The Shape of Learning: Anisotropy and Intrinsic Dimensions in Transformer-Based Models

Anton Razzhigaev[1,2], Matvey Mikhalchuk[2,4], Elizaveta Goncharova[2,5], Ivan Oseledets[1,2], Denis Dimitrov[2,3,4], and Andrey Kuznetsov[2,3,6]
[1]Skoltech, [2]AIRI, [3]SberAI,
[4]Lomonosov Moscow State University,
[5]HSE University,
[6]Samara National Research University
razzhigaev@skol.tech

Abstract

In this study, we present an investigation into the anisotropy dynamics and intrinsic dimension of embeddings in transformer architectures, focusing on the dichotomy between encoders and decoders. Our findings reveal that anisotropy profile in transformer decoders exhibits a distinct bell-shaped curve, with the highest anisotropy concentrations in the middle layers. This pattern diverges from the more uniformly distributed anisotropy observed in encoders. In addition, we found that the intrinsic dimension of embeddings increases during the initial phases of training, indicating an expansion into the higher-dimensional space. Which is then followed by a compression phase towards the end of the training with dimensionality decrease, suggesting a refinement into more compact representations. Our results provide fresh insights on the understanding of encoders and decoders embedding properties.

1 Introduction

Introduced by Vaswani et al. (2017), the transformers have underpinned many breakthroughs, ranging from language modeling to text-to-image generation. As the adoption of transformers has grown, so has the pursuit to understand the intricacies of their internal mechanisms, particularly in the realm of embeddings.

Embeddings in transformers are intricate structures, encoding vast amounts of linguistic nuances and patterns. Historically, researchers have mainly examined embeddings for their linguistic capabilities (Ettinger et al., 2016; Belinkov et al., 2017; Pimentel et al., 2022). Yet, more nuanced properties lie beyond these traditional scopes, like anisotropy and intrinsic dimensionality, which can offer critical insights into the very nature and behavior of these embeddings.

Anisotropy, essentially representing the non-uniformity of a distribution in space, provides a lens, through which we can study orientation and concentration of the embeddings (Ethayarajh, 2019; Biś et al., 2021). A higher degree of anisotropy suggests that vectors are more clustered or directed in specific orientations. In contrast, the intrinsic dimension offers a measure of the effective data dimensionality, highlighting the essence of information that is captured by the embeddings. Together, these metrics can serve as pivotal tools to probe into the black-box nature of transformers.

Our investigation uncovers the striking contrast in the anisotropy dynamics between transformer encoders and decoders. By analyzing the training phases of various transformer models, we shed light on the consistent yet previously unrecognized patterns of the anisotropy growth. Even more, our analysis reveals a unique dynamic of the averaged intrinsic dimension across layers in decoders: an initial growth during the early stages of training is followed by a decline towards the end. This suggests a two-phase learning strategy, where the model initially tries to unfold information in higher dimensional spaces and subsequently compresses it into more compact concepts, possibly leading to more refined representations.

Main Contributions:

- Uncovered a distinct bell-shaped curve for the anisotropy profile[1] in transformer decoders, contrasting with the uniformly distributed anisotropy in encoders.
- Confirmed that anisotropy increases progressively in the decoders as the training proceeds.
- Identified a two-phase dynamic in the intrinsic dimension of decoder embeddings: an initial expansion into higher-dimensional space, followed by a compression phase indicating a shift towards compact representations.

