Geometry of Decision Making in Language Models

Abhinav Joshi[◇] **Divyanshu Bhatt**[¶†*] **Ashutosh Modi**[◇]
[◇]Indian Institute of Technology Kanpur (IIT Kanpur)
[†]Indian Institute of Technology Hyderabad (IIT Hyderabad)
[¶]Samsung Research & Development Institute, Bangalore
divyanshu.bh@samsung.com,
{ajoshi,ashutoshm}@cse.iitk.ac.in

Abstract

Large Language Models (LLMs) show strong generalization across diverse tasks, yet the internal decision-making processes behind their predictions remain opaque. In this work, we study the geometry of hidden representations in LLMs through the lens of *intrinsic dimension* (ID), focusing specifically on decision-making dynamics in a multiple-choice question answering (MCQA) setting. We perform a large-scale study, with 28 open-weight transformer models and estimate ID across layers using multiple estimators, while also quantifying per-layer performance on MCQA tasks. Our findings reveal a consistent ID pattern across models: early layers operate on low-dimensional manifolds, middle layers expand this space, and later layers compress it again, converging to decision-relevant representations. Together, these results suggest LLMs implicitly learn to project linguistic inputs onto structured, low-dimensional manifolds aligned with task-specific decisions, providing new geometric insights into how generalization and reasoning emerge in language models.

1 Introduction

Large Language Models (LLMs) have exhibited impressive generalization across diverse natural language tasks [Radford et al., 2019, Brown et al., 2020]. Despite their success, how these models internally arrive at decisions, particularly in tasks requiring structured reasoning, remains underexplored. Understanding this process is central to interpretability and may yield insights into model generalization, failure modes, and capabilities. Recent work in mechanistic interpretability has highlighted specific circuits or components underlie LLM reasoning [Elhage et al., 2021, Olsson et al., 2022]. In parallel, probing-based approaches have tracked how task-relevant information flows across layers [Tenney et al., 2019, Hewitt and Manning, 2019]. However, these techniques often focus on how the information is represented and where it resides, rather than how the representation geometry evolves to support decision-making. To complement these perspectives, we study decision-making by analyzing geometric properties of underlying manifolds. We specifically make use of the *Intrinsic Dimension* (ID), which quantifies the minimal degrees of freedom required to describe a distribution in high-dimensional space [Bishop, 2006]. Prior work has demonstrated that neural representations often lie on low-dimensional manifolds [Gong et al., 2019, Valeriani et al., 2023], with ID fluctuations signalling transitions in learning and abstraction [Cheng et al., 2025]. Yet, the connection between these geometric changes and model decisiveness, i.e., the commitment to a specific prediction, has not been explored extensively.

Our primary focus is to understand how internal decision-making unfolds within transformer-based LLMs, particularly in tasks requiring symbolic reasoning and choice commitment. To this end, we

[*]Work primarily done at IIT Hyderabad

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

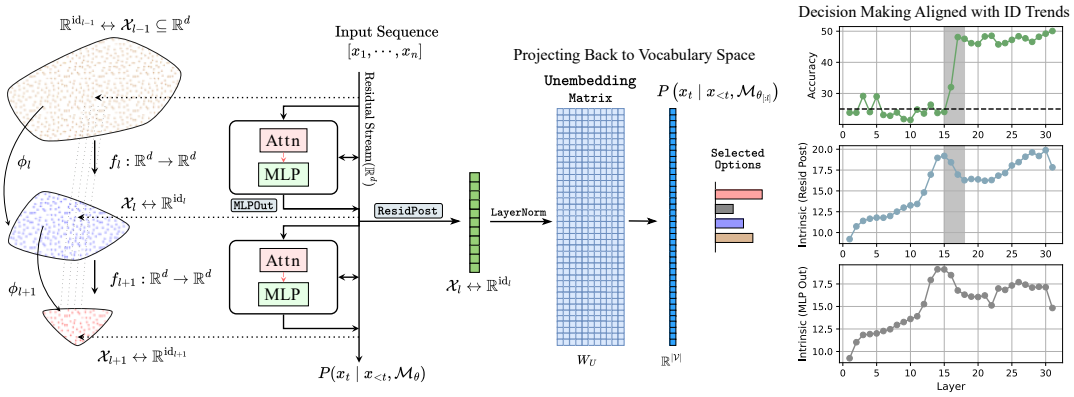

Figure 1: In the transformer-based architectures, a vector (latent features) of the same hidden dimensions d, is transformed by transformer blocks f_l. Though the extrinsic dimension remains the same, we find that the feature space lies on low-dimensional manifolds of different intrinsic dimensions $\mathbb{R}^{\mathrm{id}_l}$. Intrinsically, there exists a mapping ϕ_l corresponding to each f_l, from $\mathbb{R}^{\mathrm{id}_{l-1}} \to \mathbb{R}^{\mathrm{id}_l}$. We study how these compressed manifolds align with the decision-making process in middle layers. We project the internal representations back to the vocabulary space to inspect the decisiveness. There is a sudden shift in performance that is aligned with the follow-up of a sharp peak observed in the residual-post ID estimates.

are guided by three key questions: **1) How** does ID evolve across layers, and how does this reflect the model's progression from contextual encoding to decision-making? **2) Can** geometric markers, such as ID peaks, serve as interpretable indicators of decisiveness and confidence in model predictions? **3) Are** these ID dynamics consistent across different model families and tasks, and what role does model size, training stage, or prompt conditioning (e.g., few-shot examples) play in shaping these trajectories? We aim to bridge representational geometry with functional behavior in LLMs through these questions, providing a complementary perspective to circuit-based or probing-based analyses. Our findings reveal that ID can act as a proxy for representational focus and task commitment, helping identify critical layers that solidify/freeze model decisions, and provide insights that may guide future interpretability and intervention strategies.

In this work, we study the evolution of hidden representations that develop during decision making in LLMs using ID estimates by experimenting with reasoning-based multiple-choice question answering (MCQA)-style prompts. We conduct an extensive investigation into the internal representations of LLMs, analyzing 28 open-weight transformer models spanning multiple architectures and sizes (list of models in App. D). We build upon classical estimators such as Maximum Likelihood Estimation (MLE) [Levina and Bickel, 2004] and Two Nearest Neighbors (TwoNN) [Facco et al., 2017], and incorporate the recently proposed Generalized Ratios Intrinsic Dimension Estimator (GRIDE) [Denti et al., 2022] which demonstrates improved robustness to sampling noise and curvature distortions (see §3). Fig. 1 outlines our approach (details in §4). Our primary findings are as follows:

- **Emergence of Decision Geometry:** Across models and tasks, we observe a characteristic hump-shaped trend in intrinsic dimension estimates (notably at the MLP output layers), where ID increases, peaks, and then declines. This reflects an early phase of abstraction followed by convergence toward decision-specific subspaces.
- **ID Peaks Coincide with Decisiveness:** For most models, the peak in intrinsic dimension aligns closely with the onset of confident predictions (as revealed via projection to vocabulary space). This suggests a geometric marker of decisiveness within the model's forward pass.
- **Layer-Specific Dynamics Differ by Component:** We distinguish between MLP output and residual post-activations. While MLP outputs exhibit clear ID peaks and sharper reductions in later layers, residual post activations show more gradual trends, providing complementary views on when and how information/decisions solidify.
- **Few-Shot Prompting Sharpens Representations:** Increasing few-shot examples leads to steeper ID transitions, implying more efficient compression and faster convergence to decision-ready states.

- **Model Scale and Architecture Matter:** Larger models tend to reach ID peaks earlier in the layer stack and maintain lower terminal ID, hinting at more efficient abstraction and early decisiveness. Notably, model families like LLaMA and Pythia show distinct ID trends, underscoring architectural influence on representational geometry.

In a nutshell, our study covers both real-world benchmarks and template-based tasks, enabling us to characterize representational dynamics across a diverse range of reasoning and language understanding skills. We release the codebase and results at `https://github.com/Exploration-Lab/dim-discovery-archive`.

2 Related Works

The manifold hypothesis posits that high-dimensional data often lie on low-dimensional manifolds [Ruderman, 1994, Brand, 2002, Fefferman et al., 2013, Goodfellow et al., 2016]. In this context, the intrinsic dimension (ID) has emerged as a useful geometric lens for studying neural representations. Prior work has shown that deep networks learn low-ID features [Gong et al., 2019, Ansuini et al., 2019, Aghajanyan et al., 2020, Pope et al., 2021], with lower ID correlating with better generalization [Gong et al., 2019, Nakada and Imaizumi, 2020, Aghajanyan et al., 2020]. These trends have been well-explored in computer vision, where ID is linked to optimization geometry [Li et al., 2018, Ma et al., 2018, Zhu et al., 2018, Zhang et al., 2021] and dataset complexity [Pope et al., 2021, Deng, 2012, Krizhevsky et al., 2009, Deng et al., 2009, Lin et al., 2015, Liu et al., 2015]. In transformers, early studies demonstrated similar ID patterns across layers in models trained on non-text domains like proteins and images [Valeriani et al., 2023]. More recent work has brought these geometric insights into NLP. Cheng et al. [2025] identifies a high-ID abstraction phase in transformers, predictive of generalization and linguistic transfer. Antonello and Cheng [2024] provides complementary fMRI evidence for a two-phase abstraction process, with ID peaks corresponding to the most brain-like representations. Cheng et al. [2023] further bridges geometric and information-theoretic compression, showing that lower ID predicts faster adaptation in LMs. A few studies apply ID to practical NLP tasks. Tulchinskii et al. [2024] use ID to differentiate LLM-generated and human-written text, while Yin et al. [2024] introduce local ID as a metric for hallucination detection. On the other hand, Cheng et al. [2024] leverages intrinsic dimension to refine word embeddings, improving model performance and explainability. However, how ID evolves across layers during reasoning and decision-making remains underexplored. Similarly, while Doimo et al. [2024] examines differences in internal geometry induced by fine-tuning vs. in-context learning, they focus on semantic clustering and representation alignment, not decision making or ID-based trends. Our work fills this gap by analyzing how ID relates to decisiveness during inference, across multiple open-weight LLMs, considering reasoning a central theme. Complementary to recent works [Cheng et al., 2025, Valeriani et al., 2023, Doimo et al., 2024], our study shifts focus from abstraction alone to how LLMs geometrically transition from context encoding to decision formation. More specifically, our study reveals specific trends across multiple models for different reasoning tasks, where the model concretizes its decision throughout layers. We find consistent geometric trends that reflect model predictions and decisiveness (see §5). In particular, we show that ID peaks often coincide with, or slightly precede, the layer at which the model becomes most semantically committed to an answer; we validate this by projecting the mid-layer representations (resid-post and MLP-out) back to the vocabulary space. These findings provide a new geometric perspective on LLM reasoning, linking representational compression to decision formation.

3 Internal Reperesentations in Language Models

Transformer-based Language Modeling: A Language Model (LM) can be modeled as a function f (parameterized by a neural network based architecture) that helps map a sequence of input tokens (prompt) to output a vector of logits, where each entry corresponds to a token in a pre-defined vocabulary. In our study, we primarily focus on the transformer-based decoder-only architectures that are trained in an autoregressive fashion, that are widely adopted by most language models [OpenAI et al., 2024, Gemini et al., 2024]. Given a vocabulary $\mathcal{V}$, an autoregressive language model $\mathcal{M}_\theta$ (θ denotes the model parameters) learns a parameterized function that maps an input space $\mathcal{X}$, containing a sequence of tokens $x = [x_1, \ldots, x_{t-1}] \in \mathcal{X} \subseteq \mathcal{V}^{t-1}$ to an output probability distribution $P_{\mathcal{M}_\theta} : \mathcal{V} \rightarrow [0, 1]$, that helps predicting the next token (x_t) given the sequence of

previous tokens $P(x_t \mid [x_1, \ldots, x_{t-1}])$. Internally, the transformer-based language models consist of transformer blocks/layers $(f_{\theta_1}, f_{\theta_2}, \ldots f_{\theta_L})$ stacked together that read information from and write onto the residual stream (connected by residual connections, see Elhage et al. [2021] for more details), i.e. for an input sequence $[x_1, \ldots, x_{t-1}]$, the model considers representations corresponding to each token (x_i) and finally predicts the distributions corresponding to the next tokens $[x_2, \ldots, x_t]$. For our study, we only consider representations corresponding to the last token in the input prompt, i.e., the token responsible for answering the query present in a prompt. After the last transformer block, the final state of the residual stream is passed through a LayerNorm, which is further then projected onto the vocabulary space via a weight matrix $\mathbf{W}_U \in \mathbb{R}^{|\mathcal{V}| \times d_{model}}$ (also known as Unembedding layer [Elhage et al., 2021]). The final probability distribution $P_{\mathcal{M}_\theta}(x_t)$ is obtained by passing the obtained logits to a softmax, leading to the final prediction. Our goal is to study the geometry of representations learned by these transformer blocks, we consider the representations corresponding to the last token for each layer's output/transformer blocks (i.e. the output corresponding to the MLP layers present in each transformer block) that writes onto the residual stream (§4) (also see Figure 1). Note, unlike CNN-based vision models, the extrinsic dimensions after each layer remain the same $(\mathbb{R}^{d_{model}})$ for the transformer-based models, making the comparison between the layers more reliable.

Intrinsic Dimension: Intrinsic Dimension is defined as the minimum number of dimensions required to describe the data manifold with minimal information loss [Bishop, 2006]. More formally, a set of (data/feature/vector) points $\mathcal{D} \subseteq \mathbb{R}^N$ are said to have intrinsic dimension (ID) equal to d if its elements lie entirely, without information loss, within a d-dimensional manifold of $\mathbb{R}^d$, where $d < N$ [Camastra and Staiano, 2016]. The problem of estimating the dimensions of the underlying manifold, considering a data-generating process, has been an area of interest for the last two decades [Levina and Bickel, 2004, Facco et al., 2017, Bac et al., 2021]. Specifically, the DL community usually prefers estimators based on the scale of the distances between the data points due to their robustness and reliability [Ansuini et al., 2019, Gong et al., 2019, Pope et al., 2021]. A common approach to computing intrinsic dimensions given a set of points is to investigate the space around each point and assume a constant density within the local neighborhoods; the data generation process can be modeled using a homogeneous Poisson Point Process (PPP) [Streit and Streit, 2010]. For our setup, we consider three ID estimators (App. Fig. 6 summarizes the estimators):

MLE [Levina and Bickel, 2004]: Assuming the data generation processing as a PPP, the MLE estimator formulates a likelihood expression as a function of the local intrinsic dimension specific to a data point x, resulting in the following maximum likelihood estimate

$$\hat{d}_k(x) = \left(\frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{T_k(x)}{T_j(x)} \right)$$

where $T_i(x)$ represents the distance of the i^{th} nearest neighbor from the data point x, and k is a hyperparameter for the estimator, making MLE a local intrinsic dimension estimator, i.e., it estimates the intrinsic dimension in the neighborhood of a particular point. For calculating the global intrinsic dimension of the datasets, these local dimensions are aggregated using either the arithmetic or the harmonic mean operator [MacKay and Ghahramani, 2005]:

$$\hat{d} = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \hat{d}_k(x) \quad \text{OR} \quad \hat{d} = \left(\frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \frac{1}{\hat{d}_k(x)} \right)^{-1}$$

TwoNN [Facco et al., 2017]: is a global intrinsic dimension estimator that builds upon the same assumption of points coming from a PPP and formulates ID estimate using a relationship between the cumulative distribution of the random variable μ, defined as the ratio of the distance between the second and the first nearest neighbor and the intrinsic dimension of the dataset and prove it to be Pareto distributed, i.e., $\mu = \frac{T_2(x)}{T_1(x)} \sim \texttt{Pareto}(1, d)$,

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

where d, is the intrinsic dimension, $F(\mu)$ is the cumulative density function. The real-world datasets being i.i.d, the cumulative density function can be estimated given a set of data points as

$$F_{\text{emp}}(x) = \frac{1}{|\mathcal{D}|} \sum_{y \in \mathcal{D}} \mathbb{I}\{\mu(y) \leq \mu(x)\}$$

GRIDE [Denti et al., 2022]: is a recent work that mitigates the sensitivity of TwoNN towards noisy datasets, and provides a generalization over TwoNN. Instead of taking the ratio of the second and the first nearest neighbor, GRIDE uses higher order distances, i.e., the ratio between the n_2^{th} and the n_1^{th} nearest neighbor, where n_1 and n_2 are the hyperparameters of the estimator, i.e., $\mu = \frac{T_{n_2}(x)}{T_{n_1}(x)}$ which is distributed as

$$f(\mu) = \frac{d(\mu^d - 1)^{n_2 - n_1 - 1}}{\mu^{(n_2-1)d+1}\beta(n_2 - n_1, n_1)}$$

where $\beta(\cdot, \cdot)$ is the beta function. The log-likelihood is maximized for the above distribution, resulting in the following optimization problem in d,

$$\max_d \log \mathcal{L}(d) \equiv \max_d (n_2 - n_1 + 1) \sum_{i=1}^{|\mathcal{D}|} \log(\mu_i^d - 1) + |\mathcal{D}| \log d - (n_2 - 1)d \sum_{i=1}^{|\mathcal{D}|} \log \mu_i$$

where μ_i is the ratio of distances corresponding to the i^{th} data point and d is the intrinsic dimension, leading to the above concave optimization problem (see App. A). Some of the other recent extensions include Hidalgo [Allegra et al., 2020], which is more robust when the generated datasets have multiple underlying manifolds; we leave this for future analysis and assume representations coming from a single manifold for this study.

