2	26.52	1	37.16	nsubj	55.28	0	89.92
ADP	17557	116	6	6.01	2	64.49	2	34.57	case	92.33	0	99.09
DET	16123	36	1	7.35	3	60.41	1	57.10	det	96.52	0	98.18
ADJ	12648	2156	4	6.19	4	22.45	1	52.92	amod	68.24	0	65.24
AUX	11526	90	2	14.86	2	31.25	1	46.55	aux	51.22	0	98.13
PROPN	11203	3808	1	6.47	5	17.75	1	22.98	compound	20.15	0	46.26
ADV	9913	716	1	8.59	4	36.17	1	39.39	advmod	93.20	0	85.62
CCONJ	6627	21	7	5.93	3	86.27	1	43.40	cc	98.25	0	99.26
PART	4241	7	4	8.32	2	76.85	1	81.66	mark	76.42	0	99.15
SCONJ	3815	62	1	11.66	2	40.58	2	24.25	mark	98.43	0	98.56
NUM	3675	728	1	8.49	3	22.67	1	29.33	nummod	41.06	0	56.35
SYM	651	38	1	9.37	1	87.40	1	37.02	cc	21.04	0	49.92
INTJ	556	66	1	50.54	6	45.14	1	39.57	discourse	95.50	0	94.78
X	301	154	4	9.63	4	17.28	-1	27.57	flat	48.17	0	85.71

Table 3: Per-POS distribution of words in the UD-EWT training set.

	Counts		POS		Index		Length		Relation type		Arity	
Head dist.	n_{tokens}	n_{types}	top	%	top	%	top	%	top	%	top	%
-6	3251	1701	NOUN	37.47	8	10.34	1	18.61	punct	19.26	3	23.81
-5	4545	2268	NOUN	41.41	8	9.42	4	17.29	obl	21.30	3	23.06
-4	6684	3050	NOUN	46.95	7	8.95	4	18.73	obl	21.47	3	31.57
-3	10487	4117	NOUN	55.21	6	7.96	4	19.61	nmod	20.32	2	46.96
-2	15690	5483	NOUN	39.99	5	7.39	4	21.22	obj	22.54	1	49.94
-1	13038	2647	PUNCT	26.24	2	8.18	1	26.86	punct	26.24	0	85.46
0	10046	3032	VERB	67.99	2	17.29	4	28.58	root	100.0	3	25.46
1	57036	5218	DET	16.14	4	5.69	3	26.16	det	16.06	0	95.13
2	28588	2747	ADP	21.23	1	8.36	3	26.21	case	21.48	0	92.87
3	13592	1840	ADP	19.41	1	10.53	2	24.01	case	19.81	0	85.98
4	6439	1171	ADP	14.35	1	11.18	3	22.43	nsubj	16.40	0	83.20
5	3249	795	PUNCT	16.65	1	11.27	3	20.71	punct	16.65	0	75.96
6	1673	575	PUNCT	16.92	1	12.43	4	19.43	punct	16.92	0	68.98

Table 4: Per-head-distance distribution summary of words in the UD-EWT training set.

	Counts		POS		Index		Length		Head dist.		Relation	
Arity	n_{tok}	n_{types}	top	%	top	%	top	%	top	%	top	%
0	129064	7581	PUNCT	17.14	1	6.53	3	22.93	1	42.04	punct	17.14
1	20239	6441	NOUN	45.31	6	5.68	4	19.97	-2	38.72	obj	15.83
2	18037	5846	NOUN	51.82	2	6.32	4	22.05	-3	27.30	obl	16.97
3	13546	4621	VERB	44.92	3	8.25	4	23.67	0	18.88	root	18.88
4	7856	3169	VERB	54.88	4	8.95	4	26.03	0	31.73	root	31.73

Table 5: Per-arity distribution of words in the UD-EWT training set.

