# Tracing the complexity profiles of different linguistic phenomena through the intrinsic dimension of LLM representations

Anonymous ACL submission

## Abstract

We explore the intrinsic dimension (ID) of LLM representations as a marker of linguistic complexity, asking if different ID profiles across LLM layers differentially characterize *formal* and *functional* complexity. We find the formal contrast between sentences with multiple coordinated or subordinated clauses to be reflected in ID differences whose onset aligns with a phase of more abstract linguistic processing independently identified in earlier work. The functional contrasts between sentences characterized by right branching vs. center embedding or unambiguous vs. ambiguous relative clause attachment are also picked up by ID, but in a less marked way, and they do not correlate with the same processing phase. Further experiments using representational similarity and layer ablation confirm the same trends. We conclude that ID is a useful marker of linguistic complexity in LLMs, that it allows to differentiate between different types of complexity, and that it points to similar stages of linguistic processing across disparate LLMs.

## 1 Introduction

Given LLMs' remarkable linguistic skills, there is widespread interest in understanding how language processing unfolds in their inner layers, both because a better understanding of their inner workings could lead to more efficient and controllable models (Balestriero et al., 2024; Gromov et al., 2025), and because LLMs are imperfect but practical in silico models of the human faculty of language (Futrell and Mahowald, 2025; Levy et al., 2025).

In this context, the *Intrinsic Dimension* (ID) of LLMs' inner representations has been used to gather insights on their internal processing stages (e.g., Valeriani et al., 2023; Doimo et al., 2024; Cheng et al., 2025). When a set of data is mapped to a high-dimensional representation (e.g., sentences embedded in an LLM layer space), i) they will typically lie near a manifold that effectively occupies a much smaller number of dimensions than the full space, i.e., the intrinsic dimension is much lower than the so-called ambient dimension; and ii) the more complex the data are for the representational system (in our case, the LLM layer), the higher this intrinsic dimension will be. Consequently, we can use ID as a probe for complexity of linguistic processing across LLM layers.

It has been shown that LLMs have consistent ID profiles across their layers, and that these profiles cue phases of more or less abstract linguistic processing. Earlier studies, however, only looked at average model behaviour in response to generic corpus data (Cai et al., 2021; Cheng et al., 2025) or at very coarse distinctions between data types (Tulchinskii et al., 2023; Yin et al., 2024; Lee et al., 2025). We take a more granular view of linguistic complexity and use ID to characterize how LLMs handle three complexity contrasts that have long been acknowledged in the (psycho)linguistic literature. In particular, we study, in English, i) the distinction between sentences with subordinated vs. coordinated clauses; ii) the contrast between center-embedding structures with long-distance agreement and equivalent sentences with right branching; and, finally, iii) sentences with attachment ambiguities compared to sentences where meaning disambiguates the attachment site. We classify the first contrast as *formal*, and the other two as *functional*, as discussed in Section 2.

We find a remarkable degree of consistency in how different LLMs process data characterized by these phenomena, and that their ID profiles point at different processing strategies for each of them. Our evidence from ID is further supported by experiments tracking representational similarities and ablation effects. Overall, our results suggest that sufficiently powerful systems converge to similar ways to handle language, even when they lack explicit priors for it, and consequently that LLMs could provide us with new insights on the nature of

linguistic complexity, that can in turn complement human sentence processing research.

## 2 Related work

**Intrinsic dimension** While naturalistic language data appear high-dimensional, the *manifold hypothesis* posits they actually lie near a low-dimensional manifold (Goodfellow et al., 2016). The *intrinsic dimension* of the data is then the dimension of this possibly nonlinear manifold, that is, the number of degrees of freedom that explain it under minimal information loss (Campadelli et al., 2015). The manifold hypothesis holds not only for LLM parameter spaces (Aghajanyan et al., 2021; Zhang et al., 2023), but also for their activations: no matter the model or dataset, existing work shows that LLM representation manifolds have an ID orders-of-magnitude lower than their ambient dimension (Cai et al., 2021; Valeriani et al., 2023; Cheng et al., 2025). Similar to our work, Lee et al. (2025) and Cheng et al. (2023) show that the mean representational ID over layers correlates to formal or psycholinguistic notions of linguistic complexity, in particular, $n$-gram diversity (Lee et al., 2025) and surprisal and learnability (Cheng et al., 2023). We build directly on the work by Cheng et al. (2025) who found, across different LLMs, a characteristic per-layer ID profile, whereby there is an ID peak in intermediate layers. Through probing and downstream tasks, they show that this ID peak coincides with the phase where the model is first able to perform complex linguistic tasks, suggesting that the peak cues a stage of deep linguistic processing.

**Linguistic probing of LLMs** A rich literature has probed the internal representations of linguistic structures in LLMs, demonstrating that they encode aspects of syntactic knowledge and provide insights into where and how this information is stored (Belinkov and Glass, 2019; Linzen and Baroni, 2021; Rogers et al., 2020; Ferrando et al., 2024; Li and Subramani, 2025; Simon et al., 2025). Early work by Hewitt and Manning (2019) introduced a method based on linear probing, showing that bidirectional models represent information about syntactic dependencies and the relation between the involved constituents. Further research on more recent models and other probing techniques revealed a more nuanced picture of syntactic representation, showing that although LLMs do encode syntactic information, they are also sensitive to the interference of local cues, like closely occurring words (Agarwal et al., 2025; Simon et al., 2025). However, models can still pick up structural differences in sentences that look similar on the surface, and capture the subtle meaning changes they produce, showing an ability to integrate syntactic and semantic processing (Kennedy, 2025). He et al. (2024) compared how models process minimal pairs in which a well-formed sentence is paired with an otherwise identical sentence which includes a targeted grammatical violation: through this method, they showed that some information about syntactic competence is represented in early layers, and that as sentences become more complex, the models need more layers to evaluate their grammaticality. Crucially, features at the interface between syntax and semantics are more difficult for models to learn than purely syntactic patterns.

**Linguistic complexity** Linguistic complexity can be examined from two complementary perspectives. Formal complexity pertains to the structural and computational properties of grammar. This involves, for instance, the number of rules, the degree of hierarchical embedding, the number of idiosyncrasies in the system, or the length and depth of constituent and sentence structure (Culicover, 2014; Hawkins, 2014; Trotzke and Zwart, 2014). Functional or processing complexity, in turn, concerns how such structures are implemented and experienced by language users, shaping parsing, memory, acquisition, and neural activity (Hawkins, 2014; Menn and Duffield, 2014). The two perspectives are interrelated: formal descriptions delimit what structures a grammar makes available, while functional accounts reveal how these structures are deployed and constrained in real-time language processing. Importantly, functional complexity extends beyond formal properties, encompassing factors such as frequency, predictability, communicative efficiency, and cognitive limitations on processing and learning. Yet, formal and functional complexity do not always align: two structures may differ formally without exhibiting processing differences, and conversely, formally similar structures may impose very different functional loads. For the purposes of the present work, we adopt a simplified view in which certain linguistic phenomena primarily tap into one or the other dimension of complexity, while recognizing that the ultimate relationship between formal and functional factors is far more intricate.

| Complexity | Contrast | Example |
|---|---|---|
| Formal | Coordination | (1) The blacksmith is babbling and the politicians are doubting and the tutor is writing and the banker is listening |
| | **Subordination** | (2) The blacksmith is babbling that the politicians are doubting that the tutor is writing that the banker is listening |
| Functional | Right branching | (3) The politicians advised the potters that were waiting |
| | **Center embedding** | (4) The potters that the politicians advised were waiting |
| Functional | Unambiguous | (5) The mother of the infant who lost their first tooth stayed close |
| | **Ambiguous** | (6) The playmate of the infant who lost their first tooth stayed close |

Table 1: Examples of the experimental manipulations in the three datasets.

## 3 Experimental materials

In order to assess the explanatory potential of the ID of LLMs' representations, we designed three controlled datasets that probe formal and functional dimensions of linguistic complexity, exemplified in Table 1: coordination vs. subordination, right branching vs. center embedding, and unambiguous vs. ambiguous relative clause attachment. These contrasts are implemented through classic minimal pairs, allowing us to isolate the contribution of each linguistic manipulation.

**Coordination vs. subordination** These are clause-combining operations that yield syntactic configurations differing, among others, in hierarchical depth. Coordinated sentences are the result of an iterative process generating a flat structure. Subordinated sentences result from a recursive process generating an embedded structure (examples (1) and (2) in Table 1). In this sense, subordination corresponds to a higher level of formal structural complexity (Trotzke and Zwart, 2014). In spite of the near-total lexical overlap, the hierarchical differences between coordinated and subordinated structures lead to clear meaning differences. For example, since coordination links clauses of equal syntactic status, swapping them leaves meaning essentially unchanged. In contrast, the hierarchical dependency created by embedding one clause within another becomes evident when the clauses are swapped. Our datasets consist of sentences which only vary in the conjunction used (*and* vs. *that*) and the number of clauses that are combined (from 2 to 4). Details are in App. B.1.

**Right-branching vs. center-embedding** These structures (examples (3) and (4) in Table 1) differ sharply in their functional, processing-related complexity, despite being formally similar in terms of hierarchical depth and the type of underlying dependency relations: the two contain a relative clause (RC) modifying one of the NPs in the main clause. Yet, while right-branching sentences extend linearly by placing the RC at the end, the RC in center-embedding sentences appears between the subject and the verb, creating a long distance dependency. Even though this alternation leaves sentence meaning essentially intact (differing only in information-structural choices about which noun phrase is in focus), a vast amount of psycholinguistic literature has shown that center-embedding sentences are harder to process due to increased integration and memory storage cost (e.g., Gibson, 1998; Lewis and Vasishth, 2005). Our dataset creation procedure is detailed in App. B.2.

