Extended Abstract Track

Geometric Signatures of Compositionality Across a Language Model's Lifetime

Editors: List of editors' names

Abstract

Compositionality, the notion that the meaning of an expression is constructed from the meaning of its parts and syntactic rules, permits the infinite productivity of human language. For the first time, artificial language models (LMs) are able to match human performance in a number of compositional generalization tasks. However, much remains to be understood about the computational mechanisms underlying these abilities. We take a geometric approach to this problem by relating the degree of compositionality in data to the intrinsic dimensionality of their representations under an LM, a measure of feature complexity. We show that the degree of dataset compositionality is reflected in representations' intrinsic dimensionality, and that the relationship between compositionality and geometric complexity arises due to learned linguistic features over training. Overall, our results highlight that linear and nonlinear dimensionality measures capture different and complementary views of data complexity.

1. Introduction

By virtue of linguistic compositionality, few syntactic rules and a finite lexicon can generate an unbounded number of sentences (Chomsky, 1957). That is, language, though seemingly high-dimensional, can be explained using relatively few degrees of freedom. If an LM is a good model of language, we expect its internal representations to exhibit the low-dimensional and compositional structure of the latter. That is, representations should reflect the *manifold hypothesis*, or the notion that real-life, high-dimensional data lie on a low-dimensional manifold (Goodfellow et al., 2016). The dimension of this manifold, or *intrinsic dimension* (ID), is then the minimal number of degrees of freedom required to describe it without information loss (Campadelli et al., 2015).

The manifold hypothesis has been attested for linguistic representations: LMs compress inputs to an ID orders-of-magnitude lower than their extrinsic dimension (Cai et al., 2021; Cheng et al., 2023; Valeriani et al., 2023), yet, it is unknown whether ID reflects linguistic compositionality. In controlled experiments on the Pythia family of language models (Biderman et al., 2023) and a carefully designed synthetic dataset, we provide the first experimental insights into the relationship between linguistic compositionality and representational ID. We show, over the course of LM training, that (1) LMs represent their inputs on low-dimensional manifolds, (2) representational ID reflects the degree of input compositionality over training, and (3) the time-evolution of ID tracks a phase transition in linguistic competence.

2. Setup

Models We evaluate pre-trained Transformer-based LMs of sizes $\in$ {70m, 140m, 1.4b, 6.9b, 12b} from the Pythia family (Biderman et al., 2023). Models were trained on the causal language modeling task on The Pile, a natural language corpus comprising encyclopedic text, books, social media, code, and reviews (Gao et al., 2020).

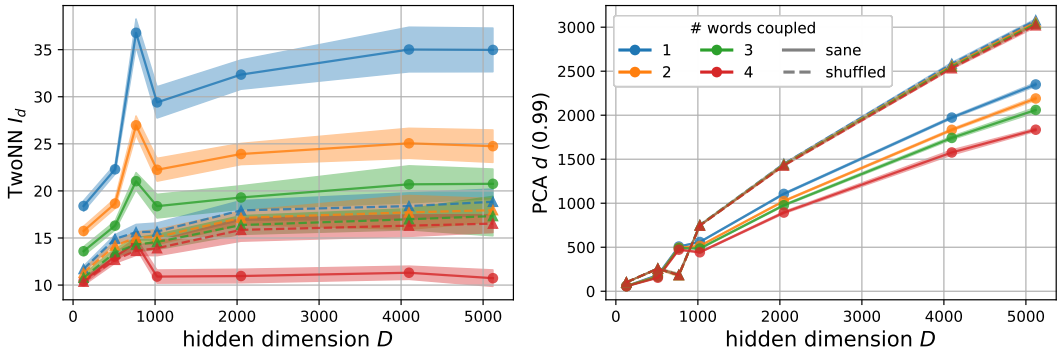

Figure 1: **Mean dimensionality over model size.** Mean nonlinear I_d (left) and linear d (right) over layers is shown for increasing LM hidden dimension D. While the nonlinear I_d does not depend on extrinsic dimension D (flat lines), PCA d scales roughly linearly in D. Curves are averaged over 5 random splits, shown $\pm$ 1 SD.