	Counts		POS		Index		Length		Head dist.		Arity	
Relation	n_{tok}	n_{types}	top	%	top	%	top	%	top	%	top	%
punct	22123	94	PUNCT	100.0	10	4.48	1	96.14	-1	15.46	0	100.0
case	16577	128	ADP	97.79	6	6.01	2	63.67	2	37.04	0	98.68
nsubj	15648	2562	PRON	64.49	1	21.20	4	19.67	1	35.85	0	71.04
det	15562	31	DET	100.0	1	7.36	3	60.36	1	58.85	0	99.92
advmod	10307	704	ADV	89.64	1	7.98	4	33.07	1	43.64	0	90.67
root	10046	3032	VERB	67.99	2	17.29	4	28.58	0	100.0	3	25.46
obj	9685	2978	NOUN	70.18	5	7.64	4	22.28	-2	36.52	1	33.07
amod	9312	1960	ADJ	92.69	5	5.52	4	18.98	1	75.83	0	89.68
obl	8743	3228	NOUN	68.20	7	5.94	4	21.22	-3	21.25	2	35.01
conj	7557	3810	VERB	39.31	11	5.27	4	21.19	-2	26.17	1	28.04
mark	7097	80	SCONJ	52.91	5	6.97	2	67.66	1	50.08	0	98.72
compound	7018	2758	NOUN	62.20	6	5.77	4	18.18	1	80.61	0	79.91
nmod	6794	2980	NOUN	66.50	9	5.78	4	17.28	-2	34.56	2	36.78
cc	6694	26	CCONJ	97.27	7	5.92	3	85.55	1	44.07	0	98.98
aux	5904	71	AUX	100.0	2	17.45	3	29.86	1	48.49	0	99.88

Table 6: Per-relation-type distribution of words (top 15 relations by token count) in the UD-EWT training set.

Appendix B. Token pooling

Because our linguistic annotations are defined at the word level, whereas the models operate over tokenized inputs, we construct one representation per word by aggregating the hidden states of the tokens aligned with that word.

Figure 12 reports Spearman rank correlations between the class-by-layer profiles of each geometric metric obtained with MEAN pooling and those obtained with the main pooling convention. The correlations are consistently high across metrics, with $\rho = .92$–$.997$ for BERT and $\rho = .816$–$.842$ for GPT-2, all with $p < .001$. This indicates that the main geometric profiles are robust to the choice of token-pooling strategy.

Appendix C. Metrics

Research on representation geometry often distinguishes between two broad families of metrics: (i)

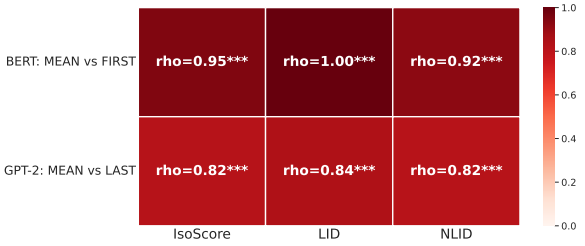


Figure 12: Spearman rank correlations compare class-by-layer metric profiles computed with MEAN token pooling against the original pooling convention: FIRST for BERT and LAST for GPT-2. Correlations are pooled across token length, token index, dependency relation type, and dependency arity. All reported correlations are significant at $p < .001$.

isotropy (or anisotropy) measures, which assess how uniformly representations occupy the ambient embedding space, and (ii) *intrinsic dimensionality* (ID) estimators, which quantify the effective number of degrees of freedom of a set of representations. These approaches can be organized along two main axes: (a) *global* metrics return a single summary for an entire dataset or layer, whereas *local* metrics produce neighbourhood-wise or point-wise estimates and (b) for ID specifically, the assumed geometry: *linear* (subspace) structure versus *nonlinear* (manifold) structure.

In the isotropy literature, most metrics are global. For example, IsoScore (Rudman et al., 2022) quantifies how evenly variance is distributed across principal directions and highlights limitations of earlier heuristics (e.g., cosine-similarity- and partition-based scores). Earlier global metrics also include eigenvalue flatness ratios, and approaches based on fitting a von Mises-Fisher distribution to ℓ_2-normalized vectors, where the estimated concentration parameter reflects directional anisotropy (Sra, 2012). Overall, global isotropy measures ask whether the spread of representations is roughly uniform across directions or dominated by a small number of axes.

Turning to intrinsic dimensionality, methods similarly split between linear and nonlinear approaches. Linear ID estimators assume representations lie approximately in a lower-dimensional linear subspace and summarize the covariance spectrum via the number of principal components needed to explain a target fraction of variance (e.g., 99%) or via continuous surrogates such as the effective rank (Roy and Vetterli, 2007) or the participation ratio (Recanatesi et al., 2019). By contrast, nonlinear ID estimators allow for curved manifolds and nonuniform sampling density. Global methods in this class include the TwoNN estimator (Facco et al., 2017), which infers dimension from ratios between the distances to each point's first and second nearest neighbours, and its multi-scale extension GRIDE (Denti et al., 2022), which generalizes these distance-ratio statistics to estimate ID as a function of neighbourhood scale. Many manifold-

