# **Emergence of a High-Dimensional Abstraction Phase** in Language Transformers

# **Emily Cheng**

Universitat Pompeu Fabra emilyshana.cheng@upf.edu

#### Diego Doimo

Area Science Park diego.doimo@areasciencepark.it

#### **Corentin Kervadec**

Universitat Pompeu Fabra corentin.kervadec@upf.edu

#### Iuri Macocco

Universitat Pompeu Fabra iuri.macocco@upf.edu

#### Jade Yu

University of Toronto jadeleiyu@cs.toronto.edu

# Alessandro Laio

SISSA laio@sissa.it

#### Marco Baroni

Universitat Pompeu Fabra and ICREA marco.baroni@upf.edu

#### **Abstract**

A language model (LM) is a mapping from a linguistic context to an output token. However, much remains to be known about this mapping, including how its geometric properties relate to its function. We take a high-level geometric approach to its analysis, observing, across five pre-trained transformer-based LMs and three input datasets, a distinct phase characterized by high intrinsic dimensionality. During this phase, representations (1) correspond to the first full linguistic abstraction of the input; (2) are the first to viably transfer to downstream tasks; (3) predict each other across different LMs. Moreover, we find that an earlier onset of the phase strongly predicts better language modelling performance. In short, our results suggest that a central high-dimensionality phase underlies core linguistic processing in many common LM architectures.

# 1 Introduction

Compression is thought to underlie generalizable representation learning [Deletang et al., 2024, Yu et al., 2023]. Indeed, language models compress their input data to a manifold of dimension orders-of-magnitude lower than their embedding dimension [Cai et al., 2021, Cheng et al., 2023, Valeriani et al., 2023]. Still, the *intrinsic dimension* (ID) of input representations may fluctuate over the course of processing: we ask, what does the evolution of ID over layers reveal about representational generalizability, and, broadly, about linguistic processing?

The layers of an autoregressive language model (LM) transform the LM's input into information useful to predict the next token. In this paper, we characterize the geometric shape of this transformation across layers, uncovering a profile that generalizes across models and inputs: (1) there emerges a distinct phase characterized by a peak in the intrinsic dimension of representations; (2) this peak is significantly reduced in presence of random text and nonexistent in untrained models; (3) the layer at which it appears correlates with LM quality; (4) the highest-dimensional representations of different networks predict each other, but, remarkably, neither the initial representation of the input nor representations in later layers; (5) the peak in dimension marks an approximate borderline between representations that perform poorly and fairly in syntactic and semantic probing tasks, as well as in transfer to downstream NLP tasks.

Taken together, our experiments suggest that all analyzed transformer architectures develop, in the intermediate layers, a high-dimensional representation encoding complex and abstract linguistic information. The results of this processing are stored in representations which are then used, possibly through a process of incremental refinement, to predict the next token.

#### 2 Related work

The remarkable performance of modern LMs, combined with the opacity of their inner workings, has spurred a wealth of research on interpretability. At one extreme, there are studies that benchmark LMs treated as black boxes (e.g., Liang et al. [2023]). At the other extreme, researchers are "opening the box" to mechanistically characterize how LMs perform specific tasks (e.g., Meng et al. [2022], Conmy et al. [2023], Geva et al. [2023], Ferrando et al. [2024]). We take the middle ground, using geometric tools to characterize the high-level activation profiles of LMs, and we relate these profiles to their processing behaviour. In particular, we draw inspiration from the *manifold hypothesis*, or the idea that real-life high-dimensional data often lie on a low-dimensional manifold [Goodfellow et al., 2016]: we estimate the *intrinsic dimension* of the representational manifold at each LM layer to gain insight into how precisely layer geometry relates to layer function.

The notion that nominally complex, high-dimensional objects can be described using few degrees of freedom is not terribly new: it, for instance, underlies a number of popular dimensionality reduction methods such as PCA [Jolliffe, 1986]. But, while PCA is linear, the data manifold need not be: as such, general nonlinear methods have been proposed to estimate the *topological dimension*, or manifold dimension, of point clouds (see Campadelli et al. [2015] for a survey). As neural representations tend to constitute nonlinear manifolds across modalities [Pope et al., 2021, Cai et al., 2021], we use a state-of-the-art nonlinear ID estimation method, the Generalized Ratios Intrinsic Dimension Estimator (GRIDE) [Denti et al., 2022], which we further describe in Section 3.4.

Deep learning problems tend to be high-dimensional. But, recent work reveals that these ostensibly high-dimensional problems are governed by low-dimensional structure. It has, for instance, been shown that common learning objectives and natural image data lie on low-dimensional manifolds [Li et al., 2018, Pope et al., 2021, Psenka et al., 2024]; that learning occurs in low-dimensional parameter subspaces [Aghajanyan et al., 2021, Zhang et al., 2023]; that modern neural networks learn highly compressed representations of images, protein structure, and language [Ansuini et al., 2019, Valeriani et al., 2023, Cai et al., 2021]; and moreover, that lower-ID tasks and datasets are easier to learn [Pope et al., 2021, Cheng et al., 2023].

In the linguistic domain, considerable attention has been devoted to the ID of LM *parameters*. Zhang et al. [2023] showed that task adaptation occurs in low-dimensional parameter subspaces, and Aghajanyan et al. [2021] that low parameter ID facilitates fine-tuning. In turn, the low effective dimensionality of parameter space motivates parameter-efficient fine-tuning methods such as LoRA [Hu et al., 2022], which adapts pre-trained transformers using low-rank weight matrices.

Complementary to parameter ID, a number of works focus on the ID of *representations* in LM activation space. In particular, Cai et al. [2021] were the first to identify low-dimensional manifolds in the contextual embedding space of (masked) LMs. Balestriero et al. [2023] linked representational ID to the scope of attention, and showed how toxicity attacks can exploit this relationship. Tulchinskii et al. [2023] demonstrated that representational ID can be used to differentiate human- and AI-generated texts. Cheng et al. [2023] established a relation between representational ID and information-theoretic compression, also showing that dataset-specific ID correlates with ease of fine-tuning. Closer to our aims, Valeriani et al. [2023] studied the evolution of ID across layers for vision and protein transformers, with a preliminary analysis of a single language transformer tested on a single dataset. Like us, they found that ID develops in different phases, with a consistent early peak followed by a valley, and less consistent markers of a second peak. We greatly extend their analysis of linguistic transformers by investigating the functional role of the main ID peak in five distinct LMs.

# 3 Methods

#### 3.1 Models

We consider five causal LMs of different families, namely OPT-6.7B [Zhang et al., 2022], Llama-3-8B [Meta, 2024], Pythia-6.9B [Biderman et al., 2023], OLMo-7B Groeneveld et al. [2024], and Mistral-7B (non-instruction-tuned) [Jiang et al., 2023], hereon referred to as OPT, Llama, Pythia, OLMo, and Mistral, respectively.

