Geometric Signatures of Compositionality Across a Language Model's Lifetime

Jin Hwa Lee*[1], Thomas Jiralerspong*[2,3], Lei Yu[4], Yoshua Bengio[2,3], Emily Cheng[5]

[1]UCL, [2]Université de Montréal, [3]Mila, [4]UToronto, [5]Universitat Pompeu Fabra

***Equal contribution, co-first authors.**

Correspondence: jin.lee.22@ucl.ac.uk and thomas.jiralerspong@mila.quebec and emilyshana.cheng@upf.edu

Abstract

By virtue of linguistic compositionality, few syntactic rules and a finite lexicon can generate an unbounded number of sentences. That is, language, though seemingly high-dimensional, can be explained using relatively few degrees of freedom. An open question is whether contemporary language models (LMs) reflect the intrinsic simplicity of language that is enabled by compositionality. We take a geometric view of this problem by relating the degree of compositionality in a dataset to the intrinsic dimension (I_d) of its representations under an LM, a measure of feature complexity. We find not only that the degree of dataset compositionality is reflected in representations' I_d, but that the relationship between compositionality and geometric complexity arises due to learned linguistic features over training. Finally, our analyses reveal a striking contrast between nonlinear I_d and linear dimensionality, showing they respectively encode semantic and superficial aspects of linguistic composition.

jinhl9/llm-compositionality-lifetime

1 Introduction

Compositionality, the notion that the meaning of an expression is constructed from that of its parts and syntactic rules (Szabó, 2024), enables linguistic atoms to locally combine to create global meaning (Frege, 1948; Chomsky, 1999). Then, a rich array of meanings at the level of a phrase may be explained by simple rules of composition. If a language model is a good model of language, we expect its internal representations to reflect the latter's relative simplicity. That is, representations should reflect the *manifold hypothesis*, or that real-life, high-dimensional data lie on a low-dimensional manifold (Goodfellow et al., 2016).

The dimension of this manifold, or *nonlinear intrinsic dimension* (I_d), is the minimal number of degrees of freedom needed to describe it without information loss (Goodfellow et al., 2016; Campadelli et al., 2015). The manifold hypothesis is attested for linguistic representations: LMs are found to compress inputs to an I_d orders-of-magnitude lower than their ambient dimension (Cai et al., 2021; Cheng et al., 2023; Valeriani et al., 2023).

A natural question is whether the inherent simplicity of linguistic utterances, enabled by compositionality, manifests in representation manifolds of low complexity, described by the manifolds' I_d. For example, consider the set of possible phrases generated by the syntactic structure `[DET][N][V]`. While the possible utterances have combinatorial complexity, they are generated by only three degrees of freedom. An LM that can handle this syntactic structure should be able to both (1) represent all combinations; (2) encode the few degrees of freedom underlying them, or their *degree of compositionality*. We ask, which geometric aspects of the representation space encode (1) and (2)?

The manifold hypothesis predicts the I_d of LM representations to encode (2) (Recanatesi et al., 2021). But, so far in the literature, while some linguistic categories have been found to occupy low-dimensional linear subspaces (Hernandez and Andreas, 2021; Mamou et al., 2020), an explicit link between degree of compositionality and representational dimensionality has not been established. To bridge this gap, in large-scale experiments on causal LMs and a custom dataset with tunable compositionality, we provide first insights into the relationship between the degree of compositionality of inputs and geometric complexity of their representations over the course of training.

Outline On both our controlled datasets and naturalistic stimuli, we reproduce the finding (Cai et al., 2021) that LMs represent linguistic inputs on low-dimensional nonlinear manifolds. We also show for the first time that LMs expand representations into high-dimensional *linear* subspaces, concretely, that

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5292–5320
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

(I) nonlinear I_d and linear dimensionality scale differently with model size. We show that feature geometry is relevant to behavior over training, in that **(II)** LMs' feature complexity tracks a phase transition in linguistic competence. Different from past work, we consider two different measures of dimensionality, nonlinear and linear, showing that they encode complementary information about compositionality. Our key result is that *representations' geometric complexity reflects dataset complexity*. However, *nonlinear* I_d encodes (2) meaningful compositionality while *linear* dimensionality encodes (1) superficial input complexity, in a way that arises over training: **(III)** nonlinear I_d reflects superficial complexity as an inductive bias of LM's architecture, but meaningful compositional complexity by the end of training. Instead, **(IV)** linear dimensionality, not I_d, encodes superficial complexity throughout training. Overall, results show a contrast between linear and nonlinear feature complexity that suggests their relevance to form and meaning in how LMs process language.

2 Background

Compositionality Compositionality is the notion that *the meaning of a complex expression can be constructed from that of its parts and how they are structured* (Frege, 1948; Partee, 1995). While compositionality, per this definition, is a general property of *language*, it spans various interpretations that we review here. In linguistic analysis, compositionality is typically measured at the level of a *single sample*: for instance, "silver table", whose meaning can be constructed intersectively from "silver" and "table", is more compositional than the idiom "silver spoon" (Labov and Weinreich, 1980). Beyond single instances, compositionality has been characterized at the *system level*, for instance, to measure the extent to which English realizes all possible word combinations (Sathe et al., 2024) or to compare the structures of different communication systems (Brighton and Kirby, 2006; Chaabouni et al., 2020; Elmoznino et al., 2025). Beyond language itself, researchers have considered what it means for *users or models of language* to be compositional (Hupkes et al., 2019) by evaluating compositionality in terms of model behavior (Lake and Baroni, 2018; Bahdanau et al., 2019) and representations (Smolensky, 1990; Andreas, 2019).

We are interested in a system-level notion of compositionality that is quantified on a linguistic dataset, further detailed in Section 3.1. Then, rather than explicitly describe the LM's composition function (Smolensky, 1990), we ask simply whether LMs preserve *inputs'* degree of compositionality in their representations' geometric complexity.

Manifold hypothesis and low-dimensional geometry Deep learning problems appear high-dimensional, but research suggests that they have low-dimensional intrinsic structure. In computer vision, image data and common learning objectives live on low-dimensional manifolds (Li et al., 2018; Valeriani et al., 2023; Psenka et al., 2024; Ansuini et al., 2019). Similarly, LMs' learning dynamics occur in low-dimensional parameter subspaces (Aghajanyan et al., 2021; Zhang et al., 2023). The parsimony that governs these models' latent spaces reduces learning complexity (Cheng et al., 2023; Pope et al., 2021) and may arise from the training objective, seen in classification (Chung et al., 2018) and sequential prediction (Recanatesi et al., 2021).

For language, the geometry of representations has been explored in several contexts. Prior work describes how concepts organize in representation space (Engels et al., 2025; Park et al., 2025; Balestriero et al., 2024; Doimo et al., 2024); representational geometry can encode tree-like syntactic structures (Andreas, 2019; Murty et al.; Alleman et al., 2021; Hewitt and Manning, 2019); and certain linguistic categories like part-of-speech lie in low-dimensional linear subspaces (Mamou et al., 2020; Hernandez and Andreas, 2021). Most similar to our work, Cheng et al. (2023) report the I_d of representations over layers as a measure of feature complexity for natural language datasets, finding an empirical link between information-theoretic and geometric compression. But, our work is the first to explicitly relate the compositionality of inputs, a critical feature of language, to the degrees of freedom, or I_d, of their representation manifold.

Language model training dynamics Most research on LMs focuses on their final configuration at the end of training. Yet, recent work shows that learning dynamics provide cues as to how behavior arises (Chen et al., 2024; Singh et al., 2024; Tigges et al., 2024). It has been found that, over training, LM weights become higher-rank (Abbe et al., 2023) and their gradients more diffuse (Weber et al., 2024). Phase transitions are attested for some, but not all, aspects of language learning in LMs. Negative evidence includes that circuits involved in linguistic tasks are stable (Tigges et al., 2024) and gradually reinforced (Weber et al., 2024) over training. Positive evidence includes that BERT's feature complexity tracks sudden syntax acquisition and drops in training loss (Chen et al., 2024), with similar findings for Transformers trained on formal languages (Lubana et al., 2025). Our work complements these results, exploring how the relationship between compositionality of inputs and complexity of their representations arises over training.

3 Setup

3.1 Compositionality of a system

We focus on a *system-level* view of compositionality (Sathe et al., 2024): the extent to which a language system expresses all possible compositional meanings given a vocabulary. Consider, for example, systems **A** and **B** with the same syntactic structure `[DET][ADJ][N][V]` and vocabulary items:

A	**B**
`the blue dog runs`	`the blue dog runs`
`the red cat runs`	`the red cat talks`
`the red dog talks`	`the blue dog talks`
`the blue cat talks`	`the red cat runs`

In **A**, all adjectives, nouns, and verbs can co-occur, while in **B**, some adjective-noun pairs (“blue cat" and “red dog") can never co-occur. Then, since more meanings can be created from the same vocabularies, we consider system **A** more compositional than system **B**.

We see that a system's degree of compositionality is intimately linked to its *degrees of freedom*. In the more compositional system **A**, there are three degrees of freedom: `[ADJ]`, `[N]` and `[V]`. In **B**, there are two: `[ADJ][N]` and `[V]`. If a language model adheres to the manifold hypothesis, then these degrees of freedom in the inputs should be reflected in their representation manifold. Then, we expect “system **A** is more compositional than **B**" to imply “under an LM, **A**'s representations are more complex than **B**'s". Testing this hypothesis requires controlled datasets with different degrees of compositionality, detailed in the next section, and a measure of feature complexity, see Section 3.4.

3.2 Datasets

Following the intuition outlined in the previous section, we design several datasets who differ solely in their degree of compositionality. We later replicate key findings on The Pile (Gao et al., 2020), a dataset of naturalistic language.

3.2.1 Controlled grammar

Our datasets contain sentences generated from a synthetic grammar. To create it, we set 12 semantic categories and uniformly sample a 50-word vocabulary for each category; categories' vocabularies are disjoint. We fix the syntactic structure by ordering the word categories:

The [quality$_1$.ADJ] [nationality$_1$.ADJ] [job$_1$.N] [action$_1$.V] the [size$_1$.ADJ] [texture.ADJ] [color.ADJ] [animal.N] then [action$_2$.V] the [size$_2$.ADJ] [quality$_2$.ADJ] [nationality$_2$.ADJ] [job$_2$.N].

This produces sentences that are 17 words long. The order is chosen so that generated sentences are grammatical and that the adjective order complies with the accepted ordering for English (Dixon, 1976). Vocabularies are chosen so that sentences are semantically coherent: for instance, if the agent is a person and patient is an animal, then verbs are constrained to permit, e.g., “pets", but not “types". To explore the effect of sequence length on feature complexity, we design four additional, similar grammars of lengths $\in \{5, 8, 11, 15\}$ words. Details for the sampling procedure, vocabularies, and syntactic structures are found in Appendix E.

Controlling compositionality We modify the grammar to vary the dataset's degree of compositionality. To do so, we couple the values of k contiguous word positions to produce a different dataset for each $k \in \{1 \cdots 4\}$. That is, during data generation, k-grams are independently sampled, which constrains the degrees of freedom in each sentence to l/k where l is the number of categories. For instance, for the longest grammar, there are l=12 categories; in the 1-coupled setting, each category is sampled independently, hence 12 degrees of freedom; in the 2-coupled setting, each bigram is sampled independently, hence 6 degrees of freedom. Crucially, varying k maintains the dataset's unigram distribution, only changing the dataset's k-gram distributions, or compositional complexity.

Since the sentences in our datasets are *literal* (form-meaning mappings are one-to-one), the trends we find when varying k could be attributed to superficial complexity differences in wordforms, rather than in meanings. To ablate this effect, for each k, we create a parallel dataset of “meaningless" inputs by randomly shuffling the words in each sequence. Shuffling preserves shallow distributional properties like length, sentence-level co-occurrences, and unigram frequencies (Figure E.1 right). Notably, shuffled datasets are semantically vacuous, but their superficial complexities are still ordered in k; trends observed on shuffled datasets can only be due to changes in superficial complexity and not meaningful compositional complexity.

If representations encode inputs' degree of compositionality, set by k and whether sequences are shuffled, then we expect: (1) smaller k, meaning more degrees of freedom in the dataset, implies *higher* feature complexity; (2) shuffling, which breaks semantic composition, entails *lower* feature complexity. For each case in $k \in \{1 \cdots 4\} \times \{\text{coherent, shuffled}\}$, we randomly sample 5 data splits of 10000 sequences, reporting over splits.

Measuring compositionality As linguistic structure permits compression, it is common to measure the complexity of linguistic data using *information-theoretic* quantities (Levy, 2008; Mollica and Piantadosi, 2019; Elmoznino et al., 2025). Along these lines, we quantify compositionality of a dataset by its Kolmogorov complexity (KC). In algorithmic information theory, KC is the length of the shortest program in bits needed to generate the dataset. Datasets that are more compositionally complex (lower k, more degrees of freedom) are less compressible, thus have higher KC. While the true KC is intractable (Cover and Thomas, 2006), we approximate it as others have (Jiang et al., 2023b), using the lossless compression algorithm `gzip`. We estimate KC for $k \in \{1 \cdots 4\} \times \{\text{coherent, shuffled}\}$ by the `gzip`-ped dataset size in kilobytes, then correlate it to feature complexity (Section 3.4) for each layer.

