Probing Syntax in Large Language Models: Successes and Remaining Challenges

Pablo J. Diego-Simón[*1] **Emmanuel Chemla**[1,2] **Jean-Rémi King**[3] **Yair Lakretz**[1]

[1]LSCP, ENS, PSL, EHESS, CNRS, Paris, France
[2]Earth Species Project, Berkeley, California, USA
[3]Meta AI, Paris, France

Abstract

The syntactic structures of sentences can be readily read-out from the activations of large language models (LLMs). However, the "structural probes" that have been developed to reveal this phenomenon are typically evaluated on an indiscriminate set of sentences. Consequently, it remains unclear whether structural and/or statistical factors systematically affect these syntactic representations. To address this issue, we conduct an in-depth analysis of structural probes on three controlled benchmarks. Our results are fourfold. First, structural probes are biased by a superficial property: the closer two words are in a sentence, the more likely structural probes will consider them as syntactically linked. Second, structural probes are challenged by linguistic properties: they poorly represent deep syntactic structures, and get interfered by interacting nouns or ungrammatical verb forms. Third, structural probes do not appear to be affected by the LLMs' predictability of individual words. Fourth, despite these challenges, structural probes still reveal syntactic links far more accurately than the linear baseline or the LLMs' raw activation spaces. Taken together, this work sheds light on both the challenges and the successes of current structural probes and provides a benchmark made of controlled stimuli to better evaluate their performance.

1 Introduction

The autonomy of syntax. Understanding how word sequences are combined to form the meaning of a sentence is a central challenge to linguistics (Chomsky, 1957; 1995). Behavioral (Rayner, 1978; Frazier & Rayner, 1987; Gleitman, 1990) and neuroimaging research (Pallier et al., 2011; Dehaene & Cohen, 2011; Devauchelle et al., 2009; Caucheteux et al., 2021; Friederici, 2018; Santi & Grodzinsky, 2007) show that this phenomenon requires the brain to build latent structures that link the words of sentences. Critically, such "syntactic" system is thought to be *autonomous*: the infamous sentence "colorless green ideas sleep furiously" shows that words can be syntactically linked even if each word is highly unpredictable, and thus lead to a nonsensical sentence Croft (1995); Chomsky (1957). Yet, the neural and computational implementations of such *autonomous* syntactic structures remain largely unresolved (Grodzinsky & Friederici, 2006; Kaan & Swaab, 2002).

Syntactic representations in LLMs. Large Language Models (LLMs) now offer an unprecedented opportunity to revisit this historical question. Not only can these models generate coherent sentences, but their internal activations explicitly represent the syntactic structures long theorized by linguists (Belinkov et al., 2017; Tenney et al., 2019; Peters et al., 2018; Hale & Stanojević, 2024; Hewitt & Manning, 2019; Diego-Simon et al., 2024). Following earlier work on linear probing (Alain & Bengio, 2017; Conneau et al., 2018), the 'Structural Probe' (Hewitt & Manning, 2019) recently showed that LLMs spontaneously learn to build a subspace of activations, where the distance between two contextualized word embedding corresponds to the distance that separates two words in the syntactic tree. Diego-Simon et al. (2024) further showed that the angle between these two words can represent the type of syntactic relations (e.g. 'subject of', 'object of').

*Correspondance to: pablo.diego-simon@psl.eu

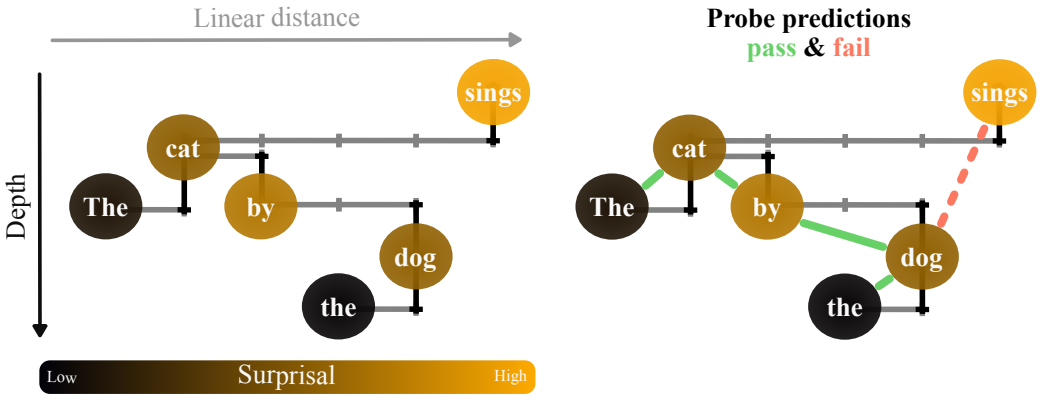


Figure 1: **Dependencies form a tree structure linking words in sentences, which the Structural Probe aims to predict.** Each dependency connects two words that are separated by a certain *linear distance*. Since the dependency tree is rooted, one word in the dependency serves as the head and the other as the child, introducing the notion of *syntactic depth*. Beyond these structural features, the LLM assigns each word a *surprisal* value based on its downstream prediction. We investigate how linear distance, depth, and surprisal influence probe predictions.

Challenge. However, the performance of such Syntactic Probes has *mostly* been assessed with aggregated scores over large naturalistic corpora (Guarasci et al., 2021). Consequently, it is unclear whether the syntactic representations revealed by these LLM probes identify abstract representations of syntax (Croft, 1995; Chomsky, 1957), or whether they rather rely on other, heuristic or statistical factors.