OPT, Pythia, and OLMo make public their pre-training datasets, which are a combination of web-scraped text, code, online forums such as Reddit, books, research papers, and encyclopedic text (see Zhang et al. [2022], Gao et al. [2020], Soldaini et al. [2024], respectively, for details). The pre-training datasets of Llama-3 and Mistral are likely similar, though they remain undisclosed at the time of submission.

All language models considered have between 6.5 and 8B parameters, 32 hidden layers, and a hidden dimension of 4096. Moreover, they all inherit the architectural design of the decoder-only transformer [Vaswani et al., 2017] with a few notable changes: 1) layer normalization operations are applied before self-attention and MLP sublayers; 2) Pythia adds in-parallel the self-attention and MLP sublayer outputs to the residual stream, while in other models, self-attention outputs are added to the residual stream before the MLP; 3) Llama/Mistral/OLMo replace the ReLU activation function with SwiGLU [Shazeer, 2020], and Pythia instead uses GeLU [Hendrycks and Gimpel, 2016]; 4) all models but OPT replace absolute positional embeddings with rotary positional embeddings [Su et al., 2022]; 5) to facilitate self-attention computation, Llama uses grouped query attention [Ainslie et al., 2023], Mistral applies a sliding window attention, and Pythia adopts Flash Attention [Dao et al., 2022]. In all cases, we use as our per-layer representations the vectors stored in the HuggingFace transformers library hidden\_states variable.<sup>1</sup>

In autoregressive models, an input sequence's last token representation is the only one to contain information about the whole sequence. Furthermore, it is the only one that is decoded at the last layer to predict the next token. For these reasons, we choose to represent input sequences with their *last token representation* at each layer.

#### 3.2 Data

Since we focus on model behavior in-distribution, we compute observables using three corpora that proxy models' pre-training data (all accessed through HuggingFace): Bookcorpus [Zhu et al., 2015]; the Pile (Gao et al. [2020]; precisely, the Pile-10k subsample available on HuggingFace) and WikiText-103 [Merity et al., 2017]. From each corpus, we randomly extract 5 non-overlapping partitions of 10k 20-token sequences (sequence length is counted according to each corpus tokenization scheme). We do not constrain the sequences to have any particular structure: in particular, their final element is not required to coincide with the end of a sentence. We additionally generated 5 partitions of 10k 20-token sequences from *shuffled* versions of the same corpora. The latter respect the source corpus unigram frequency distribution, but syntactic structure and semantic coherence are destroyed.

# 3.3 Probing and downstream tasks

To relate representations' geometry to their content, we use the probing datasets of [Conneau et al., 2018], meant to capture the encoding of surface-related, syntactic and semantic information in LM representations. For each layer, we train a lightweight MLP probe from the hidden representation to each linguistic task. Details on tasks and training procedure are given in Appendix H.

If there is a relation between ID and semantic content, then ID may also predict ease-of-transfer to downstream NLP tasks. To test this claim, we consider two such tasks, sentiment classification of film reviews [Maas et al., 2011] and toxicity classification [Adams et al., 2017], for which we train binary linear classifiers on each hidden layer representation (training details in Appendix I).

<sup>&</sup>lt;sup>1</sup>Each hidden\_state vector corresponds to the representation in the *residual stream*[Elhage et al., 2021] after one attention and one MLP update.

<sup>&</sup>lt;sup>2</sup>We replicated the Pile-based Pythia and OPT experiments with sequences extended to 128 tokens. We obtained very similar results, confirming that the sequences we are using cover the typical contextual spans encode in model representations.

#### 3.4 Intrinsic Dimension

Real-world datasets tend to show a high degree of (possibly) non-linear correlations and constraints between their features [Tenenbaum et al., 2000]. This means that, despite a very large embedding dimensionality, data typically lie on a manifold characterized by a much lower dimensionality, referred to as its intrinsic dimension (ID). This quantity may be thought of as the number of independent features needed to locally describe the data with minimal information loss.

In almost every real-world system, the ID depends on the scale at which the data is analysed. In particular, at small scales, the true dimensionality of the manifold is typically hidden by that of data noise. At very large scales, the ID can also be erroneous, due to, for instance, the curvature of the manifold [Facco et al., 2017, Denti et al., 2022]. For this reason, in order to obtain a reliable and meaningful ID estimation, a proper scale analysis is necessary. To this aim, we opted for the *generalized ratios intrinsic dimension estimator* (GRIDE) of Denti et al. [2022], which extends the commonly used<sup>3</sup> TwoNN estimator of Facco et al. [2017] to general scales.

In GRIDE, the fundamental ingredients are ratios  $\mu_{i,2k,k} = r_{i,2k}/r_{i,k}$ , where  $r_{i,j}$  is the Euclidean distance between point i and its j-th nearest neighbour. Under local uniform density assumptions, the  $\mu_{i,2k,k}$  follow a generalised Pareto distribution  $f_{\mu_{i,2k,k}}(\mu) = \frac{d(\mu^d-1)^{k-1}}{B(k,k)\mu^{d(2k-1)+1}}$ , where  $B(\cdot,\cdot)$  is the beta function. By assuming the empirical ratios  $\mu_{i,2k,k}$  to be independent for different points, one obtains the ID by numerically maximizing the mentioned likelihood.

For each (model, corpus, layer) combination, we perform an explicit scale analysis. To do so, we first estimate the ID while varying k. Then, by visual inspection, we select a suitable k that coincides with a plateau in ID estimate [Denti et al., 2022]. An example of such a scale analysis for one (model, corpus) combination is reported in Appendix C.

In order to delimit high-ID peaks across layers, we conventionally locate the end of the peak at the closest inflection point after its maximum value. The beginning of the peak corresponds, then, to the closest layer before the maximum with value equal or greater than that at the peak end.

# 3.5 Quantifying the relative information content of different representations

Similar to Valeriani et al. [2023], we wish to relate dimensional expansion or compression of representations across layers to changes in their neighborhood structure. If neighborhood structure defines a semantics in representational space [Boleda, 2020], then layers whose activations have similar neighborhood structures perform similar functions.

In particular, the layers of an LM are iterative reconfigurations of representation space. We quantify the extent of reconfiguration (conversely, stability) using a statistical measure called Information Imbalance [Glielmo et al., 2022], hereon referred to as  $\Delta$ . Given two different spaces A and B, this quantity, defined in equation (1), measures the extent to which the neighborhood ranks in space A are informative about the ranks in space B (since ID computation is based on Euclidean distance, ranks are also obtained with Euclidean distance):

$$\Delta(A \to B) = \frac{2}{N^2} \sum_{i,j|r_{ij}^A = 1} r_{ij}^B.$$
 (1)

In words,  $\Delta$  is the average rank of point j with respect to point i in space B, given that j is the first neighbour of i in space A. If  $\Delta(A \to B) \sim 0$ , space A captures full neighborhood information about space B. Conversely, if  $\Delta(A \to B) \sim 1$ , space A has no predictive power on B.