The connection between a dataset's KC and its degree of compositionality requires careful interpretation, as `gzip` is not semantics-aware: it only measures the compressibility of the superficial wordforms that make up the sequences. In coherent datasets, sequences are grammatical and literal; then, the superficial complexity measure returned by `gzip` directly proxies the degree of compositionality. For shuffled datasets, whose sequences are semantically vacuous, KC instead only measures wordforms' superficial combinatorial complexity.

3.2.2 The Pile

We focus on the synthetic grammar to tune compositionality in a surgical way. But, to ensure results are not an artifact of our prompts, we replicate experiments on The Pile, uniformly sampling 5 random splits of N=10000 sequences that are 17 words, the same length as sequences in the longest controlled grammar. We report results over splits.

3.3 Models

We consider Transformer-based causal LMs Llama-3-8B (hereon Llama) (Meta, 2024), Mistral-v0.1-7B (Mistral) (Jiang et al., 2023a), and Pythia models of sizes $\in$ {14m, 70m, 160m, 410m, 1.4b, 6.9b, 12b} (Biderman et al., 2023). Pythia is trained on The Pile, a natural language corpus spanning encyclopedic text, books, social media, code, and reviews (Gao et al., 2020); Llama and Mistral's training data are not public, but likely similar. Models are trained to predict the next token given context, subject to a negative log-likelihood loss.

Pre-training analysis Pythia's intermediate training checkpoints are public (Biderman et al., 2023). For three intermediate sizes 410m, 1.4b, and 6.9b, we report Pythia's performance throughout pre-training on the set of evaluation suites provided by (Biderman et al., 2023; Gao et al., 2024), detailed in Appendix F. This encompasses a range of higher-level linguistic and reasoning tasks, from long-range text comprehension (Paperno et al., 2016) to commonsense reasoning (Bisk et al., 2020). The evolution of task performance provides a cue for the type of linguistic knowledge learned by the LM by a certain training checkpoint.

3.4 Measuring feature complexity

We are interested in how feature complexity reflects an input dataset's degree of compositionality. We consider representations in the Transformer's *residual stream* (Elhage et al., 2021). As sequence lengths can slightly vary due to tokenization, as in prior work (Cheng et al., 2023; Doimo et al., 2024), we aggregate over the sequence by taking the last token representation, as, due to causal attention, it is the only one to attend to the entire context.

For each layer and dataset, we measure feature complexity using both a nonlinear and a linear measure of dimensionality. Nonlinear and linear measures have key differences. The nonlinear I_d is the number of degrees of freedom, or latent features, needed to describe the underlying manifold (Campadelli et al., 2015) (see Appendix C). This differs from the *linear* effective dimension d, the dimension of the minimal linear subspace that explains representations' variance. We will use *dimension-*

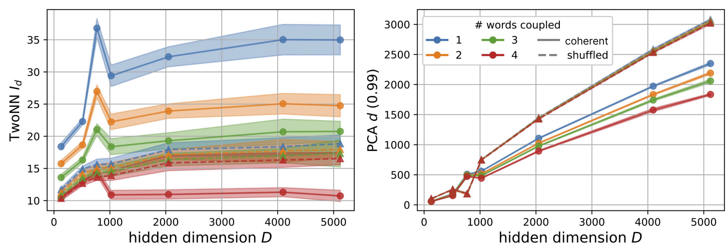


Figure 1: **Dimensionality over model size.** Mean layerwise I_d (left) and d (right) are shown for increasing hidden dimension. I_d does not depend on hidden dimension D (flat lines), but PCA d scales linearly in D. Curves are averaged over 5 data splits $\pm$ 1 SD.

ality to refer to both nonlinear and linear estimates, specifying, when needed, I_d as the nonlinear, d as the linear, and D as the LM's hidden dimension.

Intrinsic dimension We report the nonlinear I_d using the TwoNN estimator (Facco et al., 2017). We choose TwoNN for several reasons: (1) it is highly correlated to other state-of-the-art estimators (Cheng et al., 2023; Facco et al., 2017); (2) it relies on minimal assumptions of local uniformity up to the second nearest neighbor of a point, in contrast to other methods that impose stricter assumptions like global uniformity (Albergante et al., 2019); (3) TwoNN and correlated estimators are often used in the manifold estimation literature (Cheng et al., 2023; Pope et al., 2021; Chen et al., 2024; Tulchinskii et al., 2023; Ansuini et al., 2019). We focus on TwoNN in the main text, but test another estimator, Levina and Bickel, 2004's MLE, in Appendix D, confirming it is highly correlated.

TwoNN works as follows. Points on the underlying manifold are assumed to follow a locally homogeneous Poisson point process. Here, local refers to the neighborhood about each point x, up to x's second nearest neighbor. Let $r_k^{(i)}$ be the Euclidean distance between x_i and its kth nearest neighbor. Then, under the stated assumptions, distance ratios $\mu_i \triangleq r_2^{(i)}/r_1^{(i)} \in [1, \infty)$ follow the cumulative distribution function $F(\mu) = (1 - \mu^{-I_d})\mathbf{1}[\mu \geq 1]$. This yields the estimator $I_d = -\log(1 - F(\mu))/\log \mu$. Lastly, given representations $\{x_i^{(j)}\}_{i=1}^N$ for LM layer j, $I_d^{(j)}$ is fit via maximum likelihood estimation over all datapoints.

Linear effective dimension To estimate the linear effective dimension d, we use Principal Component Analysis (PCA) (Jolliffe, 1986) with a variance cutoff of 99%. We focus on PCA in the main text; for other linear methods see Appendix D.

4 Results

We find representational dimensionality to systematically reflect compositionality over pre-training and model scale. We show that LMs represent linguistic data on low-dimensional, nonlinear manifolds, but in high-dimensional linear subspaces that scale with the hidden dimension. Then, we show that, over training, feature complexity tracks an LM's linguistic competence, where both exhibit a phase transition marking the emergence of syntactic and semantic abilities. Finally, we show that feature complexity reflects degree of compositionality, but that nonlinear I_d encodes meaningful compositional complexity while linear d encodes superficial complexity. We focus on Pythia 410m, 1.4b, and 6.9b here, with full results in the Appendix.

4.1 Nonlinear and linear feature complexity scale differently with model size

We find, in line with Cai et al. (2021); Cheng et al. (2023), that LMs represent data on nonlinear manifolds of I_d orders-of-magnitude lower than the embedding dimension. For the controlled dataset (Figure 1) and The Pile (Figure H.1), I_d is $O(10)$ while linear d and D are $O(10^3)$.

Our novel finding is that d scales linearly in LM hidden dimension D, while I_d is robust to it. We fit linear regressions from $\langle d \rangle_{\text{layer}}$ and $\langle I_d \rangle_{\text{layer}}$ to D for each setting in $k \in \{1 \cdots 4\} \times \{\text{coherent, shuffled}\}$. In all cases, $d \propto D$, see Figure 1 (right); all show a highly significant linear fit with $R > 0.99$ and $p < 0.005$ (Table G.1). Instead, I_d stabilizes to a low range $O(10)$ regardless of D, see Figure 1 (left): here, effect sizes $\alpha \approx 0$, and fits are not significant (Table G.1). Results hold for The Pile, see Appendix H for details.

These results highlight key differences in how linear and nonlinear dimensions are recruited: LMs *globally* distribute representations to occupy $d \propto D$ linear dimensions of the space, but their shape is *locally* constrained to a low-dimensional (I_d) manifold. Robustness of I_d to model size (D) suggests that LMs, once sufficiently performant, recover the degrees of freedom, or compositional complexity, underlying their inputs.

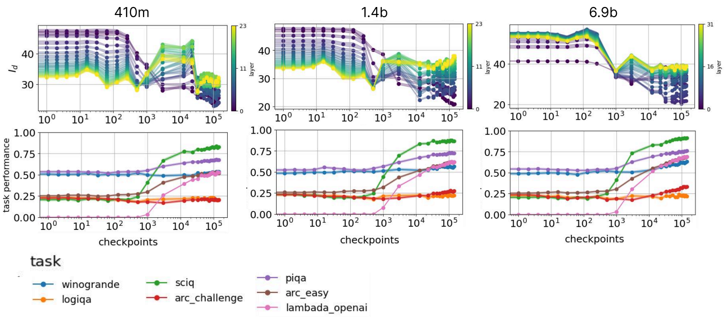


Figure 2: I_d **tracks task performance**. **Top:** Layerwise I_d over pre-training for Pythia-410m, 1.4b, and 6.9b. The phase transition of I_d at $t \approx 10^3$ holds across model size. **Bottom:** Zero-shot task performance across pre-training. Also at $t \approx 10^3$, linguistic competence measured by task performance starts to increase for all LMs.

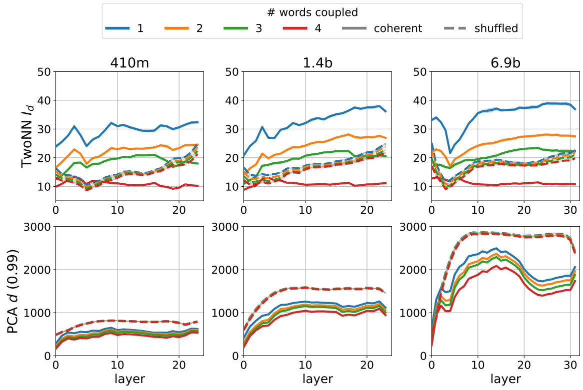


Figure 3: **Dimensionality over layers.** I_d (top) and d (bottom) over layers for Pythia 410m, 1.4b, 6.9b (left to right). Each color is a coupling factor $k \in 1 \cdots 4$. Solid lines denote coherent, and dotted curves denote shuffled, text. For all LMs, lower k implies higher I_d and d in both coherent and shuffled cases. For all LMs, shuffling collapses the I_d but raises d. Curves averaged over 5 splits ± 1 SD (shaded, SDs very small).

4.2 Feature complexity tracks emergent linguistic abilities over training

Complex linguistic abilities require, by definition, a compositional understanding of language. We use linguistic task performance (Biderman et al., 2023) as a cue for compositional understanding, finding that over training, models' feature complexity closely tracks the onset of linguistic capabilities.

Figure 2 shows the evolution of I_d on the k=1 dataset (top), where each curve is one layer, with the evolution of LM benchmark performance (bottom), where each curve plots performance on an individual task. For all LMs, the evolution of I_d closely tracks a sudden transition in task performance. In Figure 2 top, we first observe that I_d decreases sharply before checkpoint 10^3 and then redistributes. At the same time $t \approx 10^3$, task performance rapidly improves (Figure 2 bottom). Feature complexity evolution on The Pile is shown in Figure H.2, exhibiting a similar transition to that shown in Figure 2. The existence of the phase transition in representational geometry $t \approx 10^3$ is robust to the dimensionality measure, and whether the data are shuffled, see Figure G.5. Our results resonate with Chen et al. (2024), who observed in BERT a similar I_d transition on the training corpus that coincided with the onset of higher-order linguistic capabilities. Crucially, we show that the phase transition in feature geometry exists for natural language inputs *beyond* in-distribution data, which was the subject of Chen et al., and, further, beyond grammatical data (Figure G.5) as a more general property of LM processing. Taken together, results show that feature complexity can signify when LMs gain linguistic abilities that require, by definition, compositional understanding.

4.3 Feature complexity reflects input compositionality

We just established that feature complexity is informative of when models gain linguistic behaviors that require compositional understanding. Now, we establish our key result, which is that, no matter the model or grammar, **feature complexity encodes input compositionality**. We first show that this holds for fully-trained models that have reached final linguistic competency. Then, using evidence from the LM's training phase, we show that the correspondence between feature complexity and input compositionality is present first as an inductive bias of the architecture that encodes superficial input complexity; but then, that it persists due to learned features that encode compositional meaning.

Compositional complexity varying k On all fully-trained models and all datasets, representational dimensionality preserves relative dataset compositionality. Figure 3 shows for fully-trained Pythia 410m, 1.4b, and 6.9b that I_d and d increase with the compositionality of coherent inputs (solid curves): the highest curves (blue) correspond to the $k = 1$ dataset, or 12 degrees of freedom, and the lowest (red) denote the $k = 4$ dataset, or 3 degrees of freedom. The relative order of feature complexity, moreover, holds for all layers, seen by non-overlapping solid curves in Figure 3.

Grammaticality is not a precondition for representational dimensionality to reflect superficial complexity of the data: in Figure 3 (top), dashed curves corresponding to shuffled text are also ordered $k = 1 \cdots 4$ top to bottom. While the relative order of superficial complexity is preserved in the LM's feature complexity for both grammatical and agrammatical datasets, the separation is greater for grammatical text (solid curves). We hypothesize that this is due to shuffled text being out-of-distribution, such that the model cannot integrate the sequences' meaning, but nevertheless preserve superficial complexity in its representations. This tendency holds for pre-trained LMs of all sizes and families (see Figures G.1 and G.2) and for different-length sentences (see Figure J.3).

The relationship between dimensionality and superficial complexity, controlled by k, for coherent text is *not* an emergent feature over training. In Figure 4 (left), the inverse relationship between k and both I_d and d is present throughout training. But, the reason for this relation differs at the start and end: in shuffled text, where phrase-level semantics is ablated, the inverse relationship between k and dimensionality exists at the *start* and greatly diminishes by the end of training, while in coherent text it is salient throughout. Together, results show an inductive bias of the LM's architecture to preserve input complexity in its representations. Then, over training, differences in dimensionality may be increasingly explained by features beyond surface variation of inputs. We claim these features are semantic and provide evidence in what follows.

