A Geometric Approach to Understanding How Large Language Models Organize Linguistic Categories

Anonymous ACL submission

Abstract

We investigate how LLMs internally represent different parts of speech, focusing on the distinction between content and function words. We apply intrinsic dimensionality (ID) analysis to hidden state representations in three LLMs: Mistral-7B, Llama-3-8B, and Pythia-12B, computing ID across model layers for representations grouped by part of speech. Our findings reveal substantial differences in representational complexity across POS categories and model architectures. Pythia and Llama display clear geometric stratification between content and function words, whereas Mistral encodes more uniform representations. Complementary clustering analyses confirm that models differ in how clearly they encode linguistic categories. Overall, we demonstrate that ID is a useful technique to gain insights into how LLMs represent their inputs, and we report non-trivial differences in how different LLMs process linguistic categories.

1 Introduction

Large language models (LLMs) have achieved remarkable success across NLP tasks, but understanding their internal mechanisms remains challenging. One promising approach is examining the geometric properties of their internal representations using intrinsic dimensionality, which measures the effective degrees of freedom in high-dimensional data (Valeriani et al., 2023; Cheng et al., 2025). Prior work has applied ID analysis at a coarse-grained level, observing complexity changes at given layers without identifying which linguistic characteristics drive these changes. As a first step toward finer-grained understanding, we examine ID in relation to part-of-speech (POS) categories, which distinguish between content words (nouns, verbs, adjectives) carrying semantic meaning and function words (determiners, prepositions, conjunctions) providing grammatical structure.

2 Related work

Our approach builds on a growing body of work using geometric analysis to understand neural representations, with a particular focus on ID. Neural network representations often lie on low-dimensional manifolds within their high-dimensional embedding spaces, and the dimensionality of these manifolds offers insight into model behavior and generalization (Recanatesi et al., 2019; Pope et al., 2021).

Early work by Ansuini et al. (2019) found that the hidden representations of CNNs follow a "hunchback" pattern of ID across layers (first increasing, then decreasing), with lower ID in final layers correlating with better generalization. Recanatesi et al. (2019) also showed that networks first expand and then compress representations, and that the amount of compression in later layers predicts performance. These foundational studies established ID as a useful geometric tool for understanding neural network representation dynamics.

Building on these insights, researchers have applied ID analysis to transformer architectures across diverse domains.

Cheng et al. (2025) conducted a comprehensive study across multiple transformer LLMs, identifying phase transitions in ID profiles where intermediate layers exhibit increased dimensionality before representations collapse into lower-dimensional manifolds.

3 Methodology

Data We use the WikiText-103 corpus (Merity et al., 2017), containing over 100 million tokens from Wikipedia's *Good* and *Featured* articles. We annotated each token with part-of-speech tags using SpaCy (Honnibal and Montani, 2017), following the Universal Dependencies tagset (see Appendix A). From this tagged corpus, we constructed a dataset of 3 million samples covering a wide

range of POS categories.

Each non-overlapping sample consists of four consecutive sentences. From the fourth sentence, we randomly select a target word with its POS tag, ensuring sufficient context while maintaining natural linguistic variation.

For statistical robustness, we sample five non-overlapping datasets, each containing up to 10,000 examples per POS tag, resulting in approximately 50,000 contexts per category. All experiments are repeated across these partitions for stable averaging and error estimation. Low-frequency tags like `SCONJ` contain fewer examples and are distributed accordingly. We exclude very low-frequency or anomalous tags (`INTJ`, `NUM`, `PUNCT`, `SYM`, `X`), which represent special tokens like interjections, digits, punctuation, and symbols.

Models We analyze three autoregressive language models with different architectures: Pythia-12B-deduped (Biderman et al., 2023) (36 layers, hidden dimension 5120), Llama-3-8B (Meta AI, 2024) (32 layers, hidden dimension 4096), and Mistral-7B (Jiang et al., 2023) (32 layers, hidden dimension 4096).

Representation Extraction To extract the contextualized representation of target words, we feed each 4-sentence sample truncated immediately after the target word to a model and retrieve the hidden state from the final token position across all layers. We then group the extracted representations by the POS of the target word to analyze how the model encodes different syntactic categories in its internal representations.