Approach. To address this question, we evaluate the performance of two state-of-the-art syntactic probes (Hewitt & Manning, 2019; Diego-Simon et al., 2024) under a variety of linguistic conditions. To this end, we leverage both naturalistic datasets and controlled stimuli designed to manipulate specific linguistic (linear distance, depth, number interference) and statistical (word predictability) features. With these tests, we aim to (1) characterize the strengths and limitations of structural probes (2) compare them with human behavior, and ultimately (3) provide a benchmark to better evaluate structural probes.

2 Related work

Understanding syntax with linear structural probes has been an extensive area of research initiated with the seminal work of Hewitt & Manning (2019). These probes have been extended using spectral (Müller-Eberstein et al., 2022), polar (Diego-Simon et al., 2024), and orthogonal (Limisiewicz & Mareček, 2021) transformations providing insights about the underlying vectorial spaces hosting syntactic processing (Reif et al., 2019).

This suite of probes rests on the premise that linguistic structure is linearly readable from LLM representations. Recent work suggests that linear representations also encode belief states, both across pre-training (Shai et al., 2024) and during in-context learning (Park et al., 2025).

In parallel, efforts have been made to develop non-linear probes, which relax the linear constraint, offering a more flexible view of syntactic encoding (Eisape et al., 2022; White et al., 2021; Chen et al., 2021).

Probing methods, however, are subject to limitations and potential artifacts. Some have shown evidence that probe performance can be misleading, not generalizing to Jabberwocky sentences (Maudslay & Cotterell, 2021) or having marginal effects on information-theoretical scores (Pimentel et al., 2020). Revealing that, probes may exploit shallow heuristics, making controlled and linguistically motivated evaluations essential for uncovering deeper insights.

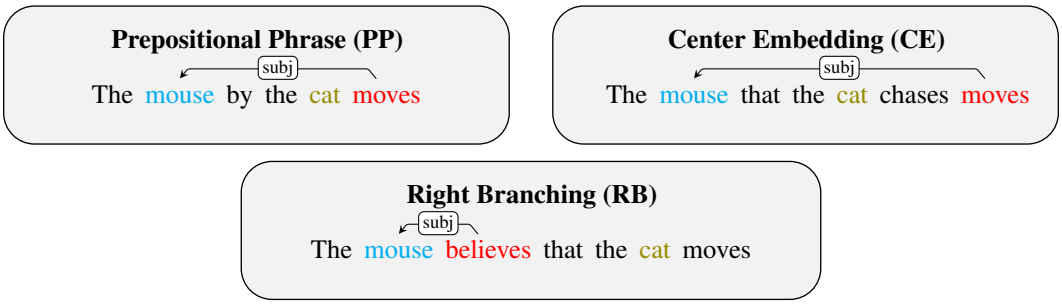
3 Methods

3.1 Data

Naturalistic sentences The Structural (Hewitt & Manning, 2019) and Polar (Diego-Simon et al., 2024) Probes were trained on the Universal Dependencies English Web Treebank (UD-EWT) corpus [*] (Silveira et al., 2014). UD-EWT contains 16,622 sentences. The corpus consists of 254,820 words and 16,622 sentences, extracted from online journals, blogs and informative websites. Every sentence in the corpus has been manually annotated following the Universal Dependencies (UD) formalism (Nivre et al., 2017).

Sentences that include email addresses or web URLs have been excluded from the dataset, as they are considered noisy and irrelevant for syntactic analysis. Also, sentences for which UD tokens and LLM tokens cannot be aligned are also excluded. We adhere to the default split provided by UD-EWT, resulting in 11,827 sentences for training, 1,851 for validation, and 1,869 for testing.

Controlled sentences We generated 3 controlled datasets one for each of the following syntactic structures:



In these syntactic structures, the attractor interferes with the subject of the sentence. Furthermore, both PP and CE structures introduce long-range subject-verb dependencies, in which the subject is separated from the main verb. Each dataset contains sentences with varying degrees of syntactic nesting, ranging from 1 to 3 levels. Adding a syntactic nesting modifies the syntactic depth of the sentence and introduces an attractor noun. For the PP and the CE case, adding more nestings implies modifying the linear distance between the head and the dependent in the subject-verb dependency. For the 2 nesting case each structure becomes:

- *The mouse by the cat beside the fox moves*
- *The mouse that the cat that the fox protects chases moves*
- *The cat believes that the fox expects that the mouse moves*

The controlled dataset is generated using a generative grammar tailored to each target syntactic structure: PP, CE, and RB. The number of nestings for each sentence is uniformly sampled from $\{1, 2, 3\}$, introducing varying levels of syntactic complexity. Three nestings are already unusual in natural language and difficult for humans to understand. A shared vocabulary is used across all structures, consisting of five lexical categories: verb, noun, transitive verb, preposition, and determiner. Lexical items are sampled from these categories for each syntactic structure, with replacement allowed only for the determiner category. To maintain dataset integrity, we ensure that all generated sentences are unique.

Finally, to dissociate the effects of additional attractors from those of increased linear distance between the subject and the verb, we augment the dataset simple sentences containing fillers. In these constructions, the fillers are adverbs that modify the main verb. Fillers extend the linear distance without introducing attractors.

- Simple sentence [2 fillers]:
  The mouse quickly and silently moves.

[*] https://universaldependencies.org/treebanks/en_ewt/

Overall, the controlled dataset consists of 80,000 sentences. To extend the analysis and generate ungrammatical sentences, the same sentences are used, but the main verb form is altered to mismatch the subject-verb agreement.