Breaking compositional structure Shuffling sequences destroys their meaning, removing dataset complexity attributed to sentence-level semantics. Figure 3 shows, for fully-trained models, dimensionality over layers for coherent and shuffled inputs. Here, nonlinear I_d and linear d show opposing trends: I_d for shuffled text *collapses* to a low range, while d *increases*, seen by the dashed curves in each plot compared to solid curves. This trend is robust to sentence length (Figure J.3) and model size (Figure G.1) and family (Figure G.2).

We refer to the phenomenon where shuffling destroys phrase-level semantics and I_d, also attested for naturalistic sentences (Cheng et al., 2025), as *shuffling feature collapse*. Evidence from train-

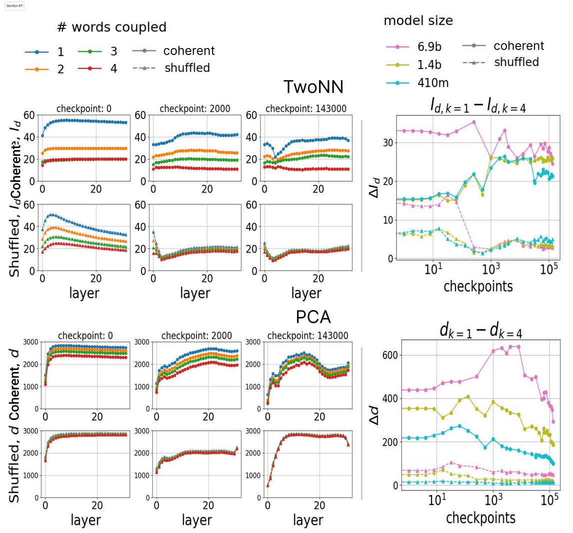


Figure 4: **Training dynamics of dimensionality.** (**Top**) TwoNN I_d; (**Bottom**) PCA d. **(Left):** For Pythia-6.9b, layerwise I_d for 3 training checkpoints for coherent vs. shuffled sequences, with different coupling k. The I_d difference on shuffled data with varying k diminishes as training persists. All curves are shown with ± 1 SD (SDs are very small). **(Right):** ΔI_d between $k = 1, k = 4$ across training for various model sizes (different colors).

ing dynamics further suggests that this feature collapse is due to semantics. We saw in Section 4.2 that training step $t = 10^3$ approximately marked a phase transition after which the LM's linguistic competencies sharply rose. Crucially, the epoch $t = 10^3$ preceding the sharp increase in linguistic capabilities is also the first to exhibit shuffling feature collapse. Figure 4 (right) shows the ΔI_d between the $k = 1$ and $k = 4$ dataset for several model sizes, across training (x-axis). Shuffling feature collapse, given by low ΔI_d, occurs around $t = 10^3$ for all models. On the other hand, ΔI_d for coherent text stabilizes to around 25 for different model sizes. This transition does not occur for linear d, see Figure 4 (right, bottom). This suggests that shuffling feature collapse for I_d is symptomatic of when the LM learns to extract meaningful semantic features. We investigate why shuffling feature collapse exists for I_d, but not for d, in the following section. For how shuffling feature collapse emerges with sequence length, see Appendix J.

5 Form-meaning dichotomy in representation learning

We interpret shuffling feature collapse using an argument from Recanatesi et al. (2021), who propose that predictive coding (i.e. the theory that the

Spearman ρ	14m	70m	160m	410m	1.4b	6.9b	12b	Llama	Mistral
I_d	-0.20	-0.06	-0.20	-0.05	0.04	0.01	0.05	-0.36	0.00
d	0.90*	0.47†	-0.50†	0.96*	0.96*	0.92*	0.86*	1.0*	1.0*

Table 1: **Spearman correlations between dimensionality and estimated Kolmogorov complexity**. Spearman correlations ρ between the gzipped dataset size (KB) and representational dimensionality (rows), averaged over layers, for all **Pythia** sizes (left 7 columns) and **Llama** and **Mistral** (right 2 columns). Values with * are significant at a p-value threshold of 0.05, and † at 0.1. Across LMs, average-layer I_d is not correlated to the estimated KC, or surface complexity, of datasets. Linear d is highly correlated to estimated KC, except the outlier 160m.

goal of any intelligent agent is to minimize prediction errors between what it expects and what is actually perceived) requires a model to satisfy two objectives: (1) encode the vast "input and output space", exerting upward pressure on feature complexity, while (2) extracting latent features to support prediction, exerting a downward pressure on complexity. Recanatesi et al. (2021) argue that pressure (1) expands the linear representation space (d), while (2) compresses representations to a I_d-dimensional manifold. Recall that we quantify a dataset's superficial complexity by its KC. We saw that shuffling words, which increases inputs' KC, also increases d, but that it destroys compositional semantics, collapsing I_d (Figure 3). These results suggest a dichotomy in what linear and nonlinear complexity encode, the former capturing superficial complexity (form), and the latter, latent compositional structure (meaning).

We hypothesize d to capture surface variation, and I_d to encode meaningful compositional variation. We saw the latter in the previous section: I_d collapses to a narrow range in the absence of compositional semantics while d does not, suggesting I_d, not d, encodes compositionality. Now, we show that d, not I_d, encodes superficial complexity. Table 1 shows Spearman correlations ρ between KC (estimated with `gzip`) and dimensionality. For all LMs, `gzip` highly correlates to average layerwise d but not to I_d; we discuss the outlier Pythia-160m in Appendix I. The high $\rho(d, \texttt{gzip})$ is surprisingly consistent across layers (Figures I.2 and I.3) and already present as an inductive bias of the architecture (Figure I.5, Appendix I for training discussion). Instead, the correlation to I_d is seldom significant for all of training. This suggests surface complexity, already present in the LM's inputs, is *preserved* by the architecture, while meaning complexity is instead *learned* over training.

6 Discussion

On extensive experiments spanning degree of compositionality, sequence length, model size and training time, we found strong links between the compositionality of linguistic inputs and the complexity of their representations. We asked what aspects of feature complexity encode inputs' (1) surface variation and (2) compositional structure, finding that linear d tracks (1) and nonlinear I_d tracks (2). We showed feature complexity to encode inputs' superficial complexity as an inductive bias of the architecture; but that, over training, the bias is shed to reflect meaningful linguistic compositionality. This change marks an onset in linguistic competence, a cue for compositional understanding.

A central throughline in our results is that LMs represent their inputs on low-dimensional nonlinear manifolds, yet, at the same time, in high-dimensional linear subspaces. This tendency suggests a solution to the curse of dimensionality that also enjoys its blessings. High-dimensional representations classically engender overfitting and poor generalization (Hughes, 1968); but, they may form low-dimensional manifolds that capture the task's latent, in our case, *compositional*, structure (De and Chaudhuri, 2023). At the same time, higher linear dimensionality has been shown to improve generalization thanks to its expressivity (Cohen et al., 2020; Elmoznino and Bonner, 2023; Sorscher et al., 2022). Intriguingly, the dual patterning of intrinsic and effective dimension, where latent structure is captured by low-I_d manifolds that live in high-d linear spaces, has been observed in the neuroscience literature for both biological and artificial systems (Jazayeri and Ostojic, 2021; Recanatesi et al., 2021; Haxby et al., 2011; Huth et al., 2012; De and Chaudhuri, 2023; Manley et al., 2024). That LMs also conform to this picture suggests that representation spaces that are both highly expressive (d) yet geometrically constrained (I_d) may be desirable in intelligent systems that recover compositional structure from data.

Limitations

- While the present work is the first to show that nonlinear and linear representational dimension in LMs correspond to a form-meaning dichotomy, further work is needed to disentangle how nonlinear and linear features causally contribute to predictive coding. This would require careful probing and feature attribution.
- Due to computational constraints, the present work is limited to several syntactic structures. Future work can apply this framework to investigate other structures like recursive embedding.
- Also due to computational constraints, the present work is limited to model sizes of up to 8B parameters. We encourage further, better-resourced research to apply our framework to larger models.
- Dimensionality measures tell us how complex the features are, but not *what* they are. How to isolate and interpret nonlinear features remains an open problem in the literature.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101019291). This paper reflects the authors' view only, and the funding agency is not responsible for any use that may be made of the information it contains. JHL thanks the Gatsby Charitable Foundation (GAT3755) and The Wellcome Trust (219627/Z/19/Z). TJ is funded by FRQNT. EC, JHL and TJ thank the Center for Brain, Minds and Machines for hosting the collaboration.

The authors thank Marco Baroni, Noga Zaslavsky, Erin Grant, Francesca Franzon, Alessandro Laio, Andrew Saxe, and members of the COLT group at Universitat Pompeu Fabra for valuable feedback.

Author Contributions

EC and LY contributed the idea of exploring ID, compression, and linguistic compositionality in the static setting. JHL greatly expanded this idea to incorporate investigation of dynamics and evolution of the features of interest over training. EC, JHL and TJ created the grammar, implemented and ran experiments; TJ led the static, and JHL led the dynamic, comparison of geometry to compositionality. EC helped with computation, did Kolmogorov complexity experiments with TJ, and led the writing. All authors contributed to the manuscript.

References

Emmanuel Abbe, Samy Bengio, Enric Boix-Adserà, Etai Littwin, and Joshua M. Susskind. 2023. Transformers learn through gradual rank increase. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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.

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)*, page 1–8.

Matteo Alleman, Jonathan Mamou, Miguel A Del Rio, Hanlin Tang, Yoon Kim, and SueYeon Chung. 2021. Syntactic perturbations reveal representational correlates of hierarchical phrase structure in pretrained language models. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 263–276, Online. Association for Computational Linguistics.

Jacob Andreas. 2019. Measuring compositionality in representation learning. In *International Conference on Learning Representations*.

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

Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron Courville. 2019. Systematic generalization: What is required and can it be learned? In *International Conference on Learning Representations*.

Randall Balestriero, Romain Cosentino, and Sarath Shekkizhar. 2024. Characterizing large language model geometry helps solve toxicity detection and generation. In *Forty-first International Conference on Machine Learning*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Henry Brighton and Simon Kirby. 2006. Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artificial Life*, 12(2):229–242.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations*.

P. Campadelli, E. Casiraghi, C. Ceruti, and A. Rozza. 2015. Intrinsic dimension estimation: Relevant techniques and a benchmark framework. *Mathematical Problems in Engineering*, 2015:e759567.

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. 2020. Compositionality and generalization in emergent languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4427–4442, Online. Association for Computational Linguistics.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L Leavitt, and Naomi Saphra. 2024. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In *The Twelfth International Conference on Learning Representations*.

Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Jade 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*.

Emily Cheng, Corentin Kervadec, and Marco Baroni. 2023. Bridging information-theoretic and geometric compression in language models. In *Proceedings of EMNLP*, pages 12397–12420, Singapore.

Noam Chomsky. 1999. Derivation by phase.

SueYeon Chung, Daniel D. Lee, and Haim Sompolinsky. 2018. Classification and geometry of general perceptual manifolds. *Phys. Rev. X*, 8:031003.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Uri Cohen, SueYeon Chung, Daniel D. Lee, and Haim Sompolinsky. 2020. Separability and geometry of object manifolds in deep neural networks. *Nature Communications*, 11(1):746.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience.

Anandita De and Rishidev Chaudhuri. 2023. Common population codes produce extremely nonlinear neural manifolds. *Proceedings of the National Academy of Sciences*, 120(39):e2305853120.

Robert Mw Dixon. 1976. Iwhere have all the adjectives gone. *Studies in Language*, 1:19–80.

Diego Doimo, Alessandro Serra, Alessio Ansuini, and Alberto Cazzaniga. 2024. The representation landscape of few-shot learning and fine-tuning in large language models. In *Advances in Neural Information Processing Systems*, volume 37, page 18122–18165. Curran Associates, Inc.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Eric Elmoznino and Michael F. Bonner. 2023. High-performing neural network models of visual cortex benefit from high latent dimensionality. *PLOS Computational Biology*, 20.

Eric Elmoznino, Thomas Jiralerspong, Yoshua Bengio, and Guillaume Lajoie. 2025. A complexity-based theory of compositionality. In *International Conference on Machine Learning*.

Joshua Engels, Isaac Liao, Eric J. Michaud, Wes Gurnee, and Max Tegmark. 2025. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations*.

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.

Gottlob Frege. 1948. Ueber sinn und bedeutung. *Philosophical Review*, 57(n/a):209.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Peiran Gao, Eric M. Trautmann, Byron M. Yu, Gopal Santhanam, Stephen I. Ryu, Krishna V. Shenoy, and Surya Ganguli. 2017. A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv*.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

James V. Haxby, J. Swaroop Guntupalli, Andrew C. Connolly, Yaroslav O. Halchenko, Bryan R. Conroy, Maria Ida Gobbini, Michael Hanke, and Peter J. Ramadge. 2011. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72:404–416.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

G. Hughes. 1968. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 14(1):55–63.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2019. Compositionality decomposed: How do neural networks generalise? *J. Artif. Intell. Res.*, 67:757–795.

Alexander G. Huth, Shinji Nishimoto, An T. Vu, and Jack L. Gallant. 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76:1210–1224.