4 Experimental Setup

Decisiveness through Multiple-Choice Prompting In this work, we stick to reasoning captured using multiple-choice question answering (MCQA)-style prompts [Robinson et al., 2023, Wiegreffe et al., 2024]. The MCQA setup provides a principled and constrained setting for investigating the internal decision-making processes of LLMs. Unlike open-ended or cloze-style generation, MCQA structures the task as a selection among discrete alternatives, thereby reducing confounding factors related to token frequency, length bias, and linguistic fluency [Brown et al., 2020]. This format enables precise analysis of the transition from contextual representation to decision, making it well-suited for studying the geometry and structure of intermediate representations. Prior work in mechanistic interpretability has focused on identifying circuits and submodules responsible for reasoning [Elhage et al., 2021, Olsson et al., 2022], while probing studies have examined layer-wise information flow [Tenney et al., 2019, Hewitt and Manning, 2019]. However, relatively little attention has been given to how these representations evolve geometrically to support discrete reasoning tasks. By leveraging MCQA, we aim to isolate and examine the structural properties (specifically intrinsic dimensions) of hidden states as they converge towards a decision, providing insight into the representational dynamics that govern model predictions. In our case, each input prompt is composed of: **1) query information** ($query$): which includes the information related to that specific instance of the dataset. **2) A Choice Set** (**A.** $o_{correct}$; **B.** o_{wrong})) consisting of two or more options from which the LLM must select the correct answer and generate as output the correct choice text: **A** or **B**, or **C**, etc. [Robinson et al., 2023, Wiegreffe et al., 2024, Joshi et al., 2025a,b]. Note that the **A.** and **B.** are for representation, and in the actual run, the correct/wrong options are shuffled to marginalize the effect of models choosing a specific option. The prediction by the LLM ($\mathcal{M}_\theta$) depends on the above two critical components. Additionally, the predictions also depend on how the query is framed, i.e., the prompt template (x_ϵ) used to frame the queries. The predicted probability/logit value of the next token can be written as:

$$P(x_t|x_{i<t}, \mathcal{M}_\theta) = P(x_t|x_{query}, x_{options}, x_\epsilon, \mathcal{M}_\theta) \quad ; \quad \mathcal{M}_\theta = \{f_{\theta_1}, f_{\theta_2}, \dots f_{\theta_L}\}$$

$$x_{query} \leftarrow s_i \sim \mathcal{D} \quad ; \quad x_{options} \leftarrow \{\textbf{A.}\ o_{correct}, \textbf{B.}\ o_{wrong}\}; \quad x_\epsilon \in \text{set of prompt templates}$$

where s_i is a sample/instance from the language-based dataset $\mathcal{D} := \{s_1, s_2, \dots, s_N\}$ of size N. In the LLM the input prompt ($x_{i<t}$) is passed through a sequence of transformer blocks/layers ($f_{\theta_1}, f_{\theta_2}, \dots f_{\theta_L}$), providing a distribution of logits over the vocabulary for the next tokens ($x_1, x_2, \dots, x_t$), we only consider the predicted distribution of the last token (x_t), i.e., the token responsible for predicting the next plausible token or answering the question query: $\mathcal{M}_\theta(x_{i<t}) = f_{\theta_L}(\mathtt{I} + f_{\theta_{L-1}}(\dots(\mathtt{I} + f_{\theta_1}(x_{i<t}))))$. These sequences of operations play a crucial role in modifying the residual stream (the $\mathtt{I}+$ denotes the update in the residual stream), leading to the final predicted token x_t. Essentially, the model $\mathcal{M}_\theta$ processes the input and predicts the next token, which is

expected to be the correct option identifier (e.g., "_A", "_B", etc.). This token serves as a clear decision point, providing a precise locus for analyzing representational geometry.

Representation Extraction and Decision Emergence As the prompt flows through the transformer layers, we collect intermediate hidden representations (h_j) corresponding to the final (decision) token position after each layer $j \in [1, L]$. These vectors trace how the model updates its beliefs through residual stream modifications. Mathematically: $h_j = f_j(x_{i<t}) = f_{\theta_j}(\mathtt{I} + f_{\theta_{j-1}}(\ldots(\mathtt{I} + f_{\theta_1}(x_{i<t}))))$. The representations correspond to the MLP module of the transformer block that writes back to the residual stream, i.e., f_{θ_j} represents the operations in the j^{th} transformer block. We extract the representations from two places in the transformer block: **1) the MLP component** (i.e., after the nonlinearity, before writing back to the residual stream), and **2) the Resid-Post** (i.e., after the residual stream is updated/written by adding the MLP representations) corresponding to the final token (see Fig. 1). These per-layer activations, collected across a dataset $\mathcal{D} = \{s_1, \ldots, s_N\}$, define the manifold structure from which we compute ID using standard estimators: MLE, TwoNN, and GRIDE. Similar to Cheng et al. [2025], we choose representations corresponding to the final token for computing the intrinsic dimension, as this token represents the model's predicted answer and is expected to encapsulate all information necessary for the prediction. Given a dataset $\mathcal{D}$, we get a space of these representations for each layer's output corresponding to text instances present in the dataset, forming a set of $|\mathcal{D}|$ features/representations of the underlying manifold. Note that we only consider the representations corresponding to the last token to form this set. We make use of Transformer-Lens [Nanda and Bloom, 2022] for saving the corresponding representations. The obtained set of vectors is further considered to estimate the ID of the underlying manifold formed by these transformer blocks. Note, in actual transformer implementation, there are two points in a single transformer block where the computational blocks read/write back from/to the residual stream (self-attention and MLP); we skip the mid-skip connection in the equations above for brevity.

Measuring Representation Quality via Logit Accuracy To quantify the semantic sharpness of each layer's representation, we take inspiration from Logit Lens [nostalgebraist, 2020, Haviv et al., 2023], and compute the accuracy of representations at different layers (see Fig. 1). Considering the stacked set of transformer blocks, writing sequentially over the residual stream, we take the representations after each transformer block (MLP and Residual) and project it to the vocabulary space by multiplying it with the unembedding matrix directly ($\mathbf{W}_U$), i.e., $\mathtt{logits}(z_t) = \mathbf{W}_U\mathtt{LayerNorm}(z_t)$, where z_t is the representation corresponding to the last token. We report accuracy using the Residual Post representations rather than the MLP outputs, as the residual stream carries the accumulated state that the model propagates forward across transformer blocks. In contrast, the MLP output contains only the delta/difference added to the residual stream, representing a more localized, high-leverage adjustment rather than the full signal that contain the context. Consequently, the Residual Post signal provides a more presentable view of the model's evolving decision state. Note that we focus on the representation corresponding to the last token, since in autoregressive transformers, this position uniquely has access to the entire context and is solely responsible for generating the next-token prediction. From a decision-making perspective, this is the point at which the model must commit to an output, making it the most informative location for understanding decision-making from a representational perspective. While analyzing intermediate tokens could yield complementary insights into how information is distributed and refined, the last-token view most directly captures where and when the model's decision solidifies. We consider the obtained logits to compute the accuracy of the representations corresponding to a particular layer. The obtained performance estimates not only help quantify the quality of representations but also provide the localization in layers where the model starts to be decisive about the decision/answer/next token. The token-level accuracy at each layer conveys how often that representation alone predicts the correct answer. This metric, coupled with ID, lets us localize the decision emergence layer: the point in the network where the model becomes sharply predictive and the representation starts collapsing into a low-dimensional manifold. In Figures 2 and 3 (see App. Table 9 for other datasets), we visualize this phenomenon across layers and models. We consistently observe that accuracy peaks and ID drops at the same layer, suggesting a tight coupling between task certainty and representational compactness.

Real-World Tasks We first examine LLM behavior on real-world, language-based tasks (MCQA format) where generalization is the key. We specifically choose tasks (and corresponding datasets) related to linguistic abilities (Dataset: CoLA), topic knowledge (Dataset: AG News), field-specific knowledge (MMLU: STEM, humanities, social sciences, other), sentiment analysis (Rotten Tomatoes, SST2), and reasoning abilities (Causal reasoning: COPA, COLD). Note that synthetic datasets

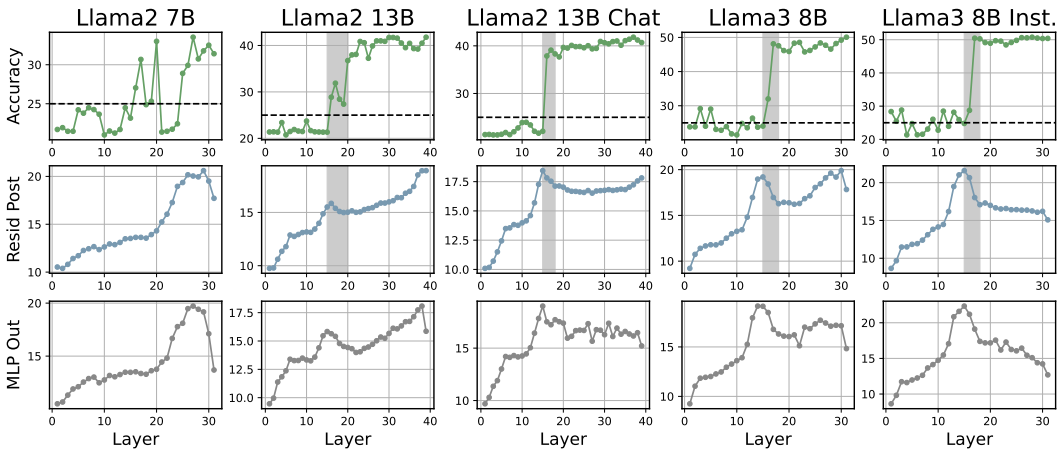

Figure 2: Accuracy along with ID trends for LLaMA model variants on the MMLU STEM dataset.

(template-based) also capture mathematical reasoning in some sense; however, they cannot be considered real-world due to template-based generation. We make minor changes to the template part of the prompt query (x_ϵ) for each dataset. We provide details of datasets in the App. B. We use different prompt templates (details in App. C) for different datasets with minimal changes, keeping the MCQA format consistent. We experiment with various models like Llama2 family, GPT-2 family, Mistral, Phi family, and Gemma family (details in App. D).

Synthetic Tasks: Controlled Learning Trajectories Another aspect of LLMs is their open-ended reasoning via generation. To complement our real-world findings on MCQA-based reasoning, we analyze simplified template-based reasoning tasks where we can observe LLMs learning from scratch under tightly controlled conditions. To monitor the improvements throughout the training trajectory, we require reasoning datasets with low complexity for a comparison that could work for both smaller as well as larger models. We choose the Greater Than (GT) task introduced by Hanna et al. [2023] for simplicity and the arithmetic task [Razeghi et al., 2022], considering its usage by the Pythia suite to monitor performance during model training. **The Greater Than (GT) task** consists of examples of the format "The war lasted from the year 1743 to 17 $\rightarrow xy$", where the language modeling objective is to assign a greater probability to continuations $44, 45, \ldots, 99$ than $00, 01, \ldots, 42$. Random accuracy is upper bounded by $99/|\mathcal{V}|$, where $|\mathcal{V}|$ is the vocab size. **The arithmetic task** also follows a template that consists of input operands $x_1 \in [0, 99]$ and $x_2 \in [1, 50]$ and an output y, i.e. "Q:What is $x_1 \# x_2$? A:" with $\#$ being "plus" for addition and "times" for multiplication. We measure the accuracy of a prompt instance by checking the model's prediction against the label y, making the random accuracy $1/|\mathcal{V}|$. These tasks provide low input complexity but require abstract reasoning, making them ideal for analyzing how internal manifolds evolve over time. We use the Pythia model suite [Biderman et al., 2023], a family of 16 autoregressive transformers (14M–6.9B parameters) (see Fig. 25, 26). With 154 publicly released training checkpoints per model, we can track the formation of decision-critical layers from early to late training (details in App. D).

5 Results and Trends

We conduct a large-scale empirical study analyzing the ID of representations across multiple LLM architectures, tasks, and input prompting settings. Due to space constraints, we describe the main results here, and the remaining ones are provided in the Appendix.

Layerwise Geometry and Task-Specific Trends Fig. 2 and 3 reveal how ID and accuracy evolve across the transformer layers for MMLU-STEM and COPA, respectively. Across models, we observe a characteristic "hunchback" shape in the MLP output's ID profile, i.e. ID increases in early layers, peaks at a mid-network depth, and then declines. Notably, this geometry emerges only in settings where model accuracy rises significantly above baseline (dashed lines), indicating that the presence of the hump is a marker of non-trivial abstraction and task-specific decision-making. The residual post-activations, in contrast, display smoother and more monotonic ID changes, consistent with

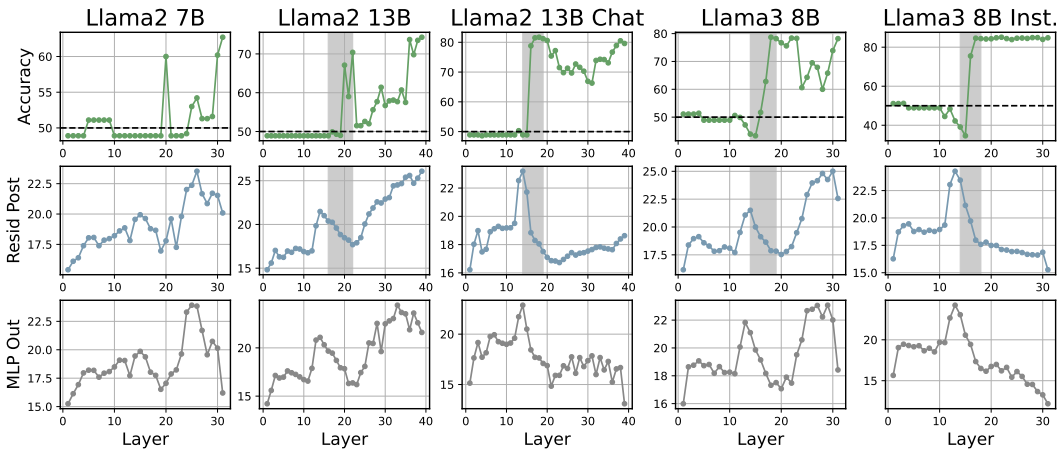

Figure 3: Accuracy along with ID trends for for LLaMA model variants on the COPA dataset.

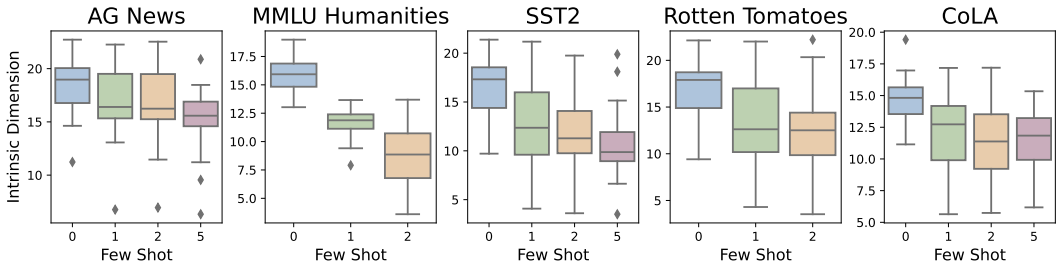

Figure 4: The figure shows the ID of the last layer (MLP Out) feature representation in the in-context learning setting. The box plot shows the distribution of ID for all the 28 open-weight models. Overall, we observe IDs decreasing as more number of examples are provided in the context.

their role in progressively aggregating signals across layers. This distinction is especially prominent in the contrast between reasoning and retrieval tasks. In COPA (Fig. 3), models must synthesize causal and contextual information; here, the ID peak is sharp, and the post-peak ID drop coincides with decisive increases in prediction accuracy. In MMLU-STEM (Fig. 2), a fact-retrieval task, the ID trend is flatter, and the accuracy increases monotonically across layers with no prominent compression phase. Interestingly, we observe a striking alignment between ID peaks and abrupt accuracy shifts in MMLU; the sharpest increase in accuracy always follows the ID peak. These observations suggest that transformer layers undergo an information compression phase just prior to forming confident predictions, supporting the idea that compression marks the onset of semantic decisiveness. While these patterns reveal a strong correlation between geometric compression and model decisiveness, we emphasize that the relationship is not strictly causal. The alignment of ID peaks with an increase in accuracy suggests that representational geometry and decision confidence co-evolve, with compression emerging just before the model commits to a prediction. This ordering provides a weak hint at an underlying causal structure that is worth investigating further; however, at present, the evidence should be interpreted as correlational rather than causal.

MLP Outputs vs. Residual Post-Activations A key finding is the distinct behavior of MLP outputs compared to residual post-activations (as shown in the zoomed-in version in App. Figure 14). While residuals reflect a smoothed integration of signals across the network, MLP outputs consistently display sharper ID transitions, highlighting their role in injecting task-specific refinements. In other words, the residual stream represents the model's continuously updated internal state, an accumulated integration of information that is propagated across layers. In contrast, the MLP output reflects a targeted modification to this stream, often acting as a high-leverage "correction" that sharpens or reorients the representation toward task-relevant directions. Consequently, the ID trajectories of these two signals reveal complementary aspects of computation, residual post-activations evolve smoothly, capturing the gradual stabilization of meaning, while MLP outputs exhibit sharper, more localized ID transitions, corresponding to points of semantic refinement or decision formation.

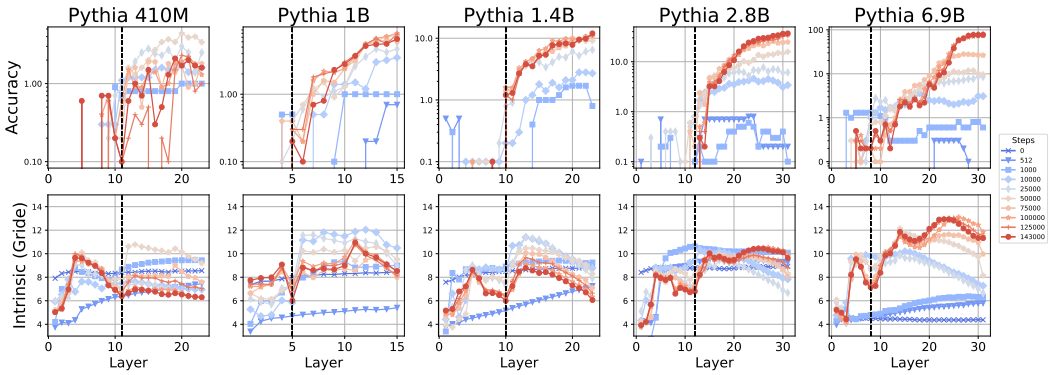

Figure 5: ID of residual post hidden layers in Pythia series models evolving throughout training for the Arithmetic dataset. The red curve shows the final checkpoint for architectures of different sizes. The top row shows the quality of layer representations in the form of accuracy (log scale). Interestingly, we observe that the model starts to be decisive about the correct token, where the ID shows a reverse peak (highlighted as black dashed vertical lines).

Few-Shot Prompting Accelerates Compression Few-shot prompting modulates model geometry in systematic ways. Fig. 4 and 16 show that increasing the number of in-context examples lowers the final-layer intrinsic dimension, especially in MLP outputs, indicating that few-shot prompts induce more efficient compression of the input space. This pattern is particularly salient in reasoning-heavy tasks, where compression accelerates with each additional example, suggesting that LLMs generalize better when they can abstract patterns across shots. Moreover, as shown in App. Fig. 16 and App. Figs. 27–30, well-performing models exhibit earlier ID peaks and steeper ID declines, reinforcing the view that efficient compression precedes confident prediction. We also observe that across datasets, models of varying sizes follow similar normalized ID trajectories when aligned by relative model depth (from 0 to 1). As shown in App. Figs. 31–40 (see App. Table 9), the ID profiles maintain high inter-model correlation, highlighting that despite architectural variation, models learn consistent representational transformations. This suggests that LLMs may converge towards a shared geometric inductive bias when trained on language data.