3.2 Probes

Structural Probe Let $W = (w_1, w_2, \ldots, w_t)$ be a sentence consisting of t words belonging to a dataset $\mathcal{D}$. Let $T_W = (V_W, E_W)$ be its corresponding syntactic dependency tree, where V_W is the set of words, and E_W is the set of edges connecting V_W in an acyclic manner. The adjacency matrix $A \in \{0,1\}^{t \times t}$ of T_W is defined as:

$$\mathbf{A}_{i,j} = \begin{cases} 1, & \text{if } \{w_i, w_j\} \in E_W, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

For any two words $w_i, w_j \in W$, the syntactic distance $d_T(w_i, w_j)$ is defined as the number of edges in the unique path connecting w_i and w_j in T_W. Formally, it can be written as:

$$d_T(w_i, w_j) = \arg\min_k \left(\mathbb{1}[(\mathbf{A}_{i,j})^k > 0] \right), \tag{2}$$

where $\mathbb{1}[\cdot]$ represents the indicator function.

Thus we can construct the pairwise syntactic distance $\mathbf{M} \in \mathbb{N}_0^{t \times t}$ between all words in a sentence.

$$\mathbf{M}_{i,j} = d_T(w_i, w_j) \tag{3}$$

Let $\mathbf{h}_i^l \in \mathbb{R}^d$ denote the contextual embedding of the word w_i obtained from layer l of an LLM with a hidden dimensionality d. In cases where a word is tokenized several, the word representation $\mathbf{h}_i^l$ is computed as the average of its subtoken embeddings to ensure a one-to-one correspondence between words and embeddings.

We follow Hewitt & Manning (2019) and introduce a Structural Probe, defined by a linear transformation $\mathbf{B}_S \in \mathbb{R}^{m \times d}$, where $m \leq d$ represents the output dimensionality of the Structural Probe.

In the original embedding space $\mathcal{H}$, the syntactic relations between two words $w_i, w_j \in W$ in a sentence W, are interpreted as the difference between their contextualized word embeddings.

$$\delta^l(w_i, w_j) = \mathbf{h}_i^l - \mathbf{h}_j^l \tag{4}$$

The Structural Probe maps word representations from $\mathcal{H}$ to a new space $\mathcal{S}$.

$$\mathbf{s}^l(w_i, w_j) = \mathbf{B}_S \delta^l(w_i, w_j) \tag{5}$$

The pairwise squared euclidean distance matrix $\hat{\mathbf{M}} \in \mathbb{R}^{t \times t}$ between probed word representations for a given sentence W can be computed as follows:

$$\hat{\mathbf{M}}_{ij} = ||\mathbf{s}^l(w_i, w_j)||_2^2, \tag{6}$$

where $||\cdot||$ represents L2 norm.

The goal of the Structural Probe is to find a subspace $\mathcal{S}$ in which $\hat{\mathbf{M}}_{ij}$ approximates $\mathbf{M}_{ij}$. For that, the Structural Probe will be trained with the following objective.

$$\begin{aligned} \mathcal{L}_S(\mathbf{B}_S) &= \frac{1}{|D|} \sum_{W \in D} \frac{1}{|W|^2} \sum_{(i,j)=1}^{|W|} \left| \mathbf{M}_{ij} - \hat{\mathbf{M}}_{ij} \right|, \\ \mathbf{B}_S^* &= \arg\min_{\mathbf{B}_S} \mathcal{L}_S. \end{aligned} \tag{7}$$

Polar Probe Complete dependency trees are directed and labeled acyclic graphs rather than undirected and unlabeled. Thus, each edge (dependency) $e \in E_W$ in the graph (dependency tree) is associated with two functions: U and C, which determine its directionality and label (dependency type) respectively.

The Polar Probe (Diego-Simon et al., 2024), extends the Structural Probe and predicts a labeled and directed dependency tree, approximating U and C for each predicted syntactic edge. We extend the study to the Polar Probe since it augments Hewitt & Manning (2019)'s framework with additional syntactic information while preserving a linearly readable, distance-based syntactic code. For more details about its implementation and training refer to Appendix A.

3.3 Training

The Structural and Polar probes were trained on activations from layer 16 of `Mistral-7B-v0.1` (Jiang et al., 2023), `Llama-2-7b-hf` (Touvron et al., 2023), and `BERT-large` (Devlin et al., 2019), as this layer has been shown to best encode dependency structures[†] (Diego-Simon et al., 2024).

Training was conducted on the filtered UD-EWT corpus (Silveira et al., 2014) using Stochastic Gradient Descent (SGD) with the Adam optimizer (Kingma & Ba, 2017). Probes were trained for 30 epochs with a batch size of 200 sentences and a learning rate of 0.005. For the Polar Probe, the hyperparameter λ was set to 10.0 as specified in (Diego-Simon et al., 2024).

Evaluation Following Hewitt & Manning (2019), for each sentence W in the test or controlled dataset, we use the trained probe $\mathbf{B}_P^*$ or $\mathbf{B}_S^*$ to compute its probed pairwise distance matrix $\hat{\mathbf{M}}$. Then, we apply Kruskal's algorithm Kruskal (1956) using networkx (Hagberg et al., 2008) to obtain the Minimum Spanning Tree (MST), $\hat{E}_W$. Ignoring edge direction and labels, the probe performance on a given gold edge $e \in E_W$ is measured by whether e appears in $\hat{E}_W$:

$$\text{Accuracy}(e) \;=\; \begin{cases} 1, & \text{if } e \in \hat{E}_W, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

3.4 Baselines

To gauge the effectiveness of the probes, we compare them against a strong baseline, modifying the definition of $\hat{\mathbf{M}} \in \mathbb{R}^{t \times t}$. For more baselines, refer to Appendix A.