[1]Layer-wise anisotropy

Findings of the Association for Computational Linguistics: EACL 2024, pages 868–874

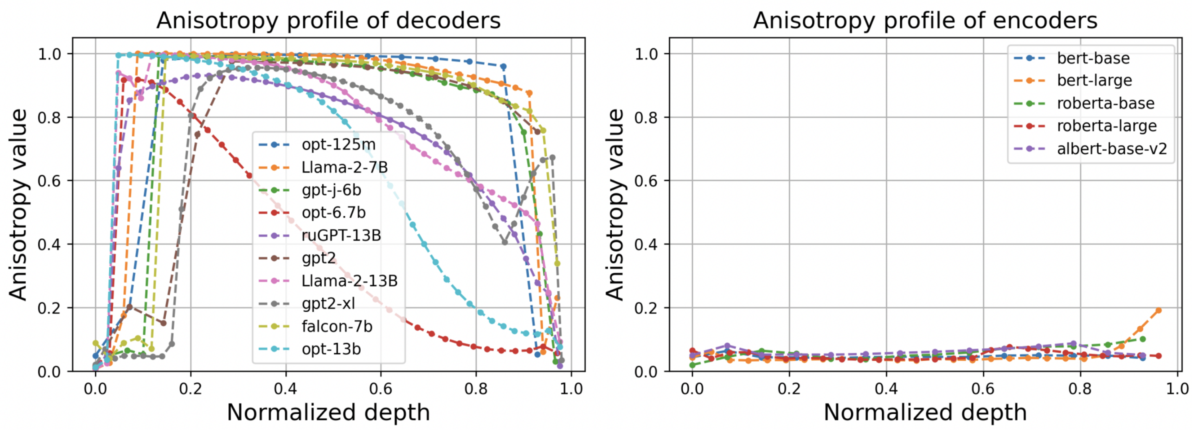

Figure 1: Different anisotropy profiles for transformer-based encoders and decoders.

2 Methodology

2.1 Datasets

As our source for embedding we chose enwik8 dataset (English Wikipedia[2]) that contains 100 million bytes of Wikipedia dump, making it a rich source of diverse textual content. It is publicly available through the Hutter Prize website[3]. The preprocessing stage includes the removal of all the code, media, and HTML tags, resulting in a clean and structured dataset with the vocabulary of 205 distinct characters.

2.2 Embeddings

The vectors are grouped into batches, each with a minimum of 4096 elements. We apply the selected method to determine anisotropy or intrinsic dimension to this batch. Prior to assessing intrinsic dimension, the embeddings are shuffled (before batching) to mitigate potential correlations. The results from individual batches are then averaged to calculate the metric for that layer, also capturing the standard deviation.

2.3 Anisotropy

To compute anisotropy, we employ the singular value decomposition (SVD).

Let $X \in \mathbb{R}^{n\_\text{samples} \times \text{emb\_dim}}$ represent the centered matrix of embeddings, where $\sigma_1, \ldots, \sigma_k$ are its singular values. The anisotropy score of X is given by:

$$\text{anisotropy}(X) = \frac{\sigma_1^2}{\sum_{i=1}^{k} \sigma_i^2}.$$

[2] https://www.wikipedia.org/

[3] http://prize.hutter1.net

Equivalently, this can be deduced using the eigenvalues $\sigma_1^2, \ldots, \sigma_k^2$ of the covariance matrix:

$$C = \frac{X^T X}{n\_\text{samples} - 1}.$$

For some models, we compare the anisotropy measurement approach based on the SVD decomposition with the average cosine (Ethayarajh, 2019; Biś et al., 2021) between embeddings for each layer.

$$\text{average\_cosine} = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \cos(X_i, X_j),$$

where X_i and X_j denote two vectors of embeddings of the same layer (these vectors can originate from different contexts and correspond to different model inputs).

We also study the effect of the centering (subtraction of average vector from embeddings before calculations) for these two types of metrics.

2.4 Intrinsic Dimension

To determine the intrinsic dimension of a set of embeddings, we utilize the approach proposed by Facco et al. (2018). This method explores how the volume of an n-dimensional sphere (representing the count of embeddings) scales with dimension d.

For each data point within our embeddings, we determine the distances r_1 and r_2 to their two closest neighboring points. This process generates a set of pairs $\{(r_1, r_2)\}$. Using this set, the intrinsic dimension d can be estimated. Firstly, we define:

$$\mu_i = \frac{r_2}{r_1},$$

for each point i.