**Unambiguous vs. ambiguous** RC attachment is a common source of structural ambiguity that arises when a RC can modify more than one preceding NP. This ambiguity can be eliminated when only one NP is semantically compatible with the content of the RC, illustrating how cues at the syntax–semantics interface guide attachment decisions. Our datasets contain minimal pairs differing only in whether the attachment is lexically disambiguated in favor of the closer NP (example (5) in Table 1, a case of so-called "low" attachment) or whether both NPs remain semantically plausible, thereby sustaining the ambiguity (example (6) in Table 1).[1] Although the two sentences are formally parallel, many psycholinguistic studies have shown that semantically unambiguous RCs permit faster attachment, whereas ambiguous ones generate processing slow-downs due to increased competition between attachment sites (e.g., Gibson et al., 1996; Carreiras and Clifton, 1999). Dataset details are in App. B.3.

[1]We also include an unambiguous condition favoring attachment to the first NP, so-called "high" attachment (App. B.3). We focus on the comparison between ambiguous and low-attachment sentences because low attachment reflects the typical attachment preference in English, and thus provides a natural baseline. High-attachment sentences behave similarly to low-attachment ones (App. D.5).

**Higher complexity elicits higher LLM surprisal** We ran a preliminary check that LLM processing aligned with the predictions from the (psycho)linguistic literature sketched above. Indeed, for all datasets, all LLMs showed a higher per-token surprisal for the more complex condition as determined by a one-sided t-test ($\alpha = 0.05$). Exact surprisal values are given in Table 5 of App. D.1.

## 4 Methods

Given an LLM and a complexity contrast (e.g., coordination vs. subordination), we want to see whether and how the evolution of representation geometry over layers differs for the "hard" and "easy" conditions. To do so, for each dataset-model pair, we first extract the layerwise last-token embeddings in the residual stream (Elhage et al., 2021). We use last-token representations as they are the only ones to attend to the entire sequence, and the ones used to predict the next token. For a dataset of size $N$ and an LLM with hidden dimension $D$, this yields a sequence of representations in $\mathbb{R}^{N\times D}$. We compute the ID and between-contrast representational similarities using these representations.

**Intrinsic dimension** In order to compare LLM processing complexity profiles in different conditions, for each model-dataset-layer combination, we compute the ID of the representations. Among the many ID estimators developed (see Campadelli et al., 2015, for a survey), we choose the TwoNN estimator (Facco et al., 2017), as it employs minimal assumptions (only local data uniformity up to the second nearest neighbor), correlates highly to other estimators (Cheng et al., 2023), and has been widely used in past work estimating the ID of data representations (Valeriani et al., 2023; Chen et al., 2024; Lee et al., 2025). Details are in App. C.1.

**Representational similarity** Because we consider data *manifolds*, we are interested in a notion of *local* representational similarity. We make use of the Information Imbalance ($\Delta$) (Glielmo et al., 2022), a measure of neighborhood divergence between representation spaces. $\Delta$ is *asymmetric*: in general, $\Delta(A \to B) \neq \Delta(B \to A)$. As such, $\Delta$ can be thought of as a *directional* distance between spaces. Intuitively, it quantifies the amount of information about one space (e.g., the space of coordinated sentences) that is also captured in another space (e.g., the space of matched subordinated sentence). Both Cheng et al. (2025) and Acevedo et al. (2025) found $\Delta$ to provide a clearer signal than the more widely used Centered Kernel Alignment measure (Kornblith et al., 2019) in the context of comparing LLM representations. Details are in App.C.2.

**Models** We test six LLMs in sizes ranging from 7 to 14 billion parameters: Gemma-2-9b (Gemma) (Riviere et al., 2024), Llama-3-8B (Llama) (Grattafiori et al., 2024), OLMo-2-13B (OLMo) (Walsh et al., 2025), Mistral-7B-v0.1 (Mistral) (Jiang et al., 2023), Pythia-12B (Pythia) (Biderman et al., 2023), and Qwen-2.5-14B (Qwen) (Yang et al., 2025). These models may differ in specific architectural choices, for instance nonlinear activation or positional embedding, but all consist of decoder layers that each include an attention and feedforward module. Our use of residual stream representations abstracts away from specific architectural choices to allow high-level comparison of the models' layerwise dynamics.

## 5 Experiments

Our observations are always based on inspecting the behavior of all six studied LLMs. However, for space reasons, we display the results for Llama, OLMo and Pythia in the main text, and Gemma, Mistral and Qwen in appendices.

### 5.1 Intrinsic dimension

**Generic sequences** We build on Cheng et al. (2025), who showed that, when fed random naturalistic input from a corpus, LLMs systematically display a profile with (at least) one ID "peak" that marks an area of deep linguistic processing. We replicate the result in App. D.2, extending it moreover to 3 LLMs that were not studied by Cheng and colleagues. In the same appendix we confirm their observation that there is a broad alignment between the ID peaks and top performance phases in three semantic and syntactic probing tasks. In all the following experiments, we thus take the generic-sequence ID-peak span across LLM layers (automatically identified as explained in App. D.2) as a reference point for a phase in which the model is focusing on deeper linguistic processing, as opposed to input reading or output generation.

Before analyzing our contrast-specific datasets, we note that, in absolute terms, the maximum ID of the various datasets is always orders of magnitudes smaller than the ambient dimension, and it largely correlates with the length of the inputs. It is

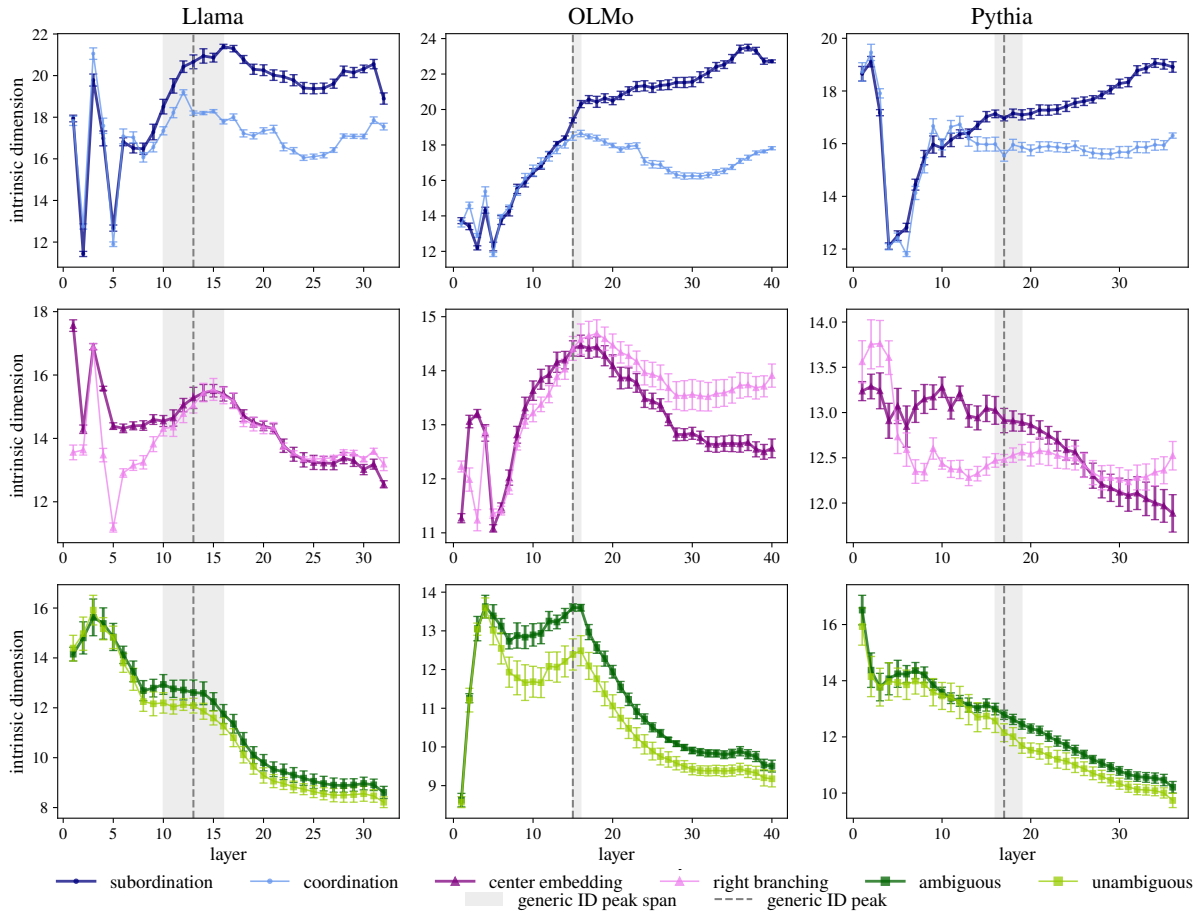


Figure 1: ID profiles through LLM layers (means and error bars across 5 partitions). Vertical dashed line marks maximum ID on generic sequences, and shaded area the corresponding span, estimated as explained in App. D.2.

highest for the 100-word-long generic sequences, next-higher for the 18-word-long 4-clause coordinated/subordinated sentences, and then lower and comparable for the right-branching/center-embedding inputs (8 words) and the unambiguous/ambiguous inputs (13 words on average).[2]

**Coordination vs. subordination** The first rows of Fig. 1 and Fig. 6 (App. D.3) show how ID evolves for this contrast across layers.[3] After some initial fluctuations possibly due to estimation noise, we observe a very consistent pattern across LLMs: the coordination and subordination ID curves first overlap, but they eventually diverge, typically under or just before the generic-ID peak (considerably earlier for Qwen only). From that point onward, ID is clearly higher for the more complex subordination condition than for coordination.

**Right-branching vs. center-embedding** This contrast is shown in the middle rows of Fig. 1 and Fig. 6 (App. D.3). After some initial fluctuation, we observe the more complex center-embedding condition to reach a higher ID than the right-branching one. However, this distinction is present only until the curves reach the generic-ID peak (for Pythia, a bit later). After the generic-ID peak, the two curves merge or, for 4/6 models, there is even a *reversal*, whereby right-branching ID is higher. Intriguingly, right-branching sentences end with a nested clause ("The politicians advised the potters [that were waiting]"), unlike the center-embedding ones ("The potters [that the politicians advised] were waiting"). This suggests that the higher right-branching ID after the generic-ID peak might be related to the presence of a nested clause at the end of the sentence, analogously to what we just saw for subordination (recall that we are measuring the ID of the *last* token, that will be inside the nested

[2]The ID of the the ambiguity pairs is comparable to that of the slightly shorter right-branching/center-embedding pairs, if not lower, probably because the ambiguity datasets contain less data-points than the others (App. B.3). In any case, our analysis focuses on the relative ID *profile* over layers, rather than on absolute ID values.