Intrinsic Dimensionality Estimation We compute ID of hidden state representations to analyze how part-of-speech categories are encoded across model layers. Higher ID values suggest distributed, complex representations, while lower values indicate compact manifolds (Ansuini et al., 2019; Pope et al., 2021). We use the Generalized Ratios Intrinsic Dimension Estimator (GRIDE) (Denti et al., 2022) for its robustness to noise. Unlike the commonly used TwoNN estimator, which is sensitive to noise, GRIDE leverages higher-order nearest neighbors for more reliable estimates. Compared to MLE, which can underestimate ID (Levina and Bickel, 2004), GRIDE offers more accurate estimation without discarding data points. GRIDE analyzes how distances between neighboring points scale. For each data point, it computes distances to its n_1-th and n_2-th nearest neighbors and calculates their ratio. The intrinsic dimension is estimated by maximizing the likelihood of observed ratios (Denti et al., 2022). We use the `return_id_scaling_gride` function from DADApy (Glielmo et al., 2022) with $n_1 = n_2/2$. The GRIDE procedures includes a scale hyperparameter we set as illustrated in Appendix B.

4 Results

Figures 1 and 2 show intrinsic dimensionality (ID) profiles for content and function words across Mistral-7B, Llama-3-8B, and Pythia-12B-deduped.

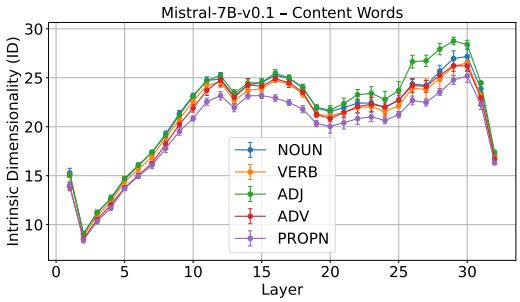
Overall ID patterns. The results confirm that, for all models and POS types, ID is orders of magnitude lower than the ambient dimensions, i.e., the models compress the data so that they lie on a low-dimensional manifold.

We confirm moreover the presence of the mid-layer ID peak reported by Cheng et al. (2025) and additionally note a second peak in later layers, particularly pronounced for Mistral.

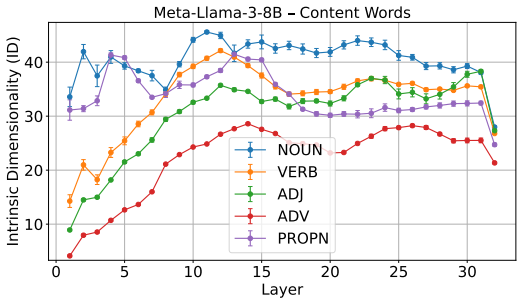
POS separability across models. Models differ substantially in POS differentiation. Mistral exhibits minimal ID divergence across POS categories, with only a 4-dimensional difference between the highest and lowest peak values observed. This narrow range suggests a flatter representational space with limited differentiation between part-of-speech categories. In contrast, Llama and Pythia show greater stratification, approximately 19 and 23 dimensions respectively, indicating more distinct internal encoding across POS types.

Error bars across Figures 1 and 2 show narrow standard errors (typically 0.5 to 2.0 dimensions), confirming the robustness of ID estimates. Variation is smallest in plateau layers and grows near ID peaks. Content words, particularly nouns and proper nouns, exhibit more variability, whereas function word ID values are more stable, possibly due to the smaller number of unique word types.

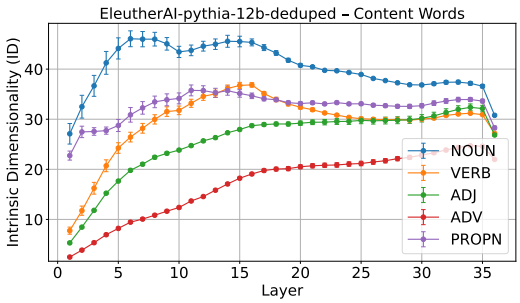
Content vs. function words. In Llama and Pythia, semantically richer content words (NOUN, VERB, ADJ, ADV, PROPN) consistently show higher ID, with nouns highest, followed by verbs and adjectives, then adverbs. Function words (DET, ADP, CCONJ, PRON, SCONJ, PART, AUX) occupy lower-dimensional subspaces, consistent with principles of efficient coding of frequent terms (Gibson et al., 2019). Prepositions (ADP)



