Geometric Signatures of Compositionality in Language Models

Thomas Jiralerspong[*]
Université de Montréal
Montreal, Canada
thomas.jiralerspong@mila.quebec

Jin Hwa Lee[*]
University College London
London, United Kingdom
jin.lee.22@ucl.ac.uk

Lei Yu
University of Toronto
Toronto, Canada
jadeleiyu@cs.toronto.edu

Emily Cheng
Universitat Pompeu Fabra
Barcelona, Spain
emilyshana.cheng@upf.edu

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 high-level geometric approach to this problem, relating the degree of compositionality in a dataset to the intrinsic dimensionality of their representations under an LM, a measure of feature complexity. We find that the degree of dataset compositionality is reflected in the intrinsic dimensionality of data representations, where greater combinatorial complexity of the data results in higher representational dimensionality. Finally, we compare linear and nonlinear methods of computing dimensionality, showing that they capture different but complementary aspects of compositional complexity.

1 Introduction

By virtue of compositionality, few syntactic rules and a finite lexicon can generate an unbounded number of sentences [11]. That is, language, though seemingly high-dimensional, can be explained using relatively few degrees of freedom. A great deal of effort has been made to test whether neural language models (LMs) exhibit human-like compositionality [23, 4]. We take a geometric perspective towards this question, asking how an LM's representational structure reflects and supports compositional understanding over the course of training.

If an LM is a good model of language, we expect its internal representations to exhibit the low-dimensional 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 [20]. 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 [20, 8].

[*]Equal contribution

Submitted to NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward. Do not distribute.

The manifold hypothesis has been attested for linguistic representations: LMs indeed compress inputs to an ID orders-of-magnitude lower than their extrinsic dimension [7, 9, 34]. However, despite their conceptual similarity, no work has explicitly linked the degree of linguistic compositionality to representational ID. To bridge this gap, we provide initial experimental insights into the relationship between compositional complexity of inputs and the ID of their representations. In a series of controlled experiments on the Pythia family of language models [6] and a carefully designed synthetic dataset, we confirm that (1) LMs represent linguistic inputs on low-dimensional, nonlinear manifolds, and (2) representational ID predictably reflects degree of input compositionality.

2 Background

Compositionality It has long been a topic of debate whether neural networks also exhibit human-like compositionality when processing natural language [16, 33, 28]. This debate has fueled an extensive line of empirical exploration aimed at assessing the compositionality of neural networks in language modeling via synthetic data [5, 25, 3]. After the recent introduction of large language models with human-level linguistic capability, researchers have shown via mechanistic interpretability analyses that LMs often extract individual word meanings from early layer multi-layer perceptron modules, and compose them via upper-layer self-attention heads to construct semantic representations for multi-word expressions [21, 19]. Our work takes a different approach to understand language model compositionality by connecting it with the geometric properties of a model's embedding space.

Manifold hypothesis Deep learning problems are often considered high-dimensional, but research suggests that they are governed by low-dimensional structures. In computer vision, studies have demonstrated that common learning objectives and natural image data reside on low-dimensional manifolds [27, 30, 34, 31]. Similarly, the learning dynamics of neural LMs have been shown to occur within low-dimensional parameter subspaces [1, 35]. The nonlinear, low-dimensional structure that emerges in the semantic space of these models likely follows from the training objective of predicting sequential observations [32], which can simplify transfer learning to new tasks and datasets [9].

3 Setup

Models We evaluate pre-trained Transformer-based LMs of sizes $\in$ {70m, 140m, 1.4b, 6.9b, 12b} from the Pythia family [6]. 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 [17].

Dataset As we investigate compositional generalization of the LM, we construct a dataset consisting of nonce sentences from a toy grammar. To create the grammar, we set 12 semantic categories and randomly sample a 50-word vocabulary for each category, where the categories' vocabularies are disjoint. The categories include 6 adjective types (quality, nationality, size, color, texture), 2 noun types (job, animal) and 1 verb type. We use a simple, fixed syntax 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].