[3]Figure 7 (App. D.4) shows ID profiles for different numbers of clauses being combined.

clause in right-branching sentences only).

**Unambiguous vs. ambiguous** Results are in the last rows of Fig. 1 and Fig. 6.[4] We observe a general downward trend of ID, that in most cases becomes clearer after the generic-ID peak. While this pattern is robust for the two conditions of this dataset, we do not attribute it directly to the attachment manipulation. Unlike in the previous datasets (coordination vs. subordination and right-branching vs. center-embedding), where sentence structure is tightly controlled and the final token is always a present participle, the sentences in the present dataset display more variation in the final tokens (e.g., they may be nouns, adverbs, or adjectives, see examples in Table 4, App. B.3). Although such factors are controlled within each minimal pair, allowing us to interpret differences in ID profiles between conditions, they are not controlled across pairs. As a result, global properties of a dataset may shape the overall ID trajectory independently of the phenomenon under investigation. In this sense, the shared descending pattern observed at the end of both the ambiguous and unambiguous datasets may reflect dataset-specific characteristics unrelated to relative clause attachment per se.

More directly relevant to our investigation, we observe a general tendency for the ID of the more complex ambiguous condition to dominate the simpler unambiguous condition. In this case, however, the generic-ID-peak phase is not acting as a clear delimiter between different ID patterns, and the difference between the curves tends to be small.

In sum, we found that ID reflects our predictions based on the (psycho)linguistic literature, but in very distinct ways for the three contrasts considered. Formal complexity emerges as the linguistic phenomenon most consistently affecting ID profiles: processing deeply nested structure plausibly requires more complex features, leading to higher ID profiles for subordinated structures. Further, the differentiation between the two sentence types emerges over the layers that coincide with the generic-ID peak, which was proposed as a signature of deeper linguistic processing. It figures that nested structure processing is part of this phase.

ID also responds to long-distance agreement resolution, but in this case the complexity effect is visible earlier, typically before the generic-ID peak. From a strictly structural point of view, center-embedded and right-branching sentences are identical, and their meaning only differs in terms of which element is foregrounded. Thus, the early ID differentiation might be a marker of a lower-level, word-bound kind of processing, linked to matching agreement features between the main subject and its predicate. Intriguingly, for most models we observe higher post-generic-peak ID for right-branching sentences, which we linked to their final nesting, thus relating it to a formal property similar to that of subordinated sentences above.

Finally, yet another ID pattern appears for ambiguity resolution. Ambiguous and unambiguous profiles are similar, but the ambiguous constructions have (slightly) higher ID across the board. This pattern is open to multiple interpretations, but it might suggest a contrast that is not linked to specific phases of linguistic processing, but rather to the instability triggered by ambiguity (e.g., the higher complexity could be triggered by keeping both sentence interpretations open until the end).

### 5.2 Representation similarity across layers

Our claim that ID profiles reflect a genuine difference in LLM processing of formal vs. functional complexity is supported by a follow-up experiment in which we tracked representational similarity between matched subordinated vs. coordinated and right-branching vs. center-embedding sentences.[5]

The top rows of figures 2 (and 9 in App. D.6) show the evolution of the Information Imbalance $\Delta$ across layers for the coordinated and subordinated sentence datasets. $\Delta$ is generally extremely low in both directions, i.e., not surprisingly, the model representations of sentence pairs such as those in Table 2 (App. B.1), that share word order and virtually all lexical materials, are extremely similar. To put this degree of similarity into perspective, according to the theoretical simulations of Acevedo et al. (2025), a $\Delta$ of 0.1 (the largest value on our $y$ axes) already correspond to a 90% of shared features. There is also, however, a clear asymmetric divergence in $\Delta$ profiles for 5/6 models. Just under the generic-ID peak, and thus in the same point where the ID difference between the coordination and subordination sets emerges, $\Delta(coord \rightarrow subord)$ becomes progressively *larger* than $\Delta(subord \rightarrow coord)$, except for Qwen (Fig. 9),

[4] Fig. 8 (App. D.5) shows ID profiles for sentences with relative clauses that have unambiguous high attachment.

[5] The sentences in the ambiguity datasets feature important differences in lexical material, and thus they cannot be naturally paired to measure representational similarity.

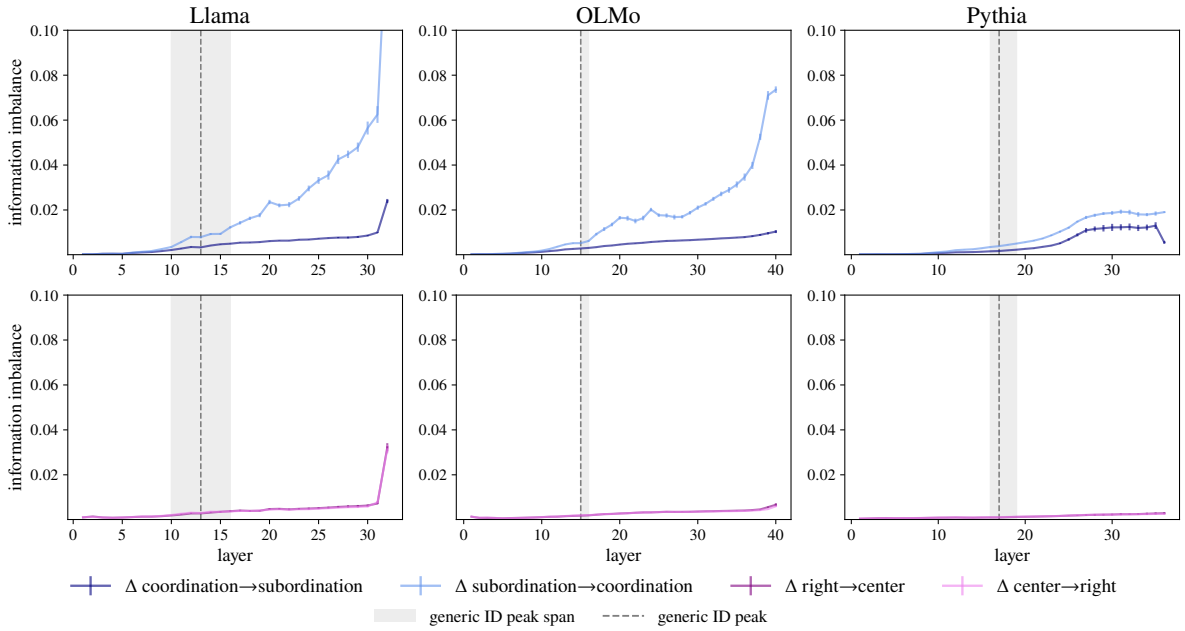


Figure 2: Information Imbalance Δ between coordinated/subordinated sentences (top) and right-branching/center-embedding sentences (bottom): means across 5 partitions with error bars (often invisible). Shaded area marks generic ID-peak span, with a vertical dashed line at the generic-ID maximum. Higher Δ means lower similarity.

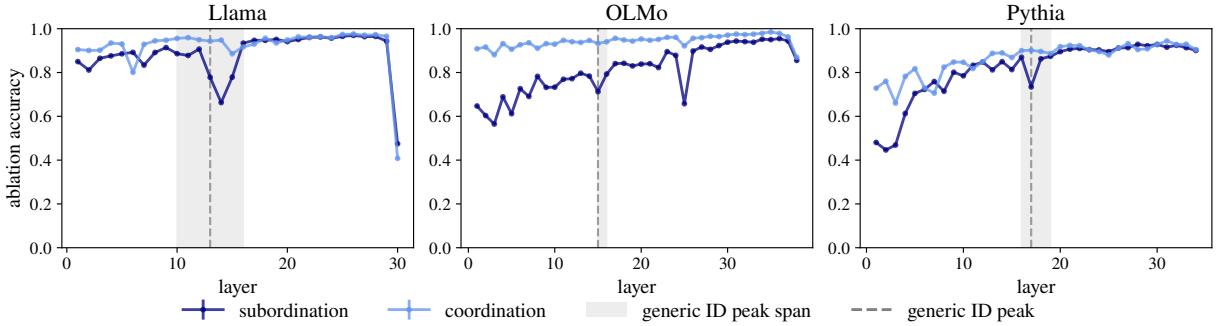


Figure 3: Ablation accuracy (proportion of cases in which ablating a layer did not change next-token prediction) of **coordinated**/**subordinated** sentences. Means and (invisible) standard error bars over 5 partitions.

where the generic-ID peak only marks a phase in which representations diverge, but there is no asymmetry. Thus, in general, the representations of the two sentence types grow apart, and the information contained in the coordinated sentences predict increasingly *less* that in the corresponding subordinated sentences. This makes intuitive sense: if I say that "*The blacksmith is babbling that the politicians are doubting*", I am approximately providing the same information than in the corresponding coordinated sentence ("*The blacksmith is babbling and the politicians are doubting*"), but on top of that I am also specifying *what* it is that the blacksmith is babbling, a piece of information that is not contained in the coordinated sentence.

Looking next at the bottom rows of the figures (center-embedding vs. right-branching), we again observe very strong similarity, decreasing only slightly across the layers. Importantly, in this case there is no asymmetry. This suggests that the LLMs capture the fact that pairs such as "*The potters that the politicians advised were waiting*" and "*The politicians advised the potters that were waiting*" carry the same denotational meaning.

Finally, in both cases, we observe that Δ sometimes spikes on the very last layers. This is possibly due to the fact that, as the models must switch to predicting a concrete next token on these layers, and this is likely different for the compared conditions, they must represent different superficial information at the very end.