(a) Mistral-7B

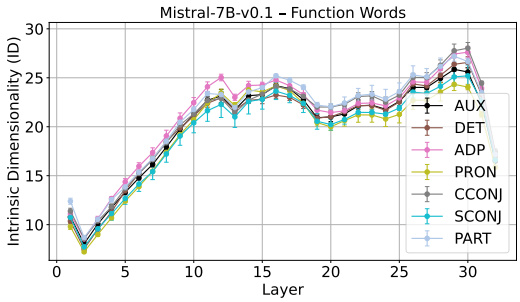


(b) Llama-3-8B

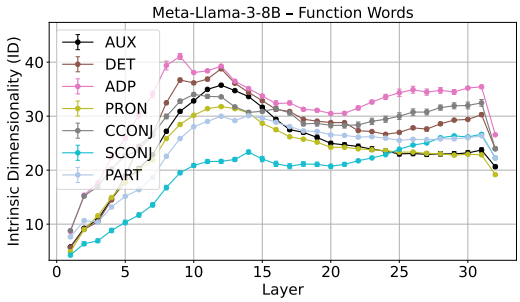


(c) Pythia-12B-deduped

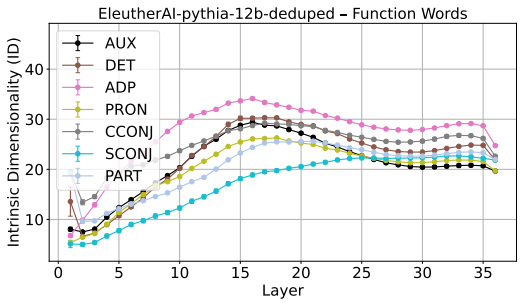
Figure 1: Intrinsic dimensionality (ID) across layers for **content words**.



(a) Mistral-7B



(b) Llama-3-8B



(c) Pythia-12B-deduped

Figure 2: Intrinsic dimensionality (ID) across layers for **function words**.

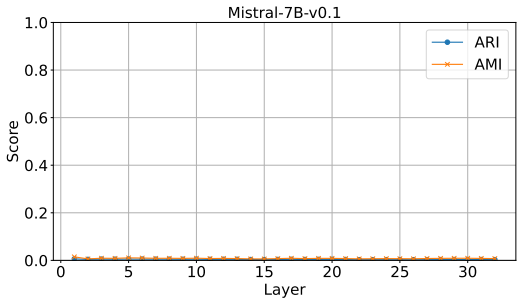
show the highest ID among function words, possibly reflecting their syntactic structuring role.

5 Clustering Analysis

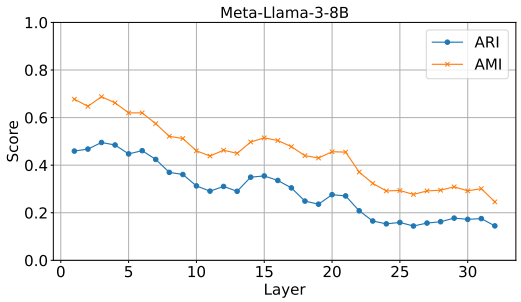
To validate the results we obtain with ID in terms of how much models separate the representations of different syntactic categories, we performed both supervised and unsupervised analyses. In this section, we focus on the unsupervised setting, evaluating whether POS categories form coherent clusters in representation space. Results from a complementary supervised classification experiment are reported in Appendix C.

We conducted K-Means clustering at each layer with the number of clusters set equal to the number of POS tags, and evaluated alignment using Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI) (Figure 3). ARI measures similarity between predicted clusters and POS labels, ranging from -1 to 1, where 1 indicates perfect agreement and 0 random labeling (Hubert and Arabie, 1985). AMI quantifies agreement between clusters and POS categories, ranging from 0 to 1 (Nguyen et al., 2010).

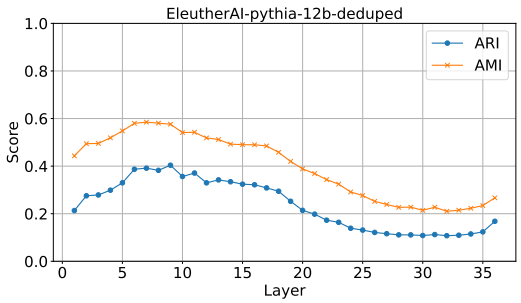
Mistral exhibits near-zero ARI and AMI across layers, confirming its representations do not cluster by POS. Llama achieves the highest scores (ARI 0.5, AMI 0.7 in early–mid layers), decreasing in deeper layers. Pythia shows slightly lower peaks (ARI 0.4, AMI 0.6) but retains moderate separability. To visualize these patterns, we generated confusion matrices at layer 10 (Figure 4). Mistral-7B (Figure 4a) shows POS tags distributed evenly across clusters with no clear alignment. Llama-3-8B (Figure 4b) reveals clusters strongly aligned with tags such as `CCONJ`, `DET`, and `SCONJ`. Pythia-