The vocabularies are found in Appendix D. The syntax is chosen so that sentences are grammatical and that adjective order complies with the accepted order for English [12]. 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.

Controlling compositionality 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.

First, 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, constraining

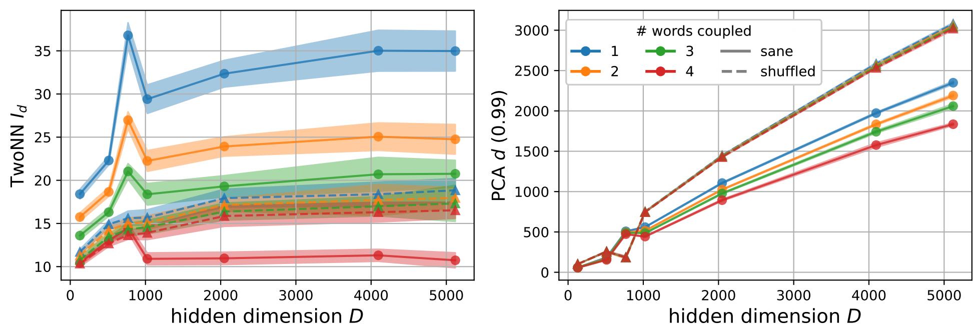

Figure 1: **Mean dimensionality over model size.** Mean nonlinear I_d (left) and linear d (right) over layers is shown for increasing LM hidden dimension D. While the nonlinear I_d does not depend on extrinsic dimension D (flat lines), the PCA d scales roughly linearly in D. Curves are averaged over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

the number of degrees of freedom to l/k where $l = 12$ is the number of variable words in the sequence. For instance, in the 1-coupled setting, words are sampled independently, thus 12 degrees of freedom; if 2-coupled, bigrams are sampled independently, hence 6 degrees of freedom. Increasing k maintains the dataset's unigram distribution, but constrains its combinatorial complexity.

Second, to investigate *compositional semantics*, we randomly shuffle the words in each sequence. This destroys syntactic coherence, and in turn, the composed meaning of the sentence; it instead preserves distributional properties like sequence length and unigram frequencies. Then, LM behavior on grammatically sane vs. shuffled sequences proxies compositional vs. lexical-only semantics.

Dimensionality estimation We are interested in whether the geometry of representations reflects their underlying degree of compositionality. In particular, we consider representations in the *residual stream* of the Transformer [14]. Because sequence lengths vary, in line with prior work [9], we aggregate over the sequence by taking the last token representation, as it is the only to attend to the entire context. For each layer and dataset, we compute both a nonlinear and a linear measure of dimensionality, which have key conceptual differences. The nonlinear I_d is the number of degrees of freedom, or latent features, needed to describe the underlying representation manifold [8, 2, 15]. This differs from the *linear* effective dimensionality d, or the dimension of the minimal linear subspace needed to contain the set of representations. Throughout, we will use *dimensionality* to refer to both nonlinear and linear estimates. When appropriate, we will specify I_d as the nonlinear ID, d as the *linear* effective dimension, and D as the extrinsic dimension, or hidden dimension of the model.

We report the nonlinear I_d using the popular TwoNN estimator of 15, and we estimate the linear effective dimensionality d using Principal Component Analysis [24] 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 [26] and the Participation Ratio [32]. For mathematical details, see Appendix C.

4 Results

We find representational dimensionality to reflect compositionality in ways that are predictable across model scale. First, we demonstrate that linear and nonlinear dimensionality measures behave differently across model scale. Then, we show that dimensionality reflects the degree of compositionality of its inputs, highlighting the difference between nonlinear and linear measures. For brevity, we focus on model sizes 410m, 1.4b, and 6.9b in the main text, with full results in the appendix.