Dataset We construct a stimulus dataset of sentences from an artificial grammar (details in Appendix D). To do so, we set 12 semantic categories and randomly sample a 50-word vocabulary for each, where categories' vocabularies are disjoint. Categories include 5 adjective types (quality, nationality, size, color, texture), 2 noun types (job, animal) and 1 verb type. We use a fixed syntax by ordering the 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].

We modify the grammar to control the degree of compositionality. In particular, we are interested in two types of compositionality: (1) combinatorial dataset complexity, where a dataset is more compositional if it contains more unique word combinations; (2) sentence-level compositional semantics, where sentence meaning is composed, via syntax, from word meanings. To control for dataset compositionality, we couple the values of k word positions for $k = 1 \cdots 4$. When k positions are coupled, the sequence's atomic units are sets of k contiguous words, thus, lower k produces a more compositional dataset. Second, to investigate *compositional semantics*, we randomly shuffle the words in each sequence. LM behavior on syntactically sane vs. shuffled sequences then proxies compositional vs. lexical-only semantics.

Dimensionality estimation We are interested in whether the geometry of representations reflects the degree of input compositionality. Because sequence lengths may vary, we consider only the last token representation, as it is the only to attend to the entire context, in the Transformer's *residual stream* (Elhage et al., 2021). Then, for each layer's representation, we compute both a nonlinear and a linear measure of dimensionality, which have key conceptual differences. The nonlinear I_d, computed with the TwoNN estimator (Facco et al., 2017), is the number of degrees of freedom or latent features needed to describe the underlying representation manifold (Campadelli et al., 2015; Ansuini et al., 2019). This differs from the *linear* effective dimension d, computed with PCA (Jolliffe, 1986), which is that of the

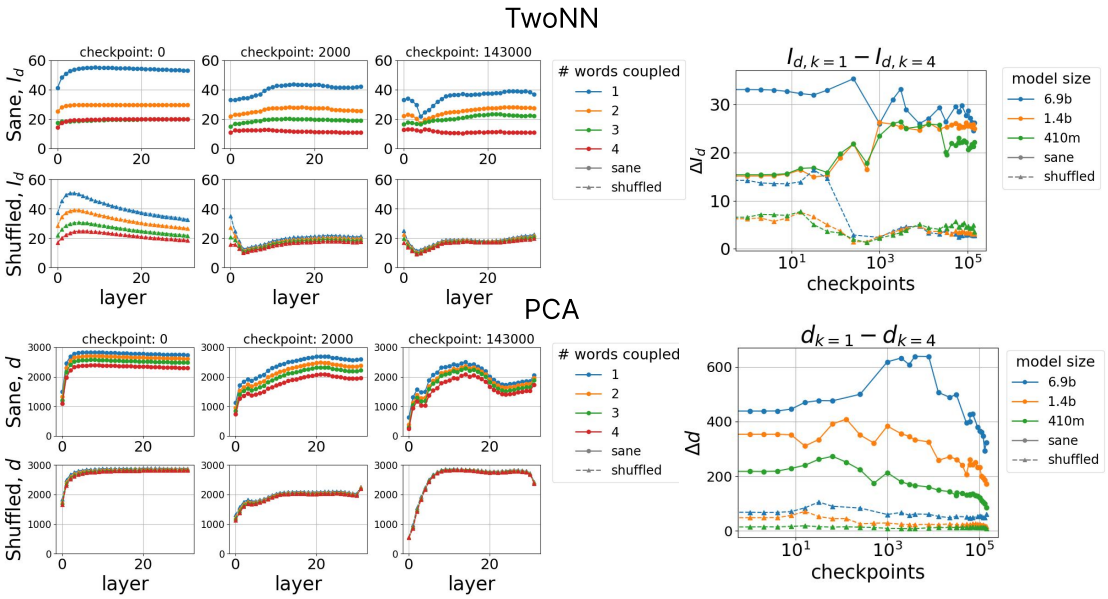

Figure 2: **Training dynamics of dimensionality.** For TwoNN I_d (top) and PCA d (bottom); left: Mean I_d at different timepoints of training for sane vs. shuffled examples with different coupling k over the layers (6.9b model). I_d difference of shuffled examples with varying k diminishes as the training persists. Right: I_d difference between coupling $k = 1, k = 4$ across training, and varying model size.

minimal linear subspace which explains representations' variance up to a threshold. See Appendix C for details about the dimensionality measures used.