(a) Mistral-7B

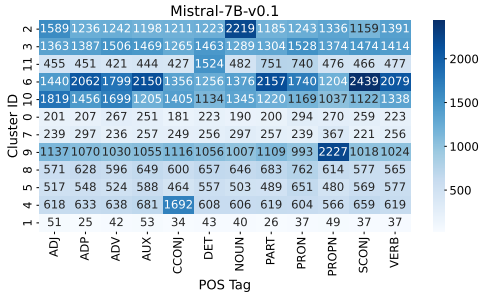


(b) Llama-3-8B

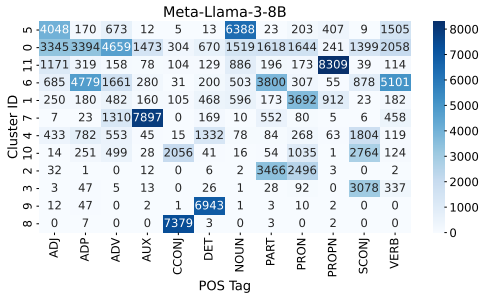


(c) Pythia-12B-deduped

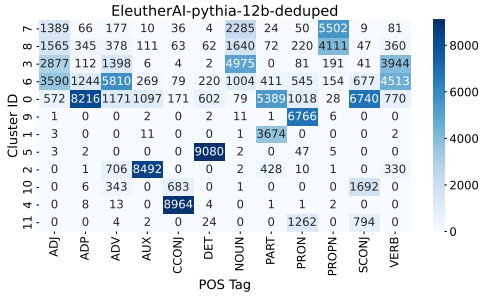
Figure 3: ARI and AMI scores across layers for each model.



(a) Mistral-7B



(b) Llama-3-8B



(c) Pythia-12B-deduped

Figure 4: Confusion matrix for clustering at layer 10.

12B (Figure 4c) also shows strong alignment, particularly for DET, PART, and CCONJ.

6 Conclusion

We analyzed intrinsic dimensionality (ID) of hidden state representations across Mistral-7B, Llama-3-8B, and Pythia-12B, revealing fundamental differences in how models organize linguistic categories. Llama and Pythia exhibit clear geometric stratification where content words occupy higher-dimensional manifolds than function words, while Mistral maintains uniform complexity across POS categories. Interestingly, Llama's geometry aligns more closely with Pythia than with the architecturally more similar Mistral, suggesting POS encoding is influenced by factors beyond architecture.

Overall, our findings confirm ID as a powerful framework for understanding language model organization. Future work should extend this approach to multilingual settings, other linguistic categories beyond POS, and investigate how models like Mistral achieve strong performance without clearly distinguishing word classes.

7 Limitations

Our analysis focused on three autoregressive LLMs using English text: this limits the generalizability to other architectures (e.g., encoder-based models), languages, and domains. The use of WikiText-103 may introduce dataset-specific biases, and, perhaps most importantly, the underlying causes of the observed dimensionality patterns–in particular the question of how Mistral encodes linguistic data without apparently distinguishing between parts of speech–will have to be explored in future work.

References

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. 2019. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.

Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Jade Yu, Alessandro Laio, and Marco Baroni. 2025. Emergence of a high-dimensional abstraction phase in language transformers. In *Proceedings of ICLR*, Singapore. Published online: `https://openreview.net/group?id=ICLR.cc/2025/Conference`.

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

Edward Gibson, Richard Futrell Steven Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How efficiency shapes human language. *Trends in Cognitive Science*, 23(5):389–407.

Aldo Glielmo, Iuri Macocco, Diego Doimo, Matteo Carli, Claudio Zeni, Romina Wild, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. 2022. Dadapy: Distance-based analysis of data-manifolds in python. *Patterns*, 3(10):100589.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2(1):193–218.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh, Fangyu Zhang, Benoît Sagot, and Jenia Jitsev. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Elizaveta Levina and Peter J. Bickel. 2004. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, pages 777–784.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

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

Xuan Vinh Nguyen, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.*, 11:2837–2854.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. 2021. The intrinsic dimension of images and its impact on learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Stefano Recanatesi, Megan Farrell, Madhu S Advani, Subhaneil Lahiri, and Surya Ganguli. 2019. Dimensionality compression and expansion in deep neural networks. abs/1906.00443.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. 2023. The geometry of hidden representations of large transformer models. In *Proceedings of NeurIPS*, pages 51234–51252, New Orleans, LA.

A Universal POS Tag Reference

Table 1 provides definitions for all POS tags used in this study, following the Universal Dependencies framework.