Nonlinear and linear ID scale differently with model size Like in previous work [7, 34, 10, 22, 13], we confirm that inputs are represented in a nonlinear manifold with orders-of-magnitude lower dimension than the ambient dimension. In particular, we find that $I_d \sim O(10)$ across models sizes (see Figure 1 left). We find, moreover, that larger models tend to have higher representational dimensionality, but that the scaling is not uniform. Figure 1 shows that while the linear d scales linearly with hidden dimension D, nonlinear I_d instead stabilizes to the mentioned range $\sim O(10)$

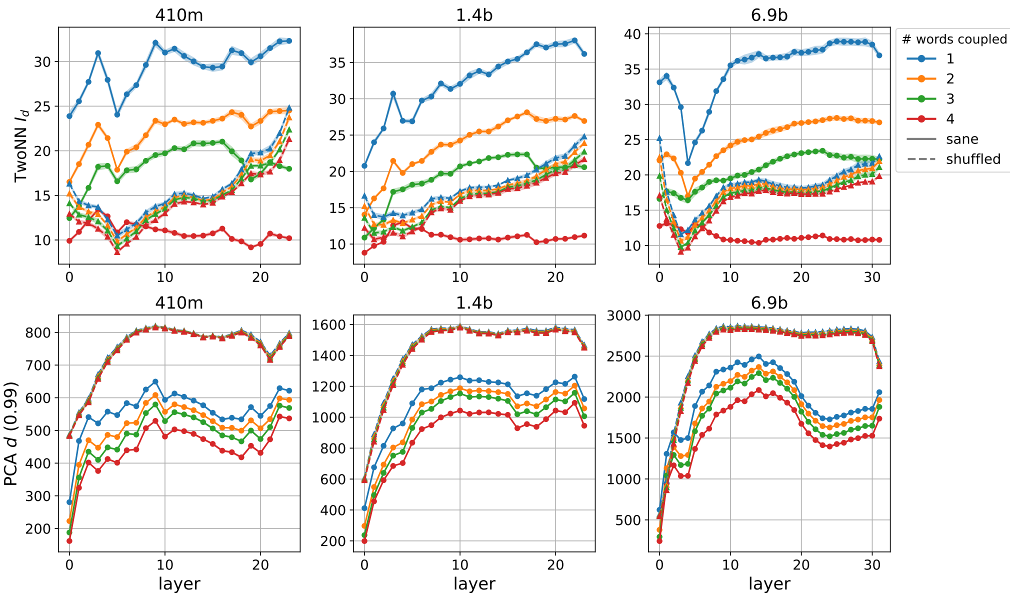

Figure 2: **Dimensionality over layers.** Nonlinear I_d (top) and linear d (bottom) over layers are shown for three sizes: 410m, 1.4b, and 6.9b (left to right). Each color corresponds to a coupling length $k \in 1 \cdots 4$. Solid curves denote sane sequences, and dotted curves denote shuffled sequences. 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.

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. Figure 2 shows I_d and d over LM layers for $k = 1...4$ coupling lengths (different colors). For both sane and shuffled 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, while the lowest correspond to the 4-coupled dataset, or 3 degrees of freedom.

Now, we consider sequence-level compositional semantics. See Figure 2 again for the dimensionality over layers in sane (solid curves) and shuffled (dotted curves) settings. Intriguingly, nonlinear and linear dimensionalities of shuffled examples show opposing patterns: compared to the sane text, shuffled text I_d generally decreases and is compressed to a small range, while d *increases*. These diverging patterns do not necessarily contradict each other, however. We interpret the discrepancy in line with Recanatesi et al. [32]. Predictive coding requires an LM to encode the vast space of inputs and outputs, as well as extract latent semantic features to support the former. Recanatesi et al. [32] argue that encoding all possible sequences makes use of the *global* representation space $\mathbb{R}^D$; instead, encoding semantic relationships between sequences, i.e., latent features, occurs via *local* correlations that give rise to a I_d-dimensional manifold. In our setting, randomly permuting words in a length-l sequence increases the implied input space by a factor of $\sim l!$, which puts an upward pressure on d. But, permuting words destroys the semantics of the sequence, exerting a downward pressure on I_d.