- **Activation Space prediction**
  In this baseline, we use the raw LLM activations directly, without training or applying any additional linear transformation.

$$\hat{\mathbf{M}}_{ij} \;=\; ||\delta^l(w_i, w_j)||_2^2$$

3.5 Surprisal calculation

Surprisal, in causal language models, quantifies the information content of a word w_i via its conditional probability given its previous context (Shannon, 1948). We compute word probabilities $p(w_i)$ (Pimentel & Meister, 2024), from which surprisal I is derived.

$$I(w_i) = -\log(p(w_i)) \tag{9}$$

3.6 Classifiers

Ridge Regression and Random Forest models were trained to predict the probe's output for ground truth positive instances $(e \in E_W)$ using linguistic features such as head depth, head surprisal, child

[†] `Mistral-7B-v0.1` and `Llama-2-7b-hf` are transformer-based LLMs with 32 layers and 7 billion parameters. `BERT-large` is a masked language model with 24 layers and 340 million parameters.

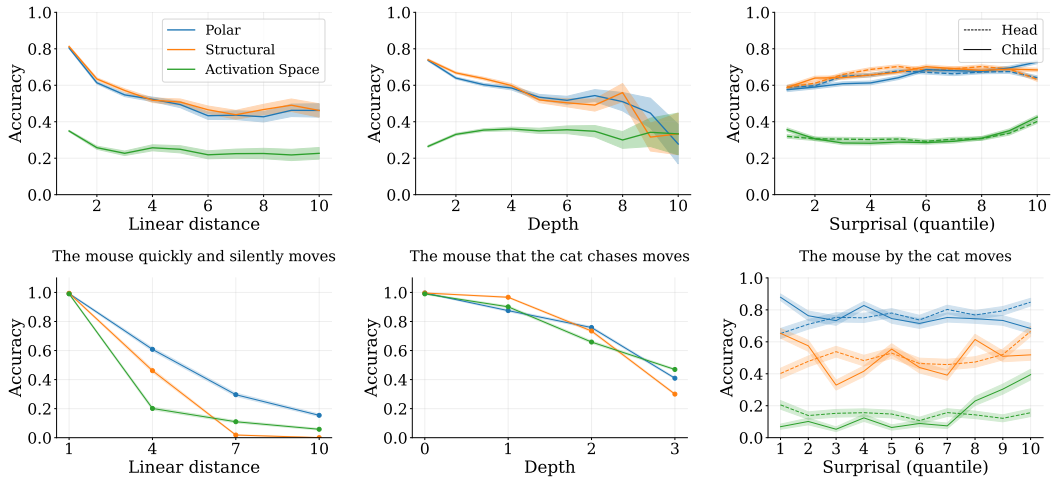


Figure 2: **The performance of trained linear probes is most affected by the linear distance and depth of syntactic dependencies in both naturalistic (Top) and controlled (Bottom) stimuli. (Top)**: Accuracy on syntactic dependencies in UD-EWT. **(Top Left)** By the linear distance between head and child. **(Top Middle)** By the depth of the head. **(Top Right)** By the surprisal quantile of the head and the child. **(Bottom):** Accuracy on the subject-verb dependency in controlled sentences. **(Bottom Left):** Simple sentences with a varying number of fillers. **(Bottom Middle)**: In CE sentences with varying number of nestings, modifying the syntactic depth of the innermost subject-verb dependency . **(Bottom Right):** In sentences with 1 PP by the surprisal quantile for head and child. Shaded areas indicate the standard error of the mean in all plots.

surprisal, and linear distance. These classifier models reveal which properties of an existing syntactic dependency are most relevant for probes to predict it correctly.

The Ridge classifer model was implemented via scikit-learn's (Pedregosa et al., 2011) `RidgeClassifier` with a regularization parameter $\alpha = 100.0$. In parallel, the Random Forest classifier was configured using scikit-learn's `RandomForestClassifier` with 100 estimators, a maximum tree depth of 10, and a minimum of 500 samples per split to mitigate overfitting.

Both models were evaluated using 5-fold cross-validation. For each fold, feature importance values were extracted from the Random Forest model and subsequently signed using the corresponding coefficients from the Ridge Regression model.

4 Results

This section is structured as follows: First, we study which linguistic and statistical factors best predict probe performance (section 4.1), using both naturalistic (section 4.1.1) and manually-crafted controlled (section 4.1.2) datasets. Next, we investigate whether sentence parsing by structural probes exhibits processing phenomena similar to those observed in humans (section 4.2).

4.1 Probe performance is impaired by longer and deeper dependencies, but not by the surprisal of its elements.

4.1.1 Naturalistic data

We first study the Structural and Polar Probes on a naturalistic dataset, the Universal Dependencies English Web Treebank (UD-EWT) dataset. For each analysis, we evaluate whether the distances between contextualized word embedding within the probe's subspace, accurately represents the syntactic distance in the annotated dependency tree. We evaluate the probe accuracy as a function of three linguistic features: (1) the linear distance that seperate two syntactically related words (i.e. the head and the dependent, fig. 2, top left), (2) the syntactic depth of the head (fig. 2, top middle), and (3) the surprisal of the head (fig. 2, top right).

We find that both linear distance and syntactic depth negatively impact probe accuracy, indicating that long-range and deeply-nested dependencies still challenge current probes. Remarkably, for very deep structure (depth=10), the probe's accuracy is similar to what can be read out from the original space of activations of the LLM (fig. 2, top middle). By contrast, word surprisal shows little to no impact on probe accuracy. These results mean that the predictability of a word given its context does not affect the syntactic representation identified by the syntactic probes (fig. 2, top right). Overall, these three results suggest that probe performance is more sensitive to structural properties of language than to statistical predictability. Suggesting a better performance generalization to nonsensical sentences rather than to sentences containing long-range or deep dependencies.