In sum, just at the generic-ID-peak benchmark,

the LLMs start building structures that make subordinated and coordinated sentences diverge in representation, with the representations of the more richly structured subordinated sentences becoming more predictive of the simpler coordinated sentences than the reverse. On the other hand, no asymmetry is observed for the center-embedding vs. right-branching contrast, and only a very weak divergence in $\Delta$ across the layers, in line with the strong semantic and structural similarity between the conditions. Again, this suggests a stronger sensitivity of the LLMs to formal vs. functional contrasts in their deeper layers.

### 5.3 Ablation experiments

We finally perform an intervention experiment by ablating LLM layers, looking for a significant effect at the layers that the ID and $\Delta$ tests singled out as crucial for formal and functional processing. In practice, we remove in turn each layer of an LLM, we feed it one of our datasets, and we compute an *ablation accuracy* score by counting the proportion of inputs for which the ablated LLM predicts the same next token as the intact network. Note that, when the ablated-model prediction is different from the original one, there is no guarantee that the latter is better. However, the measure still captures the extent to which the ablation impacts the LLM by changing its behaviour.

Fig. 3 (and Fig. 10 in App. D.7) show ablation profiles for the coordination/subordination conditions. For all models, we observe a tendency for ablation to have a stronger effect on subordinated than coordinated sentences, in line with the other markers of extra-complexity for this condition (the effect is typically clearer in the earlier layers). Strikingly, for 4/6 LLMs there is a clear dip in accuracy during the generic-ID peak phase, bringing further support for the hypothesis that this is a stage in which different LLMs are performing structure-building operations that are especially important for formally complex sentences. Of the two outliers (Fig. 10 in the appendix), Qwen is showing a subordination-specific dip just after the generic-ID peak, whereas in the case of Gemma there is a dip late in the peak, but it's actually stronger for coordination.

Moving to the functional contrasts, figures 11 and 12 (App. D.7) do not show differential ablation effects across conditions (right-branching vs. center-embedding and unambiguous vs. ambiguous), and there are no layer-specific dips, except for strong initial- and final-layer effects.

To sum up, the ablation experiments confirm the special status of formal complexity, and again the crucial role that the layers under the generic-ID peak have in determining how a LLM processes a formally complex input.

## 6 Discussion

We showed that different pre-trained LLMs display similar ID profiles, and that these profiles are signatures of consistent processing stages. We further refined this observation by showing that generic-ID peaks, previously conjectured to be the locus of abstract linguistic processing, correspond to areas where a contrast in *formal* complexity, such as the one between coordinated and subordinated sentences, emerges. Another robust result pertains to contrasts associated with more *functional* types of complexity (right branching vs. center embedding and unambiguous vs. ambiguous relative clause attachment). The latter also have consistent ID signatures, but these patterns i) differ from those dependent on formal complexity and ii) are less clearly related to the generic-ID peak stage.

From an LLM interpretability perspective, our results shed light on the mechanisms behind the consistent ID profiles that are observed across model layers, and contribute to recent debates on the "universality" of LLM representations (Huh et al., 2024), suggesting that the conjectured universality is at least in part due to similarities in the ways in which LLMs organize linguistic processing.

From the point of view of the language sciences, where the precise measurement of linguistic complexity is an open problem (Newmeyer and Preston, 2014), the insights we were able to precisely extract from LLM representations through ID estimation could contribute to the ongoing debate. In particular, our results stress the empirical robustness of the distinction between more formal and more functional types of linguistic complexity, and we intend, in future work, to search for footprints of this distinction in human language processing data.

## Limitations

Due to compute restrictions, we cannot extend our experiments to larger language models than the ones we studied. However, future work should ascertain the extent to which the observed patterns depend on model size, for families of related LLMs, and their emergence during training, for LLMs that make their intermediate checkpoints available.

Our results should also be reproduced in languages other than English, and with more naturalistic data than the one we used here.

Finally, while the ablation experiments provide a proof of concept that there is a causal link between ID peaks and model behavior, the most important aim of future work should be to establish a clear mechanistic explanation of the relation between intrinsic dimension of representations and the linguistic phenomena we are studying.

## References


Santiago Acevedo, Andrea Mascaretti, Riccardo Rende, Matéo Mahaut, Marco Baroni, and Alessandro Laio. 2025. A quantitative analysis of semantic information in deep representations of text and images. https://arxiv.org/abs/2505.17101.

Ananth Agarwal, Jasper Jian, Christopher D Manning, and Shikhar Murty. 2025. Mechanisms vs. outcomes: Probing for syntax fails to explain performance on targeted syntactic evaluations. *arXiv preprint arXiv:2506.16678*.

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of ACL*, pages 7319–7328, Online.

Randall Balestriero, Romain Cosentino, and Sarath Shekkizhar. 2024. Characterizing large language model geometry helps solve toxicity detection and generation. In *Forty-first International Conference on Machine Learning*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of ICML*, pages 2397–2430, Honolulu, HI.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations*.

P. Campadelli, E. Casiraghi, C. Ceruti, and A. Rozza. 2015. Intrinsic dimension estimation: Relevant techniques and a benchmark framework. *Mathematical Problems in Engineering*, 2015:e759567.

Manuel Carreiras and Charles Clifton. 1999. Another word on parsing relative clauses: Eyetracking evidence from spanish and english. *Memory and Cognition*, 27(5):826–833.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L Leavitt, and Naomi Saphra. 2024. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In *The Twelfth International Conference on Learning Representations*.

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

Emily Cheng, Corentin Kervadec, and Marco Baroni. 2023. Bridging information-theoretic and geometric compression in language models. In *Proceedings of EMNLP*, pages 12397–12420, Singapore.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings ACL*, pages 2126–2136, Melbourne, Australia.

Peter W. Culicover. 2014. Constructions, complexity, and word order variation. In *Measuring Grammatical Complexity*. Oxford University Press.

Diego Doimo, Alessandro Serra, Alessio Ansuini, and Alberto Cazzaniga. 2024. The representation landscape of few-shot learning and fine-tuning in large language models. In *Advances in Neural Information Processing Systems*, volume 37, page 18122–18165. Curran Associates, Inc.

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, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. 2017. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta Costa-jussá. 2024. A primer on the inner workings of transformer-based language models. https://arxiv.org/abs/2405.00208.

Richard Futrell and Kyle Mahowald. 2025. How linguistics learned to stop worrying and love the language models. https://arxiv.org/abs/2501.17047.

Edward Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Edward Gibson, Neal J. Pearlmutter, Enriqueta Canseco-Gonzalez, and Gregory Hickok. 1996. Recency preference in the human sentence processing mechanism. *Cognition*, 59:23–59.

Aldo Glielmo, Claudio Zeni, Bingqing Cheng, Gábor Csányi, and Alessandro Laio. 2022. Ranking the information content of distance measures. *PNAS Nexus*, 1(2).

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press, Cambridge, MA.

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, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Dan Roberts. 2025. The unreasonable ineffectiveness of the deeper layers. In *The Thirteenth International Conference on Learning Representations*.

John A. Hawkins. 2014. Major contributions from formal linguistics to the complexity debate. In *Measuring Grammatical Complexity*. Oxford University Press.

Linyang He, Peili Chen, Ercong Nie, Yuanning Li, and Jonathan Brennan. 2024. Decoding probing: Revealing internal linguistic structures in neural language models using minimal pairs. In *Proceedings of LREC-COLING*, pages 4488–4497, Torino, Italy.

John Hewitt and Christopher Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of NAACL*, pages 4129–4138, Minneapolis, MN.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The Platonic representation hypothesis. In *Proceedings of ICML*, pages 20617–20642, Vienna, Austria.

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. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Mary Kathryn Kennedy. 2025. Evidence of hierarchically-complex syntactic structure within bert's word representations. *Society for Computation in Linguistics*, 8(1).

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of ICML*, pages 3519–3529, Long Beach, CA.

Jin Hwa Lee, Thomas Jiralerspong, Lei Yu, Yoshua Bengio, and Emily Cheng. 2025. Geometric signatures of compositionality across a language model's lifetime. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5292–5320, Vienna, Austria. Association for Computational Linguistics.

Roger Levy, Yoon Kim, and Danny Fox. 2025. The science of language in the era of generative AI. https://mit-genai.pubpub.org/pub/ak3evnmm.

Richard L Lewis and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.*, 29(3):375–419.

Michael Li and Nishant Subramani. 2025. Echoes of BERT: Do modern language models rediscover the classical NLP pipeline? https://arxiv.org/abs/2506.02132.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from Deep Learning. *Annual Review of Linguistics*, 7:195–212.

Iuri Macocco, Nora Graichen, Gemma Boleda, and Marco Baroni. 2025. Outlier dimensions favor frequent tokens in language models. https://arxiv.org/abs/2503.21718.

Lise Menn and Cecily Jill Duffield. 2014. Looking for a 'gold standard' to measure language complexity: what psycholinguistics and neurolinguistics can (and cannot) offer to formal linguistics. In *Measuring Grammatical Complexity*. Oxford University Press.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *Proceedings of ICLR Conference Track*, Toulon, France. Published online: https://openreview.net/group?id=ICLR.cc/2017/conference.

Frederick Newmeyer and Laurel Preston, editors. 2014. *Measuring Grammatical Complexity*. Oxford University Press, Oxford, UK.

Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 178 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Pablo Simon, Emmanuel Chemla, Jean-Remi King, and Yair Lakretz. 2025. Probing syntax in large language models: Successes and remaining challenges. In *Proceedings of COLM*, Montreal,

Canada. Published online https://openreview.net/forum?id=nrZysNmJ0n.

Andreas Trotzke and Jan-Wouter Zwart. 2014. The complexity of narrow syntax: Minimalism, representational economy, and simplest merge. In *Measuring Grammatical Complexity*. Oxford University Press.

Eduard Tulchinskii, Kristian Kuznetsov, Kushnareva Laida, Daniil Cherniavskii, Sergey Nikolenko, Evgeny Burnaev, Serguei Barannikov, and Irina Piontkovskaya. 2023. Intrinsic dimension estimation for robust detection of AI-generated texts. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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

Evan Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, and 23 others. 2025. 2 OLMo 2 furious (COLM's version). In *Second Conference on Language Modeling*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Fan Yin, Jayanth Srinivasa, and Kai-Wei Chang. 2024. Characterizing truthfulness in large language model generations with local intrinsic dimension. In *Proceedings of ICML*, pages 57069–57084, Vienna, Austria.