Predictive Utility of ID performance We observe that final-layer ID values negatively correlate with model accuracy across multiple datasets in the LLaMA family (App. Table 7) but the magnitude of the correlation varies a lot for different datasets, hence, the final-layer ID estimates corresponding to the last token offers a very weak unsupervised and architecture-agnostic proxy for model generalization. We also observed that when all the models that perform better than the baseline accuracy are considered, the correlation disappears (App. Table 8). In contrast to results reported by prior arts Ansuini et al. [2019], Pope et al. [2021], Birdal et al. [2021], we observe that the last layer ID (showing weak correlation) can not always be used as a strong proxy for accuracy/error across tasks.

Understanding Training Dynamics via Synthetic Tasks To better understand how intrinsic dimension (ID) evolves during learning, we turn to the Pythia model family (also see App. D). Fig. 5 tracks ID estimates and accuracy across layers and checkpoints. During training, we observe the emergence of a hunchback-like trend in ID, i.e., first increasing as representations diversify during early training, then peaking mid-network, and eventually decreasing toward the output layers. Interestingly, the relationship between ID and accuracy diverges from what we observe in MCQA tasks. In the higher-capacity models, accuracy begins to rise immediately after ID reaches a low (reverse peak). However, unlike MCQA tasks, where decisive ID transitions are tightly aligned with accuracy jumps, the transitions in generative reasoning tasks are more gradual, suggesting a more continuous integration of symbolic structure. These findings highlight two key insights. First, ID evolution during training can reveal whether a model is generalizing or merely memorizing, providing a geometric lens into the learning process. Second, the ID-based compression trends observed in real-world MCQA tasks are not artifacts of option-based formats; they also emerge, although more gradually, in open-ended generative reasoning tasks, especially in models with sufficient capacity. Thus, ID provides a unified view of learning geometry that applies both during training and across diverse reasoning paradigms.

Scaling In-Context Learning: Geometry of Few-Shot Adaptation To study how transformer representations evolve under extreme in-context learning (ICL) conditions, we analyze up to 50-shot prompting in arithmetic reasoning tasks using Pythia models of varying sizes (410M to 6.9B) (see App. Fig. 15). Interestingly, we found that the accuracy decreases after some examples with smaller models; this could be due to the fact that the model starts extracting surface patterns instead of generalizing and predicting the output according to the inherent arithmetic operation. Overall, we find that increasing the number of few-shot examples consistently improves accuracy while reducing the ID of final-layer representations, especially in larger models (2.8B being an exception). These trends highlight a tight coupling between few-shot generalization and latent space compression, as models condition on more examples, they restructure their internal geometry to form more compact, decision-relevant manifolds. Notably, this is the first study to probe up to 50-shot prompting in this context, positioning intrinsic dimension as a promising unsupervised proxy for evaluating ICL efficiency and saturation in LLMs.

Overall, the ID often relies on a smaller range of (5, 37) when compared to the extrinsic dimensions (aka model hidden dimensions (768, 4096)), irrespective of size and number of layers present in these models. We found this trend to be consistent for both template-based synthetic datasets as well as real-world datasets, pointing toward the language-specific tasks being present in low-dimensional manifolds. We provide additional results, discussion, and future directions in the App. E.

Limitations Though our work considers a wide range of open-weight models along with synthetic as well as real-world datasets, there is still room for experimentation with more language data sources. In our work, we primarily considered a setting where only a token is used for prediction (computing performance and intrinsic dimensions) and not the generative modeling setting, where multiple tokens are generated in an open-ended autoregressive fashion. Extending this analysis to Natural Language Generation (NLG) tasks becomes difficult due to the inherent autoregressive nature of these models. Another major limitation comes from the ID estimates that we use. In general, though prior arts have considered them for estimating ID estimates of features, these estimators often provide a noisy ID estimation of the underlying manifold, providing only a weaker estimate. Though we make use of a recently improved ID estimator (GRIDE), there still remains some scope for improvement. In the future, it would be interesting to revalidate these estimations via more advanced ID estimators. Moreover, in this work, the primary focus was to observe the trends across a wide range of models, and we only considered the features transformed by the transformer blocks for analysis, leaving the hidden representation inside these models aside. However, the proposed experimental setup could be utilized to study the spaces learned by each submodule of the transformer blocks (MLP-heads/Attention/LayerNorm/etc).

Further, on a broader level, exploring the relationship between intrinsic dimension and entropy provides a promising bridge between geometric and information-theoretic perspectives. Recent studies (e.g., Skean et al., 2025, Stolfo et al., 2025) have tried linking entropy to decision making, but examining this connection throughout the network could reveal how decision-making and representational geometry co-evolve. This line of work may ultimately lead to a unified framework where the activation geometry helps characterize the emergence of reasoning and understanding decision-making across layers.

6 Conclusion

In this work, we find that LLMs (stacked transformer layers) project the datasets into low-dimensional manifolds with ID estimates considerably lower than the actual latent dimensions. With a detailed analysis of 28 open-weight models, we find that the ID peaks are strongly coupled with decision-making happening inside models. This coupling between representational compression and model decisiveness provides geometric evidence of how abstract reasoning and prediction confidence co-evolve within the transformer architecture. We believe this study will open up new avenues for research in understanding the low-dimensional manifolds learned by the Language models. More broadly, we hope this study encourages the development of geometric interpretability tools that move beyond surface-level investigation, moving toward a deeper understanding of the internal topologies that support reasoning/decision-making in LLMs. By viewing activations through the dual lenses of geometry and information, we can begin to map not just what language models know, but how their internal structure gives rise to understanding and decision formation.

Acknowledgments

We would like to thank the anonymous reviewers and the meta-reviewer for their insightful comments and suggestions. This research work was partially supported by the Research-I Foundation of the Department of CSE at IIT Kanpur.

References

Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020. [Cited on page 3.]

Michele Allegra, Elena Facco, Francesco Denti, Alessandro Laio, and Antonietta Mira. Data segmentation based on the local intrinsic dimension. *Scientific Reports*, 10(1), 10 2020. doi: 10.1038/s41598-020-72222-0. URL https://doi.org/10.1038/s41598-020-72222-0. [Cited on page 5.]

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. [Cited on pages 3, 4, and 9.]

Richard Antonello and Emily Cheng. Evidence from fMRI supports a two-phase abstraction process in language models. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*, 2024. URL https://openreview.net/forum?id=VZipjFlBpl. [Cited on page 3.]

Jonathan Bac, Evgeny M. Mirkes, Alexander N. Gorban, Ivan Tyukin, and Andrei Zinovyev. Scikit-dimension: A python package for intrinsic dimension estimation. *Entropy*, 23(10):1368, October 2021. ISSN 1099-4300. doi: 10.3390/e23101368. URL http://dx.doi.org/10.3390/e23101368. [Cited on page 4.]

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. [Cited on pages 7, 30, and 33.]

Tolga Birdal, Aaron Lou, Leonidas Guibas, and Umut Şimşekli. Intrinsic dimension, persistent homology and generalization in neural networks, 2021. URL https://arxiv.org/abs/2111.13171. [Cited on page 9.]

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738. [Cited on pages 1 and 4.]

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, March 2021. URL https://doi.org/10.5281/zenodo.5297715. If you use this software, please cite it using these metadata. [Cited on page 29.]

Matthew Brand. Charting a manifold. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'02, page 985–992, Cambridge, MA, USA, 2002. MIT Press. [Cited on page 3.]

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165. [Cited on pages 1 and 5.]

J. Bruske and G. Sommer. *An algorithm for intrinsic dimensionality estimation*, pages 9–16. 11 2006. ISBN 978-3-540-63460-7. doi: 10.1007/3-540-63460-6_94. [Cited on page 34.]

Francesco Camastra. Data dimensionality estimation methods: A survey. *Pattern Recognition*, 36: 2945–2954, 12 2003. doi: 10.1016/S0031-3203(03)00176-6. [Cited on page 34.]

Francesco Camastra and Antonino Staiano. Intrinsic dimension estimation: Advances and open problems. *Information Sciences*, 328:26–41, 2016. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2015.08.029. URL https://www.sciencedirect.com/science/article/pii/S0020025515006179. [Cited on page 4.]

Emily Cheng, Corentin Kervadec, and Marco Baroni. Bridging information-theoretic and geometric compression in language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12397–12420, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.762. URL https://aclanthology.org/2023.emnlp-main.762/. [Cited on page 3.]

Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Lei Yu, Alessandro Laio, and Marco Baroni. Emergence of a high-dimensional abstraction phase in language transformers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=0fD3iIBhlV. [Cited on pages 1, 3, and 6.]

Zhenxiao Cheng, Jie Zhou, Wen Wu, Qin Chen, and Liang He. Learning intrinsic dimension via information bottleneck for explainable aspect-based sentiment analysis. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10274–10285, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.897/. [Cited on page 3.]

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848. [Cited on page 3.]

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. doi: 10.1109/MSP.2012.2211477. [Cited on page 3.]

Francesco Denti, Diego Doimo, Alessandro Laio, and Antonietta Mira. The generalized ratios intrinsic dimension estimator. *Scientific Reports*, 12(1):20005, Nov 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-20991-1. URL https://doi.org/10.1038/s41598-022-20991-1. [Cited on pages 2 and 5.]

Diego Doimo, Alessandro Pietro Serra, Alessio ansuini, and Alberto Cazzaniga. The representation landscape of few-shot learning and fine-tuning in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=nmUkwoOHFO. [Cited on page 3.]

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. [Cited on pages 1, 4, and 5.]

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), September 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-11873-y. URL http://dx.doi.org/10.1038/s41598-017-11873-y. [Cited on pages 2, 4, and 27.]

Mingyu Fan, Nannan Gu, Hong Qiao, and Bo Zhang. Intrinsic dimension estimation of data by principal component analysis, 2010. URL https://arxiv.org/abs/1002.2050. [Cited on page 26.]

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis, 2013. URL https://arxiv.org/abs/1310.0425. [Cited on page 3.]

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling. *CoRR*, abs/2101.00027, 2021. URL https://arxiv.org/abs/2101.00027. [Cited on page 30.]

Gemini, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen

Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill,

Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie

Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2024. URL https://arxiv.org/abs/2312.11805. [Cited on page 3.]

Gemma-Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian

Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology, 2024. URL https://arxiv.org/abs/2403.08295. [Cited on page 29.]

Sixue Gong, Vishnu Naresh Boddeti, and Anil K Jain. On the intrinsic dimensionality of image representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3987–3996, 2019. [Cited on pages 1, 3, and 4.]

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org. [Cited on page 3.]

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In Eneko Agirre, Johan Bos, Mona Diab, Suresh Manandhar, Yuval Marton, and Deniz Yuret, editors, **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics. URL https://aclanthology.org/S12-1052/. [Cited on pages 28, 29, and 32.]

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie,

Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi,

Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783. [Cited on page 29.]

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023. [Cited on page 29.]

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=p4PckNQR8k. [Cited on page 7.]

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. Understanding transformer memorization recall through idioms, 2023. URL https://arxiv.org/abs/2210.03588. [Cited on page 6.]

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300. [Cited on pages 28, 29, and 32.]

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419/. [Cited on pages 1 and 5.]

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20617–20642. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/huh24a.html. [Cited on page 38.]

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, Suriya Gunasekar, Piero Kauffmann, Yin Tat Lee, Yuanzhi Li, Anh Nguyen, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Michael Santacroce, Harkirat Singh Behl, Adam Taumann Kalai, Xin Wang, Rachel Ward, Philipp Witte, Cyril Zhang, and Yi Zhang. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 2023. [Cited on page 29.]

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825. [Cited on page 29.]

Abhinav Joshi, Areeb Ahmad, and Ashutosh Modi. COLD: Causal reasoning in closed daily activities. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=7Mo1NOosNT. [Cited on pages 28, 29, and 32.]

Abhinav Joshi, Areeb Ahmad, and Ashutosh Modi. Calibration across layers: Understanding calibration evolution in LLMs. In *The 2025 Conference on Empirical Methods in Natural Language Processing*, 2025a. URL https://openreview.net/forum?id=gDeWm1j5lH. [Cited on page 5.]

Abhinav Joshi, Areeb Ahmad, Divyaksh Shukla, and Ashutosh Modi. Towards quantifying commonsense reasoning with mechanistic insights. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9633–9660, Albuquerque, New Mexico, April 2025b. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.487. URL https://aclanthology.org/2025.naacl-long.487/. [Cited on page 5.]

Balázs Kégl. Intrinsic dimension estimation using packing numbers. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002. URL https://proceedings.neurips.cc/paper_files/paper/2002/file/1177967c7957072da3dc1db4ceb30e7a-Paper.pdf. [Cited on page 26.]

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 and cifar-100 datasets. *URl: https://www.cs. toronto. edu/kriz/cifar. html*, 6(1):1, 2009. [Cited on page 3.]

Elizaveta Levina and Peter Bickel. Maximum likelihood estimation of intrinsic dimension. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004. URL https://proceedings.neurips.cc/paper_files/paper/2004/file/74934548253bcab8490ebd74afed7031-Paper.pdf. [Cited on pages 2, 4, 26, and 27.]

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *CoRR*, abs/1804.08838, 2018. URL http://arxiv.org/abs/1804.08838. [Cited on page 3.]

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report, 2023. URL https://arxiv.org/abs/2309.05463. [Cited on page 29.]

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312. [Cited on page 3.]

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. doi: 10.1109/ICCV.2015.425. [Cited on page 3.]

Xingjun Ma, Bo Li, Yisen Wang, Sarah M. Erfani, Sudanthi N. R. Wijewickrema, Michael E. Houle, Grant Schoenebeck, Dawn Song, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. *CoRR*, abs/1801.02613, 2018. URL http://arxiv.org/abs/1801.02613. [Cited on page 3.]

David J.C. MacKay and Zoubin Ghahramani. Comments on 'maximum likelihood estimation of intrinsic dimension' by e. levina and p. bickel (2004). https://www.inference.org.uk/mackay/dimension/, 2005. [Cited on pages 4 and 26.]

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174):1–38, 2020. [Cited on page 3.]

Neel Nanda and Joseph Bloom. Transformerlens. https://github.com/TransformerLensOrg/TransformerLens, 2022. [Cited on page 6.]

nostalgebraist. interpreting GPT: the logit lens, LessWrong. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2020. [Accessed 26-01-2025]. [Cited on page 6.]

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads, 2022. URL https://arxiv.org/abs/2209.11895. [Cited on pages 1 and 5.]

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen,

Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774. [Cited on page 3.]

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, 2005. [Cited on pages 28, 29, and 32.]

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. URL https://openreview.net/forum?id=XJk19XzGq2J. [Cited on pages 3, 4, and 9.]

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL https://api.semanticscholar.org/CorpusID:160025533. [Cited on pages 1 and 29.]

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. Impact of pretraining term frequencies on few-shot numerical reasoning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.59. URL https://aclanthology.org/2022.findings-emnlp.59/. [Cited on page 7.]

Joshua Robinson, Christopher Michael Rytting, and David Wingate. Leveraging large language models for multiple choice question answering, 2023. [Cited on page 5.]

Daniel L. Ruderman. The statistics of natural images. *Network: Computation In Neural Systems*, 5:517–548, 1994. URL https://api.semanticscholar.org/CorpusID:2793971. [Cited on page 3.]

Gargi Singh, Dhanajit Brahma, Piyush Rai, and Ashutosh Modi. Fine-grained emotion prediction by modeling emotion definitions. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2021. [Cited on page 28.]

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=WGXb7UdvTX. [Cited on page 10.]

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013. [Cited on pages 28, 29, and 32.]

Alessandro Stolfo, Wes Gurnee, Ben Wu, Yonatan Belinkov, Xingyi Song, Mrinmaya Sachan, and Neel Nanda. Confidence regulation neurons in language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2025. Curran Associates Inc. ISBN 9798331314385. [Cited on page 10.]

Roy L Streit and Roy L Streit. *The poisson point process*. Springer, 2010. [Cited on page 4.]

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://aclanthology.org/P19-1452/. [Cited on pages 1 and 5.]

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. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288. [Cited on page 29.]

Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Sergey Nikolenko, Evgeny Burnaev, Serguei Barannikov, and Irina Piontkovskaya. Intrinsic dimension estimation for robust detection of ai-generated texts. *Advances in Neural Information Processing Systems*, 36, 2024. [Cited on pages 3 and 38.]

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=cCYvakU5Ek. [Cited on pages 1 and 3.]

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. *Advances in Neural Information Processing Systems*, 36, 2024. [Cited on page 38.]

Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax, May 2021. [Cited on page 29.]

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*, 2018. [Cited on pages 28, 29, and 32.]

Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hanna Hajishirzi, and Ashish Sabharwal. Answer, assemble, ace: Understanding how lms answer multiple choice questions. In *International Conference on Learning Representations*, 2024. [Cited on page 5.]

Ke Ye and Lek-Heng Lim. Schubert varieties and distances between subspaces of different dimensions, 2016. URL https://arxiv.org/abs/1407.0900. [Cited on page 38.]

Fan Yin, Jayanth Srinivasa, and Kai-Wei Chang. Characterizing truthfulness in large language model generations with local intrinsic dimension. *arXiv preprint arXiv:2402.18048*, 2024. [Cited on page 3.]

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, February 2021. ISSN 0001-0782. doi: 10.1145/3446776. URL https://doi.org/10.1145/3446776. [Cited on page 3.]

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification, 2016. URL https://arxiv.org/abs/1509.01626. [Cited on pages 28, 29, and 32.]

Wei Zhu, Qiang Qiu, Jiaji Huang, Robert Calderbank, Guillermo Sapiro, and Ingrid Daubechies. Ldmnet: Low dimensional manifold regularized neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2743–2751, 2018. doi: 10.1109/CVPR.2018.00290. [Cited on page 3.]