The cumulative distribution function (CDF) of $\{\mu_i\}$ is provided by:

$$F(\mu) = (1 - \mu^{-d})\mathbf{1}_{[1,+\infty)}(\mu).$$

This expression for F is based on the derivations and proofs presented by the authors of the referenced paper. From the CDF, we deduce:

$$\frac{\log(1 - F(\mu))}{\log(\mu)} = d.$$

To estimate d, linear regression $y = kx$ is applied on the plane (x, y), with:

$$x_i = \log(\mu_i) \quad \text{and} \quad y_i = 1 - F_{\text{emp}}(\mu_i),$$

where F_{emp} signifies the empirical CDF for $\{\mu_i\}$.

For some models, we also measure the intrinsic dimension by other local methods. We use Manifold-adaptive dimension estimation (Farahmand et al., 2007) and Method of Moments (Amsaleg et al., 2018).

All three local methods show correlating results in our experiments.

3 Related Work

3.1 Isotropy of Hidden Representations

Gao et al. (2019) introduce the *representation degeneration problem*. This is the phenomenon of degenerating in the representation of learned embeddings in the generative models, particularly when they are tied. The authors conclude that, unlike fixed word embeddings (e.g., word2vec (Mikolov et al., 2013)), vanilla transformer embeddings are clustered within the narrow cone.

Recent research revealed that global anisotropy is a common trait among all transformer-based architectures (Ait-Saada and Nadif, 2023; Godey et al., 2023; Tyshchuk et al., 2023). However, within the local subspaces, isotropy prevails, enhancing model expressiveness and contributing to high performance in the downstream tasks.

Ding et al. (2022) conducted an extensive empirical evaluation of modern anisotropy calibration methods, showing no statistically significant improvements in the downstream tasks. They conclude that the local isotropy of the hidden space of transformers may lead to the high level of model's expressiveness (Cai et al., 2021). While most isotropy findings are observed in encoder-only or encoder-decoder architectures, Cai et al. (2021) brought an interesting variation to light. The authors conducted experiments on various architectures, evaluating the reduced effective embedding dimension using PCA, and observed high cosine values across the layers, especially in models such as GPT-2 (decoder).

The work (Ait-Saada and Nadif, 2023) supports previous research through extensive experimental evaluation. This study arose from the presence of local isotropy in hidden representations, suggesting that anisotropy does not necessarily compromise the expressiveness of these representations.

Godey et al. (2023) investigated the potential causes of anisotropy, particularly its connection to rare words in the model's vocabulary. They explored character-level models to eliminate the influence of rare tokens, but these models did not show any significant improvements in the experiments. The authors also uncovered that adding common bias term to the inputs can lead to the increased attention score variance, promoting the emergence of categorical patterns in self-attention softmax distributions. Increasing input embeddings norm shows signs of anisotropy based on the query and key values.

3.2 Intrinsic Dimensionality

Following the idea of local isotropy of the hidden representations, the investigation of the intrinsic task-specific subspaces offers new insights into the fine-tuning and also the potential to improve model efficiency. Li et al. (2018) suggested that the training trajectory of Transformer architectures occurs in a low-dimensional subspace. Zhang et al. (2023) demonstrated that fine-tuning engages only a small portion of the model's parameters, and it is possible to identify the principal directions of these intrinsic task-specific subspaces. Using their method of identifying the training direction they achieved performance similar to the fine-tuning in the full parameter space.

Tulchinskii et al. (2023) employed intrinsic dimension estimation to identify AI-generated texts. Specifically, they utilized the persistent homology dimension estimator (Schweinhart, 2021) as the tool for assessing dimensionality. The findings revealed that the intrinsic dimension of natural texts tends to cluster between higher values in comparison to generated texts. The latter exhibits a lower dimension, irrespective of the specific generator involved.

3.3 Training Progress

Prior research has utilized information criteria to investigate the internal regularization mechanisms of neural networks. Shwartz-Ziv and Tishby (2017) delve into simple fully connected networks and advocate for identifying a trade-off between information compression and prediction at each layer of the network. They contend that a significant portion of training epochs in deep fully-connected networks focuses on compressing the input into an efficient representation rather than fitting the training labels.