Method	Scope	Type	In a nutshell
Isotropy			
IsoScore (Rudman et al., 2022)	G	Iso	Variance uniformity across axes
Spectral Flatness (Johnston, 1988)	G	Iso	Flatness of eigen-spectrum
vMF concentration κ (Sra, 2012)	G	Aniso↑	Directional concentration
Linear intrinsic dimensionality			
Effective Rank (Roy and Vetterli, 2007)	G	ID-L	Entropy-based effective rank
Participation Ratio (Recanatesi et al., 2019)	G	ID-L	Effective # active modes
Stable Rank (Tropp, 2015)	G	ID-L	Robust rank proxy
Global PCA (PCA_{FO}) (Fukunaga and Olsen, 1971)	G/L	ID-L	Count PCs with $\lambda_i \geq \alpha \cdot \lambda_1$
Global PCA (PCA_{99}) (Cangelosi and Goriely, 2007)	G/L	ID-L	# PCs for $\alpha\%$ variance
Nonlinear intrinsic dimensionality			
TwoNN (Facco et al., 2017)	G	ID-NL	Global ID from 2nd/1st NN distance ratio
GRIDE (Denti et al., 2022)	G	ID-NL	Global ID from generalized kNN ratios
MLE (Levina and Bickel, 2005)	L→G	ID-NL	Local ID via MLE on kNN distance growth
MOM (Amsaleg et al., 2018)	L→G	ID-NL	Local ID via extreme-value moment estimator
TLE (Amsaleg et al., 2019)	L→G	ID-NL	Stable local ID for very small neighbourhood
CorrInt (Grassberger and Procaccia, 1983)	G	ID-NL	Correlation-integral slope
FisherS (Albergante et al., 2019)	G	ID-NL	ID from Fisher linear separability probability
ESS (Johnsson et al., 2014)	L→G	ID-NL	Pointwise LID from simplex-skewness
MADA (Farahmand et al., 2007)	L→G	ID-NL	Manifold-adaptive aggregation of local IDs

Table 7: Chosen metrics. *Scope*: G = global (dataset-level), L = local (neighbourhood-based), L→G = local estimates aggregated globally. *Type*: ID-L = linear intrinsic dimension, ID-NL = nonlinear intrinsic dimension.

based estimators are naturally local: for example, the maximum-likelihood estimator of Levina and Bickel (Levina and Bickel, 2005) yields per-point dimension estimates from k-NN distance spacings that are then averaged to obtain a dataset-level ID. More generally, local intrinsic dimensionality methods compute point-wise IDs and aggregate them (e.g., by averaging or robust aggregation), emphasizing "interior" points can reduce the negative bias that affects global ID estimates, which is often driven by boundary effects and finite-sample issues (Carter et al., 2009).

Appendix C.1. Implementation details and hyperparameters

Unless stated otherwise, we use the reference implementations provided by libraries `IsoScore`, `DADApy`, and `scikit-dimension` with default settings. To quantify uncertainty, for each layer and linguistic class we compute metrics on bootstrap resamples drawn *with replacement* and report the bootstrap mean and a 95% percentile interval (2.5-97.5). We use R_{fast}=50 resamples for spectral/isotropy and linear metrics, and R_{heavy}=200 resamples for kNN-based (nonlinear) estimators. For the latter we cap each resample at $M_{\max}$=5000 tokens per class. All bootstrap sampling uses a fixed random seed. For TwoNN we use DADApy's default configuration (`algorithm=base`, `mu_fraction=0.9`). For GRIDE we compute neighbour distances up to the 64th neighbour using Euclidean distance (`compute_distances(maxk=64)`) and estimate multi-scale intrinsic dimensionality with `return_id_scaling_gride(range_max=64)`, we take the final (largest scale) GRIDE value as a single per- layer summary.

Appendix C.2. Isotropy metrics

Let $X \in \mathbb{R}^{N\times D}$ be a matrix of token representations for a given layer (N tokens, hidden size D). We center X row-wise: $X_c = X - \mathbf{1}\mu^\top$ with $\mu = \frac{1}{N}\sum_i X_{i\cdot}$. For spectral metrics we compute singular values $\{s_i\}$ of X_c and set $\lambda_i = s_i^2$ (eigenvalues of $X_c^\top X_c$ up to a constant), sorted $\lambda_1 \geq \cdots \geq \lambda_D$. This scaling constant cancels in all ratios below.

IsoScore. IsoScore measures how close this variance profile is to *uniform* by comparing a normalized variance vector to the all-ones vector $\mathbf{1}$:

$$\widehat{\lambda} = \sqrt{D}\,\frac{\lambda}{\|\lambda\|_2}, \qquad \delta(X) = \frac{\|\widehat{\lambda} - \mathbf{1}\|_2}{\sqrt{2\big(D - \sqrt{D}\big)}}.$$

Here $\delta(X) = 0$ iff $\lambda_1 = \cdots = \lambda_D$ (perfectly flat spectrum) and $\delta(X) = 1$ for maximally concentrated variance (rank-1 spectrum). IsoScore then applies a fixed monotone rescaling $\iota(X) = g(\delta(X)) \in [0,1]$ so that $\iota(X) = 1$ corresponds to perfect isotropy (Rudman et al., 2022). Because it depends only on the covariance spectrum after centering, ι is invariant to translations, global rescalings, and orthogonal rotations of the embeddings.