3. Results

Nonlinear and linear ID scale differently with model size Models represent inputs on a nonlinear manifold with orders-of-magnitude lower dimension than the ambient dimension ($I_d \sim O(10)$, see Figure 1 left). Moreover, larger models exhibit higher representational dimensionality, but the scaling is not uniform. Figure 1 shows the linear d to scale linearly with hidden dimension D, but nonlinear I_d to instead stabilize to the mentioned range $O(10)$ regardless of extrinsic dimension. This result highlights key differences in how linear and nonlinear dimensions are recruited: LMs *globally* distribute representations to occupy $d \propto D$ dimensions of the space, but *locally* constrains their shape to a low-dimensional (I_d) manifold.

Representational ID reflects input compositionality Representational dimensionality preserves relative data combinatorial complexity. In Figure 1, for both sane (solid curves) and shuffled (dotted curves) settings, both I_d and d increase predictably with input complexity: the highest curves correspond to the 1-coupled dataset, or 12 degrees of freedom, and the lowest denote the 4-coupled dataset, or 3 degrees of freedom. Now, turning to sequence-level compositional semantics, Figure 1 also shows the mean dimensionality over layers for sane and shuffled settings (layerwise results in Figure F.1). Nonlinear and linear dimensionalities show opposing patterns: compared to sane text, shuffled text I_d *collapses* to a low range,

while d *increases*. We interpret this opposition as in Recanatesi et al. (2021), who argue that predictive coding requires the model to both encode the vast space of inputs, exerting upward pressure on representational complexity, as well as extract latent semantic features to support prediction, exerting downward pressure on complexity. Recanatesi et al. (2021) as well as our results suggest that the first pressure expands the global *linear* representation space $\mathbb{R}^d$, while the second compresses representations to a I_d-dimensional *nonlinear* manifold. In our setting, shuffling words in a length-l sequence increases the implied input space as $\propto l!$, increasing d. But, shuffling destroys sequence semantics, exerting a downward pressure on I_d.

Geometry reflects learned semantic features The relationship between dimensionality and data combinatorial complexity, controlled by k, for sane text is *not* an emergent feature over training. In Figure 2 (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 sequence-level semantics are not present, the relationship between k and dimensionality is salient at the *beginning* and greatly diminishes by the end. Together, these demonstrate an inductive bias of the initialized LM architecture to preserve input complexity in its representations. Then, over training, differences in dimensionality may be increasingly explained by linguistic features beyond the surface distribution of inputs. We claim that higher-level semantic processing explains the correspondence between representational and input complexity by the end of training. Figure 3 plots the I_d on the $k = 1$ dataset and the zero-shot performance on linguistic tasks requiring complex semantic understanding (see Appendix E for task details). Resonating with Chen et al. (2024), the I_d (Figure 3 top) decreases sharply around checkpoint 10^3 and then re-distributes, marking a phase of rapidly improving linguistic competence (bottom). It is also at checkpoint 10^3 that, in Figure 2 right, (1) ΔI_d between the $k = 1$ and $k = 4$ dataset collapses in the shuffled setting; (2) I_d stabilizes for different model size. These previously mentioned markers of semantic feature extraction coincide with increased linguistic competence during training.

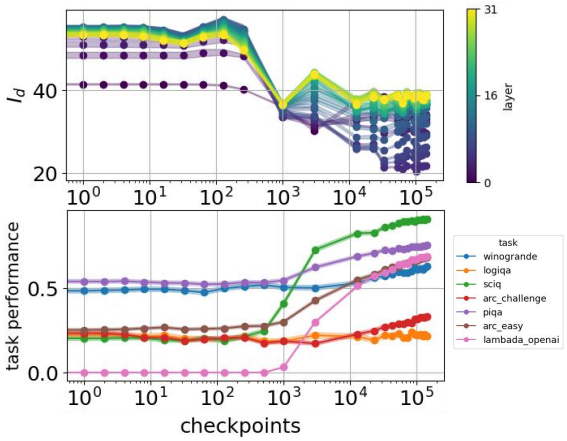

Figure 3: **Phase transition in ID and task performances**. Top: I_d development of Pythia-6.9B over pretraining across layers. Bottom: Zero-shot task performance across pretraining.