In (Achille et al., 2019), the authors found that the training process of deep neural networks is not monotonic with respect to information memorization. They identified two distinct stages in the training process. The initial stage is marked by rapid information growth, resembling a memorization procedure, while the subsequent stage involves a reduction of information — referred to as "reorganization" or "forgetting" by the authors.

This findings is on par with our observations regarding the two-phase training of the language models, where the intrinsic dimension experiences initial growth followed by a subsequent decline. Notably, during this phase, the model's performance exhibits steady improvement (see Section 4.3 and Figure 5).

3.4 Encoder and Decoder Architectures

The original transformer architecture consists of both encoder and decoder blocks, and each of these blocks can operate independently. The self-attention mechanism is a shared key feature, with decoders utilizing causal self-attention. Decoders are typically trained for language modeling tasks, focusing on generating coherent sequences of the text. In contrast, encoders are aimed to produce contextual representations (i.e., embeddings), from the input text.

Taking limited previous research on the distinctions between the inner representations of encoders and decoders into account, our study analyzes multiple encoder-based models (such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2020)), and decoder-based models (including OPT 125M-13B (Zhang et al., 2022), Llama-2 7B-13B, Llama-2 7B Chat (Touvron et al., 2023), GPT2 (Radford et al., 2019), GPT-J (Wang and Komatsuzaki, 2021), Falcon-7B, and Falcon-7B-Instruct (Almazrouei et al., 2023)) to offer a comprehensive comparison of their behavior.

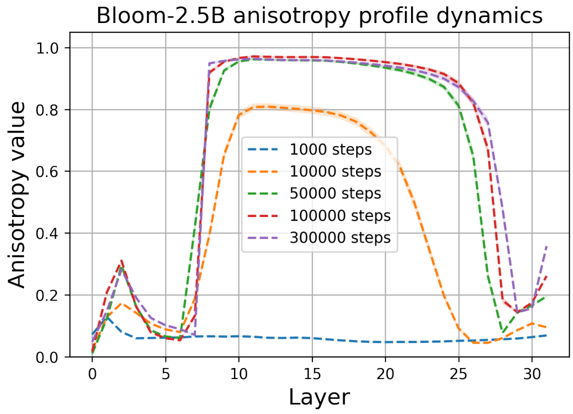

Figure 2: Anisotropy profile for Bloom-3B at different number of pretraining steps.

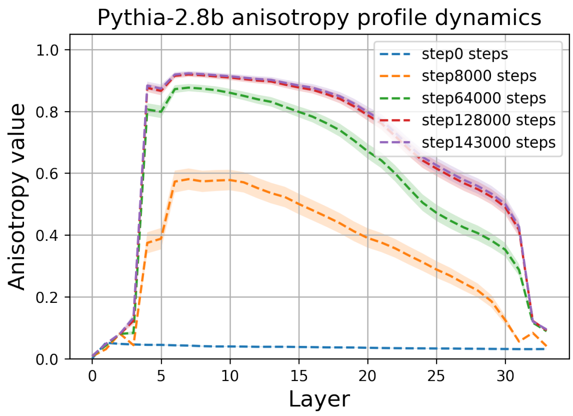

Figure 3: Anisotropy profile for Pythia-2.8B at different number of pretraining steps.

4 Results

In this section, we present our empirical findings concerning the anisotropy dynamics and intrinsic dimensionality of transformer embeddings at different layers. Our results span various pretrained transformer models, showcasing clear patterns in the behavior of encoders versus decoders, and illuminating the transformation of their properties during training.

4.1 Anisotropy Across Pretrained Transformers

We began by comparing the anisotropy levels across various pretrained transformers, analyzing both encoder and decoder models. Their anisotropy profiles can be found in the Figure 1.

Encoders: Anisotropy levels remain relatively consistent across the models, with minor variations based on the model size and training data.