Spectral flatness Spectral flatness is a simple global isotropy proxy based on the eigenvalue spectrum of the covariance matrix. Let $\lambda_1, \ldots, \lambda_D$ denote the eigenvalues of the centered covariance. The spectral flatness measure computes the ratio between the geometric and arithmetic means:

$$\mathrm{SF}(X) = \frac{\exp\Big(\frac{1}{D}\sum_{i=1}^{D}\log(\lambda_i + \varepsilon)\Big)}{\frac{1}{D}\sum_{i=1}^{D}\lambda_i + \varepsilon} \in (0,1],$$

where higher values indicate a flatter (more isotropic) spectrum. This measure is widely used in signal processing as a "tonality" versus "noise-like" indicator and has been adopted as an isotropy heuristic in embedding analysis (Johnston, 1988; Madhu, 2009).

Von Mises–Fisher concentration κ (anisotropy↑) When vectors are ℓ_2 normalized to the unit hypersphere, isotropy can also be probed by fitting a von Mises–Fisher (vMF) distribution. Let $\tilde{x}_i = X_{i\cdot}/\|X_{i\cdot}\|$ be unit vectors, $R = \big\|\frac{1}{N}\sum_i \tilde{x}_i\big\|$ their mean resultant length, and $d = D$ the ambient dimension. Following Sra (2012); Hornik and Grün (2014), we approximate the vMF concentration parameter by

$$\hat{\kappa} \;\approx\; \frac{R\,(d - R^2)}{1 - R^2 + \varepsilon},$$

so larger κ implies stronger concentration of the point cloud around a preferred direction, i.e., higher anisotropy.

Appendix C.3. Linear intrinsic dimensionality (spectral)

Let $X \in \mathbb{R}^{N\times D}$ be the (row centered) representation matrix for a layer and $\{\lambda_i\}_{i=1}^{D}$ the eigenvalues of $X^\top X$ (up to a constant). Define $p_i = \lambda_i / \sum_j \lambda_j$.

Effective rank The effective rank is defined as

$$\mathrm{eRank}(X) = \exp\Big(-\sum_{i=1}^{D} p_i \log(p_i + \varepsilon) \Big),$$

i.e., the exponential of the Shannon entropy of the normalized spectrum, and can be interpreted as the "effective number of modes" contributing to the variance (Roy and Vetterli, 2007).

Participation ratio The participation ratio,

$$\mathrm{PR}(X) = \frac{\left(\sum_i \lambda_i\right)^2}{\sum_i \lambda_i^2},$$

equals the ratio of the squared ℓ_1 norm to the ℓ_2^2 norm of the spectrum and serves as a smooth proxy for dimensionality: it attains its maximum when all eigenvalues are equal and decreases as variance concentrates in a few directions (Recanatesi et al., 2019).

Stable rank The stable rank is defined as

$$\mathrm{srank}(X) = \frac{\sum_i \lambda_i}{\lambda_{\max}} = \frac{\|X_c\|_F^2}{\|X_c\|_2^2},$$

and provides a scale invariant surrogate of matrix rank that is robust to small eigenvalues (Tropp, 2015).

Global PCA We define the global PCA dimension at variance level α as

$$d_{\mathrm{PCA}_\alpha}(X) \;=\; \min\left\{ m : \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{j=1}^{D} \lambda_j} \geq \alpha \right\}.$$

In experiments we use $\alpha = 0.99$. In the FO variant (PCA_{FO}), the local dimension is

$$k = \big|\{\, i : \lambda_i \geq \alpha_{\mathrm{FO}} \lambda_1 \,\}\big|,$$

In our experiments we compute these quantities on the full class/layer point cloud, yielding a single global estimate per class and layer. (Fukunaga and Olsen, 1971; Cangelosi and Goriely, 2007)

Appendix C.4. Nonlinear intrinsic dimensionality

All nonlinear estimators below operate on kNN distance statistics after deduplication and the addition of a small jitter to break ties.

TwoNN The TwoNN estimator of Facco et al. (2017) infers a global ID from the ratios of first and second nearest-neighbour distances. For each point i, it forms the ratio $\mu_i = r_{i,2}/r_{i,1}$, where $r_{i,1}$ and $r_{i,2}$ are distances to the first and second neighbour. Under mild assumptions, these ratios follow a Pareto distribution with CDF $F(\mu) = 1-\mu^{-d}$, where d is the intrinsic dimension. A linear fit of $y_i = -\log(1 - \hat{F}(\mu_i))$ versus $x_i = \log \mu_i$ yields an estimate $\hat{d}$.