Discussion We have studied LM compositionality from a geometric and dynamic perspective. Using a carefully designed synthetic dataset, we found representational complexity to reflect input compositionality superficially at the start of training and semantically by the end. Crucially, nonlinear complexity measures have been underexplored in the literature compared to linear ones; we demonstrate their empirical differences, highlighting a need to further investigate nonlinear measures to proxy feature learning in deep neural models.

References

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

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. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR, 2023.

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

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

P. Campadelli, E. Casiraghi, C. Ceruti, and A. Rozza. Intrinsic dimension estimation: Relevant techniques and a benchmark framework. *Mathematical Problems in Engineering*, 2015:e759567, Oct 2015. ISSN 1024-123X.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L Leavitt, and Naomi Saphra. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=MO5PiKHELW.

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

Noam Chomsky. *Syntactic Structures*. Mouton and Co., The Hague, 1957.

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

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

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. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1): 12140, Sep 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-11873-y.

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

Peiran Gao, Eric M. Trautmann, Byron M. Yu, Gopal Santhanam, Stephen I. Ryu, Krishna V. Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv*, 2017. URL https://api.semanticscholar.org/CorpusID:19938440.

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

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

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

Elizaveta Levina and Peter Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004. URL https://papers.nips.cc/paper_files/paper/2004/hash/74934548253bcab8490ebd74afed7031-Abstract.html.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*, 2020.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*, 2016.

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

Stefano Recanatesi, Matthew Farrell, Guillaume Lajoie, Sophie Deneve, Mattia Rigotti, and Eric Shea-Brown. Predictive learning as a network mechanism for extracting low-dimensional latent space representations. *Nature Communications*, 12(1):1417, March 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-21696-1.

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

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. (arXiv:2302.00294), Feb 2023. doi: 10.48550/arXiv.2302.00294. URL http://arxiv.org/abs/2302.00294. arXiv:2302.00294 [cs, stat].

Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*, 2017.

Appendix 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. ID 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.

Appendix B. Assets

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

Appendix C. ID Estimation

We report the nonlinear I_d using the popular TwoNN estimator of **?**, and we estimate the linear effective dimensionality d using Principal Component Analysis (Jolliffe, 1986) with a variance cutoff of 99%. Though in the main paper we focus on TwoNN and PCA, we also tested the Maximum Likelihood Estimator of (Levina and Bickel, 2004) and the Participation Ratio (Recanatesi et al., 2021). For mathematical details, see below:

TwoNN Estimator A number of methods have been proposed to estimate the nonlinear ID of high-dimensional point clouds (Campadelli et al., 2015). State-of-the-art ID estimators work by exploiting known relationships between points in d-dimensions, then fitting d using maximum likelihood estimation from data. We considered the commonly used TwoNN estimator of **?**, which has been found to highly correlate to other state-of-the-art estimators (Cheng et al., 2023; Campadelli et al., 2015).

The TwoNN method works as follows. In brief, points on the underlying manifold are assumed to follow a locally homogeneous Poisson point process. Local, in this case, refers to neighborhoods about each point x which encompass x's first and second nearest neighbors. Let $r_k^{(i)}$ be the Euclidean distance between point x_i and its kth nearest neighbor. Then, under the mentioned assumptions, the distance ratios $\mu_i := r_1^{(i)}/r_2^{(i)}$ follow the cumulative distribution function $F(\mu) = 1 - \mu^{-I_d}$. Finally, I_d is numerically estimated from data.

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 (**?**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 F.3 left, and Figure F.1 top to Figure F.2 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 F.3, where the PR-d for sane 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 F.3 and the dimensionality over layers in Figure F.2 for completeness, we do not attempt to interpret them.