	Bloom-560M	Bloom-1.1B	Bloom-3B	Bloom-7B	Pythia-2.8B	TinyLlama-1.1B
Architecture hyperparameters						
Layers	24	24	30	30	32	22
Hidden dim.	1024	1536	2560	4096	2560	2048
Attention heads	16	16	32	32	32	16
Activation	GELU				GELU	SwiGLU
Vocab size	250,680				50,257	32,000
Context length	2048				2048	2048
Position emb.	Alibi				RoPE	RoPE
Tied emb.	True				False	False
Pretraining hyperparameters						
Global Batch Size	256	256	512	512	1024	1024
Learning rate	3.0e-4	2.5e-4	1.6e-4	1.2e-4	1.6e-4	4.0e-4
Total tokens	341B				300B	3T
Warmup tokens	375M				3B	4B
Min. learning rate	1.0e-5				1.6e-5	4.0e-5

Table 1: Architectural and training configurations of the analyzed models.

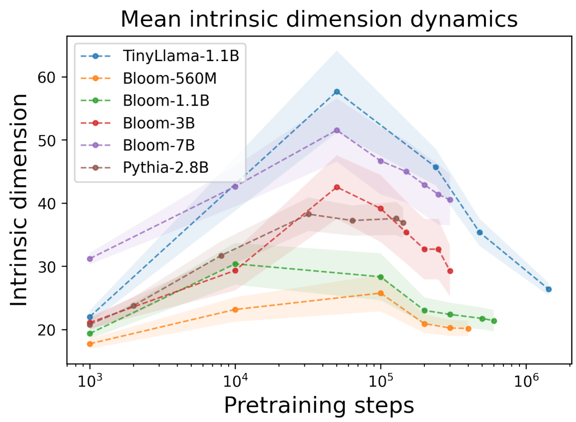

Figure 4: Intrinsic dimension averaged across layers at different pretraining steps.

Decoders: In contrast to the encoders, decoders showcase a unique bell-shaped structure, indicating that the middle layers tend to have a higher anisotropy concentration among all examined models.

4.2 Anisotropy Dynamics During Training

To further probe the evolution of anisotropy, we examine its progression through the training phases of various models.

Figure 2 and Figure 3 capture this trajectory by plotting anisotropy values for decoders at different training checkpoints at all internal layers. The consistent growth pattern, followed by stabilization, is observed across various models, suggesting an inherent characteristic of the language modeling training dynamics of decoders.

4.3 Intrinsic Dimensionality During Training

Our exploration into the intrinsic dimensionality reveals intriguing patterns: Figure 4 displays the averaged intrinsic dimension of models throughout the training process. The initial stages exhibit a sharp rise, indicating the model's attempt to map the information to higher dimensional spaces. However, as training progresses, there is a notable decline, suggesting a subsequent phase where the model compresses this information, refining more compact concepts.

4.4 Model Architecture

For the conducted research, we analyze decoder-based models with similar parameter scales but different architectural and training configurations. In Table 1, we summarize the main solutions for the models presented in Figure 4.

It is noteworthy that there is a considerable difference among models with the same number of parameters (Bloom-1.1B and TinyLlama-1.1B), each featuring distinct architectural configurations. The intrinsic dimension of the latter is higher both at the end of training and at its peak. The obtained results also leads to the conclusion that the growth and the decline of the intrinsic dimension do not show correlation with the warmup period in the learning rate scheduler.

5 Conclusion

Our exploration into the anisotropy dynamics and intrinsic dimensionality of transformer embeddings has brought significant distinctions between encoder and decoder transformers to light. Notably, the intrinsic dimensionality showcases a two-phased training behaviour, where models initially expand information into higher-dimensional spaces and then refine it into compact concepts towards the end of training. These insights not only deepen our understanding of transformer architectures but also suggest new avenues for tailoring training approaches in future NLP research.

Limitations

While our study offers valuable insights into the behavior of transformer embeddings, there are a few limitations to consider.

Model Diversity: Our findings predominantly revolve around specific transformer models, and generalization to all transformer architectures is not guaranteed.