Appendix

Table of Contents

A Estimating Intrinsic Dimensions **26**

B Details of Real World Datasets **27**

C Prompt Templates **29**

D Details of Open Weight Models **29**

E Additional Results, Discussion and Future Directions **33**

E.1 Compute Resources 33

E.2 Additional Results and Discussion 33

E.3 Future Directions 38

List of Tables

1 ID estimators' hyperparameters 27

2 The table provides details about the various real-world datasets. 29

3 The table provides reference links to the prompt templates used for different real-world datasets. 29

4 Model Description 30

5 The table shows Accuracy and corresponding last layer (MLP Out) Intrinsic Dimensions of various models on different datasets 31

6 The table shows Accuracy and corresponding last layer (MLP Out) Intrinsic Dimensions of various models on MMLU datasets 31

7 Correlation Table for LLama Models 32

8 Correlation Table for well-performing Models 32

9 Figure Reference Table 32

10 PCA Intrinsic Dimension Estimates 33

List of Figures

6 Overview of ID Estimators 28

7 Accuracy and Intrinsic trends in AGNews 34

8 Accuracy and Intrinsic trends in COLD 35

9 Accuracy and Intrinsic trends in Rotten Tomatoes 35

10 Accuracy and Intrinsic trends in SST2 35

11 Accuracy and Intrinsic trends in MMLU Social Sciences 36

12 Accuracy and Intrinsic trends in MMLU Other 36

13 Accuracy and Intrinsic trends in MMLU Humanities 36

14 Residual Post Activation vs MLP Output Layer Intrinsic Estimates 37

15 Few Shot 0 to 50 on Arithmetic Dataset, Pythia models 37

16 Few Shot trend in LLama models . . . 37
17 General Prompt Template . . . 38
18 AG News dataset Prompt Template . . . 39
19 MMLU dataset Prompt Template . . . 39
20 CoLA dataset Prompt Template . . . 39
21 Rotten Tomatoes dataset Prompt Template . . . 39
22 SST2 dataset Prompt Template . . . 40
23 COPA dataset Prompt Template . . . 40
24 COLD dataset Prompt Template . . . 40
25 Intrinsic dimension and accuracy trend of Pythia models on Arithmetic Dataset . . . 41
26 Intrinsic dimension and accuracy trend of Pythia models on Greater Than Dataset . . . 42
27 Intrinsic trend across the model (MLP Out) using MLE Estimator . . . 43
28 Intrinsic trend across the model (MLP Out) using MLE (harmonic mean) Estimator 44
29 Intrinsic trend across the model (MLP Out) using TwoNN Estimator . . . 45
30 Intrinsic trend across the model (MLP Out) using GRIDE Estimator . . . 46
31 Correlation matrix AG News dataset . . . 47
32 Correlation matrix CoLA dataset . . . 47
33 Correlation matrix COPA dataset . . . 48
34 Correlation matrix COLD dataset . . . 48
35 Correlation matrix Rotten Tomatoes dataset . . . 49
36 Correlation matrix SST2 dataset . . . 49
37 Correlation matrix MMLU STEM dataset . . . 50
38 Correlation matrix MMLU Humanities dataset . . . 50
39 Correlation matrix MMLU Social Sciences dataset . . . 51
40 Correlation matrix MMLU Other dataset . . . 51

A Estimating Intrinsic Dimensions

In higher-dimensional spaces, real-world datasets often occupy only a small portion of the ambient space, i.e. the datasets often lie on on a low dimensional manifold $\mathcal{Y}$ and there exists a mapping $f : \mathcal{Y} \to \mathcal{X}$ where $\mathcal{X}$ is the data space of higher dimensions. Typically, it is considered that the function f is smooth and continuous, which ensures that nearby points in the lower-dimensional manifold will also be close by in the higher-dimensional space. In the past, multiple methods have been proposed to estimate the dimensions of the underlying low-dimensional mapping, including classical methods like projection or geometric-based methods, PCA Fan et al. [2010] and its constrained variant CPCA Kégl [2002]. These eigenvalue-based estimators aim to uncover the subspace that captures the majority of the dataset's variability, essentially, the directions that carry meaningful information while discarding noise. PCA computes the intrinsic dimension by first computing the covariance matrix of the dataset

$$C = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} (x - \mu_{\mathcal{D}})(x - \mu_{\mathcal{D}})^T \qquad \mu_{\mathcal{D}} = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} x$$

Then, the intrinsic dimensions are estimated by satisfying one of the two following conditions

$$\min_d \frac{\lambda_d}{\lambda_{d+1}} \geq \alpha \gg 1 \qquad \min_d \frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \geq \beta \in (0, 1)$$

where λ_k respresents the k^{th} largest eigenvalue of the covariance matrix C. While such classical estimators provide valuable insights into the structure of simpler datasets, they often struggle to capture the highly nonlinear manifolds encountered in modern deep learning. This limitation has motivated the DL community to seek new ways to characterize intrinsic dimensionality, ones that go beyond linear subspaces and better reflect the complexity of learned representations.

In the DL community, the nearest neighbor-based approaches are widely accepted due to their easier scalability and robust estimates. One such approach introduced by Levina and Bickel [2004] proposes a local intrinsic dimension estimator, i.e., IDs are estimated pertaining to each point present in the manifold by looking at the k-nearest neighbors. The proposed estimator assumes that in a neighborhood of a data point, x, the density of the data distribution is approximately constant. In this region, one can model the probability of finding another sample using a homogeneous Poisson process, which helps formulate a maximum likelihood objective in terms of the intrinsic dimension, having a closed-form solution

$$\hat{d}_R(x) = \left(\frac{1}{N(R, x)} \sum_{j=1}^{N(R,x)} \log \frac{R}{T_j(x)} \right)$$

where $N(R, x)$ represents the number of sampled data points found in the neighborhood of radius R.

For real-world datasets, the approximate solution to the maximum likelihood problem can be stated as

$$\hat{d}_k(x) = \left(\frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{T_k(x)}{T_j(x)} \right)$$

where $T_k(\cdot) : \mathcal{X} \to \mathbb{R}$ computes the distance of the k^{th} closed neighbor. This approximation provides a local dimension estimator corresponding to each datapoint. Further, for computing the global estimate, Levina and Bickel [2004] propose a straightforward averaging over these values, i.e.,

$$\hat{d} = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \hat{d}_k(x)$$

MacKay and Ghahramani [2005] further modifies/improves the aggregation of local intrinsic dimension estimates by formulating another maximum likelihood problem, similar to the previous one, resulting in the global intrinsic dimension being the inverse of the average of the inverse of the local intrinsic dimension, i.e.,

$$\hat{d} = \left(\frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \frac{1}{\hat{d}_k(x)} \right)^{-1}$$

Table 1: The table shows the hyperparameters used for different Intrinsic dimension estimators.

Method	Parameter	Values
MLE	k	all values in range [12, 24]
MLE-Modified	k	all values in range [12, 24]
TwoNN	Discard Ratio	0.1
GRIDE	n_1, n_2	$20, 40$

instead of direct averaging. It is suggested to do a direct averaging over the parameter k by Levina and Bickel [2004] for real-world datasets to compute the IDs empirically.

TwoNN Facco et al. [2017] is a global intrinsic dimension estimator that uses the same assumptions as the MLE of a homogeneous Poisson process. However, it relies on the terms $T_1(x)$ and $T_2(x)$ and prove that in the intrinsic space, the ratio $\mu = \frac{T_2(x)}{T_1(x)}$ follows the distribution $\texttt{Pareto}(1, d)$, where d is the intrinsic dimension, thus, establishing the relation

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

For estimation of real-world datasets, the above term is approximated by first considering the cumulative distribution as

$$F_{\text{emp}}(x) = \frac{1}{|\mathcal{D}|} \sum_{y \in \mathcal{D}} \mathbb{I}\{\mu(y) \leq \mu(x)\}$$

where $\mu(x)$ is computed as $\frac{T_2(x)}{T_1(x)}$ for each of the given datapoint. To estimate the slope, a linear regressor is fitted on the dataset $\{(-\log \mu(x), \log(1 - F_{\text{emp}}(x))\}_{x \in \mathcal{D}}$ passing through the origin. Moreover, Facco et al. [2017] empirically suggests discarding a small $\alpha < 1$ fraction of datapoints with the highest $\mu(x)$ values, resulting in better estimates for real-world datasets.

GRIDE generalizes the idea of TwoNN by keeping the same assumptions, however, modeling the n_2^{th} and the n_1^{th} nearest neighbors, resulting in the probability distribution

$$f(\mu) = \frac{d(\mu^d - 1)^{n_2 - n_1 - 1}}{\mu^{(n_2 - 1)d + 1} \beta(n_2 - n_1, n_1)}$$

As the above distribution doesn't have a closed-form solution for the cumulative function, the intrinsic dimension is estimated by maximizing the following log-likelihood

$$\max_d \log \mathcal{L}(d) \equiv \max_d (n_2 - n_1 + 1) \sum_{x \in \mathcal{D}} \log(\mu(x)^d - 1) + |\mathcal{D}| \log d - (n_2 - 1)d \sum_{x \in \mathcal{D}} \log \mu(x)$$

which turns out to be a concave optimization problem. The GRIDE estimates using a general form are more robust to noisy observations present in the datasets.

Also, see Figure 6 for an overview of the different Intrinsic dimension estimators.

B Details of Real World Datasets

For real-world text-based tasks, we would like to investigate the overall capabilities of different sets of widely used datasets. We specifically choose linguistic abilities, topic knowledge, field-specific knowledge (STEM, humanities, other), sentiment analysis (emotional intelligence), and reasoning abilities (Causal reasoning). Note that synthetic datasets (template-based) also capture mathematical reasoning in some sense; however, they cannot be considered real-world due to template-based generation. For all these abilities, we choose specific datasets that contain text samples that help validate the task performance. We consider a common MCQA format prompt template as described in the main paper to keep the analyses comparable with each other. Another advantage that comes with the MCQA format is the transformation of the dataset query in a different format than the original text, making the predictions only work when the model is able to provide predictions based on generalized learned tasks and not the memorized examples. We make minor changes to the template part of the prompt query (x_ϵ) for each of the datasets. We provide details of the datasets below:

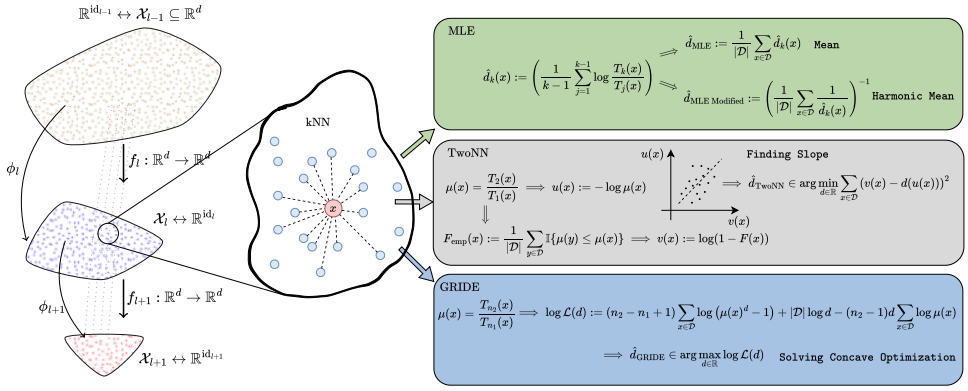

Figure 6: The figure shows an overview of the Intrinsic dimension estimators. The leftmost blobs show the underlying manifolds as the features are transformed by the transformer blocks (as explained in Figure 1). The obtained set of features is further used to compute the Intrinsic dimensions of the underlying manifold. The middle blob shows a zoomed-in version, highlighting the local neighborhood of a point $x \in \mathcal{X}_l$. All the ID estimators make use of local neighborhood estimates of the dimensionality. MLE formulates the likelihood based on k-nearest neighbors for local ID estimation, which is further averaged using the mean or harmonic mean to compute global intrinsic dimensions. TwoNN reduces this to 2 nearest neighbors and estimates the empirical cumulative distribution F_{emp} assuming the points to be i.i.d., and further uses linear regression to estimate the slope as global intrinsic dimensions. GRIDE generalizes the use of two nearest neighbors to n_1^{th} and n_2^{th} neighbors and frames a concave optimization problem to estimate the global intrinsic dimensions.

Linguistic: For linguistic abilities, we consider the widely used CoLA dataset Warstadt et al. [2018] that contains English sentences from 23 linguistics publications, expertly annotated for acceptability (grammaticality) by their original authors, making the prompts an MCQA query with 2 choices ('Accepted'/'Unaccepted') in our case.

Topic knowledge: For world topic knowledge, we make use of AG News dataset Zhang et al. [2016], which contains sentences from news articles on the web, primarily covering the 4 largest classes ("World", "Sports", "Business", "Sci/Tech"), making it a 4-choice MCQA query.

Field-specific knowledge: MMLU Hendrycks et al. [2021] is another widely used benchmark in the LLM evaluation/benchmarking community. The benchmark primarily aims to cover questions regarding world knowledge and problem-solving, including questions from different fields, including STEM, Humanities, Social Sciences, and Others. For each of the questions, the benchmark provides 4 choices/options. Note that for other datasets, the choices remain fixed; however, in MMLU queries, the option text is dynamic and keeps changing, making it a more complicated task for language understanding.

Emotional intelligence: Affective computing is another area where language plays a vital role Singh et al. [2021]. For this ability, we consider two known datasets, Rotten Tomatoes Pang and Lee [2005] and SST2 Socher et al. [2013]. Both of these datasets contain sentences annotated with the sentiment "Positive" or "Negative", making it an MCQA query with two choices in our setting.

Reasoning abilities: Real-world-based reasoning abilities are hard to capture in language benchmarks. For reasoning in real-world concepts, we found causal reasoning to be a suitable ability as it involves both real-world examples and a form of reasoning. We use COPA Gordon et al. [2012], and a small sample from the recently introduced COLD dataset Joshi et al. [2024]. Both these datasets contain a premise event and a corresponding causal query question, along with two choices, where a system is required to predict which of the two choices is the most plausible cause/effect of the premise event.

All these datasets cover a wide range of language understanding abilities, helping us quantify the generalization of the multiple open-weight models that we experimented with. We summarize various datasets in Table 2.

Table 2: The table provides details about the various real-world datasets.

Dataset Name	Task Type	# Samples	Avg. Prompt Length	# Choices
AG News	Topic Knowledge	7600	235.30	4
Rotten Tomatoes	Sentiment	1066	115.52	2
SST2	Sentiment	872	105.84	2
CoLA	Linguistic	1043	41.83	2
MMLU Stem	Field Specific Knowledge	3018	149.09	4
MMLU Humanities	Field Specific Knowledge	4705	535.10	4
MMLU Social Sciences	Field Specific Knowledge	3077	116.35	4
MMLU Others	Field Specific Knowledge	3242	163.32	4
COPA	Causal Reasoning	1000	34.89	2
COLD	Causal Reasoning	1000	29.49	2

Table 3: The table provides reference links to the prompt templates used for different real-world datasets.

Dataset Name	Templates (Ref.)
AGNews Zhang et al. [2016]	Figure 18
MMLU Hendrycks et al. [2021]	Figure 19
CoLA Warstadt et al. [2018]	Figure 20
RottenTomatoes Pang and Lee [2005]	Figure 21
SST-2 Socher et al. [2013]	Figure 22
COPA Gordon et al. [2012]	Figure 23
COLD Joshi et al. [2024]	Figure 24

C Prompt Templates

We use different prompt templates for different datasets with minimal changes, keeping the MCQA format consistent. An input prompt given to the model helps predict the probability distribution of the next token. The predicted probability/logit value of the next token can be written as:

$$
\begin{aligned}
P(x_t|x_{i<t}, \mathcal{M}_\theta) &= P(x_t|x_{query}, x_{options}, x_\epsilon, \mathcal{M}_\theta) \\
x_{query} &\leftarrow s_i \sim \mathcal{D} \\
x_{options} &\leftarrow \{\textbf{A.}\ o_{correct}, \textbf{B.}\ o_{wrong}\} \\
x_\epsilon &\in \text{set of prompt templates} \\
\mathcal{M}_\theta &= \{f_{\theta_1}, f_{\theta_2}, \ldots f_{\theta_L}\}
\end{aligned}
$$

where s_i is a sample/instance from the language-based dataset $\mathcal{D} := \{s_1, s_2, \ldots, s_N\}$ of size N. We choose a general prompt template (x_ϵ) for different datasets. Figure 17 shows a generalized prompt template used for all the datasets. All the datasets represent a different task, requiring a different generic query specific to the task. We modify only the generic query to make minimal changes to the prompt template. Figure 17, Table 3 provide the references to the prompt templates used for different datasets.

D Details of Open Weight Models

For our experiments, we consider a wide range of open-weight transformer-based LLMs. Specifically, we consider GPT-2 Small, GPT-2 Medium, GPT-2 Large, and GLT-2 XL Radford et al. [2019] from the GPT-2 family; GPT-Neo 125M, GPT-Neo 1.3B, and GPT-Neo 2.7B, from GPT-Neo family Black et al. [2021]; GPT-J 6B Wang and Komatsuzaki [2021]; Phi 1 Gunasekar et al. [2023] , Phi 1.5 Li et al. [2023], and Phi 2 Javaheripi et al. [2023], from Phi family; Gemma 2B, and Gemma 7B from Gemma family Gemma-Team et al. [2024]; Llama2 7B, Llama2 7B Chat, Llama2 13B, and Llama2 13B Chat, from Llama2 family Touvron et al. [2023]; Llama3 8B, Llama3 8B-Instruct, from Llama3 family Grattafiori et al. [2024]; and Mistral 7B Jiang et al. [2023].

All these models provide a broad spectrum of model sizes and architectural changes with different extrinsic dimensions. Table 4 provides the details of the used open-weight models.

Table 4: The table shows the list of open-weight models used for investigating the intrinsic dimensions. The list of models covers a wide range of layers with different model sizes. Note that the hidden dimension for each of the models is represented as Extrinsic Dimensions. Our experiments suggest that though these models use high extrinsic dimensions for information flow between the layers, the underlying manifold often lies in lower dimensions.

Model	**Size**	**# Layers**	**Layer Dimension**	**Vocabulary Size**
GPT-2 Small	85M	12	768	50257
GPT-2 Medium	302M	24	1024	50257
GPT-2 Large	708M	36	1280	50257
GLT-2 XL	1.5B	48	1600	50257
GPT-Neo 125M	85M	12	768	50257
GPT-Neo 1.3B	1.2B	24	2048	50257
GPT-Neo 2.7B	2.5B	32	2560	50257
GPT-J 6B	5.6B	28	4096	50400
Phi 1	1.2B	24	2048	51200
Phi 1.5	1.2B	24	2048	51200
Phi 2	2.5B	32	2560	51200
Gemma 2B	2.1B	18	2048	256000
Gemma 7B	7.8B	28	3072	256000
Llama2 7B	6.5B	32	4096	32000
Llama2 7B Chat	6.5B	32	4096	32000
Llama2 13B	13B	40	5120	32000
Llama2 13B Chat	13B	40	5120	32000
Llama3 8B	7.8B	32	4096	128256
Llama3 8B-Instruct	7.8B	32	4096	128256
Mistral 7B	7.8B	32	4096	32000
Pythia 14M	1.2M	6	128	50304
Pythia 31M	4.7M	6	256	50304
Pythia 70M	19M	6	512	50304
Pythia 160M	85M	12	768	50304
Pythia 410M	302M	24	1024	50304
Pythia 1B	805M	16	2048	50304
Pythia 1.4B	1,2B	24	2048	50304
Pythia 2.8B	2.5B	32	2560	50304
Pythia 6.9B	6.4B	32	4096	50432

Reason for Selecting Pythia Model for Synthetic Tasks: We use the Pythia model suite [Biderman et al., 2023], a family of 16 autoregressive transformers (14M–6.9B parameters), all trained on The Pile [Gao et al., 2021] using the same architecture, data order, and objective. With 154 publicly released training checkpoints per model, we can track the formation of decision-critical layers from early to late training. For feasibility, we analyze 10 evenly spaced checkpoints per model (steps: 0, 512, 1k, 10k, 25k, 50k, 75k, 100k, 125k, 143k), applying our ID and logit-based methods layer-wise.