Mehrdad Jazayeri and Srdjan Ostojic. 2021. Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. *Current Opinion in Neurobiology*, 70:113–120. Computational Neuroscience.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *Preprint*, arXiv:2310.06825.

Zhiying Jiang, Matthew Yang, Mikhail Tsirlin, Raphael Tang, Yiqin Dai, and Jimmy Lin. 2023b. "low-resource" text classification: A parameter-free classification method with compressors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6810–6828, Toronto, Canada. Association for Computational Linguistics.

Ian Jolliffe. 1986. *Principal Component Analysis*. Springer.

William Labov and Beatrice S. Weinreich. 1980. Problems in the analysis of idioms.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Elizaveta Levina and Peter Bickel. 2004. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

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

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2021. Logiqa: a challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Ekdeep Singh Lubana, Kyogo Kawaguchi, Robert P Dick, and Hidenori Tanaka. 2025. A percolation model of emergence: Analyzing transformers trained on a formal language. In *The Thirteenth International Conference on Learning Representations*.

Anemily Machina and Robert Mercer. 2024. Anisotropy is not inherent to transformers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4892–4907, Mexico City, Mexico. Association for Computational Linguistics.

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. In *Proceedings of the 37th International Conference on Machine Learning*, page 6713–6723. PMLR.

Jason Manley, Sihao Lu, Kevin Barber, Jeff Demas, Hyewon Kim, David Meyer, Francisca Martínez Traub, and Alipasha Vaziri. 2024. Simultaneous, cortex-wide dynamics of up to 1 million neurons reveal unbounded scaling of dimensionality with neuron number. *Neuron*, 112:1694–1709.e5.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date.

Francis Mollica and Steven T Piantadosi. 2019. Humans store about 1.5 megabytes of information during language acquisition. *Royal Society open science*, 6(3):181393.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher D Manning. Characterizing intrinsic compositionality in transformers with tree projections. In *The Eleventh International Conference on Learning Representations*.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. 2025. The geometry of categorical and hierarchical concepts in large language models. In *The Thirteenth International Conference on Learning Representations*.

Barbara H. Partee. 1995. *Lexical semantics and compositionality*, page 311–360. An invitation to cognitive science. The MIT Press, Cambridge, MA, US.

Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. 2021. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*.

Michael Psenka, Druv Pai, Vishal Raman, Shankar Sastry, and Yi Ma. 2024. Representation learning via manifold flattening and reconstruction. *Journal of Machine Learning Research*, 25(132):1–47.

Giovanni Puccetti, Anna Rogers, Aleksandr Drozd, and Felice Dell'Orletta. 2022. Outlier dimensions that disrupt transformers are driven by frequency. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1286–1304, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Stefano Recanatesi, Matthew Farrell, Guillaume Lajoie, Sophie Deneve, Mattia Rigotti, and Eric Shea-Brown. 2021. Predictive learning as a network mechanism for extracting low-dimensional latent space representations. *Nature Communications*, 12(1):1417.

William Rudman, Catherine Chen, and Carsten Eickhoff. 2023. Outlier dimensions encode task specific knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14596–14605, Singapore. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Aalok Sathe, Evelina Fedorenko, and Noga Zaslavsky. 2024. Language use is only sparsely compositional: The case of english adjective-noun phrases in humans and large language models. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46(0).

Aaditya Singh, Stephanie Chan, Ted Moskovitz, Erin Grant, Andrew Saxe, and Felix Hill. 2024. The transient nature of emergent in-context learning in transformers. *Advances in Neural Information Processing Systems*, 36.

Paul Smolensky. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, 46(1-2):159–216.

Ben Sorscher, Surya Ganguli, and Haim Sompolinsky. 2022. Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences*, 119(43):e2200800119.

Zoltán Gendler Szabó. 2024. *Compositionality*, fall 2024 edition. Metaphysics Research Lab, Stanford University.

Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. 2024. LLM circuit analyses are consistent across training and scale. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eduard Tulchinskii, Kristian Kuznetsov, Kushnareva Laida, Daniil Cherniavskii, Sergey Nikolenko, Evgeny Burnaev, Serguei Barannikov, and Irina Piontkovskaya. 2023. Intrinsic dimension estimation for robust detection of AI-generated texts. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio ansuini, and Alberto Cazzaniga. 2023. The geometry of hidden representations of large transformer models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Lucas Weber, Jaap Jumelet, Elia Bruni, and Dieuwke Hupkes. 2024. Interpretability of language models via task spaces. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4522–4538, Bangkok, Thailand. Association for Computational Linguistics.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Zhong Zhang, Bang Liu, and Junming Shao. 2023. Fine-tuning happens in tiny subspaces: Exploring intrinsic task-specific subspaces of pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1701–1713, Toronto, Canada. Association for Computational Linguistics.

A Computing resources

All experiments were run on a cluster with 12 nodes with 5 NVIDIA A30 GPUs and 48 CPUs each. Extracting LM representations took a few wall-clock hours per model-dataset computation. I_d computation took approximately 0.5 hours per model-dataset computation. Taking parallelization into account, we estimate the overall wall-clock time taken by all experiments, including failed runs, preliminary experiments, etc., to be of about 10 days.

B Assets

Pile `https://huggingface.co/datasets/NeelNanda/pile-10k`; license: bigscience-bloom-rail-1.0

Pythia `https://huggingface.co/EleutherAI/pythia-6.9b-deduped`; license: apache-2.0

scikit-dimension `https://scikit-dimension.readthedocs.io/en/latest/`; license: bsd-3-clause

PyTorch `https://scikit-learn.org/`; license: bsd

C Intrinsic Dimension Details

I_d estimation methods practically rely on a finite set of points and their nearest-neighbor structure in order to compute an estimated dimensionality value. The underlying geometric calculations assume that these are points sampled from a continuum, such as a lower-dimensional non-linear manifold. In our case, we actually have a discrete set of points so the notion of an underlying manifold is not strictly applicable. However, we can ask the question: if those points had been sampled from a manifold, what would the estimated I_d be? Since the algorithms themselves only require a discrete set of points, they can be used to answer that question.

D Other Dimensionality Estimators

Maximum Likelihood Estimator In addition to TwoNN, we considered Levina and Bickel (2004)'s Maximum Likelihood Estimator (MLE), a similar, nonlinear measure of I_d. MLE has been used in prior works on representational geometry such as (Cai et al., 2021; Cheng et al., 2023; Pope et al., 2021), and similarly models the number of points in a neighborhood around a reference point x to follow a Poisson point process. For details we refer to the original paper (Levina and Bickel, 2004). Like past work (Facco et al., 2017; Cheng et al., 2023), we found MLE and TwoNN to be highly correlated, producing results that were nearly identical: compare Figure 1 left to Figure G.3 left, and Figure G.1 top to Figure G.4 top.

Participation Ratio For our primary linear measure of dimensionality d, we computed PCA and took the number of components that explain 99% of the variance. In addition to PCA, we computed the Participation Ratio (PR), defined as $(\sum_i \lambda_i)^2/(\sum_i \lambda_i^2)$ (Gao et al., 2017). We found PR to give results that were incongruous with intuitions about linear dimensionality. In particular, it produced a lower dimensionality estimate than the nonlinear estimators we tested; see, e.g., Figure G.3, where the PR-d for coherent text is less than that of TwoNN. This contradicts the mathematical relationship that $I_d \leq d \leq D$. This may be because, empirically, PR-d corresponded to explained variances of $60 - 80\%$, which are inadequate to describe the bounding linear subspace for the representation manifold. Therefore, while we report the mean PR-d over model size in Figure G.3 and the dimensionality over layers in Figure G.4 for completeness, we do not attempt to interpret them.

E Controlled Grammar

Here, we describe the sampling procedure of the controlled grammar in more detail.

E.1 Sampling Procedure

Let there be l variable words in the sequence (these are the colored words in the grammar in Section 3.1). Each position $i = 1 \cdots l$ is associated with a vocabulary $\mathcal{V}^{(i)}$ which contains vocabulary items $v_j^{(i)}$, $j = 1 \cdots |\mathcal{V}^{(i)}|$. We set $|\mathcal{V}^{(i)}| = 50$ for all i. All words in $\mathcal{V}^{(i)}$ are listed in the below Table.

We state the sampling procedure, illustrated in Figure E.1, as follows.

- $\mathbf{k = 1}$ For each position i in the sentence, sample the word $w_i \sim \text{Unif}(\mathcal{V}^{(i)})$.
- $\mathbf{k = 2}$ Couples the vocabularies over bigrams in the sentence. Let $\circ$ be a concatenation operator. We construct a new *coupled* vocabulary $\mathcal{V}^{(i,i+1)}$ consisting of bigrams $v_j^{(i,i+1)} \triangleq v_j^{(i)} \circ v_j^{(i+1)}$, $j = 1 \cdots |\mathcal{V}^{(i)}|$. Then, for $i = 1, 3, 5, \cdots l-1$, sample bigrams $w_i \circ w_{i+1} \sim \text{Unif}(\mathcal{V}^{(i,i+1)})$.
- $\mathbf{k = 3}$ Couples trigrams in the sentence. Similarly, a new vocabulary over trigrams is constructed: $\mathcal{V}^{(i,i+1,i+2)}$ consists of trigrams $v_j^{(i,i+1,i+2)} := v_j^{(i)} \circ v_j^{(i+1)} \circ v_j^{(i+2)}$. For $i = 1, 4, 7, \cdots l-2$, sample trigrams $w_i w_{i+1} w_{i+2} \sim \text{Unif}(\mathcal{V}^{(i,i+1,i+2)})$...

and so on.

Although the LM is likely exposed to aspects of the syntactic structure and vocabulary items during training, primitives are sampled independently without considering their probability in relationship to others in the sentence. Therefore, generated sentences are highly unlikely to be in the training data, thus memorized by the LM.[1]

We design 5 different grammars of varying lengths $l \in \{5, 8, 11, 15, 17\}$ words.. The 17-word grammar is the one used for all controlled grammar experiments except the "Varying Sequence Length" experiments (Appendix J). The structures of the grammars can be found below.

[1]We can't verify utterances aren't in the training set, as at the time of submission, it is not possible to search The Pile.

E.2 Length: 5 words

The [job$_1$.N] [action$_1$.V] the [animal.N].

E.3 Length: 8 words

The [nationality$_1$.ADJ] [job$_1$.N] [action$_1$.V] the [color.ADJ] [texture.ADJ] [animal.N]

E.4 Length: 11 words

The [size$_2$.ADJ] [quality$_1$.ADJ] [nationality$_1$.ADJ] [job$_1$.N] [action$_1$.V] the [size$_1$.ADJ] [color.ADJ] [texture.ADJ] [animal.N]

E.5 Length: 15 words

The [quality$_1$.ADJ] [nationality$_1$.ADJ] [job$_1$.N] [action$_1$.V] the [size$_1$.ADJ] [color.ADJ] [texture.ADJ] [animal.N] then [action$_2$.V] the [size$_2$.ADJ][job$_2$.N].

E.6 Length: 17 words

The [quality$_1$.ADJ] [nationality$_1$.ADJ] [job$_1$.N] [action$_1$.V] the [size$_1$.ADJ] [texture.ADJ] [color.ADJ] [animal.N] then [action$_2$.V] the [size$_2$.ADJ] [quality$_2$.ADJ][nationality$_2$.ADJ][job$_2$.N].

Each category, colored and enclosed in brackets, is sampled from a vocabulary of 50 possible words, listed in the table on the following page.

The 50 words for each category were selected by prompting GPT4 with appropriate prompts. For example, "Please generate 50 nouns referring to jobs" or "Please generate 50 verbs referring to actions that a human can do to an animal". When there were 2 sets of words for a word category (for example action$_1$ and action$_2$), we provided GPT4 with the previous 50 words and prompted it appropriately. For example, "Please generate 50 different verbs referring to actions that a human can do to another human." When GPT4 mistakenly generated repeat words, we prompted it to replace the repeat words with other appropriate words."

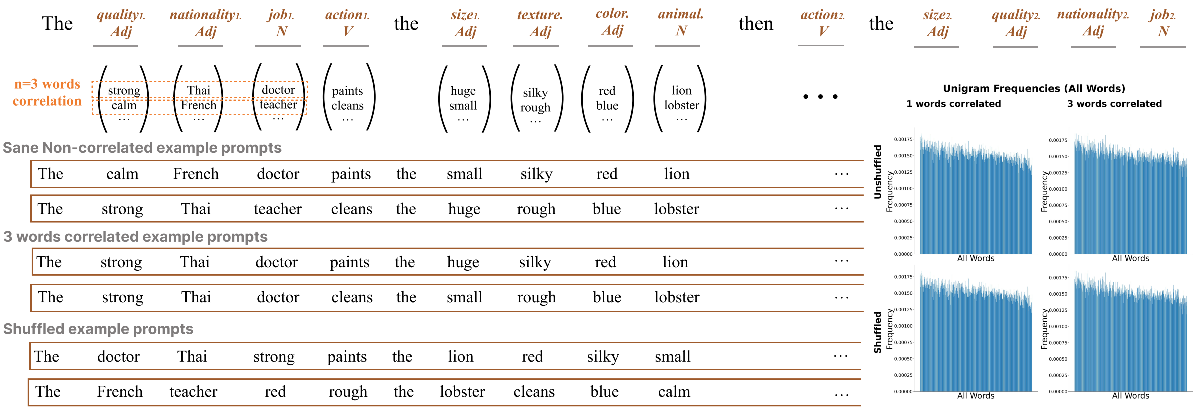


Figure E.1: **Dataset structure and distributional properties. Top:** The structure of the stimulus dataset. The top row shows the ordering of word categories, such as quality.ADJ or animal.N; below it, the vocabulary for each category, including words like "strong" (quality.ADJ) and "lion" (animal.N), respectively. When controlling the degree of dataset compositionality, contiguous word positions are coupled. For instance, when $k = 3$, the first vocabulary indices for quality$_1$.ADJ, nationality$_1$.ADJ, and job$_1$.N are tied together, such that "strong Thai doctor" or "calm French teacher" can be sampled, but "strong French doctor" cannot. **Left:** Examples of generated prompts for the normal, $k = 3$, and shuffled settings. **Right:** When controlling the compositionality across $k = 1 \cdots 4$, word unigram frequencies are preserved in the resulting datasets, shown in the distributions looking identical.