GRIDE The generalized ratios intrinsic dimensionality estimator (GRIDE) extends TwoNN to probe ID as a function of scale by considering ratios of more distant neighbours $\mu_{i,n_1,n_2} = r_{i,n_2}/r_{i,n_1}$ and generalizing the likelihood accordingly (Denti et al., 2022). In practice, GRIDE returns an ID curve over increasing neighbour scales and allows us to identify plateaus where the estimated dimension is approximately stable.

MLE The Levina–Bickel maximum-likelihood estimator uses all k nearest neighbours of each point:

$$\hat{d}(x_i) = \left[\frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{r_{i,k}}{r_{i,j}} \right]^{-1},$$

$$\hat{d} = \frac{1}{N} \sum_i \hat{d}(x_i),$$

and is derived by maximizing the likelihood of distances in a locally uniform Poisson process on a d-dimensional manifold (Levina and Bickel, 2005).

MOM / TLE The MOM and TLE estimators are extreme value theoretic approaches that infer local ID from the tail behaviour of neighbour distance distributions. MOM uses method-of-moments estimators, while TLE improves stability in "tight" localities (small k) by restricting to neighbours within a suitable radius (Amsaleg et al., 2018, 2019).

Correlation integral slope The correlation dimension D_2 is defined from the correlation integral

$$C(r) = \frac{2}{N(N-1)} \sum_{i<j} \mathbb{I}\{\|x_i - x_j\| \leq r\},$$

$$D_2 = \lim_{r \to 0} \frac{\mathrm{d} \log C(r)}{\mathrm{d} \log r}.$$

which we estimate as the slope of $\log C(r)$ versus $\log r$ over a suitable scale window (Grassberger

and Procaccia, 1983). This fractal dimension estimator captures how pairwise densities scale at small radii.

Fisher separability Fisher-separability ID estimates an *effective* dimension from how often individual points are linearly separable by Fisher discriminants after standard preprocessing: centering, PCA (to avoid tiny eigenvalues), whitening, and (optionally) projecting to the unit sphere. After preprocessing, a point x is Fisher-linearly separable from a cloud Y with parameter $\alpha \in [0, 1)$ if

$$\langle x, y\rangle \;\le\; \alpha\langle x, x\rangle \qquad \forall\, y \in Y,$$

i.e., the hyperplane $\{z : \langle x, z\rangle = \alpha\langle x, x\rangle\}$ separates x from Y. Define the empirical non-separability probability at level α as

$$\widehat{p}_\alpha(x_i) = \frac{1}{N-1}\sum_{j\neq i} \mathbb{I}\{\langle x_i, x_j\rangle > \alpha\langle x_i, x_i\rangle\},$$

$$\overline{p}_\alpha = \frac{1}{N}\sum_i \widehat{p}_\alpha(x_i).$$

For an equidistribution on the unit sphere in dimension n, $\overline{p}_\alpha$ has an explicit approximation, and one can invert it to obtain an effective dimension n_α (involving the Lambert W function); `FisherS` uses this inversion (often over a range of α) to produce a dataset-level ID.

ESS The Expected Simplex Skewness (ESS) estimator uses *angular/simplex* information in local neighbourhoods. Given a neighbourhood X with centroid c, let $\bar{x} = x - c$. For target dimension $d = 1$, ESS estimates an "expected simplex skewness" via

$$\widehat{s}^{(1)}(X) = \frac{\sum_{x,y\in X} \|\bar{x} \wedge \bar{y}\|}{\sum_{x,y\in X} \|\bar{x}\|\,\|\bar{y}\|}$$

(and similarly an "expected projection length" $\widehat{c}(X)$ by replacing $\|\bar{x}\wedge\bar{y}\|$ with $|\langle\bar{x}, \bar{y}\rangle|$). These empirical quantities are mapped to a local dimension by matching them to the corresponding *theoretical* expectations for uniform sampling in an n-ball (tabulated / closed form in the ESS derivation), using linear interpolation between integer n values:

$$\widehat{n} = n + \frac{\widehat{s}^{(d)} - s_n^{(d)}}{s_{n+1}^{(d)} - s_n^{(d)}}$$

We report ESS as the average of the resulting local $\widehat{n}$ over points

MADA The manifold-adaptive dimension estimator (MADA) exploits local volume scaling: for small radii, $\mathbb{P}(\|X - x\| \le r) \approx \eta(x)\, r^{d(x)}$, so kNN radii scale like $r_{x,k} \propto (k/n)^{1/d(x)}$ in locally homogeneous regions. This yields a log-slope estimate from two neighbour indices $k < k'$:

$$\widehat{d}(x; k, k') = \frac{\log(k'/k)}{\log(r_{x,k'}/r_{x,k})}.$$

MADA then chooses neighbourhood sizes in a *scale-adaptive* way (balancing locality bias against finite-sample variance) and aggregates the resulting local estimates to a dataset-level dimension.