Appendix D. Toy Grammar

The grammar is composed of sentences of the form

> 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].

The syntax is chosen so that sentences are grammatical and that adjective order complies with the accepted order for English (Dixon, 1976). Although the syntactic structure and vocabulary items are likely seen during training, words are sampled independently for each category without considering the sentence's global semantic coherence. Therefore, sentences are unlikely seen during training. When encountering them for the first time, a frozen LM must successfully construct their meanings from the meanings of their parts, or compositionally generalize.

Each category, colored and enclosed in brackets, is sampled from a vocabulary of 50 possible words, listed in the table below:

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

Appendix E. Benchmark tasks

Here we briefly summarize the benchmark tasks that we use to evaluate Pythia checkpoints as described in Section 4.3.

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., 2020) 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.

Appendix F. Additional Results

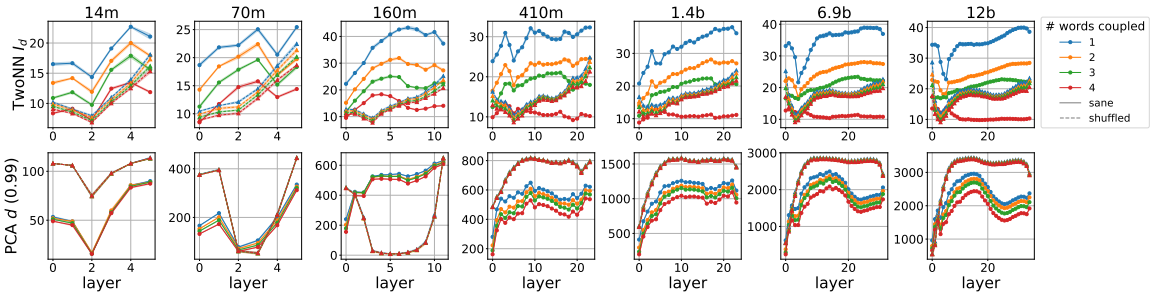

Figure F.1: **Dimensionality over layers.** TwoNN nonlinear I_d (top) and PCA linear d (bottom) over layers are shown for all sizes (left to right). Each color corresponds to a coupling length $k \in 1 \cdots 4$. Solid curves denote sane sequences, and dotted curves denote shuffled sequences. For all 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.

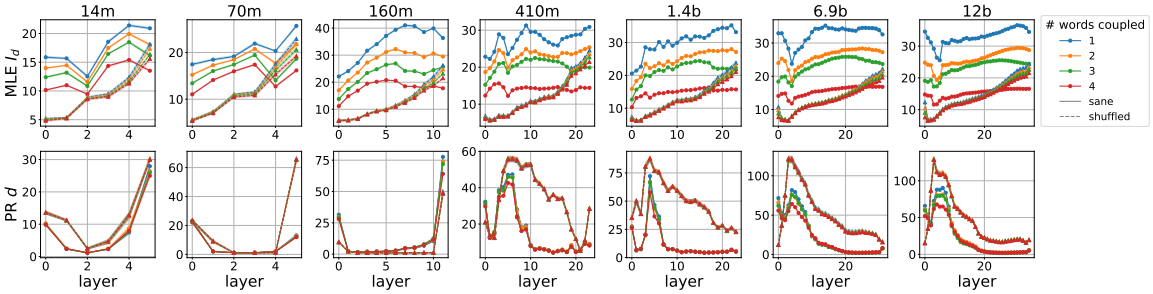

Figure F.2: **Other dimensionality metrics over layers.** MLE nonlinear I_d (top) and PR linear d (bottom) over layers are shown for all model sizes (left to right). Each color corresponds to a coupling length $k \in 1 \cdots 4$. Solid curves denote sane sequences, and dotted curves denote shuffled sequences. 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.