To better understand how intrinsic dimension (ID) evolves during learning, we turn to the Pythia model family, which uniquely provides checkpoints at regular intervals throughout training. This allows us to directly examine the temporal dynamics of representation geometry, how model manifolds emerge, expand, and compress, as the model is exposed to more data and optimizes its objective. We focus on a synthetic arithmetic reasoning task that requires symbolic computation rather than token classification. Unlike MCQA settings where the output probability distribution is conditioned on the choices (e.g., "_A", "_B"), here the whole probability distribution is considered, making the task fundamentally generative.

Table 5: The table shows Accuracy and corresponding last layer (MLP Out) Intrinsic Dimensions of various models on different datasets

Model	AG News		COPA		Rotten Tomatoes		SST2		CoLA	
	Acc	ID	Acc	ID	Acc	ID	Acc	ID	Acc	ID
Random Baseline	25	-	50	-	50	-	50	-	70	-
GPT-2	25.28	18.97	48.90	15.41	50.09	17.30	49.08	16.43	30.87	13.54
GPT-2 Medium	25.51	22.22	48.60	16.65	49.91	18.72	49.08	18.78	30.78	14.45
GPT-2 Large	24.59	21.31	48.70	16.98	49.53	20.51	48.62	20.33	31.45	15.86
GPT-2 XL	29.03	20.99	50.10	18.00	49.81	20.53	49.20	20.39	49.09	16.98
GPT-Neo 125M	25.83	15.38	48.90	15.09	49.91	18.38	49.66	17.05	30.87	12.17
GPT-Neo 1.3B	27.70	18.08	48.30	16.43	49.91	17.90	50.34	17.44	37.68	14.24
GPT-Neo 2.7B	23.16	20.03	51.20	16.73	45.97	18.15	46.79	18.30	69.03	15.01
GPT-J 6B	25.03	22.72	51.20	18.33	50.47	17.40	50.57	17.32	65.29	15.58
Phi 1	22.74	30.26	48.30	20.73	50.38	23.37	52.64	22.33	66.92	16.14
Phi 1.5	59.53	22.92	70.20	18.61	50.47	16.32	50.92	15.74	66.25	15.30
Phi 2	78.80	21.38	82.70	18.80	78.61	16.23	83.37	15.50	66.92	14.58
Gemma 2B	33.08	18.93	57.20	20.17	53.75	22.13	50.00	21.38	68.65	14.54
Llama2 7B	54.39	18.36	62.70	19.99	59.29	13.51	64.11	14.39	42.19	12.81
Llama2 7B Chat	64.46	19.20	61.60	21.37	53.85	15.52	52.98	15.43	69.03	15.65
Llama2 13B	66.34	15.10	74.30	26.29	69.61	12.56	65.37	13.87	69.32	14.82
Llama2 13B Chat	69.13	16.77	79.60	18.75	74.58	14.89	77.75	13.80	69.32	14.93
Llama3 8B	62.42	19.61	78.20	22.51	58.63	18.60	63.30	17.43	67.11	16.22
Llama3 8B Instruct	79.39	14.63	84.80	15.41	77.95	9.41	80.16	9.71	68.36	11.15
Mistral 7B	81.72	20.05	80.10	27.67	66.14	19.91	68.69	18.55	69.42	19.41

Table 6: The table shows Accuracy and corresponding last layer (MLP Out) Intrinsic Dimensions of various models on MMLU datasets

Model	STEM		Humanities		Social Sciences		Other	
	Acc	ID	Acc	ID	Acc	ID	Acc	ID
Random Baseline	25	-	25	-	25	-	25	-
GPT-2	21.37	16.53	24.25	16.69	21.81	18.55	23.60	17.80
GPT-2 Medium	21.64	17.62	24.21	17.43	21.74	19.78	23.81	19.31
GPT-2 Large	22.23	19.48	24.44	18.98	22.16	21.17	24.06	21.92
GPT-2 XL	24.62	18.50	24.31	18.13	23.46	21.24	24.46	21.46
GPT-Neo 125M	21.40	14.12	24.21	15.86	21.84	16.39	23.72	15.05
GPT-Neo 1.3B	27.87	14.06	24.82	14.32	26.91	15.02	25.17	14.83
GPT-Neo 2.7B	27.07	15.93	24.99	15.72	24.76	18.40	26.31	17.16
GPT-J 6B	25.78	15.30	26.70	14.35	26.03	18.11	27.21	17.02
Phi 1	23.56	18.97	25.89	19.99	23.63	20.53	26.31	21.05
Phi 1.5	31.84	15.16	33.62	15.75	41.57	14.97	40.38	16.16
Phi 2	44.60	18.04	47.57	15.56	63.02	17.75	58.70	20.18
Gemma 2B	28.73	13.84	30.50	16.14	33.86	16.37	35.44	16.92
Llama2 7B	31.44	13.69	36.37	14.84	42.67	15.30	43.24	15.49
Llama2 7B Chat	35.65	16.27	42.64	16.54	51.48	15.39	52.34	16.78
Llama2 13B	41.78	15.88	45.93	15.95	56.52	18.07	56.23	18.03
Llama2 13B Chat	40.69	15.21	46.93	14.81	60.81	15.35	59.22	15.49
Llama3 8B	50.07	14.83	49.22	17.65	67.31	17.21	66.13	15.70
Llama3 8B Instruct	50.40	12.70	49.99	13.03	69.78	13.63	67.80	13.09
Mistral 7B	48.05	18.22	52.22	15.91	67.86	22.86	66.78	22.25

Table 7: The table highlights the correlation between ID estimates of the last layer (MLP Out) and the corresponding model performance for LLama models. The negative correlation indicates ID being lower for better-performing models, making ID estimates a weak proxy for the model's generalization capabilities.

Dataset	Pearson	Kendall Tau	Spearman	Accuracy
COPA	−0.356	−0.467	−0.543	73.53
COLD	−0.202	−0.333	−0.486	69.16
Rotten Tomatoes	−0.673	−0.600	−0.771	65.65
SST2	−0.720	−0.867	−0.943	67.27
AG News	−0.709	−0.600	−0.771	66.02
MMLU STEM	−0.309	−0.333	−0.371	41.67
MMLU Humanities	−0.022	−0.200	−0.257	45.18
MMLU Social Sciences	−0.084	−0.067	−0.143	58.09
MMLU Other	−0.438	−0.276	−0.406	57.50

Table 8: Correlation between ID estimates of the last layer (MLP Out) and the corresponding model performances for all the models which perform better than the baseline, i.e., 25% or 50% for 4 options and 2 options datasets respectively.

Dataset	Pearson	Kendall Tau	Spearman	Accuracy
COPA	−0.02	−0.14	−0.18	74.47
COLD	−0.75	−0.60	−0.71	73.90
Rotten Tomatoes	−0.81	−0.67	−0.80	72.07
SST2	−0.66	−0.60	−0.66	69.90
AG News	−0.28	−0.14	−0.12	63.87
MMLU STEM	0.04	−0.14	−0.21	42.58
MMLU Humanities	−0.09	−0.21	−0.29	44.22
MMLU Social Sciences	0.25	0.07	0.05	56.28
MMLU Other	0.06	−0.11	−0.18	55.90

Table 9: The table provides reference links to the figures corresponding to the correlation between the trend of intrinsic dimension across the relative depth of the model and the relationship between accuracy and intrinsic dimension along the depth of the model.

Dataset	Corr. Matrix (Ref.)	Accuracy-ID (Ref.)
AGNews Zhang et al. [2016]	Figure 31	Figure 7
CoLA Warstadt et al. [2018]	Figure 32	-
COLD Joshi et al. [2024]	Figure 34	Figure 8
RottenTomatoes Pang and Lee [2005]	Figure 35	Figure 9
SST-2 Socher et al. [2013]	Figure 36	Figure 10
MMLU STEM Hendrycks et al. [2021]	Figure 37	Figure 2
MMLU Humanities Hendrycks et al. [2021]	Figure 38	Figure 13
MMLU Social Sciences Hendrycks et al. [2021]	Figure 39	Figure 11
MMLU Other Hendrycks et al. [2021]	Figure 40	Figure 12
COPA Gordon et al. [2012]	Figure 33	Figure 3

Table 10: The table compares the ID estimates across the layers of LLama3-8B on COPA dataset

Layer	PCA	MLE	MLE Corrected	TwoNN	Gride
0	32	9.6	7.92	0.97	15.58
1	41	10.31	8.7	1.12	16.16
2	43	10.82	9.3	1.11	18.38
3	48	11.04	9.76	1.19	18.95
4	45	11.86	10.35	1.3	19.13
5	43	12.11	10.54	1.38	18.58
6	40	12.85	10.97	1.55	18.29
7	36	12.92	11.07	1.67	17.82
8	37	13.55	11.62	1.89	17.87
9	41	14.14	12.22	2.22	18.21
10	45	14.43	12.49	2.4	18.1
11	45	14.66	12.73	2.59	17.71
12	68	16.52	14.24	3.08	19.51
13	83	19.03	16.37	4.16	21.07
14	69	22.06	18.98	7.76	21.5
15	54	20.87	18.13	8.68	20.0
16	47	20.08	17.5	8.94	19.12
17	37	19.84	17.26	10.75	18.64
18	31	19.18	16.75	10.55	17.87
19	28	19.12	16.69	11.11	17.83
20	30	18.37	15.99	10.16	17.53
21	33	18.68	16.26	9.7	17.8
22	42	18.94	16.37	8.95	18.24
23	62	19.2	16.6	7.14	19.49
24	88	20.13	17.24	6.02	20.75
25	129	20.8	17.7	4.93	22.91
26	145	21.37	18.04	4.46	23.96
27	158	21.27	17.89	4.11	24.15
28	160	21.09	17.62	3.85	24.81
29	147	19.59	16.58	3.58	24.26
30	160	18.0	15.56	3.02	25.01
31	116	15.69	13.8	2.64	22.56

E Additional Results, Discussion and Future Directions

E.1 Compute Resources

We perform all the experiments using a machine with 1 NVIDIA A40 GPU. We use only the open-weight models with frozen parameters to present the results for better reproducibility in the future.

E.2 Additional Results and Discussion

The intrinsic dimension often relies on a smaller range of (5 - 38) when compared to the extrinsic dimensions (aka model hidden dimensions (768 - 4096)), irrespective of size and number of layers present in these models. We found this trend to be consistent for both template-based synthetic datasets as well as real-world datasets, pointing toward the language-specific tasks being present in low-dimensional manifolds. This suggests that all these networks learn to compress the language information to a lower-dimensional space of a similar range. Table 5 and Table 6 summarize the Intrinsic dimension (of the last layer, MLP Out) estimates obtained for various datasets and compare them with the performance. We highlight the additional trends observed in our experiments below:

Manifolds evolving during Training: In our setup, we consider two template-based synthetic datasets, the arithmetic and the greater than dataset. The arithmetic dataset was also used for monitoring learning during the training of Pythia series models Biderman et al. [2023]. Figure 25

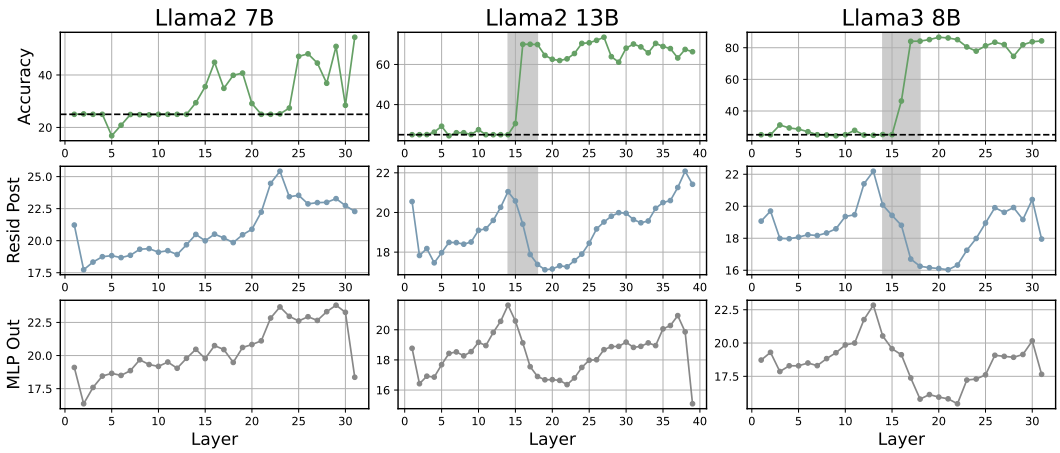

Figure 7: The figure shows accuracy along with ID trends for different variants of the LLaMA model on the **AGNews** dataset

shows ID estimates for layers in Pythia series models as the training processes for the Arithmetic Dataset. We observe all the models with different sizes following a similar trend of manifold evolution as the training progresses, finally converging to a similar characteristic, hunchback-like shape. On the Greater than dataset, we observe a similar evolution trend. However, the Greater Than dataset, being straightforward (less complex), shows the accuracy boost even from small-scale models. In Figure 26, smaller models show a hunchback-like shape with minimal changes observed for larger models that show an accuracy boost from initial layers.

Characteristic Trend in Latent Layers: In all our experiments for real-world datasets, we observe a characteristic hunchback-like trend in ID estimates across the model's MLP Out layers. The Figures 27, 28, 29, and 30 show the trends for MLE, MLE modified, TwoNN, and GRIDE estimates, respectively. All the estimators show a characteristic trend observed for a wide range of models for different real-world datasets. We observe a peak in the middle layers, highlighting a hunchback-like shape. The y-label shows the model names along with the accuracy in brackets, sorted from low-performing models to high-performing models from top to bottom. In general, we observe a similar trend being followed for similar performing models. Additionally, the relationship between accuracy and intrinsic dimension (both Residual post and MLP out) is captured by the Figures referred to in the App. 9

Correlation between Characteristic trend in different datasets: Figures 31 to 40 show a correlation between the intrinsic dimensions estimated for the hidden layers of multiple open-weight models corresponding to MLP Out. As the number of layers varies in different models, the estimated ID values were interpolated, considering the notion of relative depth in models. The figures show the correlation of IDs for all four ID estimators. (See table 9 for easier reference to figures for different datasets. Overall, we find a high correlation for similar performing models, highlighting similar trends of IDs observed for multiple models.

Comparison of trends in ID estimators: In our experiments, we compute intrinsic dimensions using 3 widely used ID estimators (MLE, MLE-Modified, TwoNN) and include a recently introduced estimator (GRIDE). Overall, we found similar trends throughout multiple ID estimators. We stick to GRIDE estimates for more reliable results in the main paper. In general, classic linear methods, such as PCA, assume a stable global eigenstructure and are inherently limited in capturing the nonlinear, locally curved manifolds that arise within deep networks. Prior studies have shown that PCA-based intrinsic dimension estimates become unreliable under high curvature or limited sampling conditions [Bruske and Sommer, 2006, Camastra, 2003]. These limitations make PCA not very well-suited for tracking how internal decision dynamics evolve layer by layer. In contrast, modern estimators like GRIDE and TwoNN explicitly account for local geometric distortions, providing a more faithful measure of the manifold structure. Table 10 compares intrinsic dimension estimates across different estimators. As noted, PCA-based estimates tend to yield considerably higher values (often exceeding 100) in later layers. This behavior arises because PCA infers dimensionality from variance along orthogonal directions in the data. As representations in deeper layers grow more dispersed and capture richer semantic variability, the explained variance spreads across more components, leading

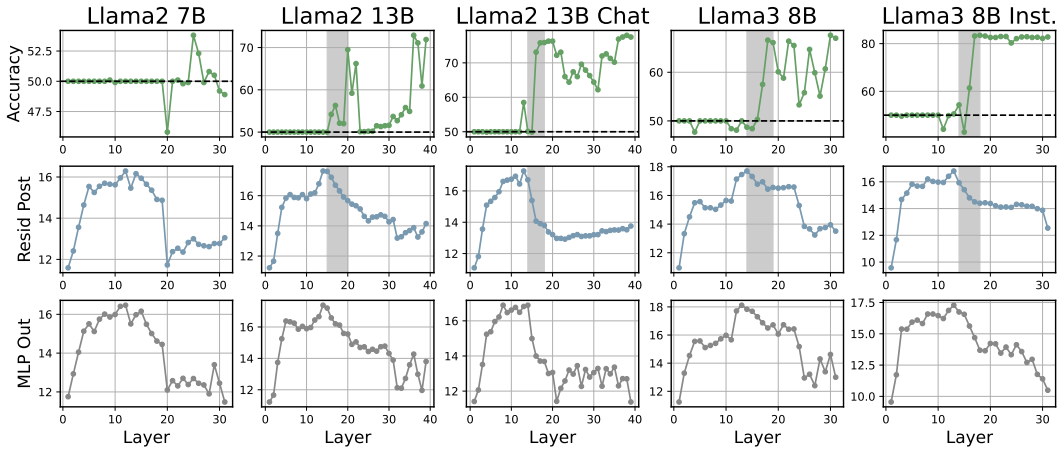

Figure 8: The figure shows accuracy along with ID trends for different variants of the LLaMA model on the **COLD** dataset