It is essential to notice that  $\Delta$  is non-commutative with respect to its arguments. That is, it is asymmetric upon a swap of spaces:  $\Delta(A \to B) \neq \Delta(B \to A)$ . This feature allows us to capture directional information containment, as we do below when considering the extent to which, e.g., a layer i contains the information present in another layer j, or to measure the reciprocal information between representations of two different networks.

As  $\Delta$  is a rank-based measure, it can be used to compare spaces of different dimensionalities and/or distance measures. In our specific case, this property implies that  $\Delta$  is robust to possible dimension misalignments between layers.

<sup>&</sup>lt;sup>3</sup>[Ansuini et al., 2019, Valeriani et al., 2023, Tulchinskii et al., 2023, Cheng et al., 2023]

In comparing the layers' representation spaces, we also considered Doimo et al. [2020]'s neighborhood overlap measure which is used in Valeriani et al. [2023], as well as Representational Similarity Analysis [Kriegeskorte et al., 2008]. These symmetric similarity measures tell a similar story to  $\Delta$ , but the signal is much weaker; therefore, we do not report them.

#### 4 Results

We find, in line with previous work, that LMs represent language on a manifold of low intrinsic dimension. Furthermore, the representational ID profile over layers reveals a characteristic phase of both geometric and functional significance, marking, respectively, a peak in ID and transition in between-layer neighborhood similarity ( $\Delta$ ), and a transition to abstract linguistic processing.

# 4.1 Emergence of a central high-dimensionality phase

Figure 1 (left) reports the evolution of the ID for all models, averaged across corpora partitions (for per-corpus results, see Appendix C). In line with previous work [Cai et al., 2021, Cheng et al., 2023, Tulchinskii et al., 2023, Valeriani et al., 2023], we first observe that the ID for all models is  $\mathcal{O}(10)$ , which lies orders of magnitude lower than the models' hidden dimension at  $4096 \sim \mathcal{O}(10^3)$ .

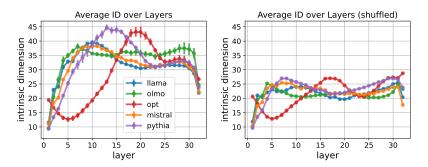


Figure 1: Average ID (over 5 random partitions) of three corpora: Bookcorpus, the Pile and Wikitext, where samples come from (left) the original corpus, (right) the shuffled corpus. All curves are shown with  $\pm$  2 standard deviations (shuffled ID SDs are very small). In the middle layers, shuffled corpus ID is *lower* than non-shuffled ID, suggesting that linguistic processing contributes to ID expansion.

All models clearly go through a phase of high intrinsic dimensionality that tends to take place relatively early (starting approximately at layer 6 or 7, and mostly being over by layer 20), except for OPT, where it approximately lasts from layer 17 to layer 23. For all models, we also observe a second, less prominent peak occurring towards the end: only for OLMo, this second peak, which we do not analyze further, is as high as the first one.

**ID** is a geometric signature of learned structure Figure 1 (right) shows that, when the models are fed shuffled corpora, ID remains low all throughout the layers. This suggests that the presence of high-ID peaks depends on the network performing meaningful linguistic processing. We further confirm this notion in Appendix D, where we analyze the evolution of the ID profile over the course of training for Pythia (the sole model whose intermediate checkpoints are public). We find that, over the course of training, the IDs' magnitude not only grows over time, but that they become more peaked (Figure D.1), indicating that the characteristic profile emerges from learned structure.

The ID peak marks a transition in layer function Figure 2 reports  $\Delta(l_i \to l_{first})$  and  $\Delta(l_i \to l_{last})$  for Llama, OPT and Pythia (with the remaining models in Appendix E). We observe that the ID peak largely overlaps with a peak in  $\Delta(l_i \to l_{first})$ . Here, a large value of  $\Delta$  implies that sequences which are nearest neighbours at the ID-peak are distant from each other in the input layer. That is, the ID peak layers no longer contain the information encoded in the initial representation of the sequence, suggesting that they instead capture higher-level information. Meanwhile, we note that the  $\Delta(l_i \to l_{last})$  profiles are not clearly related to the first ID peak, and we leave their analysis to future work.

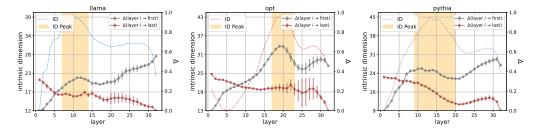


Figure 2: For Llama, OPT, Pythia (left to right), the ID is overlaid with  $\Delta(l_i \to l_{first})$  (gray) and  $\Delta(l_i \to l_{last})$  (brown). Plots are shown with  $\pm 2$  standard deviations over 5 partitions of 3 corpora. For all models, there is a peak in  $\Delta(l_i \to l_{first})$  (gray) around the ID peak.

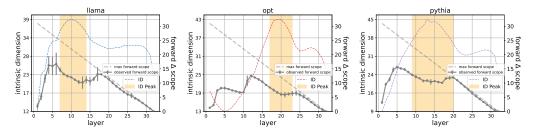


Figure 3: Forward  $\Delta$  scope (left: Llama; center: OPT; right: Pythia): continuous lines report, for each layer  $l_n$ , the number of adjacent following layers  $l_{n+k}$  for which  $\Delta(l_n \to l_{n+k}) \le 0.1$ . The dashed line represents the longest possible scope for each layer. Values are averaged across corpora and partitions, with error bars of  $\pm$  2 standard deviations.

Figure 3 shows the *forward*  $\Delta$  *scope* profile for Llama, OPT and Pythia (see Appendix F for Mistral and OLMo). In particular, the plots indicate, at each source layer  $l_n$ , for how many contiguous following layers  $l_{n+k}$  the quantity  $\Delta(l_n \to l_{n+k})$  is below a low threshold (0.1). Qualitatively, this means that the source layer  $l_n$  contains most of the information in the following k layers. Coinciding with the ID peak is a downward dip in forward-scope, compatible with the interpretation that, while high-ID layers process similar information, this information differs from that of later layers.

At the ID peak, different models share representation spaces The high-dimensionality peaks contain similar information across models, as shown in the representative comparisons of Figure 4 (see Appendix G for the other pairs), which display cross-model  $\Delta$  averaged across corpora and partitions. The high-ID layers always overlap with low- $\Delta$  layers, indicating that between models, representations at the peak have close neighborhood structures, and thus capture similar semantics.

Conversely, ID-peak layers have high cross-model  $\Delta$  with layers outside the peak, shown by a lack of points outside the shaded intersection in Figure 4. This indicates that, while ID-peak representations predict those of other models, they contain different information from other models' representations outside of the peak.