Appendix C.5. Scores tested on known manifolds

To validate the behaviour of the isotropy and intrinsic dimensionality metrics used in our analysis, we performed a quick sanity check study on synthetic manifolds with known geometric properties. We generated four canonical datasets: (i) an isotropic Gaussian in $\mathbb{R}^{16}$, with true ID $= 16$, (ii) a highly anisotropic Gaussian, (iii) a low rank Gaussian distribution confined to a 4-dimensional linear subspace of $\mathbb{R}^{64}$ and (iv) the 2-dimensional Swiss roll manifold embedded in $\mathbb{R}^3$. The isotropy scores behave as intended: IG-16 has a very high IsoScore (0.99) and spectral flatness (≈ 1), while AG-16 shows much poorer isotropy (IsoScore 0.36, SF 0.43). LR-4 appears extremely anisotropic in the ambient space (IsoScore ≈ 0.06, SF ≈ 0, spectral ratio ≈ 16.9), reflecting strong concentration onto a 4D subspace. The Swiss roll lies between these regimes (IsoScore 0.49, spectral ratio 1.64), indicating a low dimensional but curved and mildly anisotropic manifold. vMF κ increases as the manifolds become lower dimensional or more structured, from IG-16 (most isotropic) to SR-2 (most aligned). Linear intrinsic dimensionality measures (eRank, participation ratio, stable rank, PCA_{99}) recover the true ID almost exactly on IG-16, LR-4, and SR-2 (values $\approx$ 16, 4, and 2, respectively), but systematically underestimate the dimension of AG-16 (typically 8-10 instead of 16), treating strong anisotropy as a reduced "effective" dimension. Nonlinear ID estimators (TwoNN, GRIDE, MOM, TLE, CorrInt, FisherS, MLE, MADA) are very accurate on the low dimensional manifolds (LR-4, SR-2) and slightly conservative on IG-16, with a more pronounced downward bias on AG-16, again reflecting sensitivity to spectrum shape. Fish-

Family	Metric	IG-16 (ID=16, iso)	AG-16 (ID=16, aniso)	LR-4 (ID=4, 4D subsp.)	SR-2 (ID=2, curved)
Isotropy	IsoScore	0.99	0.36	0.04	0.49
	Spectral flatness (SF)	1.00	0.43	0.00	0.29
	vMF κ	0.29	0.35	1.16	0.67
Linear ID	Effective rank (eRank)	15.97	8.57	4.00	2.03
	Participation ratio (PR)	15.93	6.54	3.99	2.00
	Stable rank	14.61	3.81	3.78	1.83
	PCA_{FO}	16.00	10.00	4.00	2.00
	$\text{PCA}_{0.99}$	16.00	14.00	4.00	2.00
Nonlinear ID	TwoNN	14.38	11.59	4.04	2.50
	GRIDE	12.22	9.01	4.03	1.81
	MOM	12.51	9.19	4.13	1.79
	TLE	13.71	10.81	4.36	2.24
	CorrInt	11.39	8.89	3.90	1.92
	FisherS	16.05	7.99	3.99	1.94
	MLE	13.42	10.38	4.07	2.03
	MADA	14.10	10.71	4.38	2.09
	ESS	15.82	10.89	3.99	1.99

Table 8: Columns show four manifolds with their true ID and qualitative isotropy, rows group metrics into isotropy, linear ID, and nonlinear ID.

erS is the only method that essentially recovers the full 16D ID on IG-16.

Appendix C.6. Convergence analysis

Both isotropy and intrinsic-dimensionality metrics are computed from a *finite* point cloud of representations and can therefore vary with the number of sampled points N'.