Category	Words
job_1	teacher, doctor, engineer, chef, lawyer, plumber, electrician, accountant, nurse, mechanic, architect, dentist, programmer, photographer, painter, firefighter, police, pilot, farmer, waiter, scientist, actor, musician, writer, athlete, designer, carpenter, librarian, journalist, psychologist, gardener, baker, butcher, tailor, cashier, barber, janitor, receptionist, salesperson, manager, tutor, coach, translator, veterinarian, pharmacist, therapist, driver, bartender, security, clerk
job_2	banker, realtor, consultant, therapist, optometrist, astronomer, biologist, geologist, archaeologist, anthropologist, economist, sociologist, historian, philosopher, linguist, meteorologist, zoologist, botanist, chemist, physicist, mathematician, statistician, surveyor, pilot, steward, dispatcher, ichthyologist, oceanographer, ecologist, geneticist, microbiologist, neurologist, cardiologist, pediatrician, surgeon, anesthesiologist, radiologist, dermatologist, gynecologist, urologist, psychiatrist, physiotherapist, chiropractor, nutritionist, personal trainer, yoga instructor, masseur, acupuncturist, paramedic, midwife
animal	dog, cat, elephant, lion, tiger, giraffe, zebra, monkey, gorilla, chimpanzee, bear, wolf, fox, deer, moose, rabbit, squirrel, raccoon, beaver, otter, penguin, eagle, hawk, owl, parrot, flamingo, ostrich, peacock, swan, duck, frog, toad, snake, lizard, turtle, crocodile, alligator, shark, whale, dolphin, octopus, jellyfish, starfish, crab, lobster, butterfly, bee, ant, spider, scorpion
color	red, blue, green, yellow, purple, orange, pink, brown, gray, black, white, cyan, magenta, turquoise, indigo, violet, maroon, navy, olive, teal, lime, aqua, coral, crimson, fuchsia, gold, silver, bronze, beige, tan, khaki, lavender, plum, periwinkle, mauve, chartreuse, azure, mint, sage, ivory, salmon, peach, apricot, mustard, rust, burgundy, mahogany, chestnut, sienna, ochre
$size_1$	big, small, large, tiny, huge, giant, massive, microscopic, enormous, colossal, miniature, petite, compact, spacious, vast, wide, narrow, slim, thick, thin, broad, expansive, extensive, substantial, boundless, considerable, immense, mammoth, towering, titanic, gargantuan, diminutive, minuscule, minute, hulking, bulky, hefty, voluminous, capacious, roomy, cramped, confined, restricted, limited, oversized, undersized, full, empty, half, partial
$size_2$	lengthy, short, tall, long, deep, shallow, high, low, medium, average, moderate, middling, intermediate, standard, regular, normal, ordinary, sizable, generous, abundant, plentiful, copious, meager, scanty, skimpy, inadequate, sufficient, ample, excessive, extravagant, exorbitant, modest, humble, grand, majestic, imposing, commanding, dwarfed, diminished, reduced, enlarged, magnified, amplified, expanded, contracted, shrunken, swollen, bloated, inflated, deflated

nationality_1	American, British, Canadian, Australian, German, French, Italian, Spanish, Japanese, Chinese, Indian, Russian, Brazilian, Mexican, Argentinian, Turkish, Egyptian, Nigerian, Kenyan, African, Swedish, Norwegian, Danish, Finnish, Icelandic, Dutch, Belgian, Swiss, Austrian, Greek, Polish, Hungarian, Czech, Slovak, Romanian, Bulgarian, Serbian, Croatian, Slovenian, Ukrainian, Belarusian, Estonian, Latvian, Lithuanian, Irish, Scottish, Welsh, Portuguese, Moroccan, Algerian
nationality_2	Vietnamese, Thai, Malaysian, Indonesian, Filipino, Singaporean, Nepalese, Bangladeshi, Maldivian, Pakistani, Afghan, Iranian, Iraqi, Syrian, Lebanese, Israeli, Saudi, Emirati, Qatari, Kuwaiti, Omani, Yemeni, Jordanian, Palestinian, Bahraini, Tunisian, Libyan, Sudanese, Ethiopian, Somali, Ghanaian, Ivorian, Senegalese, Malian, Cameroonian, Congolese, Ugandan, Rwandan, Tanzanian, Mozambican, Zambian, Zimbabwean, Namibian, Botswanan, New Zealander, Fijian, Samoan, Tongan, Papuan, Marshallese
action_1	feeds, walks, grooms, pets, trains, rides, tames, leashes, bathes, brushes, adopts, rescues, shelters, houses, cages, releases, frees, observes, studies, examines, photographs, films, sketches, paints, draws, catches, hunts, traps, chases, pursues, tracks, follows, herds, corrals, milks, shears, breeds, mates, clones, dissects, stuffs, mounts, taxidermies, domesticates, harnesses, saddles, muzzles, tags, chips, vaccinates
action_2	hugs, kisses, loves, hates, admires, respects, befriends, distrusts, helps, hurts, teaches, learns from, mentors, guides, counsels, advises, supports, undermines, praises, criticizes, compliments, insults, congratulates, consoles, comforts, irritates, annoys, amuses, entertains, bores, inspires, motivates, discourages, intimidates, impresses, disappoints, surprises, shocks, delights, disgusts, forgives, resents, envies, pities, understands, misunderstands, trusts, mistrusts, betrays, protects
quality_1	good, bad, excellent, poor, superior, inferior, outstanding, mediocre, exceptional, sublime, superb, terrible, wonderful, awful, great, horrible, fantastic, dreadful, marvelous, atrocious, splendid, appalling, brilliant, dismal, fabulous, lousy, terrific, abysmal, incredible, substandard, amazing, disappointing, extraordinary, stellar, remarkable, unremarkable, impressive, unimpressive, admirable, despicable, praiseworthy, blameworthy, commendable, reprehensible, exemplary, subpar, ideal, flawed, perfect, imperfect
quality_2	acceptable, unacceptable, satisfactory, unsatisfactory, sophisticated, insufficient, adequate, exquisite, suitable, unsuitable, appropriate, inappropriate, fitting, unfitting, proper, improper, correct, incorrect, right, wrong, accurate, inaccurate, precise, imprecise, exact, inexact, flawless, faulty, sound, unsound, reliable, unreliable, dependable, undependable, trustworthy, untrustworthy, authentic, fake, genuine, counterfeit, legitimate, illegitimate, valid, invalid, legal, illegal, ethical, unethical, moral, immoral

texture	smooth, rough, soft, hard, silky, coarse, fluffy, fuzzy, furry, hairy, bumpy, lumpy, grainy, gritty, sandy, slimy, slippery, sticky, tacky, greasy, oily, waxy, velvety, leathery, rubbery, spongy, springy, elastic, pliable, flexible, rigid, stiff, brittle, crumbly, flaky, crispy, crunchy, chewy, stringy, fibrous, porous, dense, heavy, light, airy, feathery, downy, woolly, nubby, textured

F Benchmark tasks

Here we briefly summarize the benchmark tasks that we use to evaluate Pythia checkpoints as described in Section 4.3. In figure Figure 2, we did not include WSC (Winogrande Schema Challenge) which was originally included in Biderman et al., as it has been proposed that WSC dataset performance on LMs might be corrupted by spurious biases in the dataset (Sakaguchi et al., 2021). Instead, we only presented the evaluation from WinoGrande task, which is inspired from original WSC task but adjusted to reduce the systematic bias (Sakaguchi et al., 2021).

WinoGrande WinoGrande (Sakaguchi et al., 2021) is a dataset designed to test commonsense reasoning by building on the structure of the Winograd Schema Challenge (Levesque et al., 2012). It presents sentence pairs with subtle ambiguities where understanding the correct answer requires world knowledge and commonsense reasoning. It challenges models to differentiate between two possible resolutions of pronouns or references, making it a benchmark for evaluating an AI's ability to understand context and reasoning.

LogiQA LogiQA (Liu et al., 2021) is an NLP benchmark for evaluating logical reasoning abilities in models. It consists of multiple-choice questions derived from logical reasoning exams for human students. The questions test various forms of logical reasoning, such as deduction, analogy, and quantitative reasoning, making it ideal for assessing how well AI can handle structured logical problems.

SciQ SciQ (Welbl et al., 2017) is a dataset focused on scientific question answering, based on material from science textbooks. It features multiple-choice questions related to science topics like biology, chemistry, and physics. The benchmark is designed to test a model's ability to comprehend scientific information and answer questions using factual knowledge and reasoning.

ARC Challenge The ARC (AI2 Reasoning Challenge) Challenge Set (Clark et al., 2018) is a benchmark designed to test models on difficult, grade-school-level science questions. It presents multiple-choice questions that are challenging due to requiring complex reasoning, inference, and background knowledge beyond simple retrieval-based approaches. It is a tougher subset of the larger ARC dataset.

PIQA PIQA (Physical Interaction QA) (Bisk et al., 2020) is a benchmark designed to test models on physical commonsense reasoning. The questions require understanding basic physical interactions, like how objects interact or how everyday tasks are performed. It focuses on scenarios that involve intuitive knowledge of the physical world, making it a useful benchmark for evaluating practical commonsense in models.

ARC Easy ARC Easy is the easier subset of the AI2 Reasoning Challenge, consisting of grade-school-level science questions that require less complex reasoning compared to the Challenge set. This benchmark is meant to evaluate models' ability to handle straightforward factual and retrieval-based questions, making it more accessible for baseline NLP models.

LAMBADA LAMBADA (Paperno et al., 2016) is a reading comprehension benchmark where models must predict the last word of a passage. The challenge lies in the fact that understanding the entire context of the passage is necessary to guess the correct word. This benchmark tests a model's long-range context comprehension and coherence skills in natural language.

G Additional Results: Controlled Grammar

Here, we present extended results on the length-17 (longest) controlled grammar, for

1. All Pythia sizes, see Figure G.1, and Llama-3-8B and Mistral-7B, see Figure G.2.
2. Other dimensionality measures MLE (Levina and Bickel, 2004) and Participation Ratio (PR) (Gao et al., 2017), shown for Pythia models, see Figures G.3 and G.4. Results are highly consistent between TwoNN and MLE, but PR produced nonsensical values $d_{PR} < I_d$ that violated the theoretical relation $I_d \leq d \leq D$.
3. Linear fits $I_d \sim D$, $d \sim D$ for settings $k \in \{1 \cdots 4\} \times \{\text{coherent}, \text{shuffled}\}$, see Table G.1.
4. Training dynamics of feature complexity (TwoNN, PCA) for three Pythia sizes 410m, 1.4b, and 6.9b, on both coherent $k = 1$ and shuffled text, see Figure G.5.

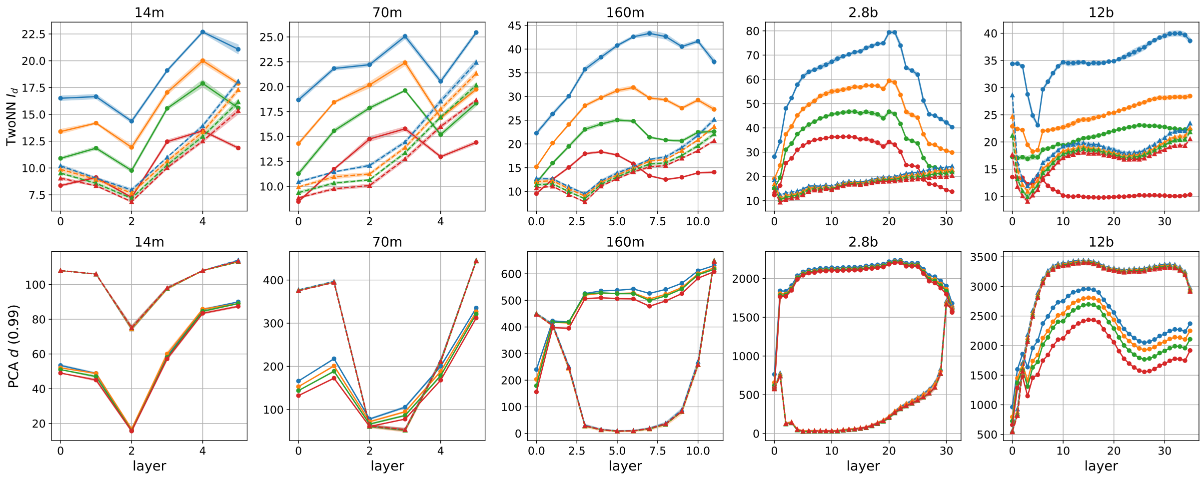


Figure G.1: **Dimensionality over layers.** TwoNN nonlinear I_d (top) and PCA linear d (bottom) over layers are shown for all sizes (left to right). Each color corresponds to a coupling length $k \in 1 \cdots 4$. solid curves denote coherent sequences, and dotted curves denote shuffled sequences. For all models, lower k results in higher I_d and d for both normal and shuffled settings. For all models, shuffling results in lower I_d but higher d. Curves are averaged over 5 random seeds, shown with ± 1 SD.