We notice, moreover, a marked asymmetry in which the ID-peak layers of Pythia and OPT, the two models with the highest absolute IDs, directionally contain other models' representations. On the other hand, when models have similar maximum IDs (such as in the Pythia vs. OPT comparison),  $\Delta$  is more symmetric and very small, implying that their representation spaces are really equivalent.

# 4.2 Language processing during the high-dimensionality phase

We just saw, via geometric evidence  $(\Delta)$ , that the high-ID phase marks a change in processing function. Now, we examine exactly *what* that function is. To do so, we look at layer-wise classification accuracies for the probing tasks described in Appendix H. We find that, in general, ID-peak representations fail at surface-form tasks but excel at semantic and syntactic tasks, indicating a functional transition from superficial to abstract linguistic processing.

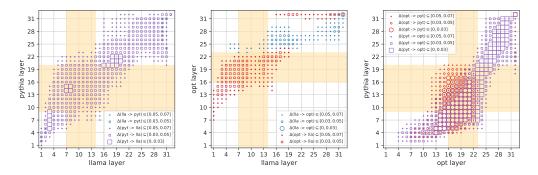


Figure 4: Cross-model  $\Delta$ . ID-peak sections are shaded in orange. Different symbols mark different  $\Delta$  levels in the two directions (lower values correspond to a stronger trend towards information containment). High  $\Delta$  scores (> 0.1), corresponding to low information containment, are not shown. Values averaged over corpora and partitions.

The ID-peak representations contain less surface-form information In Figure 5a, we consider the accuracy for two tasks, Sentence Length and Word Content, which test whether a layer retains information about superficial properties of the input. We observe that the ability to correctly reconstruct the length, measured in tokens, of the input sentence gets lost as we climb the network layers. Intriguingly, for OPT (Figure 5a, center), which is the model with the latest ID peak, initial accuracy is relatively stable, and only starts to significantly decrease at the onset of the peak.

Concerning the Word Content task, which tests the ability to detect the presence of specific words in the input, we generally observe a great decrease, particularly clear during ID expansion. Interestingly, accuracy tends to go up again after the ID peak, probably because, as the model prepares to predict the output, more concrete lexical information is again encoded in its representations. Together, the surface-form tasks confirm the evidence from  $\Delta(l_i \to l_{first})$  (Figure 2 above) that the ID peak processes a more abstract type of information that, as we are about to see, may relate to the syntactic and semantic contents of the sequence.

The ID peak marks a transition to syntactic and semantic processing Figure 5b reports accuracy for Llama, Pythia, and OPT (results for the remaining models in Appendix H) on the syntactic and semantic probe tasks. Despite variation across tasks and models, we observe that, in general, asymptotic accuracy is reached during the ID peak phase. Again, this implies that, for OPT, the asymptote is reached later. For OLMo (Appendix H), we observe in some cases a late accuracy peak, related to the second ID expansion phase that this model undergoes. In general, however, once top accuracy is reached, performance stays quite constant across layers, suggesting that the expressivity of high dimensionality permits rich linguistic representation of the inputs, but, once this information is extracted, it is propagated across subsequent layers. This is intuitive, as high-level linguistic information is useful to the network for its ultimate task of next-token prediction.

**Better LMs have higher ID peaks, earlier** Given the apparent linguistic importance of the high-ID phase, a natural question concerns the extent to which the nature of the ID peak explains the LM's performance on its original task of next-token prediction. The question was already partially answered by Figures 1 and D.1, which show an absence of peaks, respectively, when trained LMs process shuffled text and when untrained LMs process sane (unshuffled) text; in both cases, next-token prediction is impossible.

We further computed Spearman correlations between average prediction surprisal for each (model, corpus) combination and the corresponding *maximum ID values* (Figure 6, left), as well as *ID-peak onsets* (Figure 6, right). As the plots show, even when limiting the analysis to sane text, where the differences in surprisal will be smaller, there is a marginal tendency for maximum ID value to inversely correlate with surprisal: the *higher* the peak, the *better* the LM is at predicting the next token. There is, moreover, a significant positive correlation between surprisal and the onset of the ID peak: the *earlier* the peak, the *better* the model is at predicting the next token.

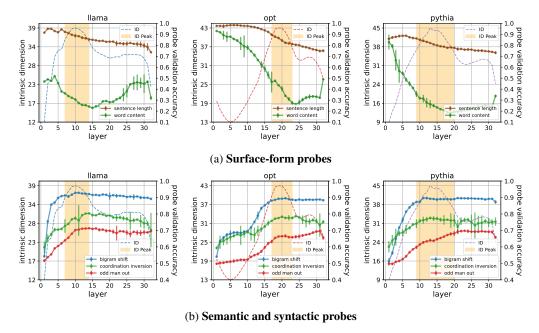


Figure 5: Linguistic knowledge probing performance  $\pm$  2 SDs across 5 random seeds is shown with the ID for Llama, OPT, and Pythia (left to right). Row (a) corresponds to surface-form tasks Sentence Length and Word Content, where probe performance decreases through the ID peak. Row (b) corresponds to syntactic and semantic tasks Bigram Shift, Coordination Inversion and Odd Man Out, where probe performance for all tasks attains maximum (or close) within the ID peak. This suggests the ID peak marks *abstract*, and not surface, representations of the input.

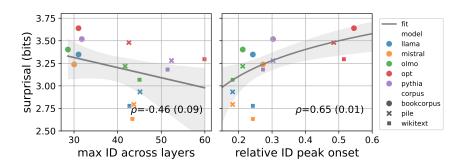


Figure 6: Surprisal plotted against the maximum ID across layers (left) and the relative ID peak onset over layers (right), where each datapoint is a (model, corpus) combination (N=50k sequences per corpus). A linear fit (left) and log-linear fit (right) are shown. (Left) Surprisal negatively correlates to maximum ID with Spearman  $\rho=-0.46, p=0.09$ , meaning that higher ID indicates better LM performance. (Right) Surprisal positively correlates to ID peak onset,  $\rho=0.65, p=0.01$ , meaning that an earlier ID peak indicates better LM performance.

We just saw that the ID peak marks the phase where the model first completes a full syntactic and semantic analysis of the input. The earlier this analysis takes place, the more layers the model will have to further refine its prediction by relying on it. The correlation between ID peak onset and surprisal thus indirectly confirms the importance of the high-dimensional processing phase for good model performance.

**ID-peak layers are the first to transfer to downstream tasks** Finally, given that the ID peak marks an abstract semantic representation of the input, representations at this peak should also viably transfer to downstream semantic tasks. We confirm this hypothesis for sentiment and toxicity classification.