Eigen-spectrum-based isotropy scores depend on the sample covariance and can be distorted when N' is small relative to the embedding dimension D. In particular, when $N' < D$, the sample covariance is rank-deficient and principal-direction statistics can spuriously suggest strong anisotropy. More broadly, this sensitivity is consistent with known finite-sample phenomena in high-dimensional PCA/covariance eigenvalues (Johnstone and Paul, 2018). Nearest-neighbour manifold estimators are inherently *scale-dependent*: the effective neighbourhood radius is set by NN distances, which shrink with N' and depend on sampling density. The ID literature documents that estimated ID can change markedly with scale/resolution, and that with limited samples many methods underestimate ID once the underlying dimension is moderately large (often reported around ID $\gtrsim 10$) (Campadelli et al., 2015). This failure mode is closely tied to *local undersampling* in high ambient dimension, which can induce systematic downward bias in small-neighbourhood regimes (Erba et al., 2019; Rozza et al., 2012). Accordingly, it is common to track $\widehat{d}(N')$ and to interpret a stable plateau region as the most reliable operating range (Facco et al., 2017; Denti et al., 2022).

In our feature-conditioned analyses, the effective sample size varies substantially across linguistic classes, and some classes contain only a few hundred tokens, so N'_c can fall below D. Rather than enforcing artificial balancing or aggressive downsampling, we preserve corpus-induced class proportions and quantify uncertainty via a within-class bootstrap: for each class c, we resample tokens with replacement, recompute each metric on the resample, and report the bootstrap mean together with percentile confidence intervals. Rare classes therefore naturally yield wider uncertainty intervals and are treated as lower-confidence than high-support categories. Since both spectrum-based and neighbourhood-based estimators can be sample-size and scale sensitive, we do not interpret any single finite-N'_c estimate as a population-level "true" geometric quantity; instead, we use the metrics comparatively: holding model, layer, estimator, and computation protocol fixed to test whether linguistic partitions induce consistent geometric *hierarchies*. Bootstrap uncertainty intervals primarily indicate when apparent separations between classes are robust versus potentially attributable to estimator noise in small N'_c regimes.

To complement these considerations, we quantify sample-size sensitivity by subsampling a fixed pool of token representations and tracking each estimator over $N' \in [200, 10{,}000]$ (18 log-spaced values, three random seeds, reporting mean $\pm$ one standard deviation across seeds). Using a simple heuristic of "stabilized" meaning within 5% of the $N' = 10{,}000$ value, PCA_{99} stabilizes by $N' \approx 2.5$k and effective rank by $N' \approx 4$k, while IsoScore is within 5% by $N' \approx 2$k and spectral flatness by $N' \approx 5$k. For nonlinear/neighbourhood-based estimators, GRIDE approaches a stable plateau at larger N', remaining within 5% of its final value from $N' \approx 4$k onward, whereas several likelihood/kNN-based estimators in our sweep (e.g., MLE, TLE, ESS) vary substantially over the full range without a clear plateau by $N' = 10{,}000$.

Appendix C.7. Metric correlations

To compare geometric diagnostics, we compute pairwise metric correlations on BERT layer 12 representations. We embed a pool of 100k tokens once, then construct $N = 200$ evaluation points by repeatedly drawing subsamples of $m = 5000$ tokens and recomputing all metrics on each subsample. This yields many partially distinct points (instead of reusing nearly identical token sets) while keeping runtime and memory manageable. We chose $m = 5000$ as a practical compromise: it is large

enough to make neighbourhood-based intrinsic-dimensionality estimates reasonably stable, yet small enough to keep the overall study tractable. We report significance stars for reference, but interpret the matrix primarily through its qualitative correlation structure.

Figure 13 reports the resulting pairwise Spearman correlations. The matrix shows a clear block structure consistent with our taxonomy. Nonlinear estimators form a looser cluster with positive within-family correlations, whereas TwoNN remains weakly related to most other metrics. Finally, CorrInt and FisherS align strongly with the global block (and less with the kNN/pointwise cluster), suggesting that in this setting they behave closer to global geometry summaries than purely local ID measures. The same analysis was performed on GPT-2 and can be found in our repository.

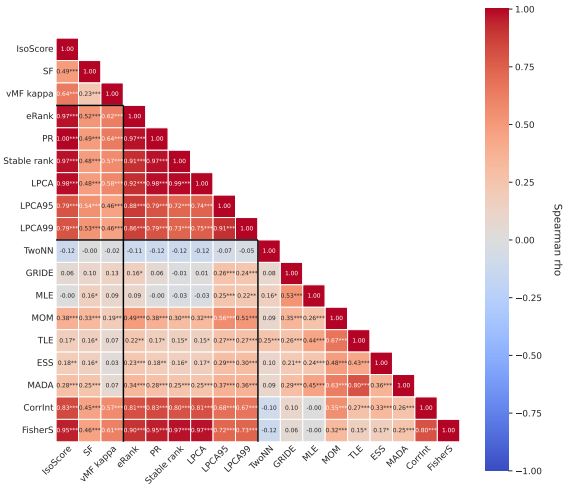


Figure 13: Spearman correlation ρ between geometric metrics. Thick separators delineate isotropy, linear intrinsic dimensionality, and nonlinear intrinsic dimensionality families.