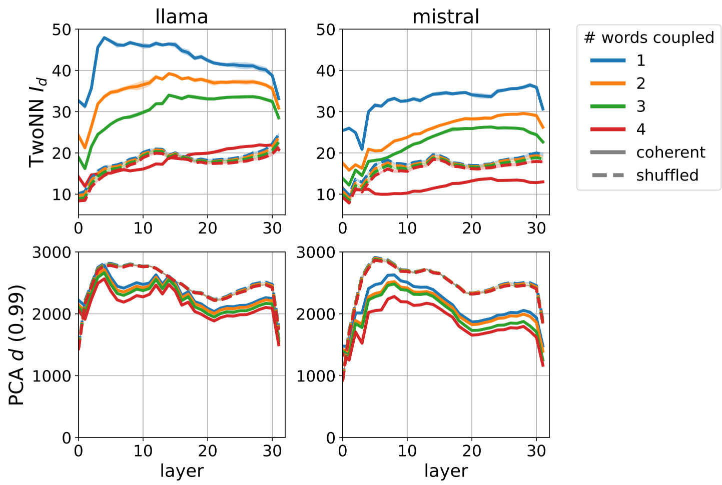


Figure G.2: **Dimensionality over layers for Llama-3-8B and Mistral-7B.** TwoNN nonlinear I_d (top) and PCA linear d (bottom) over layers are shown for all sizes (left to right). Each color corresponds to a coupling length $k \in 1 \cdots 4$. solid curves denote coherent sequences, and dotted curves denote shuffled sequences. For all models, lower k results in higher I_d and d for both normal and shuffled settings. For all models, shuffling results in lower I_d but higher d. Curves are averaged over 5 random seeds, shown with ± 1 SD. **Results mirror those of the Pythia models.**

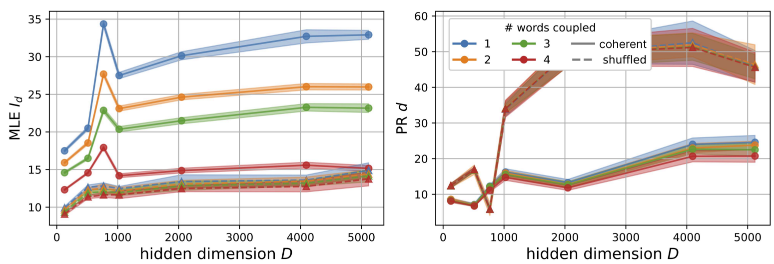


Figure G.3: **Mean dimensionality over model size (other metrics).** Mean nonlinear I_d computed with MLE (left) and linear d computed with PR (right) over layers is shown for increasing LM hidden dimension D. MLE I_d does not depend on extrinsic dimension D (flat lines). PR d produces nonsensical values, higher than the nonlinear I_d. Curves are averaged over 5 random seeds, shown with $\pm$ 1 SD.

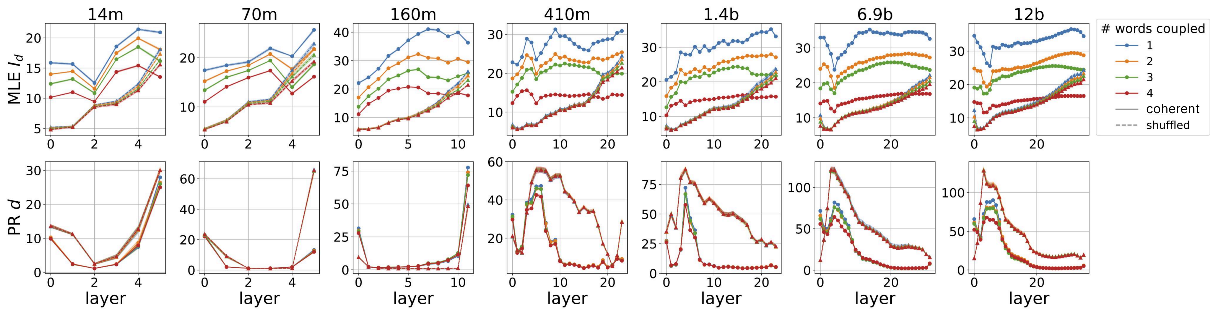


Figure G.4: **Other dimensionality metrics over layers.** MLE nonlinear I_d (top) and PR linear d (bottom) over layers are shown for all model sizes (left to right). Each color corresponds to a coupling length $k \in 1 \cdots 4$. solid curves denote coherent sequences, and dotted curves denote shuffled sequences. For all models, lower k results in higher I_d for both normal and shuffled settings. For all models, shuffling results in lower I_d. The PR-d produced nonsensical results, with linear dimensionality higher than nonlinear dimensionality. Curves are averaged over 5 random seeds, shown with ± 1 SD.

Mode	**k-coupling**	**PCA d**			**TwoNN I_d**		
		α	**R**	**p-value**	α	**R**	**p-value**
coherent	1	0.4598	0.9956	2×10^{-6}	0.0023	0.6341	0.1261
coherent	2	0.4268	0.9954	3×10^{-6}	0.0011	0.5580	0.1930
coherent	3	0.4014	0.9943	5×10^{-6}	0.0009	0.6616	0.1056
coherent	4	0.3569	0.9924	1×10^{-5}	-0.0003	-0.3523	0.4383
shuffled	1	0.6239	0.9919	1.1×10^{-5}	0.0011	0.8488	0.0157
shuffled	2	0.6193	0.9917	1.2×10^{-5}	0.0010	0.8487	0.0157
shuffled	3	0.6153	0.9916	1.2×10^{-5}	0.0010	0.8586	0.0134
shuffled	4	0.6114	0.9916	1.2×10^{-5}	0.0009	0.8559	0.0140

Table G.1: **Linear regression of average layerwise dimensionality to hidden dimension,** D. For Pythia models, for each setting (Mode, k-coupling) and dimensionality measure (PCA, TwoNN), the linear effect size α along with R-value and p-value are reported. PCA linear dimension shows a consistent strong linear relationship with large effect size α to hidden dimension D ($p < 0.001$) for all settings in $k = \{1 \cdots 4\} \times$ [coherent, shuffled]. TwoNN intrinsic dimension does not scale linearly as D in all settings, showing a non-significant relationship for coherent text and a significant one for shuffled text. For all TwoNN settings, the effect size α is near-zero, showing that nonlinear I_d is robust to changes in hidden dimension D.

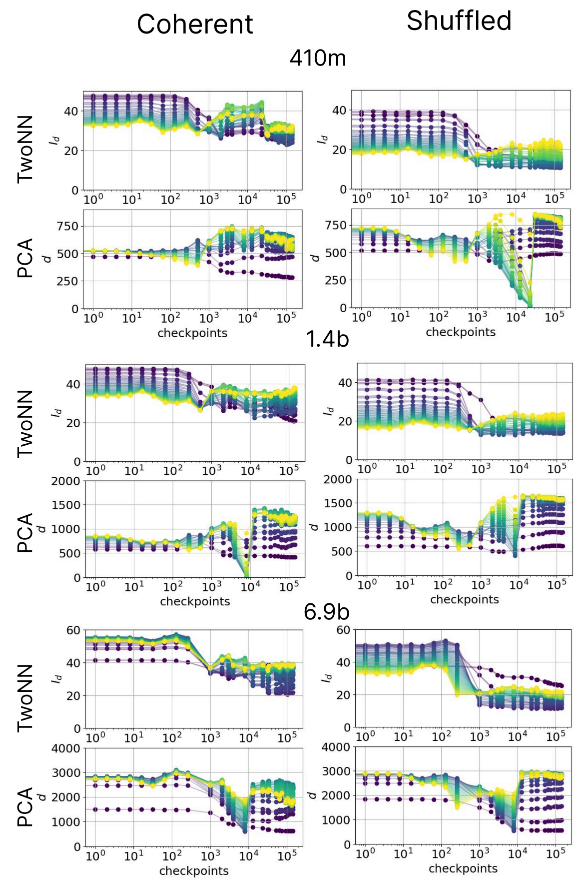


Figure G.5: **Layerwise feature complexity evolution over time, additional results.** Nonlinear I_d (top) and linear d (bottom) over training is shown for coherent (left) and shuffled (right) text, for the 1-coupled setting. Each curve is one layer of the LM (yellow is later, purple is earlier). All settings in [TwoNN, PCA]×[coherent, shuffled] exhibit a phase transition in representational dimensionality at around checkpoint 10^3, which corresponds to the sharp increase in task performance. In the nonlinear case (top row), the difference between layers' I_d is *low* at the end of training for shuffled text, and *high* for coherent text. This suggests LM learns to perform meaningful and specialized processing over layers. The difference between layers' d (bottom row) at the end of training is, conversely, *high* for shuffled and *lower* for coherent text. This is consistent with our interpretation of d as capturing implied dataset size.

H Additional Results: The Pile

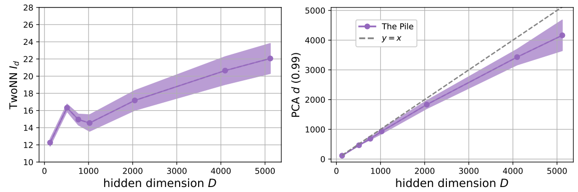


Figure H.1: **Mean dimensionality on the Pile over model size.** Mean nonlinear I_d computed with TwoNN (left) and linear d computed with PCA (right) over layers is shown for increasing LM hidden dimension D. TwoNN I_d grows very slowly with extrinsic dimension D, while PCA d grows to be nearly one-to-one with D. Curves are averaged over 5 random data splits, shown with $\pm$ 1 SD.

Similar to for the controlled grammar (Section 4.1), representational I_d and d scale differently with model ambient dimension D, where we found for the controlled grammar that $d \propto D$ while I_d is stable in D. Here, Figure H.1 and Table H.1 similarly show that $d \propto D$ with a highly significant linear fit; a large effect $\alpha = 0.81$ indicates the LM fills the ambient space such that $d \approx D$. For The Pile, $I_d \propto D$ ($R = 0.95$, $p < 0.001$) as well, but the tiny effect α=0.002 shows I_d is stable as D grows, see Figure H.1 (left).

I Additional Results: Correlation with Kolmogorov Complexity

I.1 Correlation between Kolmogorov Complexity and Feature Complexity is Robust to Sequence Length

In Table 1 we showed that, for each model, and on a single dataset ($k = 1$, $l = 17$), linear effective d highly correlates to the estimated superficial complexity (KC) using `gzip`. Table I.1 shows that this trend is robust to both model family, model size, and sequence length; average layerwise d is almost perfectly monotonic in the KC of the dataset, seen by high Spearman correlation. In contrast, for none of the sequence lengths is average layerwise I_d monotonic in KC, except for the smallest Pythia model (14m).

I.2 Kolmogorov Complexity vs. Average-Layer Feature Complexity Across Datasets

Figure I.1 shows the global correlation between feature complexity (I_d and d) and KC, estimated with `gzip`. While both nonlinear (top row) and

PCA d			TwoNN I_d		
α	R	p-value	α	R	p-value
0.8119	0.9993	2.39×10^{-8}	0.00173	0.9537	8.64×10^{-4}

Table H.1: **Linear regression of Pythia's average layerwise dimensionality on The Pile to hidden dimension,** D. For dimensionality measures (PCA, TwoNN columns), the linear effect size α along with R-value and p-value are reported. PCA linear dimension shows a statistically significant linear relationship to D, with large effect size $\alpha = 0.81$. TwoNN intrinsic dimension also shows a slightly weaker, but still highly significant, linear relationship to D. But, the effect size α is near-zero, showing that nonlinear I_d is robust to changes in hidden dimension D.