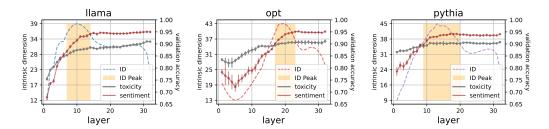


Figure 7: Validation performance of representation transfer performance on toxicity and sentiment classification for Llama, OPT, and Pythia (left to right). All validation accuracy curves are plotted with  $\pm$  2 standard deviations over 5 random seeds. For all models, classification validation accuracy converges within the ID peak.

Validation accuracy curves for both tasks are shown for Llama, Pythia and OPT in Figure 7, with more results in Appendix I. Similar to the semantic and syntactic probing results, downstream classification performance consistently converges at the ID peak across models and remains high thereafter. Note that, while the ID peak is computed from the generic input corpora Bookcorpus, the Pile, and Wikitext, it still predicts transferrability to different downstream datasets that are reasonably in-distribution.

# 5 Conclusion

It is evident that a LM needs to extract information from its input to predict the next token. The non-trivial fact we show here is that this process does not gradually refine representations, but rather undergoes a phase transition characterized by ID expansion, cross-model information sharing, and a switch to abstract information processing. Mirroring Jastrzebski et al. [2018]'s proposition for visual networks, our findings are compatible with a view of LMs in which the high-dimensional early-to-mid layers functionally specialize to analyze inputs in a relatively fixed manner, whereas later layers may more flexibly refine the output prediction, using information extracted during the high-ID phase. Interestingly, two recent papers [Gromov et al., 2024, Men et al., 2024] found that (1) late LM layers better approximate each other than earlier layers do, and (2) pruning late layers (excluding the last) affects performance less than pruning earlier layers. This fully aligns with our results and suggests a need to test the effect of pruning inside and outside the ID peak. Other studies have highlighted the importance of central layers in performing various core functions. For example, Hendel et al. [2023] find that compositional "task vectors" are formed in layers superficially consistent with the peaks we detect. Again, future work should more thoroughly study the relation between the ID profile and specific circuits detected in mechanistic interpretability work.

While our work closely relates to that of Valeriani et al. [2023], who studied the evolution of ID in vision and protein transformers, an interesting contrast is that they found crucial semantic information to coalesce during a dimensionality reduction phase, whereas we associated similar marks to a dimensionality expansion phase. In line with our results, recent work in theoretical neuroscience shows that high ID, thanks to its expressivity, underlies successful few-shot learning [Sorscher et al., 2022] and generalizable latent representations for DNNs [Elmoznino and Bonner, 2024, Wakhloo et al., 2024]. Conversely, primarily in artificial and biological vision, low ID is linked to generalization thanks to representations' robustness to noise and greater linear separability in embedding space [Amsaleg et al., 2017, Chung et al., 2018, Cohen et al., 2020]. Clearly, whether dimensionality is a curse or blessing to performance depends on the context of the learning problem. Reconciling, then, why and when high performance arises from reduced or expanded dimensionality remains an important direction for future work.

#### 6 Limitations

 While we establish an empirical connection between the ID peak and linguistic processing, we do not offer a mechanistic understanding of how the latter results in ID expansion, including which differences between models lead to different patterns, such as an earlier or later ID peak onset.

- The OLMo model displays a second ID peak and various outlying patterns that we did not attempt to interpret.
- We explored 5 language models on data extracted from three different corpora. However, due to compute constraints, the models we picked are all in the same size range (between 6 and 8 billion parameters). In the future, we would like to explore the extent to which the patterns we detected also appear at other scales.

# References

- CJ Adams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, Nithum, and Will Cukierski. Toxic comment classification challenge, 2017. URL https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge.
- Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328, Online, August 2021. Association for Computational Linguistics.
- Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 4895–4901, 2023.
- Laurent Amsaleg, James Bailey, Dominique Barbe, Sarah Erfani, Michael E. Houle, Vinh Nguyen, and Miloš Radovanović. The vulnerability of learning to adversarial perturbation increases with intrinsic dimensionality. In 2017 IEEE Workshop on Information Forensics and Security (WIFS), page 1–6, December 2017.
- Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- Randall Balestriero, Romain Cosentino, and Sarath Shekkizhar. Characterizing large language model geometry solves toxicity detection and generation. https://arxiv.org/abs/2312.01648, 2023.
- Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. (arXiv:2304.01373), Apr 2023. URL http://arxiv.org/abs/2304.01373.
- Gemma Boleda. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6(Volume 6, 2020):213–234, January 2020. ISSN 2333-9683, 2333-9691. doi: 10.1146/annurev-linguistics-011619-030303.
- Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations*, 2021.
- P. Campadelli, E. Casiraghi, C. Ceruti, and A. Rozza. Intrinsic dimension estimation: Relevant techniques and a benchmark framework. *Mathematical Problems in Engineering*, 2015:e759567, Oct 2015. ISSN 1024-123X.
- Emily Cheng, Corentin Kervadec, and Marco Baroni. Bridging information-theoretic and geometric compression in language models. In *Proceedings of EMNLP*, pages 12397–12420, Singapore, 2023.
- Sue Yeon Chung, Daniel D. Lee, and Haim Sompolinsky. Classification and geometry of general perceptual manifolds. *Phys. Rev. X*, 8:031003, Jul 2018.

- Uri Cohen, SueYeon Chung, Daniel D. Lee, and Haim Sompolinsky. Separability and geometry of object manifolds in deep neural networks. *Nature Communications*, 11(1):746, Feb 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-14578-5. URL https://www.nature.com/articles/s41467-020-14578-5.
- Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Proceedings of NeurIPS*, volume 36, pages 16318–16352, New Orleans, LA, 2023.
- Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single \$\&!\#\* vector: Probing sentence embeddings for linguistic properties. In *Proceedings ACL*, pages 2126–2136, Melbourne, Australia, 2018.
- Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.
- Gregoire Deletang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, Marcus Hutter, and Joel Veness. Language modeling is compression. In *The Twelfth International Conference on Learning Representations*, 2024.
- Francesco Denti, Diego Doimo, Alessandro Laio, and Antonietta Mira. The generalized ratios intrinsic dimension estimator. *Scientific Reports*, 12(11):20005, Nov 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-20991-1.
- Diego Doimo, Aldo Glielmo, Alessio Ansuini, and Alessandro Laio. Hierarchical nucleation in deep neural networks. *Advances in Neural Information Processing Systems*, 33:7526–7536, 2020.
- 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.
- Eric Elmoznino and Michael F. Bonner. High-performing neural network models of visual cortex benefit from high latent dimensionality. *PLOS Computational Biology*, 20(1):e1011792, January 2024. ISSN 1553-7358.
- Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):12140, Sep 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-11873-y.
- Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta Costa-jussá. A primer on the inner workings of transformer-based language models. https://arxiv.org/abs/2405.00208, 2024.
- 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. http://arxiv.org/abs/2101.00027, 2020.
- Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of EMNLP*, pages 12216–12235, Singapore, 2023.
- Aldo Glielmo, Claudio Zeni, Bingqing Cheng, Gábor Csányi, and Alessandro Laio. Ranking the information content of distance measures. *PNAS nexus*, 1(2):pgac039, 2022.
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. MIT Press, Cambridge, MA, 2016.

- Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. Olmo: Accelerating the science of language models, 2024.
- Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Daniel Roberts. The unreasonable ineffectiveness of the deeper layers. https://arxiv.org/abs/2403.17887, 2024.
- Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of EMNLP*, pages 9318–9333, Singapore, 2023.
- Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415, 2016.
- John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings EMNLP*, pages 2733–2743, Hong Kong, China, 2019.
- Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of ICLR*, Online, 2022. Published online: https://openreview.net/group?id=ICLR.cc/2022/Conference.
- Stanisław Jastrzebski, Devansh Arpit, Nicolas Ballas, Vikas Verma, Tong Che, and Yoshua Bengio. Residual connections encourage iterative inference. In *Proceedings of ICLR Conference Track*, Vancouver, Canada, 2018. Published online: https://openreview.net/group?id=ICLR.cc/2018/Conference.
- 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.
- Ian Jolliffe. Principal Component Analysis. Springer, 1986.
- Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis: Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2(4):1–28, 2008.
- Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher Manning, Christopher Ré, Diana Acosta-Navas, Drew Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 8:1–162, 2023.
- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

- Xin Men, Mingyu Xu, Qingyu Zhang, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. ShortGPT: Layers in large language models are more redundant than you expect. https://arxiv.org/abs/2403.03853, 2024.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Proceedings of NeurIPS*, volume 35, pages 17359–17372, New Orleans, LA, 2022.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.
- Meta. Introducing meta llama 3: The most capable openly available llm to date, 2024. URL https://ai.meta.com/blog/meta-llama-3/.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. *PyTorch: an imperative style, high-performance deep learning library*. Curran Associates Inc., Red Hook, NY, USA, 2019.
- Fabian Pedregosa, Gaël Varoquaux, , Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- 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.
- Michael Psenka, Druv Pai, Vishal Raman, Shankar Sastry, and Yi Ma. Representation learning via manifold flattening and reconstruction. *Journal of Machine Learning Research*, 25(132):1–47, 2024.
- Noam Shazeer. Glu variants improve transformer. arXiv preprint arXiv:2002.05202, 2020.
- Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research. (arXiv:2402.00159), January 2024. doi: 10.48550/arXiv.2402.00159. URL http://arxiv.org/abs/2402.00159 arXiv:2402.00159 [cs].
- Ben Sorscher, Surya Ganguli, and Haim Sompolinsky. Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences*, 119(43): e2200800119, October 2022. doi: 10.1073/pnas.2200800119.
- Yusheng Su, Xiaozhi Wang, Yujia Qin, Chi-Min Chan, Yankai Lin, Huadong Wang, Kaiyue Wen, Zhiyuan Liu, Peng Li, Juanzi Li, Lei Hou, Maosong Sun, and Jie Zhou. On transferability of prompt tuning for natural language processing. In *Proceedings of NAACL*, pages 3949–3969, Seattle, WA, 2022.
- Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.
- Eduard Tulchinskii, Kristian Kuznetsov, Kushnareva Laida, Daniil Cherniavskii, Sergey Nikolenko, Evgeny Burnaev, Serguei Barannikov, and Irina Piontkovskaya. Intrinsic dimension estimation for robust detection of AI-generated texts. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=8u0Z0kNji6.

- Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. In *Proceedings of NeurIPS*, pages 51234–51252, New Orleans, LA, 2023.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of NIPS*, pages 5998–6008, Long Beach, CA, 2017.
- Albert J. Wakhloo, Will Slatton, and SueYeon Chung. Neural population geometry and optimal coding of tasks with shared latent structure. (arXiv:2402.16770), February 2024. doi: 10.48550/arXiv.2402.16770. URL http://arxiv.org/abs/2402.16770. arXiv:2402.16770 [cond-mat, q-bio].
- Yaodong Yu, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Benjamin Haeffele, and Yi Ma. White-box transformers via sparse rate reduction. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 9422–9457. Curran Associates, Inc., 2023.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models. https://arxiv.org/abs/2205.01068, 2022.
- Zhong Zhang, Bang Liu, and Junming Shao. Fine-tuning happens in tiny subspaces: Exploring intrinsic task-specific subspaces of pre-trained language models. In *Proceedings of ACL*, pages 1701–1713, Toronto, Canada, 2023.
- Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of ICCV*, pages 19–27, Santiago, Chile, 2015.

# **A** Computing resources

All experiments were run on a cluster with 12 nodes with 5 NVIDIA A30 GPUs and 48 CPUs each.

Extracting LM representations took a few wall-clock hours per model-dataset computation. ID computation took approximately 0.5 hours per model-dataset computation. Information imbalance computation took about 2 hours per model-dataset computation. Probing/transfer classifiers took up to 2 days per task.

Taking parallelization into account, we estimate the overall wall-clock time taken by all experiments, including failed runs, preliminary experiments, etc., to be of about 20 days.

# **B** Assets

Bookcorpus https://huggingface.co/datasets/bookcorpus; license: unknown

Pile-10k https://huggingface.co/datasets/NeelNanda/pile-10k; license: bigscience-bloom-rail-1.0

Wikitext https://huggingface.co/datasets/wikitext; license: Creative Commons Attribution Share Alike 3.0

Llama https://huggingface.co/meta-llama/Meta-Llama-3-8B; license: llama3

Mistral https://huggingface.co/mistralai/Mistral-7B-v0.1; license: apache-2.0

**OLMo** https://huggingface.co/allenai/OLMo-7B; license: apache-2.0

OPT https://huggingface.co/facebook/OPT-6.7b; license: OPT-175B license

Pythia https://huggingface.co/EleutherAI/pythia-6.9b-deduped; license: apache-2.0

DadaPy https://github.com/sissa-data-science/DADApy; license: apache-2.0

scikit-learn https://scikit-learn.org/; license: bsd

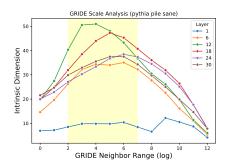


Figure C.1: GRIDE ID estimation neighborhood scale analysis example for Pythia on the Pile on a single random seed, where each line is a layer's ID estimate at different scales. All layers shown reach a plateau in the highlighted range.