Training Dynamics: The observed two-phased behavior in intrinsic dimensionality might be influenced by the datasets or specific training configurations.

Anisotropy Interpretation: While we identified distinct anisotropy patterns in encoders and decoders, the direct implications of these patterns on downstream tasks remain to be fully explored.

Ethics Statement

Our research focuses on analyzing transformer embeddings and does not involve human subjects or sensitive data. All findings are derived from publicly available models and datasets. We strive for transparency and reproducibility in our methods and analyses.

References

Alessandro Achille, Matteo Rovere, and Stefano Soatto. 2019. Critical learning periods in deep neural networks.

Mira Ait-Saada and Mohamed Nadif. 2023. Is anisotropy truly harmful? a case study on text clustering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1194–1203, Toronto, Canada. Association for Computational Linguistics.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Laurent Amsaleg, Oussama Chelly, Teddy Furon, Stéphane Girard, Michael Houle, Ken-ichi Kawarabayashi, and Michael Nett. 2018. Extreme-value-theoretic estimation of local intrinsic dimensionality. *Data Mining and Knowledge Discovery*, 32:1–38.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Daniel Biś, Maksim Podkorytov, and Xiuwen Liu. 2021. Too much in common: Shifting of embeddings in transformer language models and its implications. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5117–5130, Online. Association for Computational Linguistics.

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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yue Ding, Karolis Martinkus, Damian Pascual, Simon Clematide, and Roger Wattenhofer. 2022. On isotropy calibration of transformer models. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 1–9, Dublin, Ireland. Association for Computational Linguistics.

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

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139, Berlin, Germany. Association for Computational Linguistics.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. 2018. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *CoRR*, abs/1803.06992.

Amir Massoud Farahmand, Csaba Szepesvári, and Jean-Yves Audibert. 2007. Manifold-adaptive dimension estimation. In *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, volume 227 of *ACM International Conference Proceeding Series*, pages 265–272. ACM.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models.

Nathan Godey, Éric de la Clergerie, and Benoît Sagot. 2023. Is anisotropy inherent to transformers? *CoRR*, abs/2306.07656.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. 2018. Measuring the intrinsic dimension of objective landscapes. *CoRR*, abs/1804.08838.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Tiago Pimentel, Josef Valvoda, Niklas Stoehr, and Ryan Cotterell. 2022. Attentional probe: Estimating a module's functional potential. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11459–11472, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Benjamin Schweinhart. 2021. Persistent homology and the upper box dimension. *Discret. Comput. Geom.*, 65(2):331–364.

Ravid Shwartz-Ziv and Naftali Tishby. 2017. Opening the black box of deep neural networks via information.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Serguei Barannikov, Irina Piontkovskaya, Sergey Nikolenko, and Evgeny Burnaev. 2023. Intrinsic dimension estimation for robust detection of ai-generated texts.

Kirill Tyshchuk, Polina Karpikova, Andrew Spiridonov, Anastasiia Prutianova, Anton Razzhigaev, and Alexander Panchenko. 2023. On isotropy of multimodal embeddings. *Inf.*, 14(7):392.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

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 Alternative ID and Anisotropy Estimation Methods

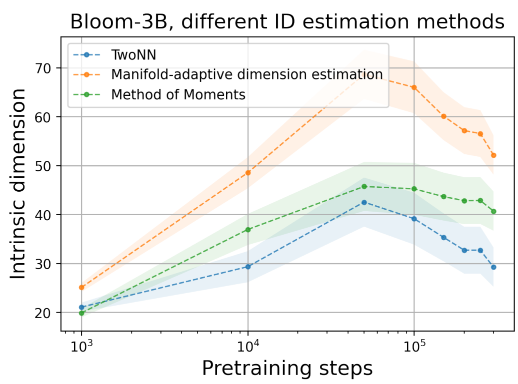

Figure 5: Intrinsic dimension (ID) averages across layers at different pretraining steps estimated via 3 different algorithms.