Linear distance, syntactic depth and word surprisals may not be fully independent from one another. To verify that the above results, we trained a Random Forest classifier to predict the probe decisions, depending on these three linguistic features (linear distance, syntactic depth and (head and child) surprisal).

fig. 3 shows the signed feature importance for the different input variables. Overall, the results confirm that probe accuracy are most affected by the linear distance and the syntactic depth. By contrast, surprisals lead to weaker and less consistent impact (fig. 2, top right). These results reinforce our earlier findings, suggesting that probe performance is most impacted by both syntactic depth and linear distance.

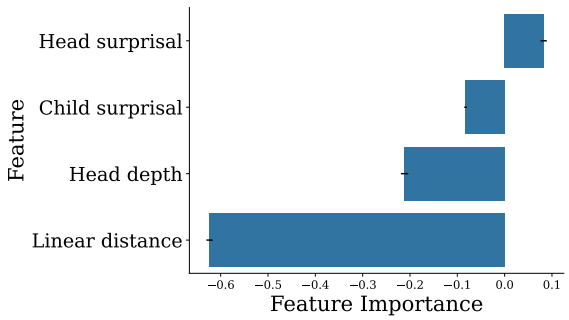


Figure 3: **Linear distance and depth of the syntactic dependency are the best predictors of probe performance** Signed mean feature importance in a Random Forest classifier trained to predict whether the Polar Probe detects existing dependencies in the UD-EWT dataset. The error bars represent the standard deviation of the feature importance values across 5 cross-validation folds.

4.1.2 Controlled sentences

The above analyses are based on syntactically annotated naturalistic sentences. To independently assess the impact linear distance, syntactic depth and word surprisal on the probes' accuracy, we extend our analyses to a set of controlled sentences, designed such that their syntactic dependencies systematically vary in linear distance and syntactic depth (see section 3.1). In all cases, probe accuracy is evaluated on the subject-verb dependency to ensure consistency across conditions. To manipulate linear distance (fig. 2, bottom left), we used main phrase sentences with a varying number of adverbs as fillers (e.g., "The cat quickly and silently walks"). For syntactic depth (fig. 2, bottom middle), we evaluate the innermost subject-verb dependency in CE sentences with different levels of nesting (e.g., "The cat that the fox chases moves"). Where the subject and verb lie at a linear distance of 1 for the different numbers of nesting. For surprisal (fig. 2, bottom right), we analyzed sentences with a single PP to avoid fillers (favoring more natural constructions) while introducing a long-range agreement.
Once again, our findings align with those from the UD-EWT dataset: linear distance and syntactic depth strongly impact probe accuracy. For linear distance, fig. 2, bottom left we find a linear distance of 3 words is enough to decrease probe accuracy by 0,4 points. A similar effect happens for syntactic depth fig. 2, bottom middle. Notably, for CE sentences, the 'Activation Space' baseline matches with the Structural and Polar probes. We suggest that this is due to the subject and the verb being located at adjacent locations. Finally, for surprisal (fig. 2, bottom right), we observe opposing trends, as reflected in fig. 3, where accuracy slightly increases with head surprisal, whereas child surprisal exhibits the opposite effect. When comparing the Structural and Polar probes, we find that the Polar Probe exhibits a slower decay in accuracy with increasing linear distance fig. 2, bottom left. We hypothesize that its more constrained objective, including additional syntactic information, fosters learning more robust syntactic representations, that rely less on surface heuristics.