5 Discussion

We have studied the computational mechanism of LM compositionality from a geometric perspective. Using a carefully designed synthetic dataset, we found strong relationships between the compositionality of linguistic expressions and the geometric complexity of their representations. In particular, dataset combinatorial compositionality is positively correlated to both nonlinear and linear dimensionality. On the other hand, sequences with high semantic compositionality exhibit high nonlinear I_d but a low linear d. 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. We hypothesize that linear d proxies a dataset's implied size, and nonlinear I_d its meaningful semantic variability.

Limitations Our analysis is limited to the Pythia family of models. Though it has been suggested that causal LMs have similar representational geometry [29, 10], experiments on a wider range of LMs and grammars, as well as theoretical work, will be necessary to draw general conclusions.

References

[1] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 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, August 2021. Association for Computational Linguistics.

[2] 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.

[3] Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron Courville. Systematic generalization: what is required and can it be learned? *arXiv preprint arXiv:1811.12889*, 2018.

[4] Marco Baroni. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, 375, 2019. URL `https://api.semanticscholar.org/CorpusID:90260325`.

[5] Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics (ACL), 2016.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Jade Yu, Alessandro Laio, and Marco Baroni. Emergence of a high-dimensional abstraction phase in language transformers, 2024. URL `https://arxiv.org/abs/2405.15471`.

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

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

[13] Diego Doimo, Alessandro Serra, Alessio Ansuini, and Alberto Cazzaniga. The representation landscape of few-shot learning and fine-tuning in large language models, 2024. URL `https://arxiv.org/abs/2409.03662`.

[14] 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.

[15] 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.

[16] Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

[17] 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.

[18] 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`.

[19] Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, 2023.

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

[21] Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. Understanding transformer memorization recall through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 248–264, 2023.

[22] Evan Hernandez and Jacob Andreas. The low-dimensional linear geometry of contextualized word representations. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 82–93, Online, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.7. URL `https://aclanthology.org/2021.conll-1.7`.

[23] Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? *J. Artif. Intell. Res.*, 67:757–795, 2019. URL `https://api.semanticscholar.org/CorpusID:211259383`.

[24] Ian Jolliffe. *Principal Component Analysis*. Springer, 1986.

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

[26] 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`.

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

[28] Gary F Marcus. *The algebraic mind: Integrating connectionism and cognitive science*. MIT press, 2003.

[29] Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=SrC-nwieGJ`.

[30] 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.

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

[32] 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.

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

[34] 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].

[35] Zhong Zhang, Bang Liu, and Junming Shao. 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, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.95. URL `https://aclanthology.org/2023.acl-long.95`.

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.

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

C ID Estimation

TwoNN Estimator A number of methods have been proposed to estimate the nonlinear ID of high-dimensional point clouds [8]. 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 15, which has been found to highly correlate to other state-of-the-art estimators [9, 8].

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 [26]'s Maximum Likelihood Estimator (MLE), a similar, nonlinear measure of I_d. MLE has been used in prior works on representational geometry such as [7, 9, 30], 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 [26]. Like past work [15, 9], we found MLE and TwoNN to be highly correlated, producing results that were nearly identical: compare Figure 1 left to Figure E.3 left, and Figure E.1 top to Figure E.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 $(\sum_i \lambda_i)^2/(\sum_i \lambda_i^2)$ [18]. 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 E.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 E.3 and the dimensionality over layers in Figure E.2 for completeness, we do not attempt to interpret them.

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].

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

E Additional Results

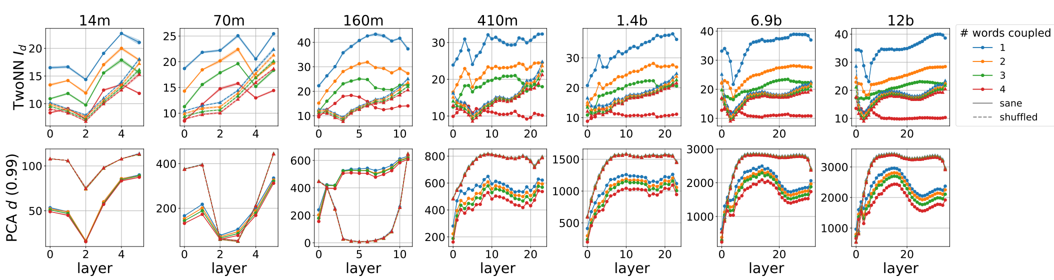

Figure E.1: **Dimensionality over layers.** TwoNN nonlinear I_d (top) and PCA linear d (bottom) over layers are shown for all sizes (left to right). Each color corresponds to a coupling length $k \in 1 \cdots 4$. Solid curves denote sane sequences, and dotted curves denote shuffled sequences. For all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