Appendix D. Extra linguistic features

Relation type Figure 14 reports results by dependency relation, grouping words according to the ten most frequent types. *In both models*, the global shape of the curves is similar to what we observed for the other features: isotropy and both notions of intrinsic dimensionality increase from the input toward intermediate layers, then flatten and, especially in GPT-2, contract again in the top layers. *In the encoder*, the different relations are clearly separated. Punctuation (*punct*) is the most isotropic relation at almost all depths, while *obj* and *obl* also

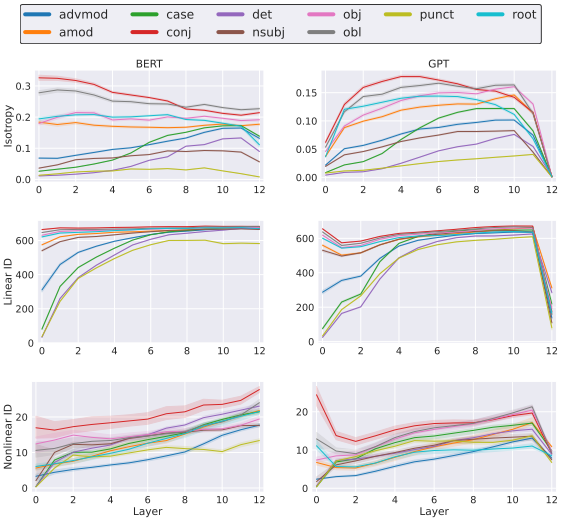


Figure 14: Encoder and Decoder layerwise geometry by tokens grouped by typed dependency relation.

become relatively isotropic and high-dimensional in the upper layers. By contrast, function-word relations that typically precede their heads, *det*, *case*, and to a lesser extent *amod* and *nsubj*, start out highly anisotropic and low-dimensional, but gain isotropy and dimensionality as the model integrates information from their heads. The *root* relation behaves differently again, since it maintains the smallest linear dimensionality and its isotropy declines toward the top, suggesting that predicates act as a kind of readout hub where sentence-level semantics are concentrated into a few dominant directions. These patterns align well with the corpus statistics in Table 6. *The decoder* shows a similar ordering of relations, but modulated by the causal architecture.

Arity Figure 15 groups tokens by *arity*, the number of syntactic dependents attached to a head up to 4. As for the previous features, *in both models* all three metrics exhibit a characteristic evolution with depth: isotropy and intrinsic dimensionality increase from the embedding layer into the middle layers, then plateau or slightly decline, with GPT-2 showing a sharper drop in the final layers. The interesting structure lies in the *relative* behaviour of the arity classes. Across almost all layers, arity 0 (leaf nodes, largely punctuation) forms a clear lower band. Classes with one or more dependents occupy progressively higher curves. In other words, heads that govern many dependents live in higher-dimensional, more "spread-out" regions of representation space than heads with few or no dependents. This ordering is consistent with

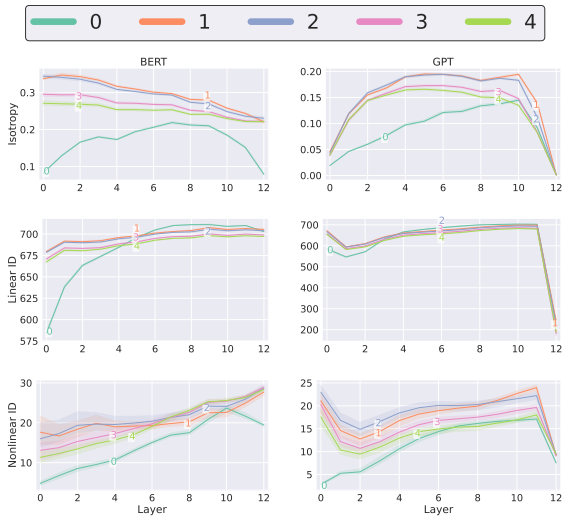


Figure 15: Encoder and Decoder layerwise geometry by arity.

the corpus statistics in Table 5. Arity 0 tokens are predominantly PUNCT and other closed-class items, with relatively few types and strongly constrained roles (mostly PUNCT and similar relations). By contrast, higher-arity heads (3–4) are mostly verbs functioning as clausal predicates (ROOT, OBJ, OBL), which are both lexically diverse and syntactically central. We hypothesize that such predicates must integrate information from multiple dependents, and the models appear to allocate more *degrees of freedom* to representing them: their representations occupy larger linear and nonlinear subspaces and are somewhat more isotropic.