Zhong Zhang, Bang Liu, and Junming Shao. 2023. Fine-tuning happens in tiny subspaces: Exploring intrinsic task-specific subspaces of pre-trained language models. In *Proceedings of ACL*, pages 1701–1713, Toronto, Canada.

## A Data, code and compute

**Data** We are releasing our datasets on github under a Creative Commons Attribution 4.0 International license. To preserve anonymity, we will only make the github link available upon publication, but we uploaded an archive containing the full datasets with the submission.

**Code** To compute intrinsic dimensionality and information imbalance, we used the DadaPy toolkit. Representation extraction was accomplished by trivially adapting the code made available by Cheng et al. (2025). Surprisal was computed using the Python `surprisal` package. In all cases, code was run with default parameters.

URLs and licenses of the used assets are provided in the following list:

**Gemma** https://huggingface.co/google/gemma-2-9b; license: gemma

**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-2-1124-13Bhttps; license: apache-2.0

**Pythia** https://huggingface.co/EleutherAI/pythia-12b-deduped; license: apache-2.0

**Qwen** https://huggingface.co/Qwen/Qwen2.5-14B; apache-2.0

**Cheng et al.'s code** https://github.com/chengemily1/id-llm-abstraction; license: MIT

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

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

**surprisal** https://github.com/aalok-sathe/surprisal; license: MIT

**Compute** Representation extraction took a few wall-clock hours for each LLM, on 1 or 2 NVIDIA A30 GPUs. The most time consuming extraction step was to run the ablation experiments, as different sets of representations had to be extracted after removing one layer at a time from each of the models, resulting in about two days of running per model. Probing experiments took between 1 and 2 days per task, running on CPUs. Compute time for the other experiments was negligible.

# B Datasets

## B.1 Coordination vs. subordination

Using OpenAI's GPT-4 followed by manual editing, we first built: i) a list of 17 propositional verbs that could also be used as intransitives, such as *babbling* and *dreaming*; ii) a list of 65 pure intransitive verbs taking human subjects (such as *shivering* and *sleeping*); and iii) a list of 74 nouns, divided between 30 proper nouns (*Michael, Sarah*) and 44 profession names (*gardener, programmer*). We randomly sampled elements from these lists to generate 50k 4-clause subordinated and coordinated sentence pairs that matched the following template:

```
NP1 PROPVERB1 CONJ NP2 PROPVERB2 CONJ
NP3 PROPVERB3 CONJ NP4 INTVERB
```

where the NPs are either proper names or noun phrases formed by *the* followed by a profession noun (randomly in singular or plural form); the PROPVERBs are propositional verbs in the present continuous (in singular or plural form in agreement with the subject); the INTVERB is a pure-intransitive verb, also in the present continuous and agreeing with its subject; and CONJ is either *that* or *and*. The words are sampled so that no noun or verb is repeated (independently of number). In order to maintain uniqueness when removing the clauses in the middle (see below), we also ensured that the tuples formed by <NP1, PROPVERB1, NP4, INTVERB> and <NP1, PROPVERB1, NP2, PROPVERB2> were unique. Five random examples from the dataset constructed in this way are shown in Table 2

In order to build the shorter 3- and 2-clause sentences analyzed in App. D.4, we simply removed one or two clauses, respectively, from the middle of the 4-clause sentences. We split the datasests into partitions of 10k sentences, and repeat all experiments 5 times, always reporting means and standard errors across the partitions.

## B.2 Right-branching vs. center-embedding

We used the same nouns and intransitive verbs as for the coordination/subordination datasets (see App. B.1). By again querying ChatGPT 4 and manually editing its outputs, we created a list of 100 transitive verbs that take human subjects and objects (*betray*, *fascinate*, *scold*...). We constructed a set of 50K matched center-embedding and right-branching sentences by randomly sampling from these lists according to the following templates:

```
CENTER EMBEDDING:
NP1 that NP2 TRVERB INTVERB
RIGHT BRANCHING:
NP2 TRVERB NP1 that INTVERB
```

where NP1 is formed by *the* followed by a profession noun in plural or singular form; NP2 is either formed in the same way or it is a proper noun; TRVERB is a transitive verb in past-simple form; and INTVERB is a past continuous form of an intransitive verb agreeing in number with NP1. Again, no sentence contains a repeated noun or verb. Five random examples are given in Table 3.

We split the datasests into partitions of 10k sentences, and repeat all experiments 5 times, always reporting means and standard errors across the partitions.

## B.3 Unambiguous vs. ambiguous

This dataset contains three attachment conditions: an ambiguous condition, in which both noun phrases are equally plausible relative-clause (RC) antecedents, and two unambiguous conditions, in which semantic biases strongly favor attachment to either NP1 (high attachment) or NP2 (low attachment). All sentences conform to the following template:

```
NP1 of NP2 who RC CONTINUATION
```

In the main text, we focus on the contrast between ambiguous and low-attachment sentences because English favors low attachment, and, thus, this condition displays the expected and least demanding interpretation, making it a natural baseline. Results for the entire datasets are in App D.5

Using ChatGPT 4, we generated a list of NPs, relative clauses, and continuations that would encode four types of semantic bias: age, gender, role, and logical contradiction. Bias-specific elements were generated using controlled prompting and subsequently filtered, ranked, and manually validated to ensure semantic clarity, grammaticality, and compatibility across conditions. To construct the final dataset, validated elements were recombined under strict constraints that prevent excessive lexical overlap across sentences, while preserving bias consistency. This procedure generated a total of 10,880 attachment triplets (32,640 sentences), where each triplet shares the same relative clause and continuation. Examples for gender- and age-based semantic disambiguations are given in Table 4.

The sentences were split into 5 equal partitions of 2176 items each. We repeat all experiments 5

| |
|---|
| Quinn is rejoicing and/that the surgeon is doubting and/that Mary is screaming and/that the driver is faltering |
| The doctors are muttering and/that the firefighter is babbling and/that Bill is complaining and/that the consultants are hesitating |
| The engineers are singing and/that Jordan is dreaming and/that the tutors are rejoicing and/that the soldier is sliding |
| The artist is remembering and/that Matthew is doubting and/that the judge is writing and/that Taylor is trembling |
| Emily is complaining and/that Casey is mumbling and/that the manager is writing and/that the blacksmith is shivering |

Table 2: Examples from the 4-clause coordination/**subordination** datasets. The 3-clause sentence derived from the first example in this table is: "Quinn is rejoicing and/that the surgeon is doubting and/that the driver is faltering". The 2-clause sentence from the same example is: "Quinn is rejoicing and/that the driver is faltering"

| *Right-branching* | *Center-embedding* |
|---|---|
| Sarah intimidated the potters that were frowning | The potters that Sarah intimidated were frowning |
| James harassed the veterinarians that were sulking | The veterinarians that James harassed were sulking |
| Bill excluded the driver that was escaping | The driver that Bill excluded was escaping |
| Elizabeth praised the foresters that were chuckling | The foresters that Elizabeth praised were chuckling |
| The gardeners applauded the blacksmith that was hurrying | The blacksmith that the gardeners applauded was hurrying |

Table 3: Examples from the right-branching/**center-embedding** datasets.

| *Age bias* | *Gender bias* |
|---|---|
| **Ambiguous**: The neighbor of the grandpa who paid a mortgage stood nearby. | **Ambiguous**: The sister of the heiress who was menstruating cooked rice. |
| **Low attachment**: The child of the comrade who paid a mortgage stood nearby. | **Low attachment**: The uncle of the maiden who was menstruating cooked rice. |
| **High attachment**: The uncle of the child who paid a mortgage stood nearby. | **High attachment**: The maiden of the uncle who was menstruating cooked rice. |

Table 4: Examples from relative-clause unambiguous/**ambiguous** attachment datasets. Low attachment is the unambiguous condition used in the main-text experiments.

times, always reporting means and standard errors across the partitions.

# C Methods

## C.1 TwoNN estimation

Assume points on the underlying manifold are distributed as a locally homogeneous Poisson point process, where "locally" means up to the second nearest neighbor of each point. Let $\delta_k^{(i)}$ be the Euclidean distance between the point $x_i$ and its $k$th nearest neighbor. Then, the distance ratios $\mu_i := \delta_2^{(i)}/\delta_2^{(i)} \in [1, \infty)$ have the cumulative distribution function $F(\mu) = (1-\mu)^d \mathbf{1}[\mu \geq 1]$. The ID, given by $d$, is estimated as $d = -\log(1 - F(\mu))/\log \mu$ via maximum likelihood estimation over all data points.

## C.2 Information Imbalance

Consider two representation spaces $A$ and $B$ consisting respectively of paired data points $\{x_i\}_{i=1}^N$ and $\{y_i\}_{i=1}^N$. Let $r_{ij}^X$ refer to the neighbor-rank of point $j$ to point $i$ in space $X$, for instance, if $x_j$ is $x_i$'s first nearest neighbor in space $X$, then $r_{ij}^X = 1$. Then, $\Delta(A \to B)$ is given by (Glielmo et al., 2022)

$$\Delta(A \to B) := \frac{2}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} r_{ij}^B \mathbf{1}[r_{ij}^A = 1]. \quad (1)$$

If nearest neighbors in $A$ are also nearest neighbors in $B$, then $\Delta(A \to B) \approx 0$. If nearest neighbors in $A$ have a uniformly-distributed neighbor rank in $B$, *i.e.*, neighbors in $A$ are uninformative about neighbors in $B$, and $\Delta(A \to B)$ is near 1.

# D Additional results

## D.1 Surprisal

For each dataset, we sampled 1,000 sentences at random to estimate the mean per-token surprisal under each LLM. We compute the mean surprisal per sentence, then average over all 1,000 sentences to obtain one surprisal value per-sentence. For each model-dataset combination, the distribution of surprisals over the 1,000 sentences was approximately normal, according to a Shapiro-Wilk test with a conservative $p$-value cutoff of $\alpha = 0.1$. We then performed a one-sided difference of means t-test between the **more** and less complex conditions, finding in all cases that the LLM has higher surprisal on the **more** complex condition, significant at $\alpha = 0.05$. Results are displayed in Table 5.