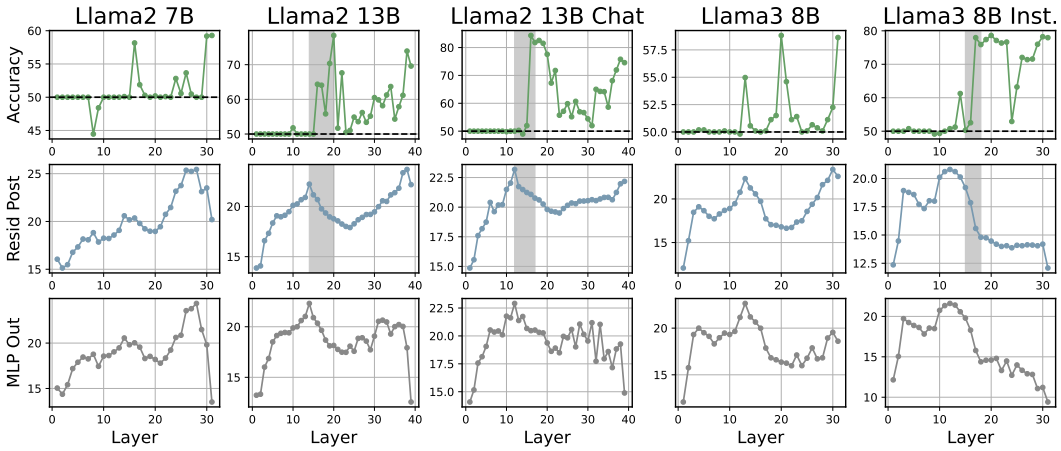

Figure 9: The figure shows accuracy along with ID trends for different variants of the LLaMA model on the **Rotten Tomatoes** dataset

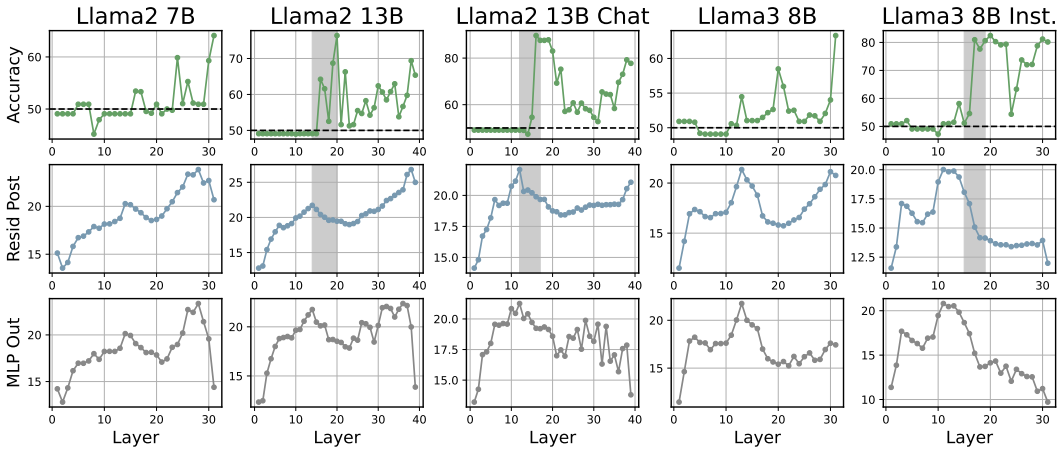

Figure 10: The figure shows accuracy along with ID trends for different variants of the LLaMA model on the **SST2** dataset

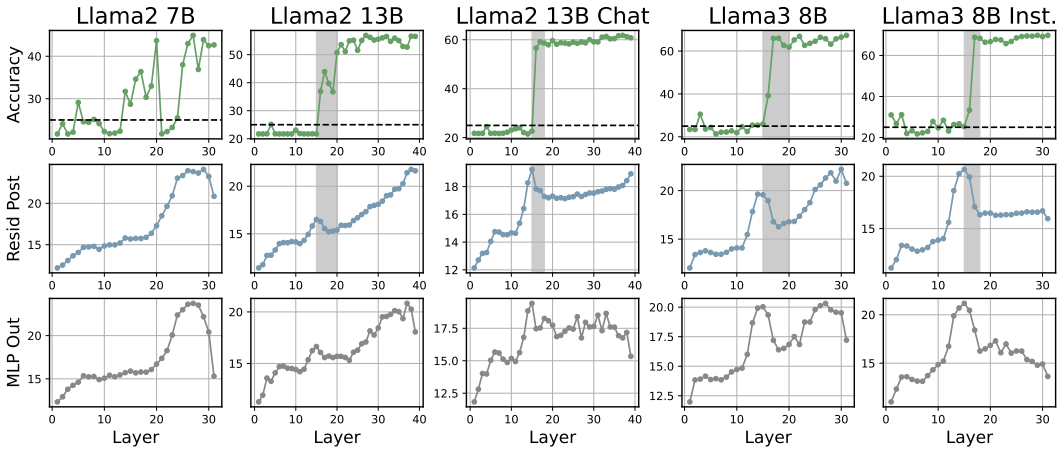

Figure 11: The figure shows accuracy along with ID trends for different variants of the LLaMA model on the **MMLU Social Sciences** dataset

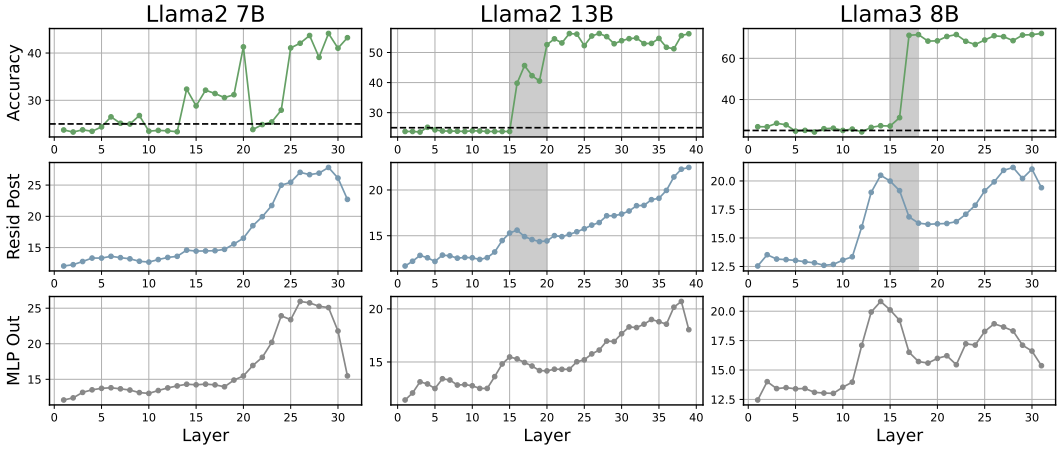

Figure 12: The figure shows accuracy along with ID trends for different variants of the LLaMA model on the **MMLU Other** dataset

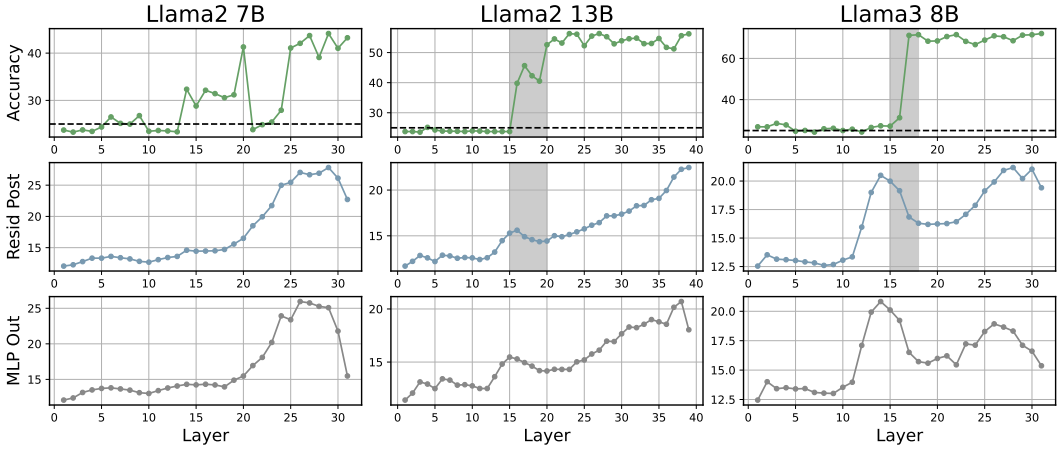

Figure 13: The figure shows accuracy along with ID trends for different variants of the LLaMA model on the **MMLU Humanities** dataset

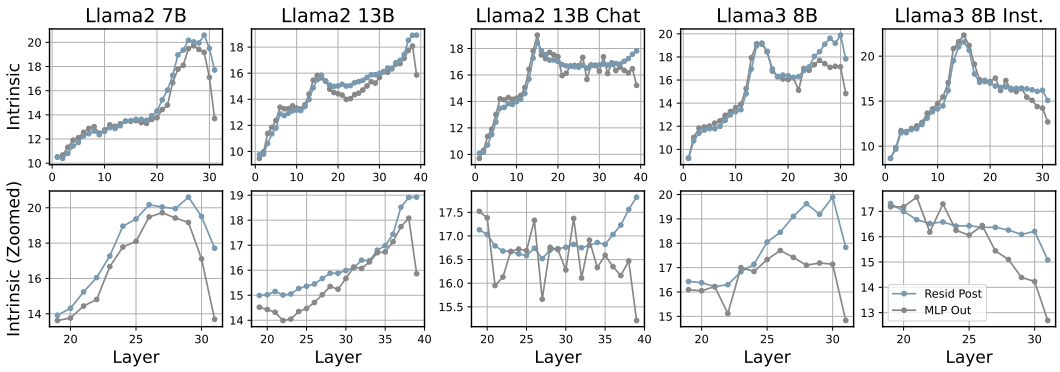

Figure 14: Comparision between ID across the model layer (MLP Out and Resid Post) on MMLU STEM Dataset

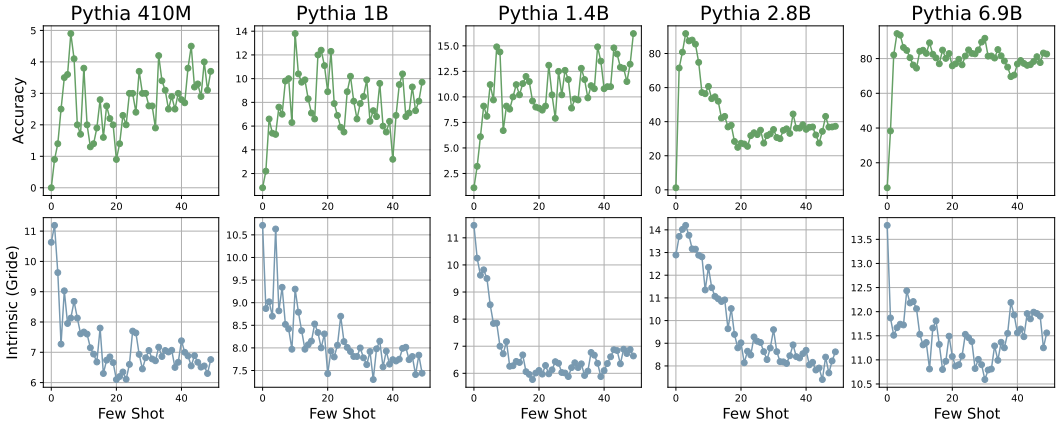

Figure 15: Relationship between accuracy and intrinsic dimensionality across Pythia models of varying sizes on Arithmetic Dataset for few shots ranging from 0 to 50.

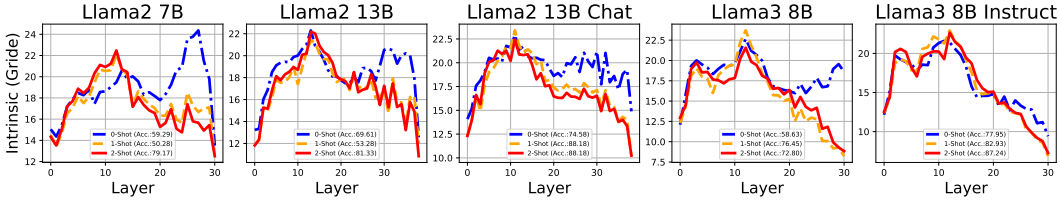

Figure 16: The figure shows ID of last token (MLP Out) throughout the stacked transformer blocks in the LLaMA family for the Rotten Tomatoes dataset. We observe that all these models show a similar characteristic curve of a hunchback-like shape, where the intrinsic dimensions first increase, reach a peak in the middle, and then further decrease. The legend includes the corresponding accuracies.

PCA to overestimate dimensionality. In contrast, nonlinear estimators such as GRIDE or TwoNN are less sensitive to global variance scaling and better capture the local geometric structure of these high-curvature manifolds.

E.3 Future Directions

The analysis we perform raises some interesting directions for future analysis. 1) Throughout all the observations, we found a space with a similar intrinsic dimension created by these models across different layers. We observe a peak arising in all these models in the middle layers. Though our findings suggest the peak pointing towards the space where the model starts to be decisive, a more detailed analysis of this space for different skills would be an interesting future venue, exploring if there lie multiple manifolds for multiple skills/tasks. 2) Provided the extrinsic to intrinsic dimension being large (found across multiple experiments), the autoregressive training objective highlights the knowledge compression capabilities of these models, making justifications for model compression by weight pruning/low-rank adaptations/finetuning/knowledge distillation. This also opens up the requirement for a more detailed analysis of manifolds evolving during low-rank finetuning objectives. 3) Though we find that the low-dimension manifolds learned by different models lie in similar ranges, little is explored about their structural similarity in low-dimensional manifolds learned by different models. In the future, it would be interesting to compare these subspaces in a more rigorous fashion, including comparison with subspace matching methods like Grassmann distance Ye and Lim [2016]. 4) Study of intrinsic dimensions changing via in-context learning examples provided in the prompt. Though the initial findings suggest the reduction in the manifold space as more in-context examples are provided, a more detailed study would be required to exploit ID estimates for choosing better in-context examples. 5) Comparison of representational space of different modalities in vision-language models. In this work, we only focused on language modality, both in terms of language models and datasets. Some of the findings have suggested that networks learning similar representations for multiple modalities Huh et al. [2024], Valeriani et al. [2024]; these findings could be reinforced by observing the manifolds learned for different modalities by the same models. 6) Though our work highlights the ID estimates showing a strong relation with model generalization, exploiting them to develop a concrete unsupervised algorithm for model comparison/task complexity comparison and generated text comparison remains open for future avenues. 7) Some of the preliminary findings also suggest ID estimates of human text be different from LLM-generated text Tulchinskii et al. [2024]. In this work, we could only explore open-weight LLMs on English datasets, leaving the extended comparison in different languages for the future.

```
Following are some multiple choice questions. You should directly answer the question by
choosing the correct option.
[ in-context examples (if few-shot/in-context learning experiment) ]
Question: A generalised statement pertaining to the task -: question/statement
A. choice1
B. choice2
Answer: A
```

Figure 17: Input prompt formats for the MCQA-based evaluation of autoregressive open-weight models (e.g., `llama(-2)`, `GPT-J`, etc.). The `black text` is the templated input for all datasets. The `orange text` is the input from the datasets, which contains either a review, a statement, or a question. The `teal text` is a template comment describing the task, which changes according to the dataset The next-token prediction probabilities of the option IDs at the `red text` are used as the observed prediction distribution.

```
Following are some multiple choice questions. You should directly answer the question by
choosing the correct option.
[ in-context examples (if few-shot/in-context learning experiment) ]
Question: Which is best fitting topic for the given news report? -: Brewers buyer expected
to step out of the shadows Monday MILWAUKEE - Paul Attanasio says the story of his brother
buying a baseball team is like a script straight out of Hollywood. He should know.
A. World
B. Sports
C. Business
D. Sci/Tech
Answer: C
```

Figure 18: Input prompt formats for the MCQA-based evaluation of autoregressive open-weight models (e.g., `llama(-2)`, `GPT-J`, etc.). The `black text` is the templated input for all datasets. The `orange text` is the input from the **AG News dataset**. The `teal text` is a template comment describing the task. The next-token prediction probabilities of the option IDs at the `red text` are used as the observed prediction distribution.

```
Following are some multiple choice questions. You should directly answer the question by
choosing the correct option.
[ in-context examples (if few-shot/in-context learning experiment) ]
Question: Mars has an atmosphere that is almost entirely carbon dioxide. Why isn't there
a strong greenhouse effect keeping the planet warm?
A: the atmosphere on Mars is too thin to trap a significant amount of heat
B: There actually is a strong greenhouse effect and Mars would be 35oC colder than it is
now without it.
C: Mars does not have enough internal heat to drive the greenhouse effect
D: the greenhouse effect requires an ozone layer which Mars does not have
Answer: A
```

Figure 19: Input prompt formats for the MCQA-based evaluation of autoregressive open-weight models (e.g., `llama(-2)`, `GPT-J`, etc.). The `black text` is the templated input for all datasets. The `orange text` is the input from the **MMLU dataset**. The next-token prediction probabilities of the option IDs at the `red text` are used as the observed prediction distribution.

```
Following are some multiple choice questions. You should directly answer the question by
choosing the correct option.
[ in-context examples (if few-shot/in-context learning experiment) ]
Question: Select the suitable option for the following statement -: The cat was bitten
the mouse.
A: Unacceptable
B: Acceptable
Answer: A
```

Figure 20: Input prompt formats for the MCQA-based evaluation of autoregressive open-weight models (e.g., `llama(-2)`, `GPT-J`, etc.). The `black text` is the templated input for all datasets. The `orange text` is the input from the **CoLA dataset**. The `teal text` is a template comment describing the task. The next-token prediction probabilities of the option IDs at the `red text` are used as the observed prediction distribution.

```
Following are some multiple choice questions. You should directly answer the question by
choosing the correct option.
[ in-context examples (if few-shot/in-context learning experiment) ]
Question: Select the suitable option for the following statement -: enchanted with
low-life tragedy and liberally seasoned with emotional outbursts . . . what is sorely
missing, however, is the edge of wild, lunatic invention that we associate with cage's
best acting .
A: Negative
B: Positive
Answer: A
```

Figure 21: Input prompt formats for the MCQA-based evaluation of autoregressive open-weight models (e.g., `llama(-2)`, `GPT-J`, etc.). The `black text` is the templated input for all datasets. The `orange text` is the input from the **Rotten Tomatoes dataset**. The `teal text` is a template comment describing the task. The next-token prediction probabilities of the option IDs at the `red text` are used as the observed prediction distribution.

```
Following are some multiple choice questions. You should directly answer the question by
choosing the correct option.
[ in-context examples (if few-shot/in-context learning experiment) ]
Question: Select the suitable option for the following statement -: this is human comedy
at its most amusing, interesting and confirming .
A: Negative
B: Positive
Answer: B
```

Figure 22: Input prompt formats for the MCQA-based evaluation of autoregressive open-weight models (e.g., `llama(-2)`, `GPT-J`, etc.). The `black text` is the templated input for all datasets. The `orange text` is the input from the **SST-2 dataset**. The `teal text` is a template comment describing the task. The next-token prediction probabilities of the option IDs at the `red text` are used as the observed prediction distribution.

```
Following are some multiple choice questions. You should directly answer the question by
choosing the correct option.
[ in-context examples (if few-shot/in-context learning experiment) ]
Question: Which of the following events (given as options A or B) is a more plausible
effect of the event -: 'The woman betrayed her friend.'?
A: Her friend sent her a greeting card.
B: Her friend cut off contact with her.
Answer: B
```

Figure 23: Input prompt formats for the MCQA-based evaluation of autoregressive open-weight models (e.g., `llama(-2)`, `GPT-J`, etc.). The `black text` is the templated input for all datasets. The `orange text` is the input from the **COPA dataset**. The `teal text` is a template comment describing the task. The next-token prediction probabilities of the option IDs at the `red text` are used as the observed prediction distribution.

```
Following are some multiple choice questions about the activity 'going grocery shopping'.
You should directly answer the question by choosing the correct option.
[ in-context examples (if few-shot/in-context learning experiment) ]
Question: Which of the following events (given as options A or B) is a more plausible
cause of the event 'drive to the nearby store.'?
A: make a list.
B: get into car.
Answer: B
```

Figure 24: Input prompt formats for the MCQA-based evaluation of autoregressive open-weight models (e.g., `llama(-2)`, `GPT-J`, etc.). The `black text` is the templated input for all datasets. The `orange text` is the input from the **COLD dataset**. The `teal text` is a template comment describing the task. The next-token prediction probabilities of the option IDs at the `red text` are used as the observed prediction distribution.