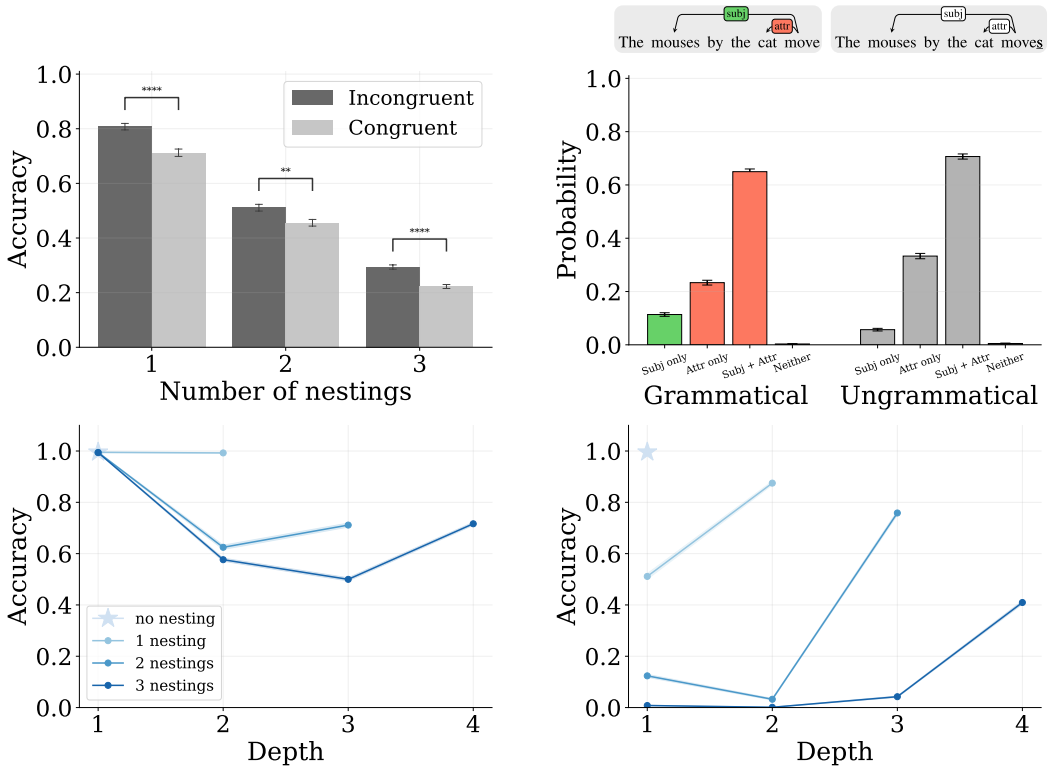


Figure 4: **Probes share some similarities with human syntactic processing but fail on simple PP sentences. (Top left):** Accuracy of the Polar Probe on the subject-verb dependency in grammatical PP sentences with varying nesting levels, comparing cases where the final attractor is congruent or incongruent with the subject. Congruent cases are harder for the probe. **(Top right):** Probability of different Polar Probe predictions in grammatical and ungrammatical with one PP. Predictions are grouped by which word; the attractor, the subject, both, or neither—binds to the verb. Because the attractor adjacent to the verb, the probes are misled and produce mostly incorrect parses (red), outnumbering correct ones (green). Ungrammatical cases increase the attraction effect. **(Bottom left):** Accuracy profile for Right Branching (RB) where the linear distance is constant across nestings, modifying syntactic depth. Probes show primacy and recency effects. **(Bottom Right):** Accuracy profile for Center Embedding (CE) sentences, inner dependencies, despite being deeper, are also closer together resulting in higher probe accuracies. Error bars represent the standard error in all plots.

4.2 Do probes and humans show similar parsing effects?

Psycholinguistic research (Wagers et al., 2009; Bock & Miller, 1991; Francis, 1986; Kimball & Aissen, 1971; Haskell & MacDonald, 2005) has shown that interfering nouns tend to bias individuals toward misinterpreting and misproducing long-range subject-verb agreements. These effects are especially pronounced in ungrammatical sentences, where attraction errors are more likely to occur. Moreover, increasing the linear distance between the subject and verb, as seen in prepositional phrase and object relative clauses, further generates confusion in syntactic agreement (Bock & Cutting, 1992).

4.2.1 Congruent nouns impact negatively on probe performance

More recent work (Lakretz et al., 2022; Lampinen, 2024; Marvin & Linzen, 2018) has shown that congruent attractor nouns reinforce the confidence of LLMs when predicting verb forms during the next-token prediction task. (refer to Appendix A for more details regarding congruence and model prediction). We hypothesize that, in contrast, incongruent attractor nouns may cause confusion in the syntactic parsing task, which is not generative.

To test this hypothesis, measure sensitivity of probes to the congruency between the subject and the last attractor in PP sentences. Therefore, in congruent cases, the last attractor functions as an interfering noun positioned adjacent to the verb.

As depicted in Figure 4, top left, the Polar Probe exhibits greater difficulty with congruent cases, suggesting a tendency to associate the verb with the attractor when surface forms are compatible. Conversely, in incongruent cases, the absence of interference aids in disambiguating the correct subject-verb dependency, leading to improved probe performance. Notably, as indicated by the star notation in Figure 4, top left, the difference between the two groups is highly significant across the different levels of nesting. Lastly, Figure 4, top left illustrates the impact of attaching additional PPs on subject-verb accuracy; as expected, accuracy decreases with the number of PPs due to both the greater linear distance and the presence of more attractors.

4.2.2 Ungrammatical verb forms impact negatively on probe performance

Analogous to the interfering effects observed in grammatical sentences, previous works (Marvin & Linzen, 2018; Hu et al., 2024; Ryu & Lewis, 2021) have shown that ungrammaticality also fosters attraction effects in LLMs when evaluated in a generative setting. Mirroring findings from psycholinguistic studies. As illustrated in Figures 4, top middle and 13, top middle, we find that in PP sentences with varying levels of nesting, the Polar Probe exhibits a consistently higher error rate when encountering ungrammatical verb forms.

4.2.3 One PP suffices to elicit attraction effects

Remarkably, we note that even in simple sentences with a single PP, the Polar Probe tends to bind the verb to both the subject and the attractor, resulting in a wrong parse (Figures 4, top middle and 13, top middle).

Such binding strategy comes at the expense of accurately capturing the *case* relation within the PP. Surprisingly, the *case* relation is challenging for the probe to parse as revealed by further dissection the UD-EWT dataset into dependency types. Refer to Appendix A for more details. Overall we find such result as an indicator of the extent to which probes are sensitive to linear distance. Such sensitiveness comes with prediction errors and points a solid separation between humans and probes.

As shown in Figure 16, top middle such failure is somewhat mitigated by the use of a masked language model, where attention heads operate bidirectionally.

4.2.4 Accuracy profiles for Right Branching and Center Embedding

Controlled sentences with Right Branching (RB) and Center Embedding (CE) structures each include one subject-verb dependency per level of nesting. In RB sentences, syntactic depth increases while the linear distance between the subject and the verb remains constant. In contrast, in CE sentences, the subject-verb linear distance decreases as syntactic depth increases. These structural differences lead to distinct accuracy profiles with respect to the number nestings and syntactic depth.

As shown in fig. 4, top right and consistent with the results presented in fig. 2, accuracy in RB sentences declines with increasing syntactic depth. Notably, a U-shaped accuracy profile emerges, suggestive of primacy and recency effects documented in human cognition (Murdock, 1962; Atkinson & Shiffrin, 1968; Baddeley & Hitch, 1974), where items at the beginning and end of a sequence tend to be processed more accurately than those in the middle.

In contrast, accuracy in CE sentences improves with deeper nesting. This finding underscores the dominant influence of linear distance over syntactic depth as shown in fig. 3.