Tag	Description
NOUN	Noun
VERB	Verb
ADJ	Adjective
ADV	Adverb
PRON	Pronoun
PROPN	Proper noun
DET	Determiner
ADP	Adposition (e.g., prepositions)
AUX	Auxiliary verb
CCONJ	Coordinating conjunction
SCONJ	Subordinating conjunction
NUM	Numeral
PART	Particle
INTJ	Interjection
PUNCT	Punctuation
SYM	Symbol
X	Other (uncategorized or foreign)

Table 1: Universal Dependencies POS Tag Definitions

B Scale Selection

The neighborhood scale in ID estimation refers to the number of nearest neighbors used to compute distance ratios (Section 3). We considered scales of 4, 8, 16, 32, 64, 128, 256, and 512.

We selected scale 128 as it balances trade-offs: smaller scales are sensitive to noise, while larger scales may obscure structure. ID profiles remain stable for scales 64 to 256, placing 128 within a consistent region. Figure 5 illustrates this stability for `Pythia`, with similar results for other models.

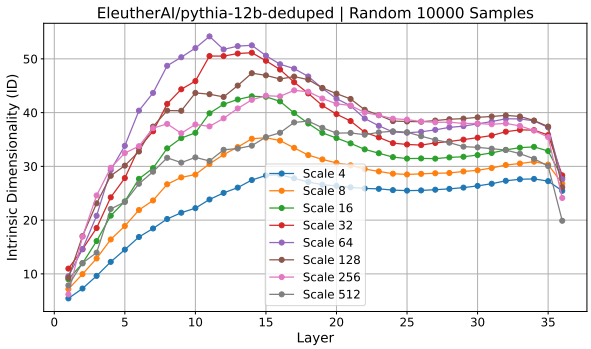


Figure 5: Intrinsic dimensionality (ID) estimated with GRIDE across layers for Pythia-12B-deduped at scales $[4, 8, 16, 32, 64, 128, 256, 512]$.

C Supervised Classification Analysis

we trained binary logistic regression classifiers using One-vs-Rest at each layer. For each POS category, its representations formed the positive class, while an equal number from other categories served as the negative class. We used an 80%/20% train/test split with up to 1000 training iterations. Mistral (Figure 6) shows low accuracy (0.51-0.61, chance=0.50), confirming minimal POS differentiation. Llama (Figure 7) and Pythia (Figure 8) achieve high accuracies (0.95-1.0), validating the clear geometric separation observed in our ID analysis.

D AI Assistant

As English is not the authors' first language, OpenAI ChatGPT was used for grammar and style refinement only. All research design, analysis, and intellectual content are entirely our own.

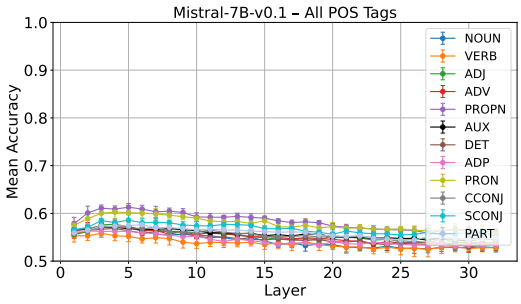


Figure 6: Mistral-7B: Binary classification accuracy across layers for all POS tags.

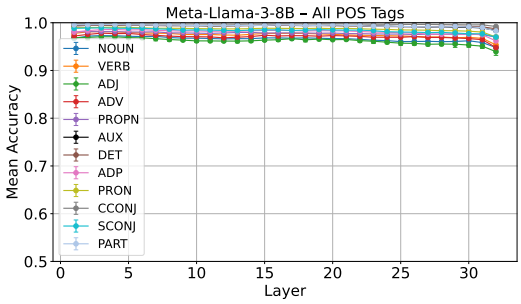


Figure 7: Llama-3-8B: Binary classification accuracy across layers for all POS tags.

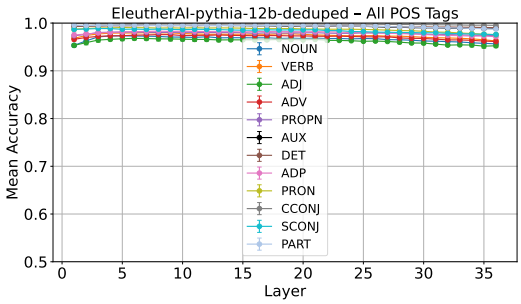


Figure 8: Pythia-12B-deduped: Binary classification accuracy across layers for all POS tags.