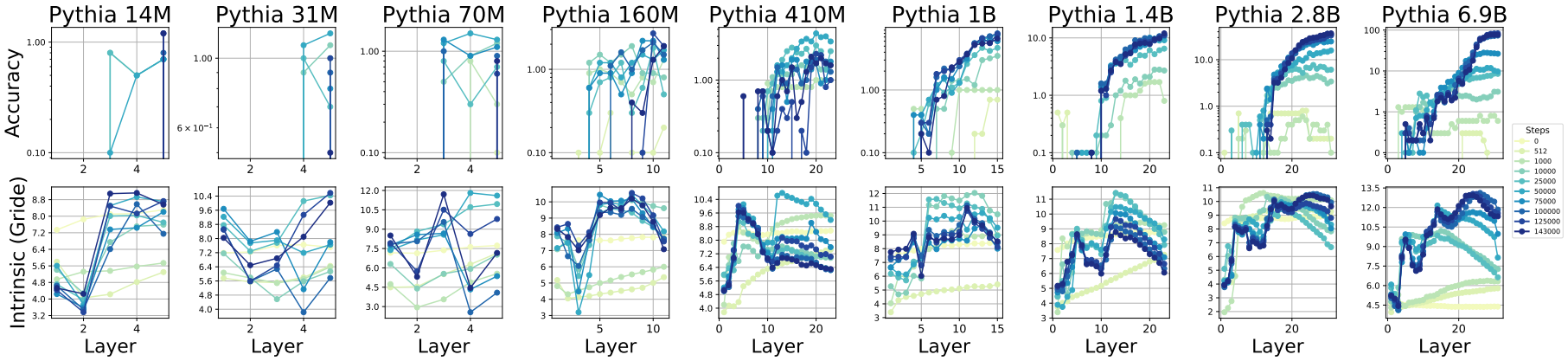

Figure 25: The figure shows ID estimates for residual post layers in Pythia series models as the training processes for the Arithmetic Dataset16-shot. We observe all the models with different sizes following a similar trend of manifold evolution as the training progresses, finally converging to a similar characteristic hunchback-like shape.

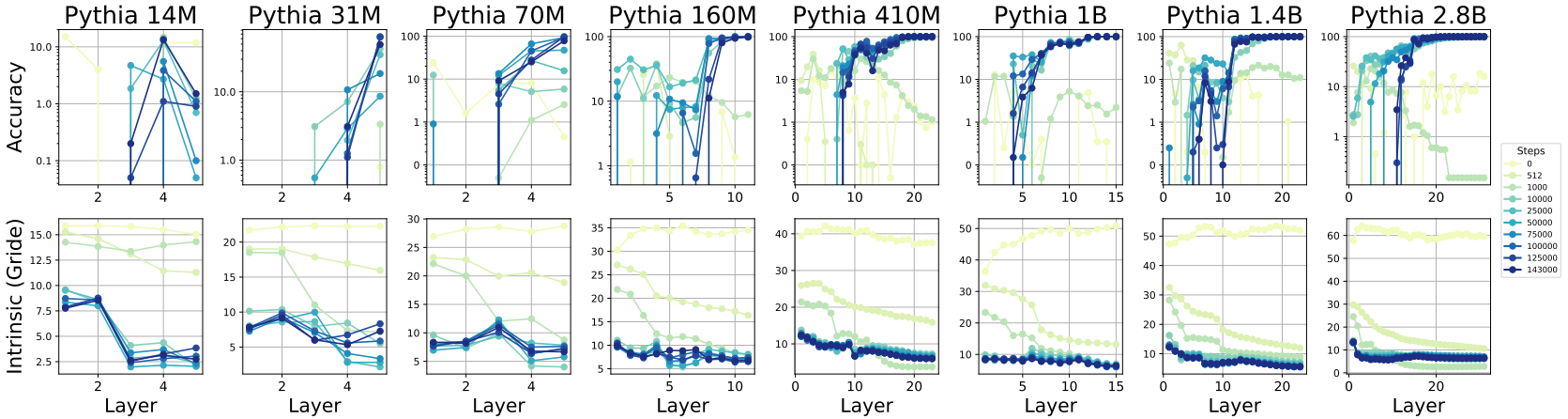

Figure 26: The figure shows ID estimates for residual post layers in Pythia series models as the training processes for the Greater Than Dataset. We observe all the models with different sizes following a similar trend of manifold evolution as the training progresses. The Greater Than dataset, being straightforward, shows the accuracy boost even from small-scale models, where smaller models show a hunchback-like shape, with minimal changes observed for larger models that show an accuracy boost from initial layers.

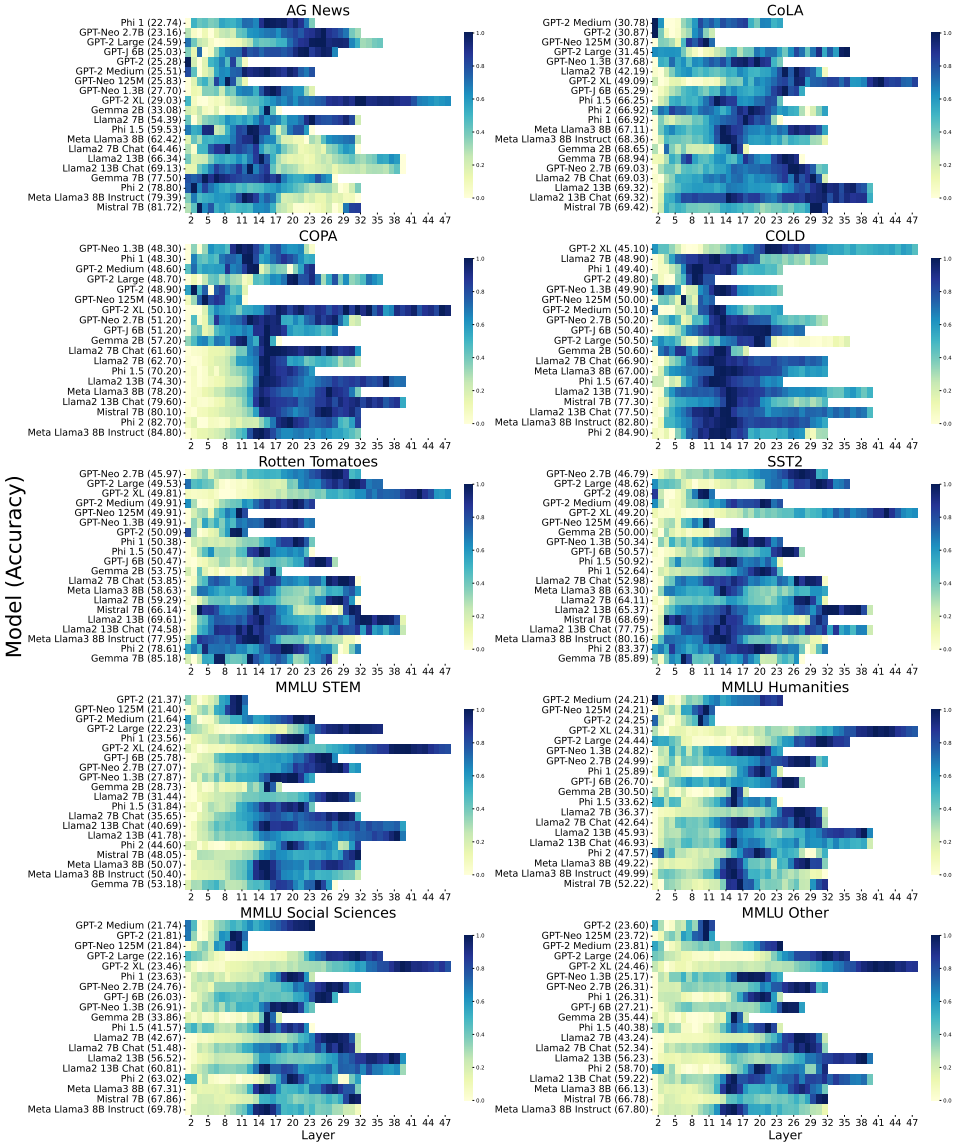

Figure 27: The figure shows a characteristic trend observed for a wide range of models for different real-world datasets. We observe a peak in the middle layers, highlighting a hunchback-like shape. The intrinsic dimensions are estimated using the **MLE** estimator. The y-label shows the model names along with the accuracy in brackets, sorted from low-performing models to high-performing models from top to bottom.

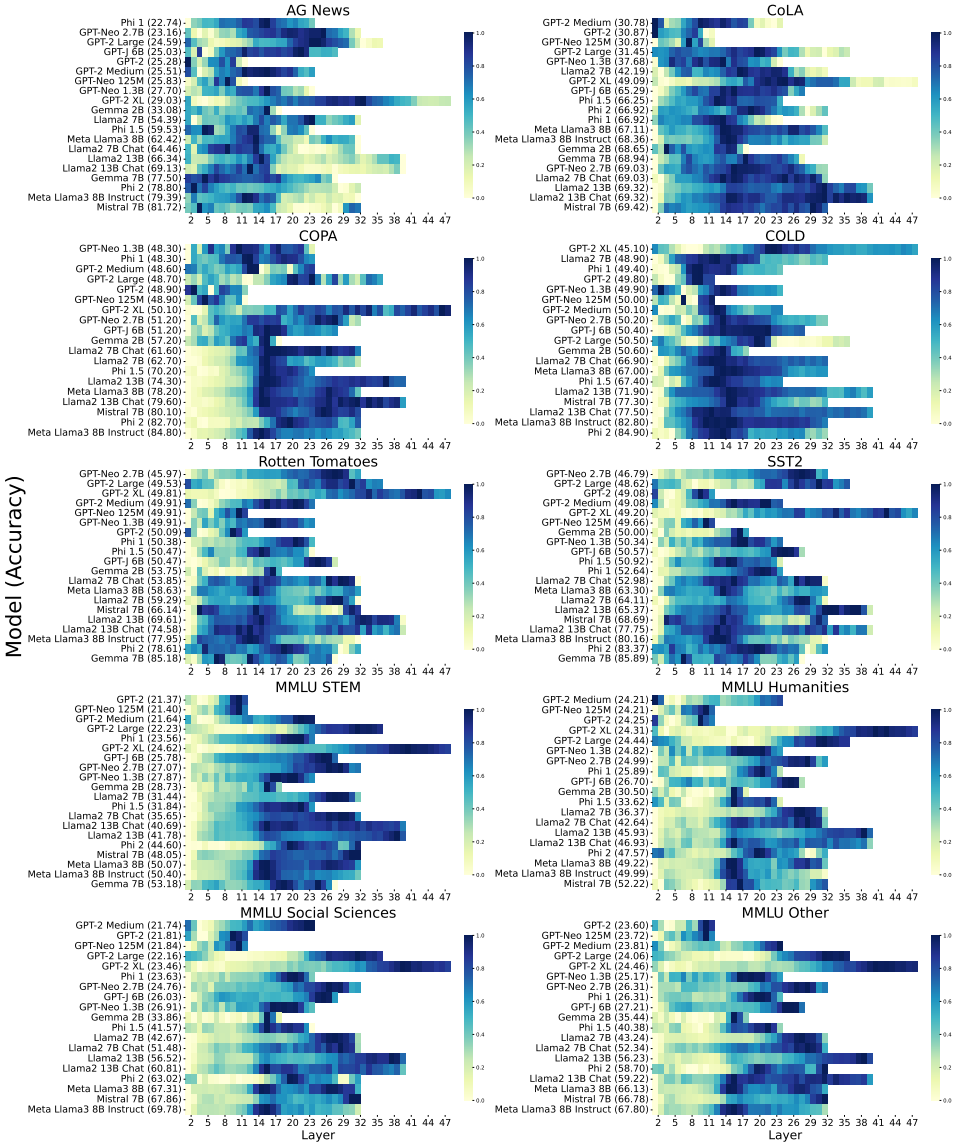

Figure 28: The figure shows a characteristic trend observed for a wide range of models for different real-world datasets. We observe a peak in the middle layers, highlighting a hunchback-like shape. The intrinsic dimensions are estimated using the modified version of the **MLE (harmonic mean)** estimator. The y-label shows the model names along with the accuracy in brackets, sorted from low-performing models to high-performing models from top to bottom.

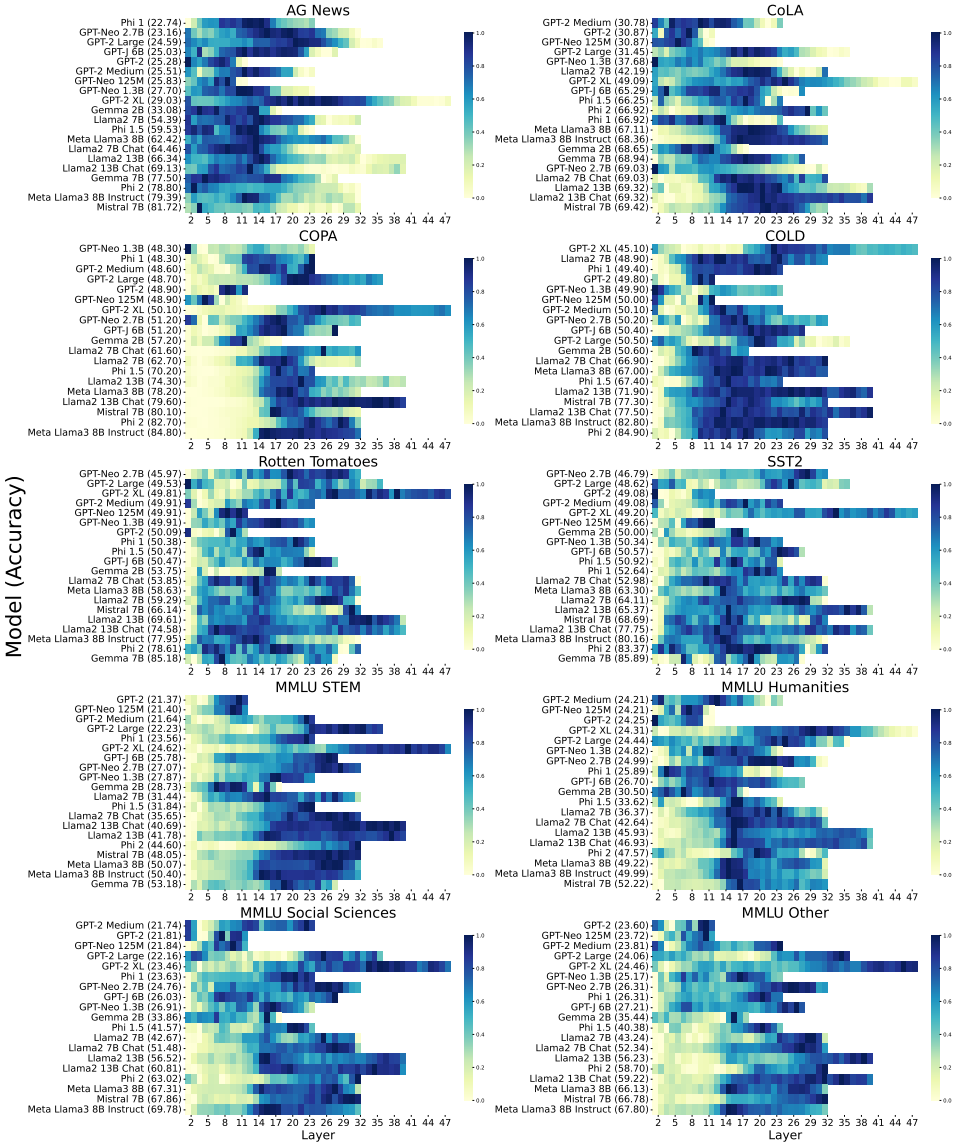

Figure 29: The figure shows a characteristic trend observed for a wide range of models for different real-world datasets. We observe a peak in the middle layers, highlighting a hunchback-like shape. The intrinsic dimensions are estimated using the **TwoNN** estimator. The y-label shows the model names along with the accuracy in brackets, sorted from low-performing models to high-performing models from top to bottom.