PyTorch https://scikit-learn.org/; license: bsd

**Probing tasks** https://github.com/facebookresearch/SentEval/tree/main/data/probing; license: bsd

Toxicity dataset https://huggingface.co/datasets/google/jigsaw\_toxicity\_pred; license: CC0

Sentiment dataset https://huggingface.co/datasets/stanfordnlp/imdb; license: un-known

# **C** Intrinsic Dimension

# C.1 Scale analysis

For each model and corpus, we perform a scale analysis to select the neighbor order. An example of such an analysis is shown in Figure C.1, performed on the Pile for Pythia. We plot the ID estimate for increasing scales (x-axis in Figure C.1) for each layer. The true ID is likely to lie at a scale in which the ID estimate plateaus [Denti et al., 2022], which is marked by the highlighted region. For simplicity, per model-corpus combination, we choose one scale for all layers: in this particular example, we choose order  $k=2^5$ . In general, the scale tends to be around k=32 for all (model, corpus) combinations, allowing us to reliably compare between them (see Tables C.1a and C.1b for all k). Once the scale is chosen for each model-corpus combination, we plot the scale-adjusted ID estimates (see, e.g., Figure C.2).

#### C.2 Additional results

Figure C.2 displays the evolution of estimated ID, scale-adjusted, per layer for all corpora and models. While the magnitude of ID differs across corpora, with ID on Bookcorpus lower than those of the Pile and Wikitext for all models, all corpora exhibit the characteristic high-ID peak, and at nearly the same onset. The dampened Bookcorpus peak IDs, which are still significantly above the corresponding shuffled ID peaks, might be explained by the fact that this corpus is entirely made of novels, and it is thus the less in-domain of the corpora we used.

# D Intrinsic dimension over training

Figure D.1 shows the evolution of the average-ID profile over the course of training for Pythia, whose checkpoints are publicly available. Of-note, the ID profile seems to converge at between 16000 and 64000 iterations, or less than halfway until training is finished. We observe that towards the beginning of training (blue line), the ID profile is quite flat, and that, over the course of training, the characteristic central ID peak emerges. Beyond shape, we see that, while the initial-layer ID remains the same across checkpoints, the IDs of other layers tend to increase in magnitude. Both observations suggest that *expansion of ID* underlies linguistic processing of the input.

model	corpus	mode	<b>GRIDE</b> $k$
llama	bookcorpus	sane shuffled	32 128
	pile	sane shuffled	64 128
	wikitext	sane shuffled	64 16
mistral	bookcorpus	sane shuffled	64 128
	pile	sane shuffled	128 256
	wikitext	sane shuffled	64 32
olmo	bookcorpus	sane shuffled	32 8
	pile	sane shuffled	32 128
	wikitext	sane shuffled	32 16
opt	bookcorpus	sane shuffled	16 16
	pile	sane shuffled	32 16
	wikitext	sane shuffled	16 16
pythia	bookcorpus	sane shuffled	32 32
	pile	sane shuffled	32 64
	wikitext	sane shuffled	32 16

corpus	Pythia step	<b>GRIDE</b> $k$
	512	32
bookcorpus	4000	32
	16000	32
	64000	32
	512	4
pile	4000	32
	16000	32
	64000	64
	512	4
:1-:44	4000	32
wikitext	16000	32
	64000	32

(b) GRIDE k for additional Pythia checkpoints.

# E Information imbalance with respect to first/last layer

Figure E.1 shows averaged  $\Delta(l_i \to l_{first})$  (gray) and  $\Delta(l_i \to l_{last})$  for Mistral and OLMo. Recall that the closer  $\Delta(A \to B)$  is to 0, the more predictive A's local neighborhood structure is of B's. As expected,  $\Delta(l_i \to l_{first})$  generally increases with i as we go deeper in the layers; the reverse is true for  $\Delta(l_i \to l_{last})$ . However,  $\Delta(l_i \to l_{first})$  appears to locally peak and plateau around the representational ID peak; that is, the ID expansion marks a phase of low predictivity from the intermediate layer to the input. OLMo's pattern is not as clear, and we might observe a second local information imbalance peak in proximity to the second ID peak characterizing this model.

# F Forward $\Delta$ scope

Figure F.1 reports the forward  $\Delta$  scope profile for the remaining two LMs not shown in the main text (Mistral and OLMo).

<sup>(</sup>a) GRIDE order k reported for each model, corpus, mode (in sane and shuffled) combination. For simplicity, we chose one k for all layers.

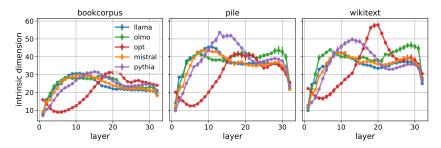


Figure C.2: ID evolution over layers, shown with one standard deviation (over corpus partitions), for all models and corpora (left to right: Bookcorpus, the Pile, and Wikitext).

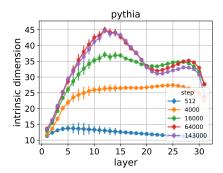


Figure D.1: Evolution of the average-ID profile over training for Pythia. Each curve corresponds to one checkpoint, at steps 512 until 143000 (the final model), and is shown with  $\pm$  2 SD over corpora and partitions.

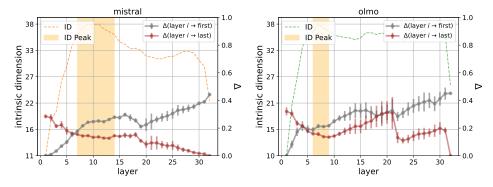


Figure E.1: For Mistral (left) and OLMo (right), the ID (hued) is overlaid with  $\Delta(l_i \to l_{first})$  (gray) and  $\Delta(l_i \to l_{last})$  (brown). Plots are shown with  $\pm$  2 standard deviations over corpora and partitions.

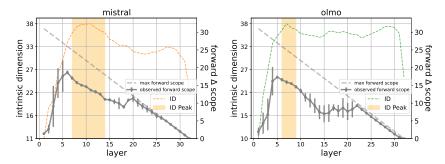


Figure F.1: Forward  $\Delta$  scope (Mistral, Olmo): continuous lines report, for each layer  $l_n$ , the number of adjacent following layers  $l_{n+k}$  for which  $\Delta(l_n \to l_{n+k}) \leq 0.1$ . The dashed line represents the longest possible scope for each layer. Values are averaged across corpora and partitions, with error bars of  $\pm$  2 standard deviations.

# **G** Cross-model $\Delta$

Figure G.1 show cross-model  $\Delta$  for the remaining combinations, confirming that there are areas of low cross-model information imbalance at the intersection of the high-ID peaks. In combinations involving OLMo, we observe a tendency for low  $\Delta$  to stretch along the other LM high-ID section, suggesting that the high-ID layers of other LMs share information with a wider range of OLMo layers.