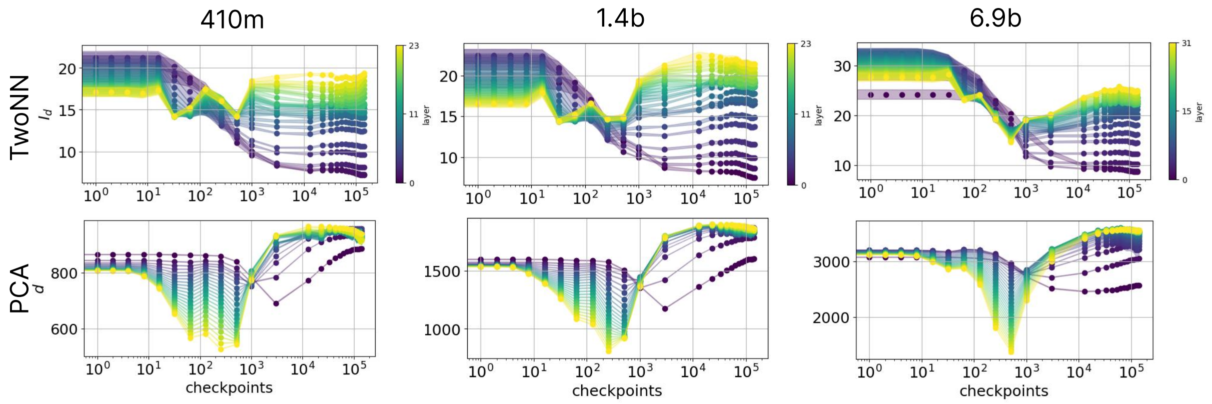


Figure H.2: I_d **phase transition in The Pile.** Nonlinear I_d (top) and linear d (bottom) over training is shown for model sizes 410m, 1.4b, and 6.9b (left to right), for The Pile. Each curve is one layer of the LM (yellow is later, purple is earlier). Representations of The Pile exhibit a phase transition in both I_d and d at slightly before checkpoint 10^3, where $t = 10^3$ corresponds to a dip and redistribution of layerwise dimensionality, and also a sharp increase in task performance in Figure 2.

linear (bottom row) dimensionality are positively Spearman-correlated to `gzip`, there are clear differences:

1. Linear d increases in the shuffled setting from the coherent setting; nonlinear I_d decreases.
2. Linear d is very highly correlated to the estimated KC, $\rho \approx 0.9$ in all cases, while nonlinear d is more weakly correlated, $\rho \in [0.5, 0.6]$.

These observations support the hypothesis that linear effective d encodes KC, while the intrinsic dimension I_d encodes sequence-level semantic complexity.

I.3 Per-layer Correlation with Kolmogorov Complexity

Linear effective d encodes KC robustly across models and datasets Figure I.4 shows, for each model, the Spearman correlation between layer dimension and KC (`gzip`). Orange boxplots correspond to d, and blue boxplots to the I_d. Each datapoint in a boxplot reports the correlation for one (model, layer, sequence length) combination; the only factor of variation in each correlation is the k-coupling factor and whether the dataset is shuffled. **With high generality across models and grammars, the linear effective d is monotonic in KC**, seen by the vast majority of layers (orange distributions) close to $\rho = 1.0$ (y-axis). Meanwhile, the I_d does not consistently encode superficial complexity, seen by the blue distributions landing about 0.0.

Outliers There was one significant outlier, 160m, in our analysis correlating layerwise dimensionality to `gzip` (KC), see Figures I.2 and I.4 and Table I.1. While other models consistently demonstrate a positive Spearman correlation between d and `gzip` across layers, 160m (and to a smaller extent, 70m) deviates from this pattern. The reason 160m displays a negative correlation is due to its behavior on shuffled corpora, see the third column in Figure G.1: for intermediate layers, PCA with a variance threshold of 0.99 yields fewer than 50 PCs. We found that this was due to the existence of so-called "rogue dimensions" (Timkey and

		sequence length (words)				
		5	8	11	15	17
14m	I_d	1.00	0.89	**0.87***	**0.87***	-0.10
	d	1.00	0.89	**0.87***	**0.87***	**0.81***
70m	I_d	1.00	0.40	0.43	0.26	-0.10
	d	1.00	**1.00***	**1.00***	**0.98***	**0.98***
160m	I_d	-1.00	0.00	0.19	0.00	-0.21
	d	-1.00	-0.60	-0.52	-0.62	-0.62
410m	I_d	-1.00	0.40	0.40	0.26	0.14
	d	1.00	**1.00***	**0.98***	**1.00***	**1.00***
1.4b	I_d	-1.00	0.40	0.40	0.43	0.14
	d	1.00	**1.00***	**0.98***	**1.00***	**1.00***
6.9b	I_d	1.00	0.40	0.40	0.36	0.48
	d	1.00	**1.00***	**0.98***	**1.00***	**0.98***
12b	I_d	1.00	0.40	0.40	0.43	0.00
	d	1.00	**1.00***	**0.98***	**1.00***	**1.00***
Llama-8b	I_d	1.00	0.40	0.19	0.00	-0.02
	d	1.00	**1.00***	**0.98***	**1.00***	**0.93***
Mistral-7b	I_d	1.00	0.40	0.40	0.00	0.29
	d	1.00	**1.00***	**0.98***	**1.00***	**0.90***

Table I.1: **Spearman correlations between dimensionality and estimated Kolmogorov complexity, varying sequence length**. The Spearman correlation ρ between the gzipped dataset size (KB) and representational dimensionality (rows), averaged over layers, is shown for all tested **Pythia** model sizes (model name omitted for readability). Values marked with a * are significant with a p-value threshold of 0.05. Values marked with † are significant with a p-value threshold of 0.1. Across models, average-layer I_d is not correlated to the estimated KC, or formal compositionality, of datasets. Average-layer **linear** d is consistently highly positively correlated to the estimated KC. Length $l = 5$ is grayed out as, due to the sequence length being too short, it was not possible to varying the coupling factor k; here, the only comparison is between coherent and shuffled ($n = 2$).

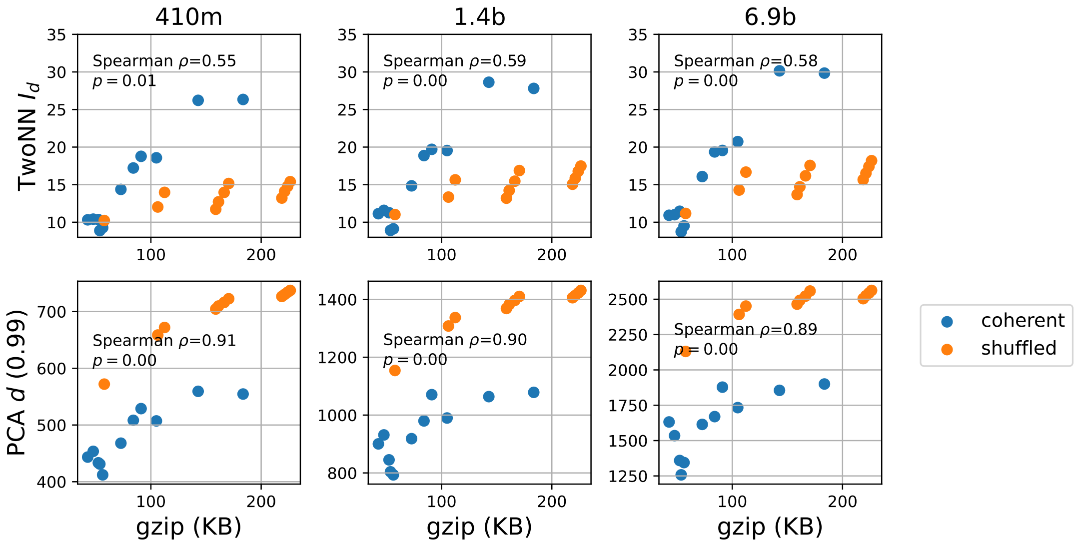


Figure I.1: **Average layerwise dimensionality vs. Estimated Kolmogorov Complexity (gzip)** for Pythia 410m, 1.4b, and 6.9b, aggregated for all grammars. For all models, PCA d highly correlates to gzip (estimated KC), with Spearman $\rho \geq 0.9$** for all models. TwoNN I_d correlates more weakly, $\rho \in [0.5, 0.6]$* for all models. Linear d and nonlinear I_d differentially encode shuffled data complexity (orange dots) compared to coherent data complexity (blue dots); where shuffled data display higher d and lower I_d. (**) Significant at $\alpha = 0.001$, (*) $\alpha = 0.01$.

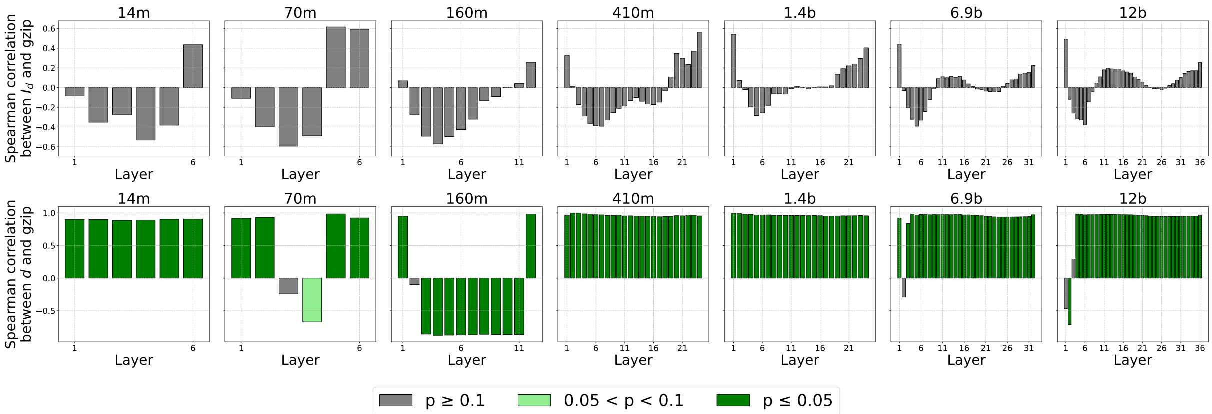


Figure I.2: **Spearman correlations between per-layer dimensionality and estimated Kolmogorov complexity, Pythia models.** The Spearman correlation between the gzipped dataset size (KB) and representational dimensionality per layer, is shown for all tested model sizes for the longest sequence length ($l = 17$). Generally across models, per-layer I_d is not correlated to the estimated KC, or formal compositionality, of datasets. Per-layer linear d is consistently highly positively correlated to the estimated Kolmogorov complexity, except one outlier (160m).

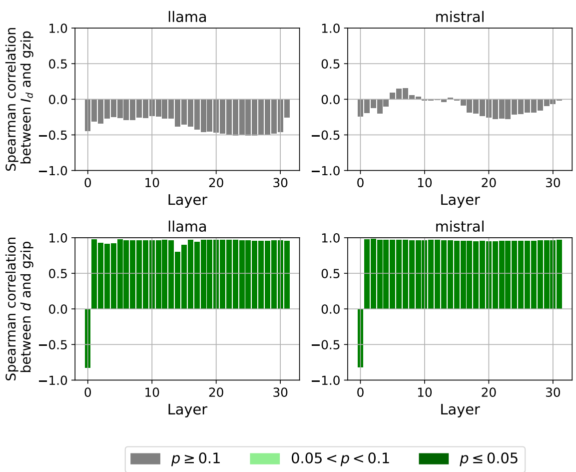


Figure I.3: **Spearman correlations between per-layer dimensionality and estimated Kolmogorov complexity, Llama-3-8B and Mistral-7B.** The Spearman correlation between the gzipped dataset size (KB) and representational dimensionality per layer, is shown for Llama (left) and Mistral (right). Consistently across models, mirroring trends for Pythia, per-layer I_d is not correlated to the estimated KC, or formal compositionality, of datasets. Per-layer linear d is consistently highly positively correlated to the estimated KC.

van Schijndel, 2021; Machina and Mercer, 2024; Rudman et al., 2023), where very few dimensions have outsized norms. Outlier dimensions have been found, via mechanistic interpretability analyses, to serve as a "sink" for uncertainty, and are associated to very frequent tokens in the training data (Puccetti et al., 2022). See Rudman et al. (2023) for exact activation profiles for the last-token embeddings in Pythia 70m and 160m. While increasing the variance threshold to 0.999 reduced the effect of rogue dimensions on PCA dimensionality estimation, we decided to keep the threshold at 0.99 for consistent comparison to other models.

Coding of superficial complexity over training The Spearman correlations between layerwise dimensionality (I_d and d) and estimated Kolmogorov complexity using `gzip`, over training steps, are shown in Figure I.5 for 410m, 1.4b, and 6.9b. Each dot in the figure is a single layer's correlation to `gzip`; each vertical set of dots is the distribution of correlations over layers, at a single timestep of training. Several observations stand out:

1. PCA encodes superficial complexity (seen by earlier dots close to $\rho = 1.0$) as an inductive bias of the model architecture. The high correlation for most layers may be unlearned during intermediate checkpoints of model training, seen by the "dip" in gray dots around steps $10^2 \sim 10^3$, but is regained by the end of training for all model sizes. This indicates that encoding superficial complexity at the end of training is a *learned behavior*.