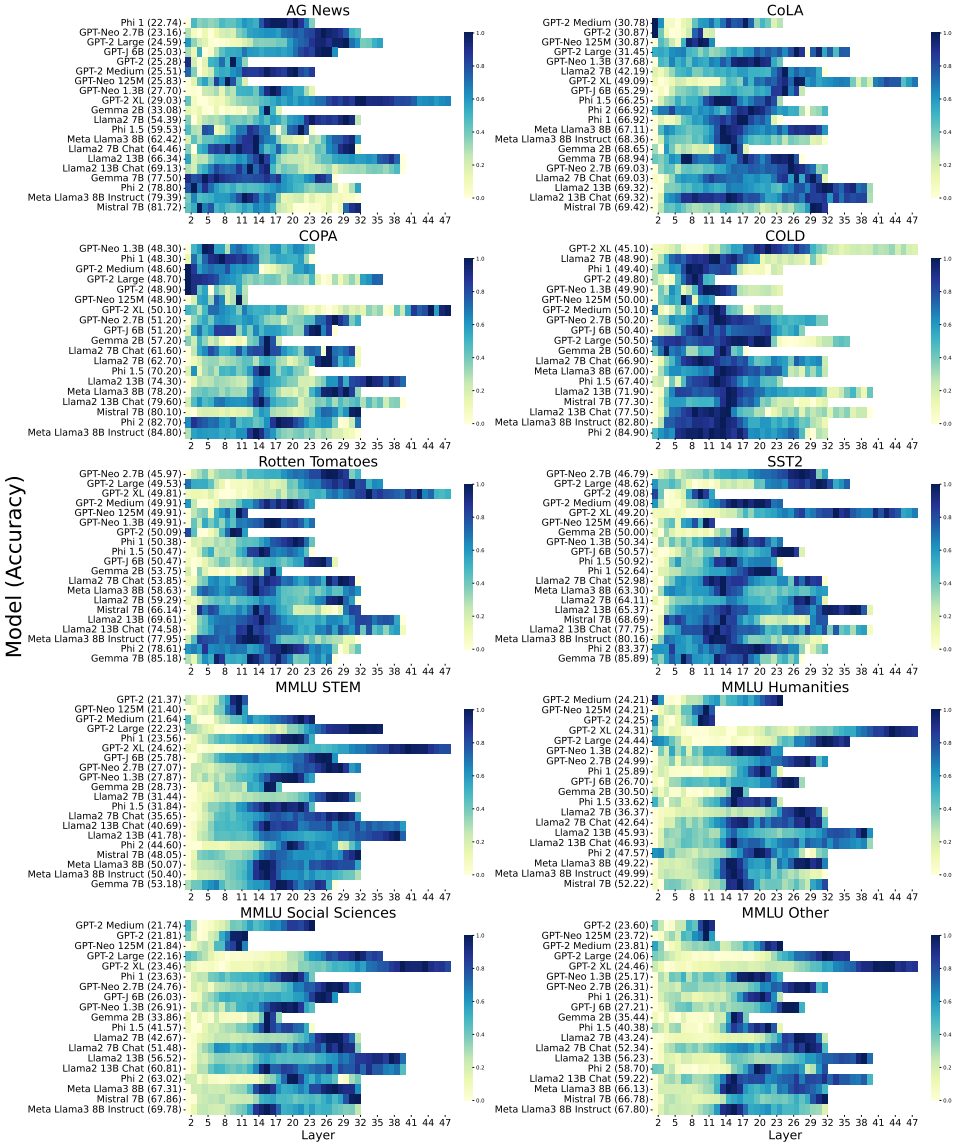

Figure 30: The figure shows a characteristic trend observed for a wide range of models for different real-world datasets. We observe a peak in the middle layers, highlighting a hunchback-like shape. The intrinsic dimensions are estimated using the **GRIDE** estimator. The y-label shows the model names along with the accuracy in brackets, sorted from low-performing models to high-performing models from top to bottom.

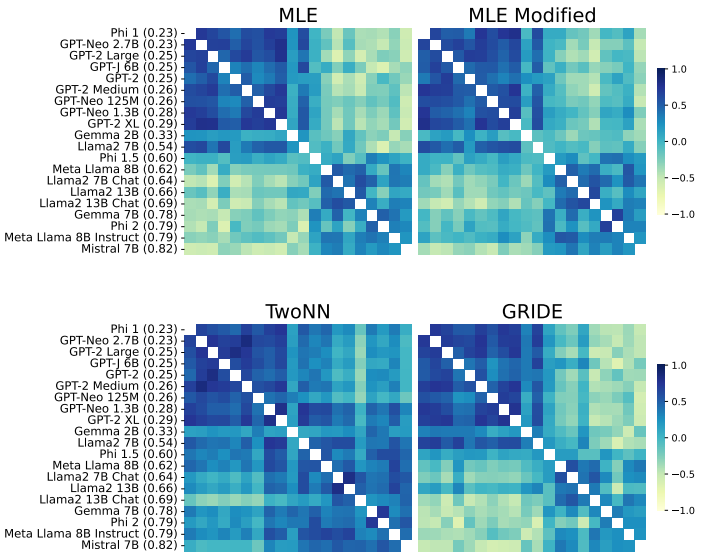

Figure 31: The figure shows a correlation between the intrinsic dimensions trajectories throughout layers of multiple open-weight models, denoting the correlation of IDs for all four ID estimators for **AG News** dataset.

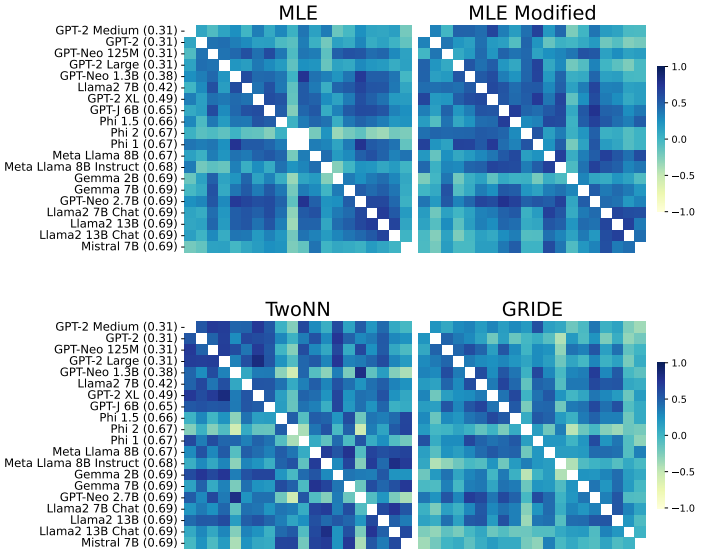

Figure 32: The figure shows a correlation between the intrinsic dimensions trajectories throughout layers of multiple open-weight models, denoting the correlation of IDs for all four ID estimators for **CoLa** dataset.

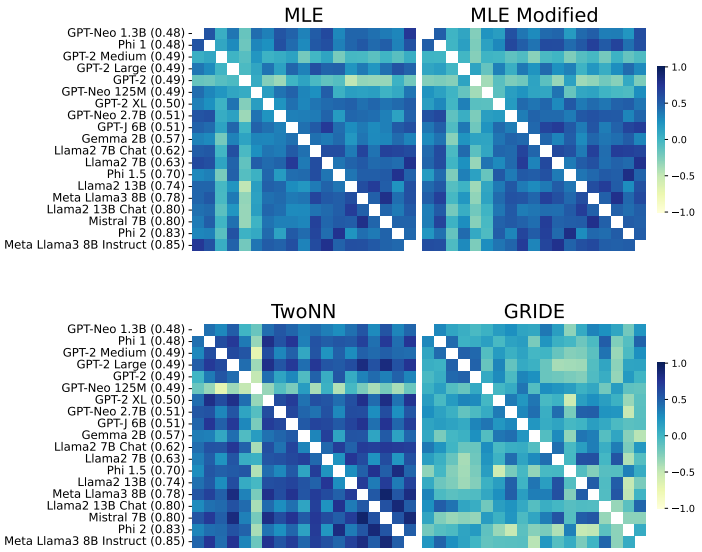

Figure 33: The figure shows a correlation between the intrinsic dimensions trajectories throughout layers of multiple open-weight models, denoting the correlation of IDs for all four ID estimators for **COPA** dataset.

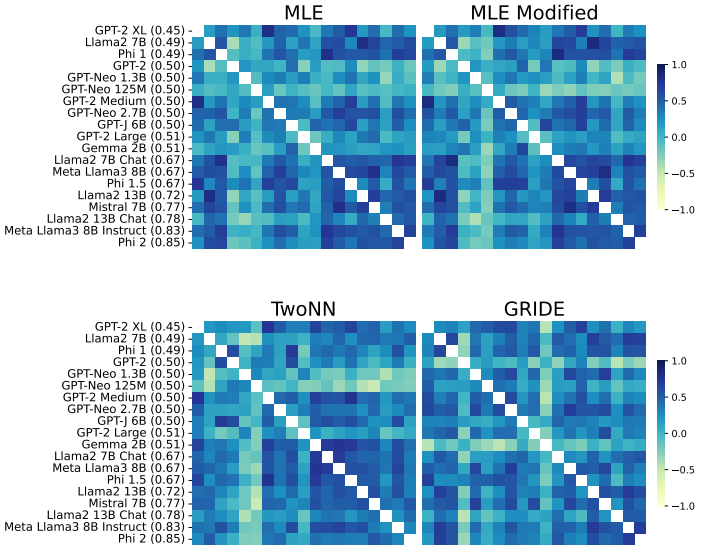

Figure 34: The figure shows a correlation between the intrinsic dimensions trajectories throughout layers of multiple open-weight models, denoting the correlation of IDs for all four ID estimators for **COLD** dataset.

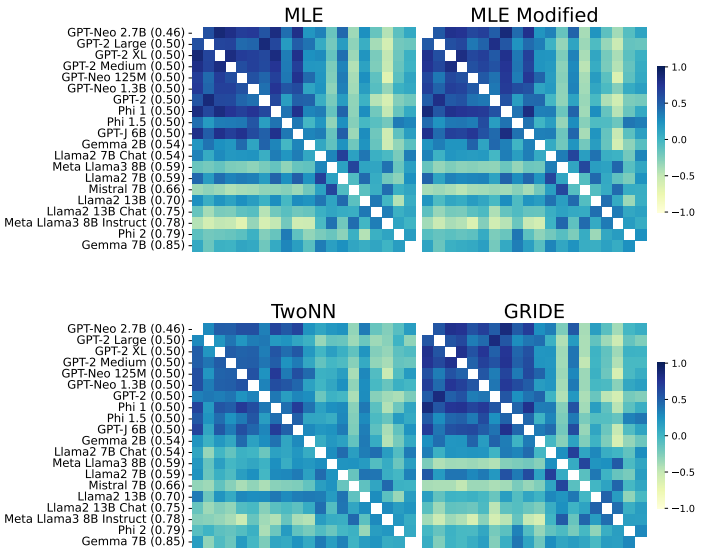

Figure 35: The figure shows a correlation between the intrinsic dimensions trajectories throughout layers of multiple open-weight models, denoting the correlation of IDs for all four ID estimators for **Rotten Tomatoes** dataset.

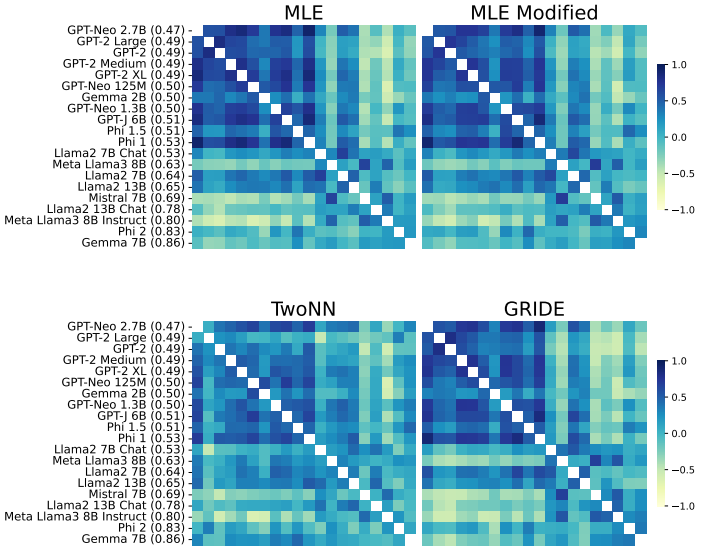

Figure 36: The figure shows a correlation between the intrinsic dimensions trajectories throughout layers of multiple open-weight models, denoting the correlation of IDs for all four ID estimators for **SST2** dataset.

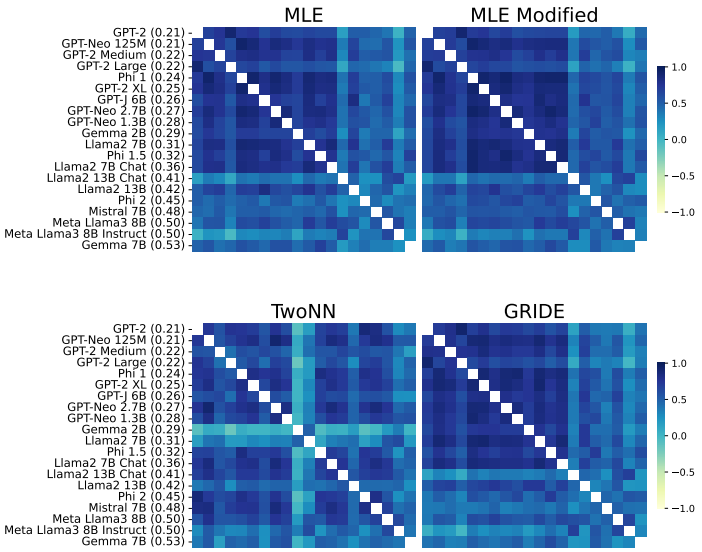

Figure 37: The figure shows a correlation between the intrinsic dimensions trajectories throughout layers of multiple open-weight models, denoting the correlation of IDs for all four ID estimators for **MMLU STEM** dataset.

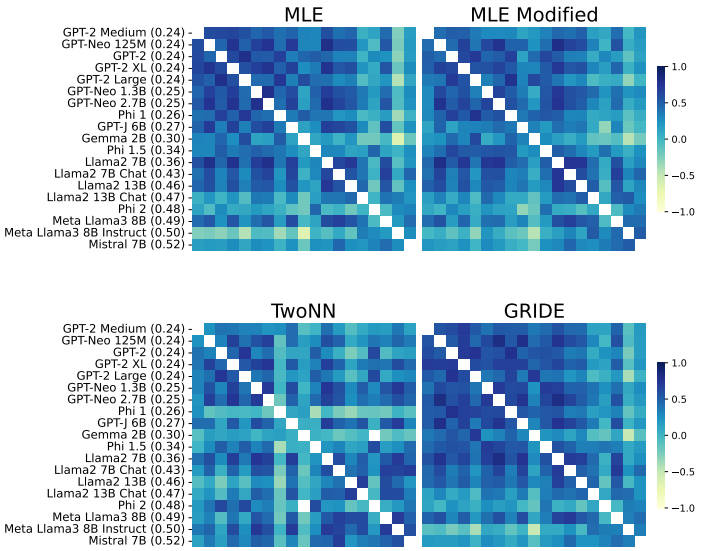

Figure 38: The figure shows a correlation between the intrinsic dimensions trajectories throughout layers of multiple open-weight models, denoting the correlation of IDs for all four ID estimators for **MMLU Humanities** dataset.

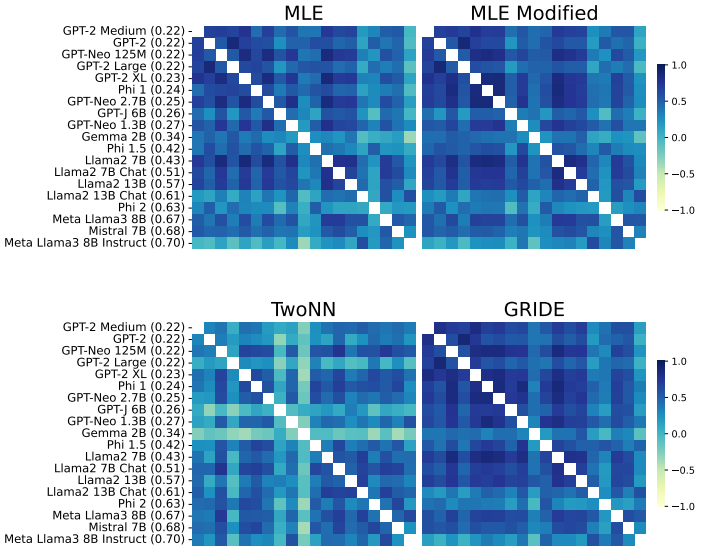

Figure 39: The figure shows a correlation between the intrinsic dimensions trajectories throughout layers of multiple open-weight models, denoting the correlation of IDs for all four ID estimators for **MMLU Social Sciences** dataset.

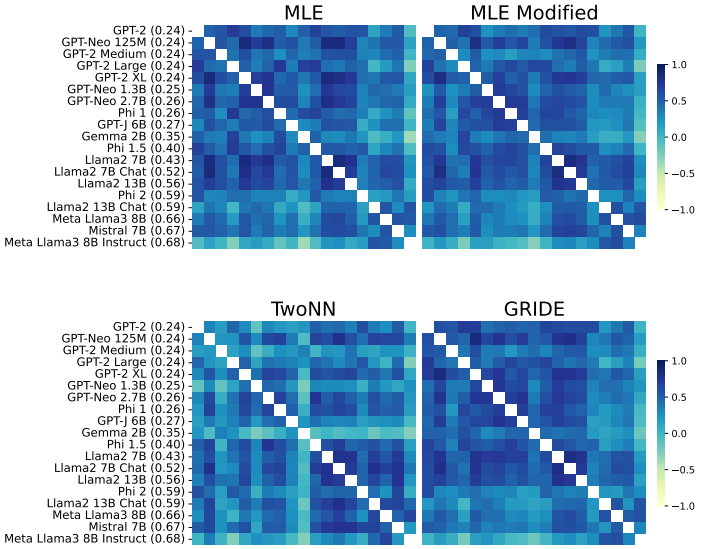

Figure 40: The figure shows a correlation between the intrinsic dimensions trajectories throughout layers of multiple open-weight models, denoting the correlation of IDs for all four ID estimators for **MMLU Other** dataset.

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: We provide details about the main claims in the Abstract and Introduction (Section 1, Section 4, and Section 5).

   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: We provide a sub-section dedicated to limitations in Section 5

   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: The paper doesn't include any theoretical results.

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: Details are provided in Section 4

   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: The code will be made available upon acceptance.

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: As discussed in Section 4, we only perform evaluation on pre-trained models and do not train/fine-tune any new models. Details for same are provided in Section 4 and App. D and App. E.

   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: We cover a wide range of open weight models and dataset to validate the trends. Details are provided in Section 5 and App. E.

   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: Details are provided in Appendix Section E.1

   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: We have reviewed the Code of Ethics, and followed these.

   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: [NA]

    Justification: To the best of our knowledge the research proposed in the paper does not have any negative social impact as we are only performing deeper analysis of existing open-weight LLMs and not training new models or creating any new technology per se.

    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: Not applicable for our paper. We are only performing deeper analysis of existing open-weight LLMs and not training new models or creating any new technology per se.

    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: We have used only open-source resources and cited relevant owners of the various resources, tools and models. Details are in App. D.

    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: We do not introduce any new asset. We are only performing deeper analysis of existing open-weight LLMs and not training new models.

    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: We do not perform any human experiments.

    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: We do not perform any 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.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: In this work we only study open weight Language Models to reveal the relationship between geometrical properties of internal representations and decision making happening inside these models. This described in detail across several sections in the paper (Section 1, Section 3, Section 4, and Section 5).

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.