## D.2 Intrinsic dimension of generic sequences

**ID profiles** We reproduce the finding by Valeriani et al. (2023) and Cheng et al. (2025) that, on generic in-distribution data, a peak in the intrinsic dimension of LLM activations emerges in the intermediate layers. To do so, we use a sample of 50k sequences from the Wikitext corpus (Merity et al., 2017) repurposed from Macocco et al. (2025). Each sequence consists of 100 words and begins with a start of sentence but is not constrained to end with a period, so that the final word can have any part of speech. Then, for each model, on five random non-overlapping data splits of 10k sequences each, we compute the ID using the TwoNN estimator on the last-token representation. This produces an ID profile across layers for each model, shown in Fig. 4. Finally, for each model, following Cheng et al. (2025), we heuristically demarcate the first ID peak-span (gray in the figures), locating the nearest inflection points around the maximum using second-order finite differences. Note that, with respect to Cheng and colleagues, we extend these experiments to Gemma, Qwen and a newer and larger version of OLMo, thus further confirming the generality of their observation.

Fig. 4 first of all confirms that the intrinsic dimension is always orders of magnitude smaller than the ambient ones (which is always $> 3.5K$). Importantly, all models have an intrinsic dimensionality peak in their mid layers, followed in some cases (OLMo, Pythia, Qwen) by another late-layer peak, a phenomenon also reported by Cheng and colleagues.

**Higher-order linguistic processing around the ID peak** Cheng et al. (2025) found that the ID peak span coincides with a phase in which higher-order linguistic information is made available, as indicated by a set of MLP-based probing tasks and downstream tests. We replicate their probing results for all tested models, using syntactic and semantic tasks from Conneau et al. (2018) (bigram shift, coordination inversion, and odd man out). While our results (in Fig. 5) are somewhat noisy, we confirm that performance on these tasks tends to grow to its maximum within the ID peak span. This confirms that the generic-sequence ID peak span provides a useful rule-of-thumb for the locus of deeper syntactic/semantic processing in the LLMs.

| Model | subordination/coordination | | center embed/right branch | | ambiguity | | |
|---|---|---|---|---|---|---|---|
| | coord. | **subord.** | right | **center** | low | high | **ambiguous** |
| Llama | $5.39_{0.01}$ | $\underline{5.65}_{0.01}$ | $7.74_{0.02}$ | $\underline{7.98}_{0.02}$ | $6.07_{0.02}$ | $6.08_{0.02}$ | $\underline{6.11}_{0.02}$ |
| OLMo | $4.87_{0.01}$ | $\underline{5.11}_{0.02}$ | $6.51_{0.02}$ | $\underline{6.75}_{0.02}$ | $5.10_{0.02}$ | $5.16_{0.02}$ | $\underline{5.18}_{0.02}$ |
| Pythia | $4.77_{0.01}$ | $\underline{5.14}_{0.01}$ | $6.66_{0.02}$ | $\underline{7.06}_{0.02}$ | $5.37_{0.02}$ | $5.39_{0.02}$ | $\underline{5.39}_{0.02}$ |
| Gemma | $6.86_{0.02}$ | $\underline{7.62}_{0.02}$ | $10.35_{0.04}$ | $\underline{10.58}_{0.04}$ | $6.79_{0.03}$ | $6.75_{0.03}$ | $\underline{6.90}_{0.03}$ |
| Mistral | $4.87_{0.01}$ | $\underline{5.22}_{0.02}$ | $6.86_{0.02}$ | $\underline{7.00}_{0.02}$ | $5.50_{0.02}$ | $5.52_{0.02}$ | $\underline{5.58}_{0.02}$ |
| Qwen | $4.77_{0.01}$ | $\underline{5.05}_{0.01}$ | $6.44_{0.02}$ | $\underline{6.61}_{0.02}$ | $5.13_{0.02}$ | $5.15_{0.02}$ | $\underline{5.17}_{0.02}$ |

Table 5: Average surprisal of contrast datasets. The mean surprisal per-token (nats) $\pm 1$ SE is shown for the six LLM rows, for each linguistic phenomenon column and contrast condition subcolumns. Each value is computed from a 1,000-sentence sample, drawn randomly from each contrast condition. For each dataset and model, the highest surprisal value across conditions is underlined. The tested LLMs consistently match expectations from the literature, where **the more (psycho)linguistically complex condition has higher mean surprisal in every case** and statistical significance was determined by a one-sided t-test ($\alpha = 0.05$). In the **subordination/coordination** datasets, the *subordination* condition, given by a more nested syntactic structure, has higher surprisal than the *coordination* condition. For the **center/right branching** datasets, the *center embedding* condition has higher surprisal than the *right branching* case. Finally, for the **ambiguity** datasets, the *ambiguous attachment* condition has higher surprisal under an LLM than the unambiguous high and low attachment cases.

### D.3 Intrinsic dimension profiles of the other models

ID profiles for the 3 models not shown in the main text (Gemma, Mistral and Qwen) are reported in Fig. 6.

### D.4 Intrinsic dimension profiles for the three levels of coordination and subordination

In the main text, we present results for coordinated/subordinated sentences that contain 4 clauses. Fig. 7 compares these results to those obtained when using 3-clause or 2-clause sentences (e.g., respectively: "Quinn is rejoicing and/that Mary is screaming and/that the driver is faltering" and "Quinn is rejoicing and/that the driver is faltering"). We observe first of all that IDs, unsurprisingly, are higher for longer sentences. More importantly, the distinction between subordinated and coordinated sentences is consistent and strong in all models for the longest 4-clause sentences. It is present for most models also for 3-clause sentences, although typically with weaker separation. It is absent of very moderate for the shortest 2-clause sentences. It thus seems that the structural differences between flatter coordinated sentences and more hierarchical subordinated ones becomes salient enough to leave a recognizable trace in the ID profiles only above a certain degree of nesting.

### D.5 Intrinsic dimension profiles for the three relative clause attachment conditions

In the main text and App. D.3, we compared the ambiguous attachment condition to relative clauses with unambiguously low attachment. Fig. 8 shows these two conditions together with a third, unambiguous *high* attachment condition. We see that there are only small and non-systematic differences between the two unambiguous attachment conditions.

### D.6 Information Imbalance profiles of the other models

Information Imbalance ($\Delta$) profiles for the 3 models not shown in the main text (Gemma, Mistral and Qwen) are reported in Fig. 9.

### D.7 Ablation profiles of the other models and conditions

Ablation results with the coordinated/subordinated datasets for the 3 models not shown in the main text (Gemma, Mistral and Qwen) are reported in Fig. 10. The different responses to ablations are more visible in early layers and a dip in accuracy specifically affecting the coordinated set is found in correspondence to the ID peak in Mistral, similar to Llama, OLMo and Pythia.

Ablation results for the other contrass are reported in figures 11 and 12. Differences across conditions are scarcely observable in these tasks, differently that for the coordination/subordination dataset.

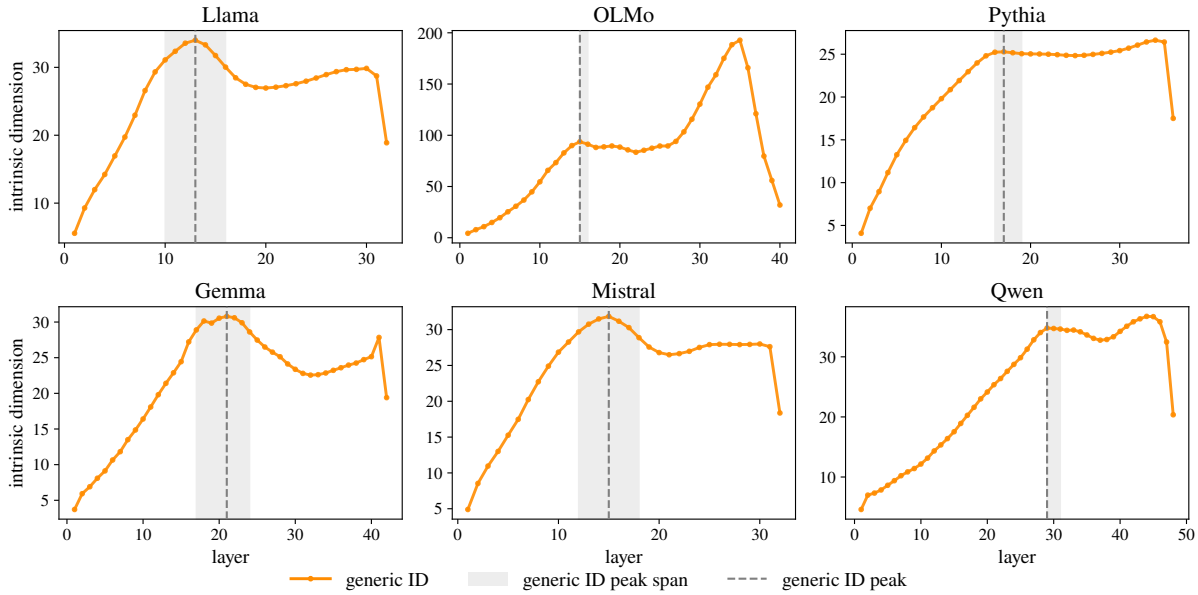


Figure 4: ID values across layers and estimated peak spans for all LLMs given an input of naturalistic corpus sequences. Mean values with (invisible) error bars.