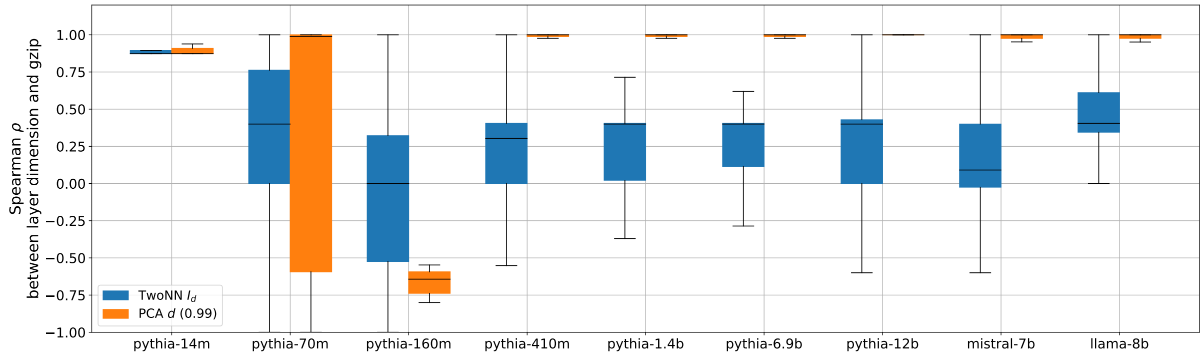
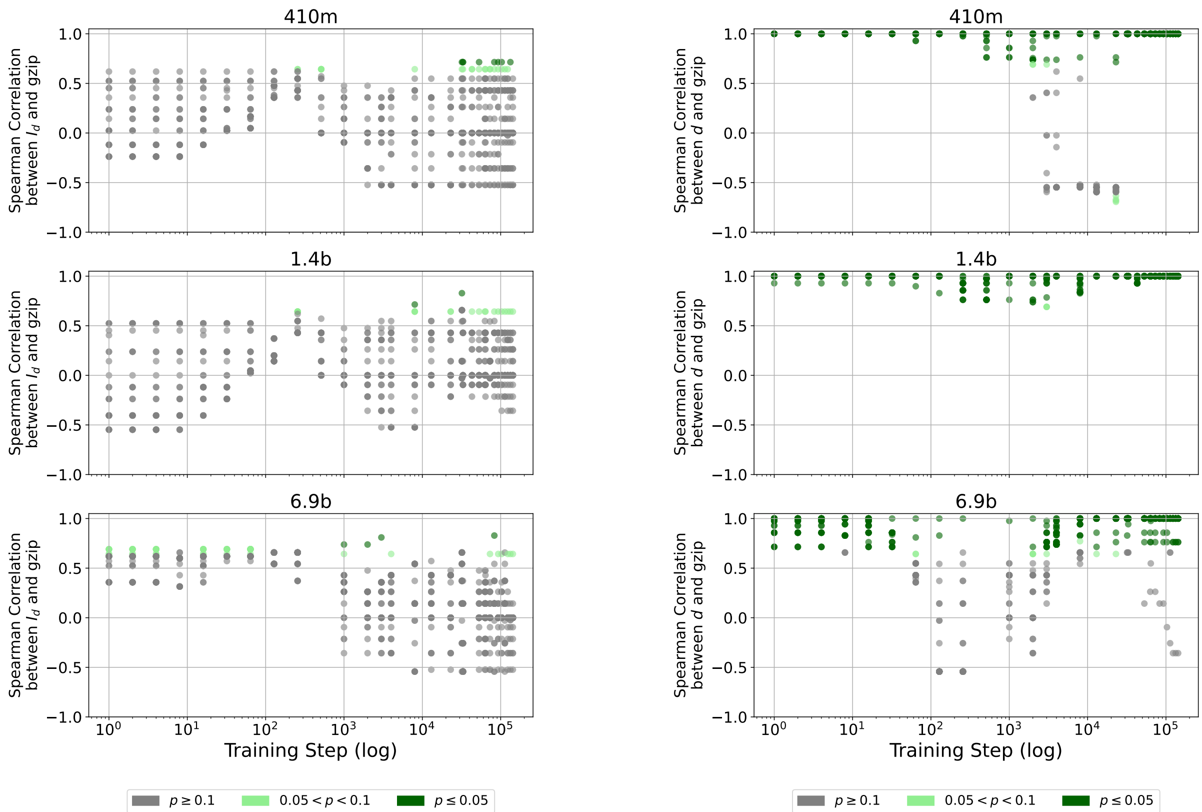


Figure I.4: **Per-layer Spearman ρ between feature complexity and KC for all models, across all tested datasets.** The layerwise Spearman ρ between KC, measured with `gzip`, and feature complexity, measured with TwoNN I_d (blue) and PCA d (orange), is shown for each model. Each datapoint in each distribution corresponds to one (model, dataset, layer) triple. Generally across models, except for the outlier Pythia-160m, the layerwise correlation between I_d and KC is low, while the correlation to d is high and close to 1.0 for the vast majority of layers, datasets, and models (orange distributions near 1.0). This shows that, with high generality across models and datasets, the vast majority of layers encode KC in linear effective d, not in the intrinsic dimension I_d. Trends are especially robust after a certain model size ($\geq$410m).



(a) TwoNN I_d

(b) PCA d

Figure I.5: **Spearman correlation of layerwise dimensionality and Kolmogorov complexity over training.** The Spearman correlations between I_d (left) and `gzip`, and d (right) and `gzip` are plotted for three models: 410m, 1.4b, and 6.9b (top to bottom) over training time (x-axis). Correlations are computed across the controlled corpora. Each vertical set of points denotes the layer distribution of Spearman correlations at a single timestep; each point is one layer's Spearman correlation, colored green if statistically significant and gray otherwise.

2. TwoNN I_d does not statistically significantly correlate to `gzip` at any point during training, for virtually all layers.

3. For I_d, the phase transition noted in Sec-

tion 4.2 is also present at slightly before $t = 10^3$; this is seen by layerwise correlations in Figure I.5a coalescing to around $\rho = 0.5$, and then redistributing. The layers that best encode superficial complexity for TwoNN at the end of trianing correspond to model-initial and model-final layers, see Figure I.2 top.

J Additional Results: Varying Sequence Length

On d and I_d computed for grammars of varying lengths, we found several behaviors that are *emergent* with sequence length:

1. Feature complexity increases with sequence length (Figure J.1). This is to be expected, as sequences that are longer contain more information that the LM needs to represent.

2. Shuffling feature collapse is an emergent property of sequence length. Figure J.2 shows the gap between coherent and shuffled I_d grows as sentence length grows. Intuitively, this suggests that phrase-level feature complexity in short coherent sentences proxies that of a bag of words (shuffled).

We also find behaviors, corresponding to main article Section 4.3, that are *robust* to sequence length:

1. Feature complexity increases as k decreases, no matter the sequence length (Figure J.3), showing that results presented in the main generalize to sequences of different length.

2. Upon shuffling, I_d collapses to a low range, while d increases, no matter the sequence length (Figure J.3).

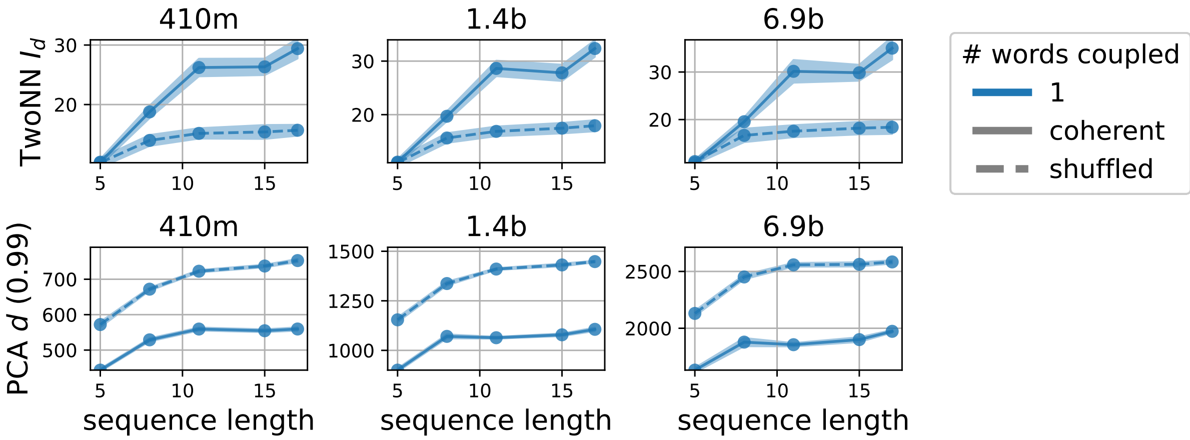


Figure J.1: **Feature complexity increases over sequence length**. The mean I_d and d over layers (y-axis) is shown for increasing sequence lengths $\in \{5, 8, 11, 15, 17\}$ (x-axis) for Pythia models $\in$ {410m, 1.4b, 6.9b} (left to right), for the $k = 1$, or the original dataset configuration. solid curves correspond to coherent, and dashed to shuffled, text. All curves are shown $\pm$1SD over 5 random seeds. Y-axes are scaled to the minimum and maximum for each plot for readability. All curves increase from left to right, evidencing that both nonlinear and linear feature complexity increase with sequence length. Moreover, all curves *saturate*, or plateau, around length=11, indicating this dependence is sublinear.

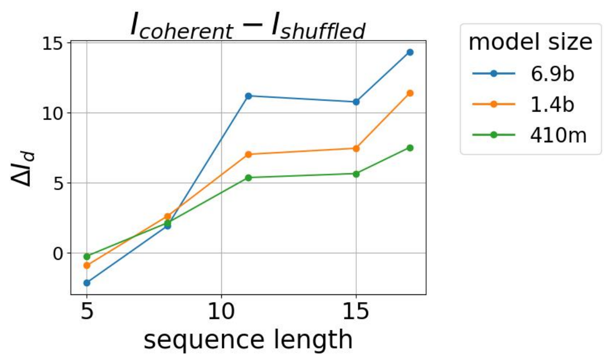


Figure J.2: **Differential coding of semantic complexity increases with sequence length**. The $\Delta(I_d)$ between coherent and shuffled text (y-axis) is shown for Pythia models $\in$ {410m, 1.4b, 6.9b} (different curves), as a function of sentence length $\in \{5, 8, 11, 15, 17\}$, (x-axis). For all models, $\Delta(I_d)$ increases as the sequence length increases. For the shortest sequence length $l = 5$, the $\Delta(I_d) \approx 0$, suggesting that at short lengths, (semantic) representational complexity proxies that of a bag of words.

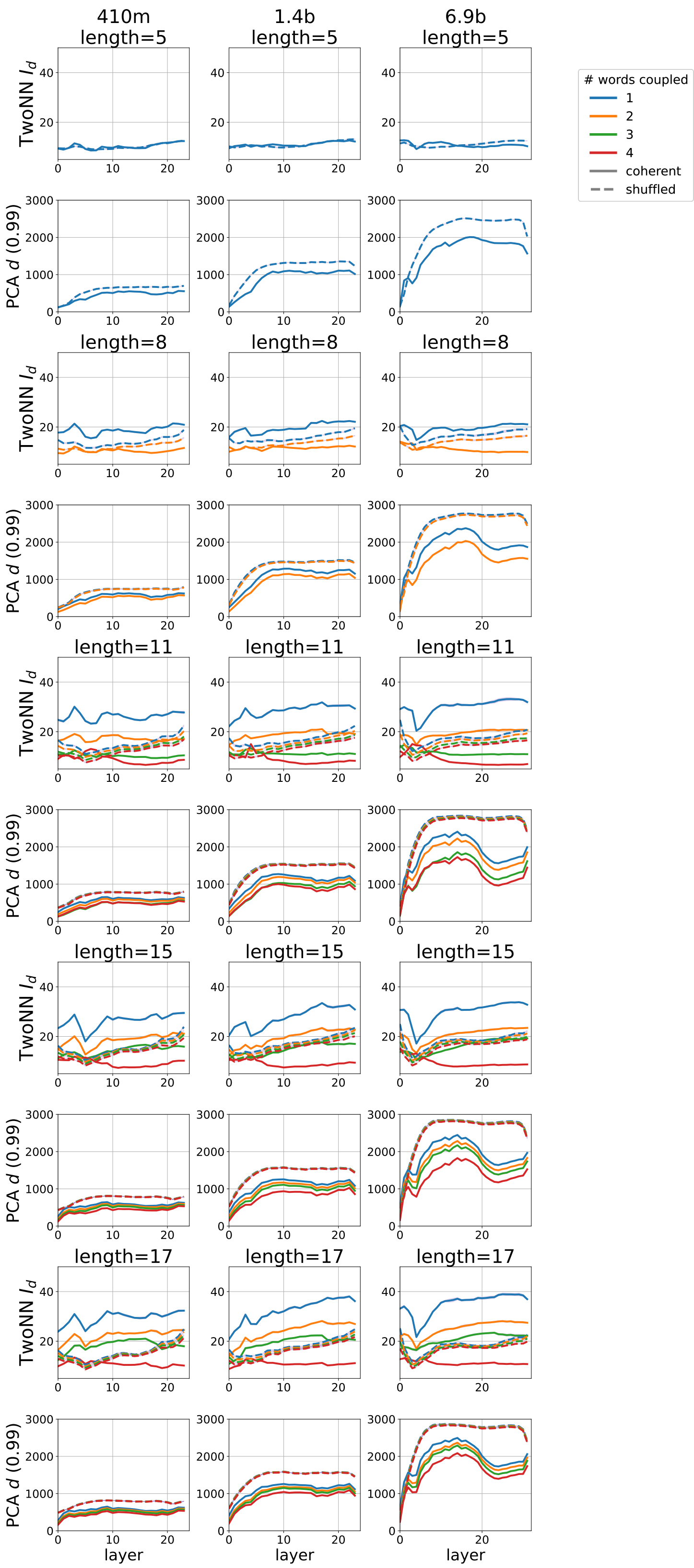


Figure J.3: **Dimensionality over layers, varying sequence length**. Pythia models $\in \{410\text{m}, 1.4\text{b}, 6.9\text{b}\}$ (left to right) and sentence lengths $\in \{5, 8, 11, 15, 17\}$, (top to bottom). In all settings, I_d and d monotonically decrease in k; upon shuffling, I_d collapses to a low range while d increases.