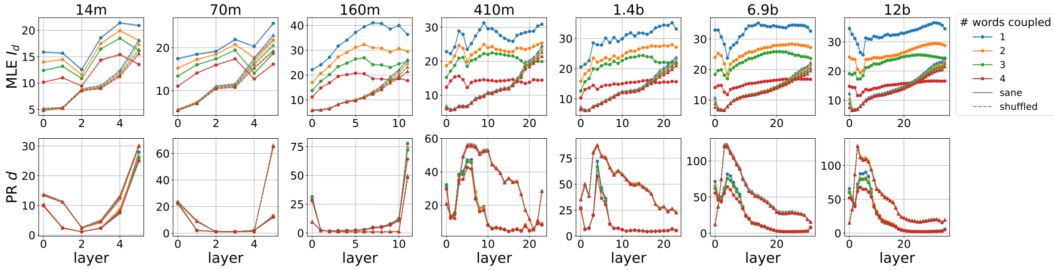

Figure E.2: **Other dimensionality metrics over layers.** MLE nonlinear I_d (top) and PR linear d (bottom) over layers are shown for all model sizes (left to right). Each color corresponds to a coupling length $k \in 1 \cdots 4$. Solid curves denote sane sequences, and dotted curves denote shuffled sequences.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

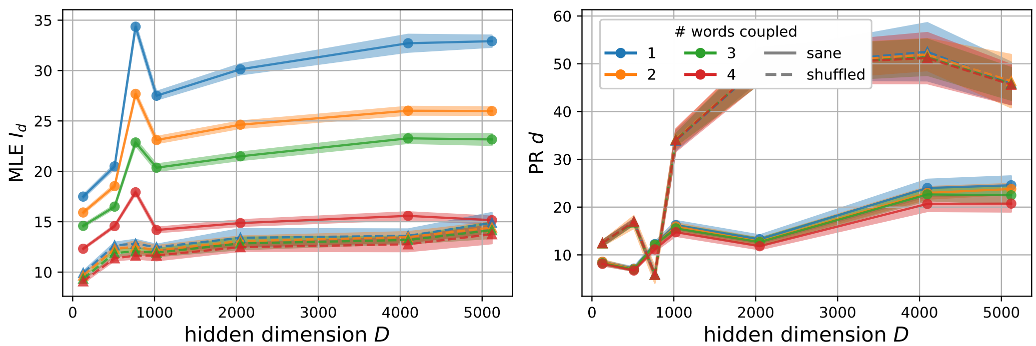

Figure E.3: **Mean dimensionality over model size (other metrics).** Mean nonlinear I_d computed with MLE (left) and linear d computed with PR (right) over layers is shown for increasing LM hidden dimension D. MLE I_d does not depend on extrinsic dimension D (flat lines). PR d produces nonsensical values,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

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Sections 4.1-3 support claims in the Abstract and Introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The "Limitations" subsection in Section 4 discusses the limitations of this work.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This is an empirical work.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Data preprocessing steps are provided in the Appendix. Code will be deanonymized upon acceptance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will release the code on github after the notification decision.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The details necessary to understand the results can be found in appendices A and B

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Standard errors are plotted for all results (although they are often too small to see).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In Appendix A.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics and made sure to follow it

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The "Broader Impacts" subsection in section 4 discusses broader impacts

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Assets used are the estimator implementations and the pretrained Pythia models Appendix B.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new assets are introduced in the paper

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No human crowdsourcing.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No human experiments.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.