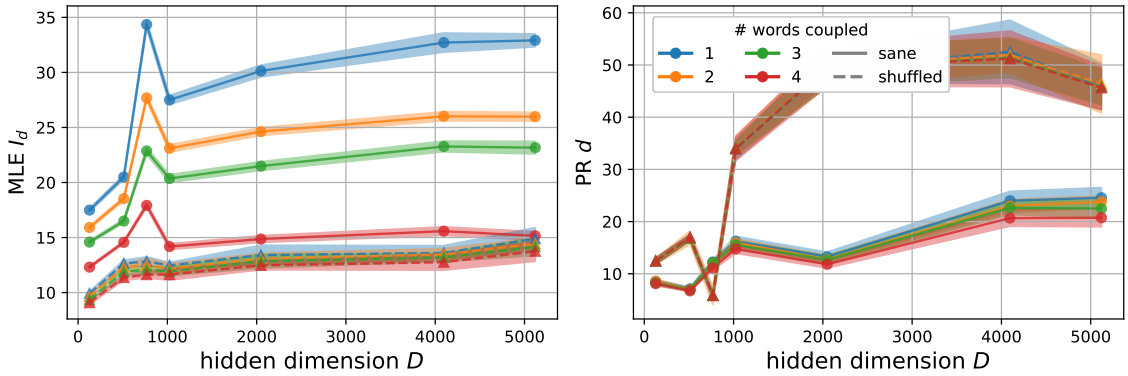

Figure F.3: **Mean dimensionality over model size (other metrics).** Mean nonlinear I_d computed with MLE (left) and linear d computed with PR (right) over layers is shown for increasing LM hidden dimension D. MLE I_d does not depend on extrinsic dimension D (flat lines). PR d produces nonsensical values,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

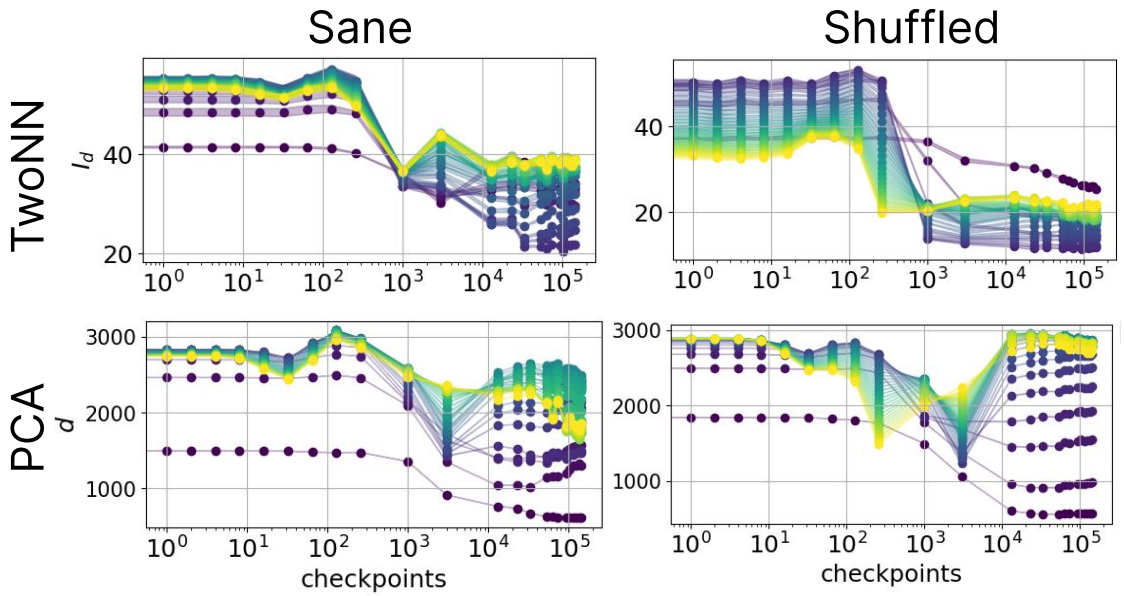

Figure F.4: **ID tracks task performance, additional results.** Nonlinear I_d (top) and linear d (bottom) over training is shown for sane (left) and shuffled (right) text, for the 1-coupled setting. Each curve is one layer of the LM (yellow is later, purple is earlier). All settings in [TwoNN, PCA]×[sane, 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 sane 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 sane text. This is consistent with our interpretation of d as capturing implied dataset size.

Extended Abstract Track