5 Discussion, Limitations and Future work

Our analyses reveal that accuracy of syntactic probes is primarily sensitive to structural rather than to statistical properties. In fact, in both naturalistic and controlled datasets, we find that the model's surprisal to the words in the dependency barely has an impact on probe accuracy. In contrast, probe accuracy decreases as the linear distance and syntactic depth of the dependency increases.

Notably, in relatively simple sentences with one Prepositional Phrase (eg: The keys to the cabinet are big.), the probe tends to bind both the subject and the attractor to the verb, yielding a wrong parse (Section 4.2.3).

Such result puts forward a key challenge for linear probes, their sensitiveness to linear distance and syntactic depth makes them prompt to errors. In the PP case, the probes' sensitiveness to linear distance becomes especially apparent— the attractor being adjacent to the main verb is enough to confuse probes to yield a wrong syntactic parse.

To address these challenges, and better evaluate probes, we introduce a set of controlled sentences to serve as a benchmark. These carefully designed stimuli enable a linguistically motivated and systematic evaluation of probe performance.

Despite these challenges, probes also exhibit notable successes. They capture syntactic structure more accurately than the model's raw activation space, and perform far beyond a linear distance baseline. Moreover, we find some similarities between probes and human behavior. Specifically, these similarities manifest as increased syntactic errors under both noun interference and presence ungrammatical verb forms (Wagers et al., 2009; Bock & Miller, 1991). However, this resemblance remains superficial due to the current failure cases observed in linear structural probes.

Several factors likely underlie the current challenges of structural probes. First, their training data is strongly skewed toward dependencies with a linear distance of one, introducing a bias into the learned linear probe during its training. Second, if LLMs encode syntax along a nonlinear manifold in hidden space, any strictly linear probe will inevitably distort that representation. Third, some errors may originate not from the probes themselves but from the LLMs, which may fail to encode syntactic structure with full fidelity. Lastly, prompting LLMs has been shown to affect their downstream behavior (Lampinen, 2024), contextualizing the sentences using a carefully designed prompt might result in linguistically richer representations, potentially enhancing the performance of syntactic probes.

We posit that exploring non-Euclidean probes that better align with the geometry of LLM representations is a promising direction for future work. Likewise, training syntactic probes on controlled stimuli to better capture long-range dependencies and deeper hierarchical structures remains an important area for continued research.

6 Acknowledgments

This project was provided with computer and storage resources by GENCI at IDRIS thanks to the grant 2023-AD011014766 on the supercomputer Jean Zay's the V100 and A100 partition (PDS).
This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 945304 (PDS).
This project received funding from PSL University under the grant agreement ANR-10-IDEX-0001-02 (EC).
This project received funding from the Département d'Études Cognitives (DEC) at ENS under the grant agreement FrontCog, ANR-17-EURE-0017 (EC).
This project received funding from the French National Research Agency (ANR) under the grant agreement ComCogMean, Projet-ANR-23-CE28-0016 (EC).

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2017. URL https://openreview.net/forum?id=ryF7rTqgl.

R.C. Atkinson and R.M. Shiffrin. *Human Memory: A Proposed System and its Control Processes*, pp. 89–195. Elsevier, 1968. doi: 10.1016/s0079-7421(08)60422-3. URL http://dx.doi.org/10.1016/S0079-7421(08)60422-3.

Alan D. Baddeley and Graham Hitch. Working memory. In Gordon H. Bower (ed.), *New York: Academic Press*, volume 8 of *Psychology of Learning and Motivation*, pp. 47–89. Academic Press, 1974. doi: https://doi.org/10.1016/S0079-7421(08)60452-1. URL https://www.sciencedirect.com/science/article/pii/S0079742108604521.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. What do neural machine translation models learn about morphology? In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 861–872, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1080. URL https://aclanthology.org/P17-1080/.

Kathryn Bock and J.Cooper Cutting. Regulating mental energy: Performance units in language production. *Journal of Memory and Language*, 31(1):99–127, 1992. ISSN 0749-596X. doi: https://doi.org/10.1016/0749-596X(92)90007-K. URL https://www.sciencedirect.com/science/article/pii/0749596X9290007K.

Kathryn Bock and Carol A Miller. Broken agreement. *Cognitive Psychology*, 23(1):45–93, January 1991. ISSN 0010-0285. doi: 10.1016/0010-0285(91)90003-7. URL http://dx.doi.org/10.1016/0010-0285(91)90003-7.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. Disentangling syntax and semantics in the brain with deep networks. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1336–1348. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/caucheteux21a.html.

Boli Chen, Yao Fu, Guangwei Xu, Pengjun Xie, Chuanqi Tan, Mosha Chen, and Liping Jing. Probing {bert} in hyperbolic spaces. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=17VnwXYZyhH.

Noam Chomsky. *Syntactic Structures*. De Gruyter, December 1957. ISBN 9783112316009. doi: 10.1515/9783112316009. URL http://dx.doi.org/10.1515/9783112316009.

Noam Chomsky. Language and nature. *Mind*, 104(413):1–61, 1995. doi: 10.1093/mind/104.413.1.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2126–2136, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1198. URL https://aclanthology.org/P18-1198/.

William Croft. Autonomy and functionalist linguistics. *Language*, 71(3):490, September 1995. ISSN 0097-8507. doi: 10.2307/416218. URL http://dx.doi.org/10.2307/416218.

Stanislas Dehaene and Laurent Cohen. The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, 15(6):254–262, June 2011. ISSN 1364-6613. doi: 10.1016/j.tics.2011.04.003. URL http://dx.doi.org/10.1016/j.tics.2011.04.003.