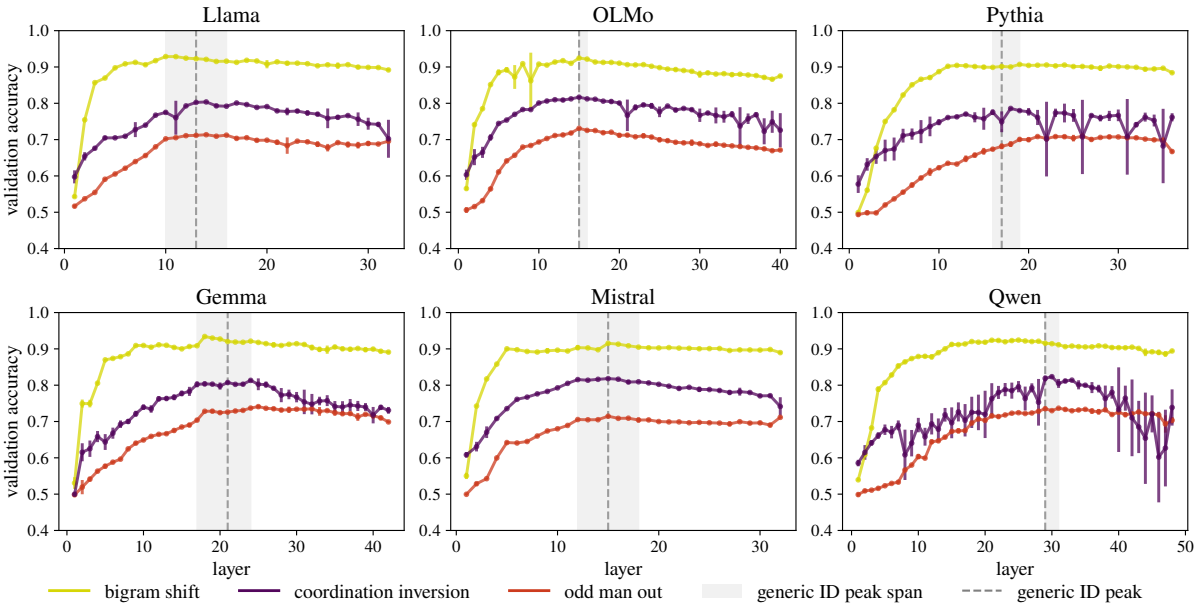


Figure 5: Accuracies in three probing tasks (bigram shift, coordination inversion and odd man out) across layers, for all models. Means and standard errors across five seeds.

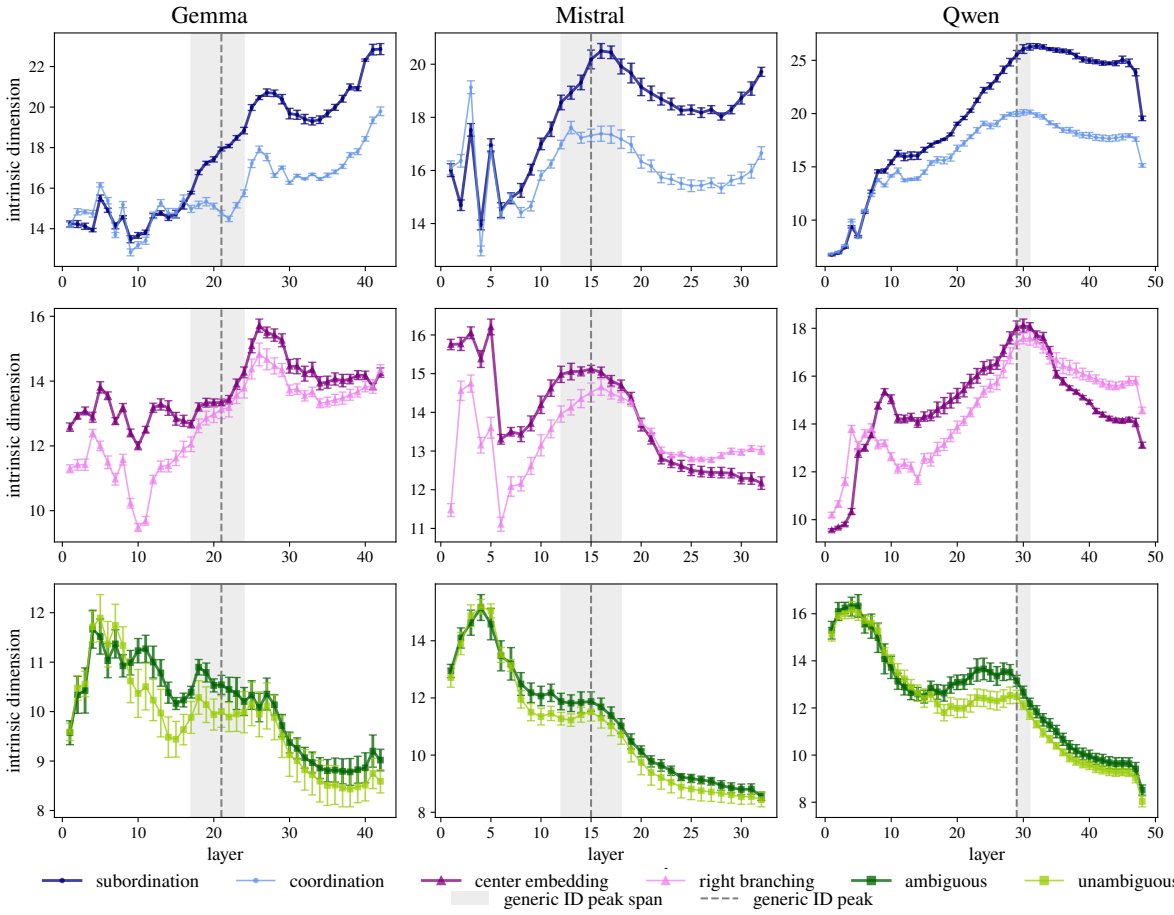


Figure 6: ID profiles through LLM layers (means and error bars across 5 partitions). Vertical dashed line marks maximum ID on generic sequences, and shaded area the corresponding span, estimated as explained in App. D.2.

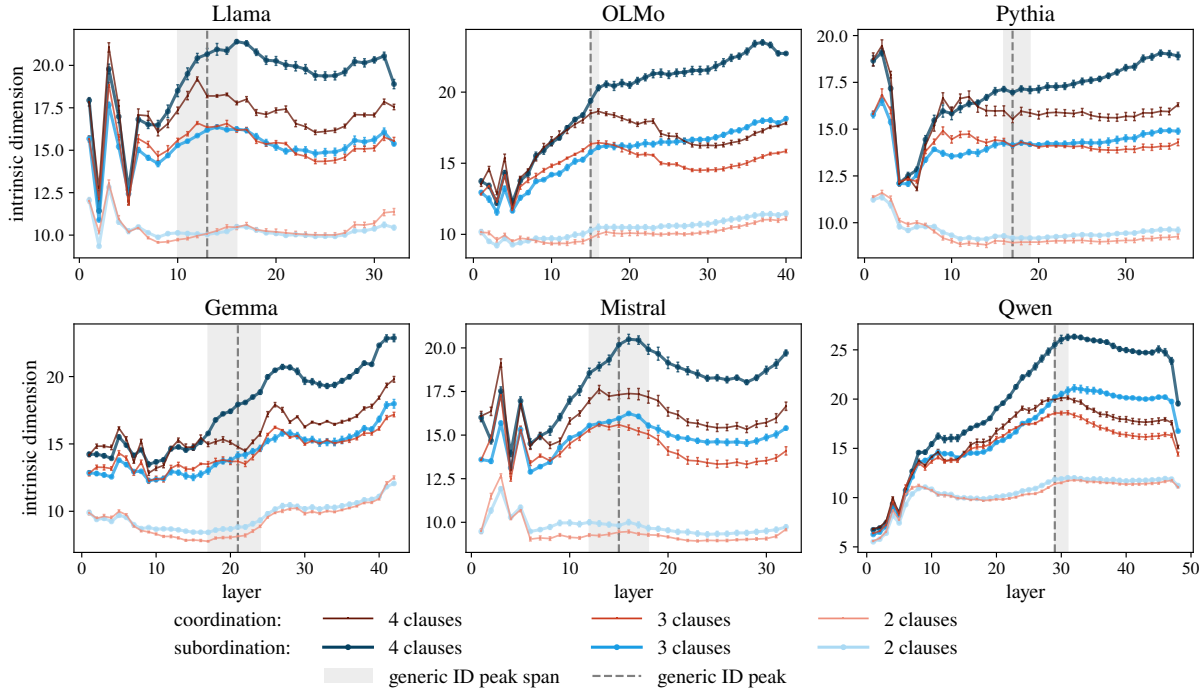


Figure 7: ID profiles through LLM layers (means and error bars across 5 partitions) for coordination/subordination sets featuring different numbers of clauses. Vertical dashed line marks maximum ID on generic sequences, and shaded area the corresponding span.

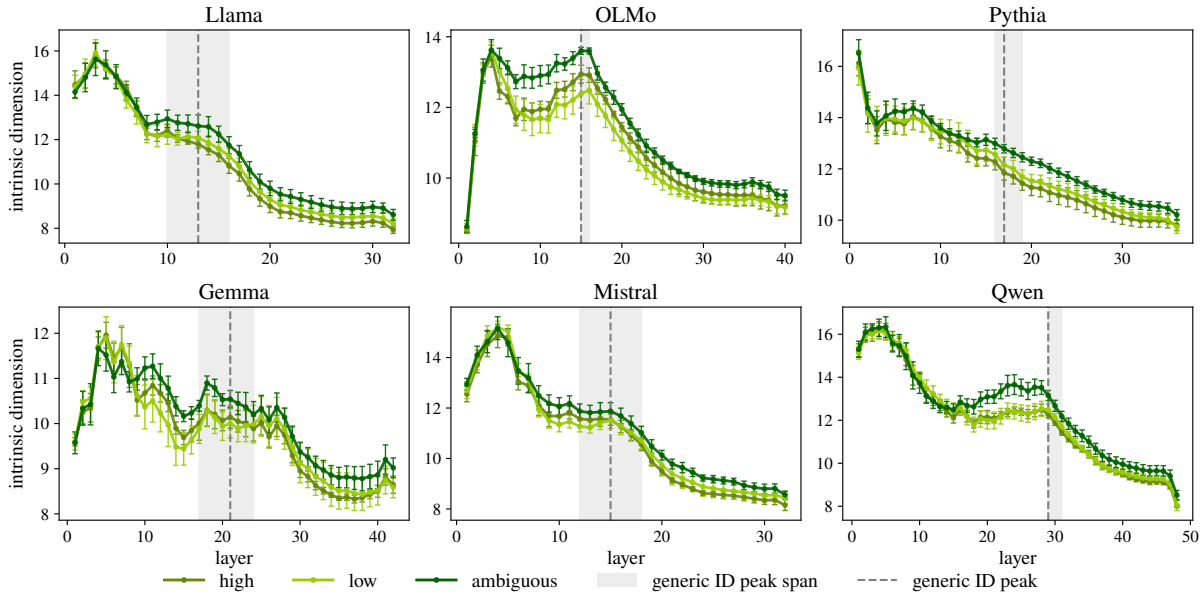


Figure 8: ID profiles through LLM layers (means and error bars across 5 partitions) for the high, low and ambiguous relative clause attachment conditions. Note that low attachment corresponds to the unambiguous condition shown in the last rows of figures 1 and 6. Vertical dashed line marks maximum ID on generic sequences, and shaded area the corresponding span.