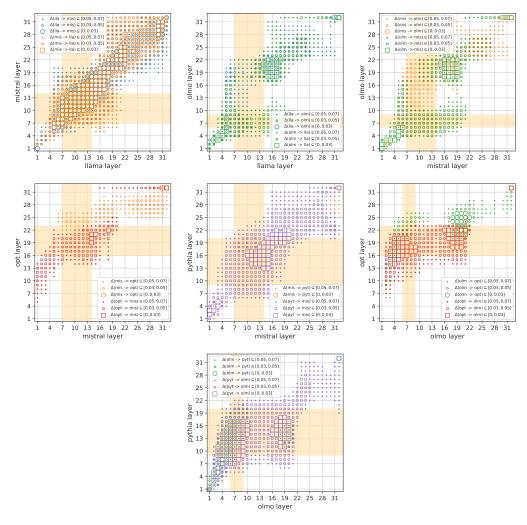


Figure G.1: Cross-model  $\Delta$ . ID-peak sections are shaded in orange. Different symbols mark different information imbalance levels in the two directions (the lower the  $\Delta(A \to B)$  values, the more the information in B is contained in A). High imbalances (> 0.1) are not shown. Values averaged across corpora and partitions.

# **H** Probing tasks

#### H.1 Tasks

We use the following classification tasks from Conneau et al. [2018]:

• Surface form

**Sentence Length** Predict input sentence length in tokens (lengths binned into 5 intervals). **Word Content** Tell which of a pre-determined set of 1k words occurs in the input sentence.

Syntax

**Bigram Shift** Tell whether the input sentence is well-formed, or it has been corrupted by inverting the order of two adjacent tokens (e.g., "They were in present droves, going from table to table and offering to buy meals, drinks or generally attempting to strike up conversations").

• Semantics

**Coordination Inversion** Tell whether a sentence is well-formed or it contains two coordinated clauses whose order has been inverted (e.g., "Then I decided to treat her just as I would anyone else, but at first she'd frightened me").

**Odd Man Out** Tell whether a sentence is well-formed, or it involves the replacement of a noun or verb with a random word with the same part of speech (e.g., "The people needed a sense of chalk and tranquility").

We exclude the following tasks because we observed ceiling effects across the layers, suggesting that the models could latch onto spurious correlations in the data: Past Present, Subject Number and Object Number. We exclude Top Constituents and Tree Depth because they produced hard-to-interpret results that we believe are due to the fact that they rely on automated syntactic parses that are not necessarily consistent with the way modern LMs process their inputs.

#### H.2 Setup

We use the training and test data provided by Conneau et al. [2018]. We train a MLP classifier for each task and each layer of each LM, repeating the experiment with 5 different seeds.

We fixed the following hyperparameters of the MLP, attempting to approximate those used in the original paper (as each task takes days to complete, we could not perform our own hyperparameter search):

• Number of layers: 1

• Layer dimensionality: 200

• Non-linearity: logistic

• L2 regularization coefficient: 0.0001

• seeds: 1, 2, 3, 4, 5

For all other hyperparameters, we used the default values set by the Scikit-learn library [Pedregosa et al., 2011].

We repeated all experiments after shuffling the example labels (both at training and test time). This provides a baseline for each task, as well as functioning as a sanity check that a probing classifier is not so powerful as to simply memorize arbitrary patterns in the representations [Hewitt and Liang, 2019].

# **H.3** Additional Results

Figure H.1 reports the probing performance for OLMo and Mistral, and for surface form (top row) as well as syntactic and semantic tasks (bottom row). As for Llama, OPT, and Pythia, we observe that the first ID peak converges to viable syntactic/semantic abstraction of inputs, while discarding information about surface form.

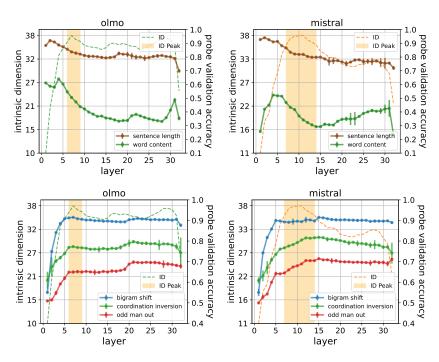


Figure H.1: Linguistic knowledge probing performance  $\pm 2$  SDs across 5 random seeds is plotted with the ID across layers for OLMo and Mistral (left to right). (Top row) Surface form tasks Sentence Length and Word Content, where probe performance decreases through the ID peak. (Bottom row) Semantic and syntactic tasks Bigram Shift, Coordination Inversion and Odd Man Out, where probe performance for all tasks attains maximum within the ID peak.

Figure H.2 shows that, on the shuffled corpora ablation, the MLP probes perform at chance. We then confirm that semantic and syntactic information is contained in the *model representations* and not in the probes.

#### I Downstream Tasks

# I.1 Tasks

**Toxicity detection.** We use a random balanced subset (N=30588) of Kaggle's jigsaw toxic comment classification challenge [Adams et al., 2017], where each data point consists of a natural language comment and its binary toxicity label.

**Sentiment classification.** We use a dataset of IMDb movie reviews [Maas et al., 2011], where each data point consists of a natural language film review and a corresponding label  $\in$  {positive, negative}. To train the linear probes, we take a sample of N=25000 corresponding to the train split on HuggingFace. We repeat the experiment with 5 distinct seeds.

# I.2 Setup

To train the linear probes, we first divide the data at random into train (80%) and validation (20%) sets. We feed each training set through each model and gather the last token hidden representations at each layer. Then, using PyTorch [Paszke et al., 2019], we train one linear probe per layer with hyperparameters as follows,

• Number of epochs: 1000

• lr: 0.0001

• seeds: 32, 36, 42, 46, 52

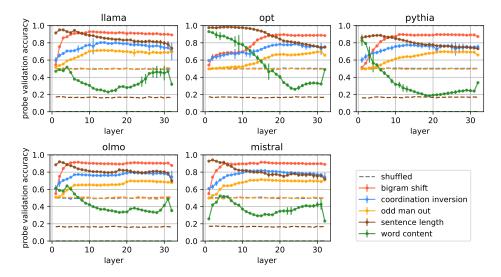


Figure H.2: Linguistic probe validation accuracy for semantic, syntactic, and surface tasks (solid lines) and their shuffled versions (dashed lines) are shown across layers for all models. The probing performance on shuffled corpora is constant at chance for all tasks and models.

and we report the best validation accuracy over 5 random seeds.

# I.3 Additional Results

Similar to Llama, OPT, and Pythia, Figure I.1 shows that validation performance for downstream tasks for Mistral and OLMo converge in the ID peak.

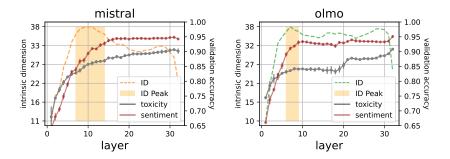


Figure I.1: Validation performance on downstream transfer tasks on toxicity and sentiment classification for Mistral (left) and OLMo (right). All validation accuracy curves are plotted with  $\pm$  2 standard deviations over 5 random seeds. For all models, classification validation accuracy converges within the ID peak.