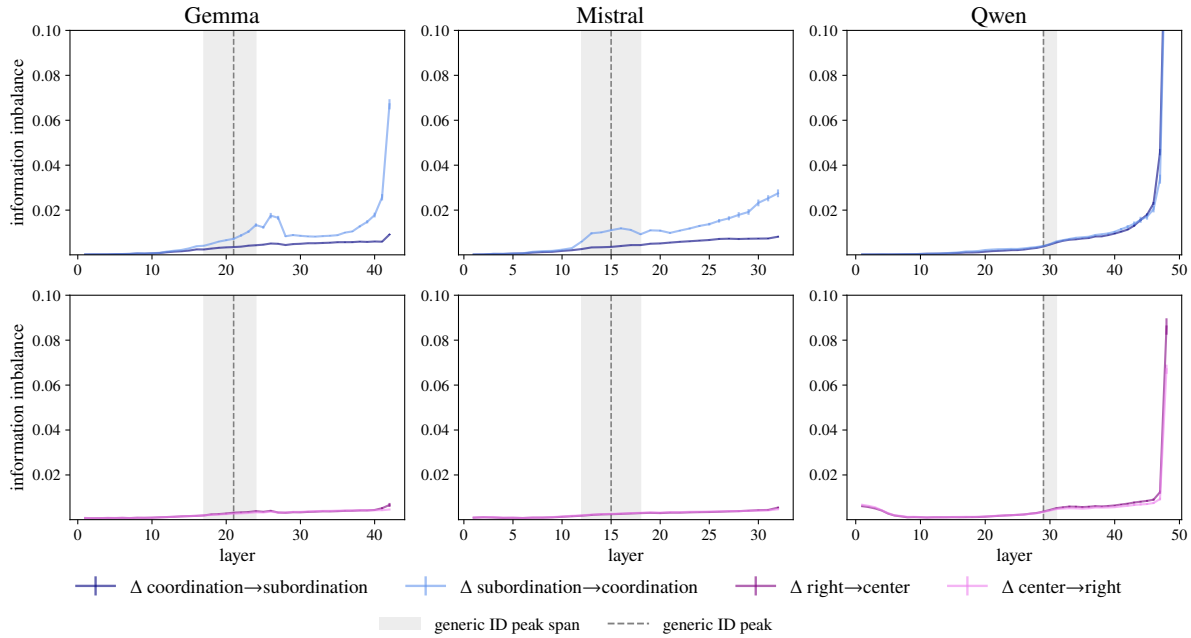


Figure 9: Information Imbalance Δ between coordinated/subordinated sentences (top) and right-branching/center-embedding sentences (bottom): means across 5 partitions with error bars (often invisible), for the Gemma, Mistral and Qwen models. Shaded area marks generic ID-peak span, with a vertical dashed line at the generic-ID maximum. Higher Δ means lower similarity.

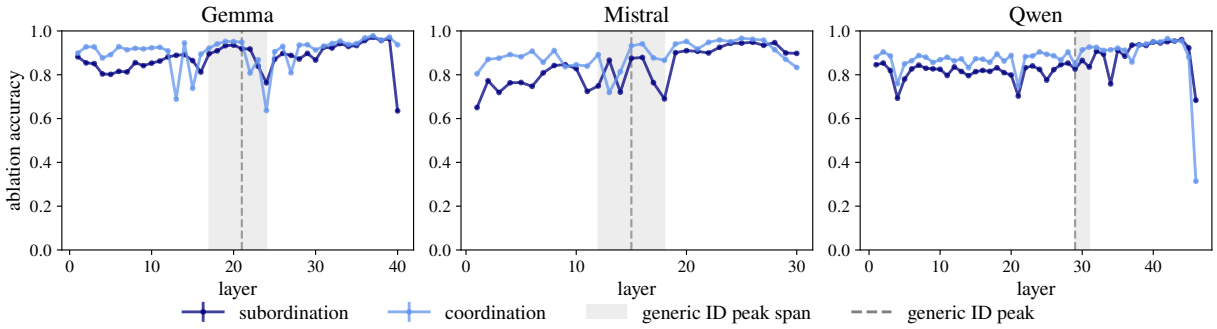


Figure 10: Ablation accuracies comparing the coordination vs. subordination conditions for Gemma, Mistral and Qwen. Means and (invisible) standard error bars over 5 partitions.

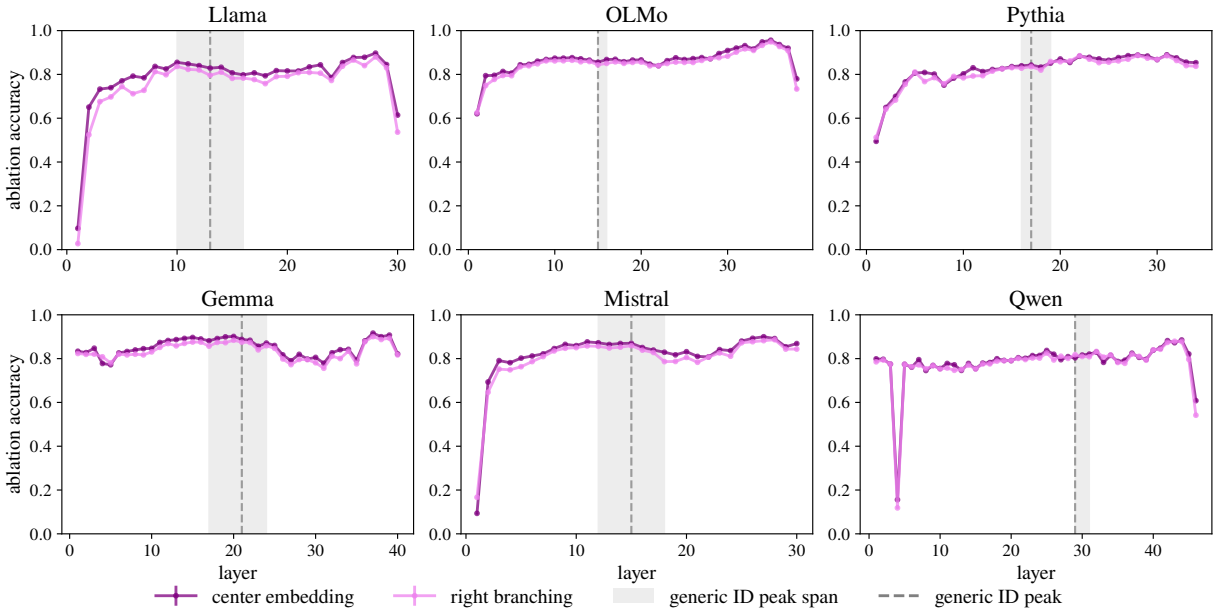


Figure 11: Ablation accuracy for right branching vs. center embedding. Means and standard errors over 5 partitions

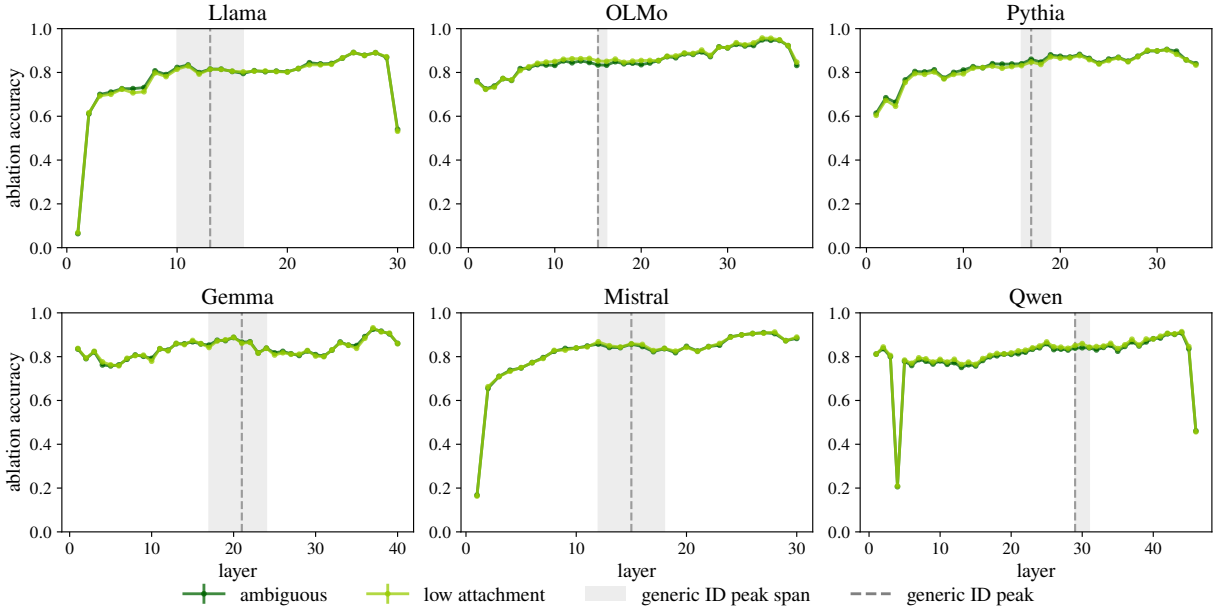


Figure 12: Ablation accuracy for the ambiguous vs. unambiguous conditions. Means and standard errors over 5 partitions.