Anne-Dominique Devauchelle, Catherine Oppenheim, Luigi Rizzi, Stanislas Dehaene, and Christophe Pallier. Sentence syntax and content in the human temporal lobe: An fmri adaptation study in auditory and visual modalities. *Journal of Cognitive Neuroscience*, 21(5):1000–1012, May 2009. ISSN 1530-8898. doi: 10.1162/jocn.2009.21070. URL http://dx.doi.org/10.1162/jocn.2009.21070.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Pablo J. Diego-Simon, Stéphane d'Ascoli, Emmanuel Chemla, Yair Lakretz, and Jean-Remi King. A polar coordinate system represents syntax in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=x2780VcMOI.

Tiwalayo Eisape, Vineet Gangireddy, Roger Levy, and Yoon Kim. Probing for incremental parse states in autoregressive language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2801–2813, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.203. URL https://aclanthology.org/2022.findings-emnlp.203/.

W. Nelson Francis. Proximity concord in english. *Journal of English Linguistics*, 19:309 – 317, 1986. URL https://api.semanticscholar.org/CorpusID:145224184.

Lyn Frazier and Keith Rayner. Resolution of syntactic category ambiguities: Eye movements in parsing lexically ambiguous sentences. *Journal of Memory and Language*, 26(5):505–526, October 1987. ISSN 0749-596X. doi: 10.1016/0749-596x(87)90137-9. URL http://dx.doi.org/10.1016/0749-596X(87)90137-9.

Angela D Friederici. The neural basis for human syntax: Broca's area and beyond. *Current Opinion in Behavioral Sciences*, 21:88–92, 2018. ISSN 2352-1546. doi: https://doi.org/10.1016/j.cobeha.2018.03.004. URL https://www.sciencedirect.com/science/article/pii/S2352154617301286. The Evolution of Language.

Lila Gleitman. The structural sources of verb meanings. *Language Acquisition*, 1(1):3–55, January 1990. ISSN 1532-7817. doi: 10.1207/s15327817la0101_2. URL http://dx.doi.org/10.1207/s15327817la0101_2.

Yosef Grodzinsky and Angela D Friederici. Neuroimaging of syntax and syntactic processing. *Current Opinion in Neurobiology*, 16(2):240–246, 2006. ISSN 0959-4388. doi: https://doi.org/10.1016/j.conb.2006.03.007. URL https://www.sciencedirect.com/science/article/pii/S0959438806000328. Cognitive neuroscience.

Raffaele Guarasci, Stefano Silvestri, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. Assessing bert's ability to learn italian syntax: a study on null-subject and agreement phenomena. *Journal of Ambient Intelligence and Humanized Computing*, 14(1):289–303, May 2021. ISSN 1868-5145. doi: 10.1007/s12652-021-03297-4. URL http://dx.doi.org/10.1007/s12652-021-03297-4.

Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

John T. Hale and Miloš Stanojević. Do LLMs learn a true syntactic universal? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17106–17119, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.950. URL https://aclanthology.org/2024.emnlp-main.950/.

Todd R. Haskell and Maryellen C. MacDonald. Constituent structure and linear order in language production: Evidence from subject-verb agreement. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5):891–904, 2005. ISSN 0278-7393. doi: 10.1037/0278-7393.31.5.891. URL http://dx.doi.org/10.1037/0278-7393.31.5.891.

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

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121, 2024. doi: 10.1073/pnas.2400917121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2400917121.

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

Edith Kaan and Tamara Y Swaab. The brain circuitry of syntactic comprehension. *Trends in cognitive sciences*, 6(8):350–356, 2002.

John Kimball and Judith Aissen. I think, you think, he think. *Linguistic Inquiry*, 2(2):241–246, 1971. ISSN 00243892, 15309150. URL http://www.jstor.org/stable/4177629.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Joseph B. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7(1):48–50, February 1956. ISSN 1088-6826. doi: 10.1090/s0002-9939-1956-0078686-7. URL http://dx.doi.org/10.1090/S0002-9939-1956-0078686-7.

Yair Lakretz, Théo Desbordes, Dieuwke Hupkes, and Stanislas Dehaene. Can transformers process recursive nested constructions, like humans? In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 3226–3232, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.285/.

Andrew Lampinen. Can language models handle recursively nested grammatical structures? a case study on comparing models and humans. *Computational Linguistics*, 50(3):1441–1476, December 2024. doi: 10.1162/coli_a_00525. URL https://aclanthology.org/2024.cl-4.8/.

Tomasz Limisiewicz and David Mareček. Introducing orthogonal constraint in structural probes. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *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)*, pp. 428–442, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.36. URL https://aclanthology.org/2021.acl-long.36/.

Rebecca Marvin and Tal Linzen. Targeted syntactic evaluation of language models. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1192–1202, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1151. URL https://aclanthology.org/D18-1151/.

Rowan Hall Maudslay and Ryan Cotterell. Do syntactic probes probe syntax? experiments with jabberwocky probing. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 124–131, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.11. URL https://aclanthology.org/2021.naacl-main.11/.

Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. Probing for labeled dependency trees. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7711–7726, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.532. URL https://aclanthology.org/2022.acl-long.532/.

Bennet B Murdock. The serial position effect of free recall. *J. Exp. Psychol.*, 64(5):482–488, November 1962.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. Universal Dependencies. In Alexandre Klementiev and Lucia Specia (eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-5001/.

Christophe Pallier, Anne-Dominique Devauchelle, and Stanislas Dehaene. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, 108(6):2522–2527, January 2011. ISSN 1091-6490. doi: 10.1073/pnas.1018711108. URL http://dx.doi.org/10.1073/pnas.1018711108.

Core Francisco Park, Andrew Lee, Ekdeep Singh Lubana, Yongyi Yang, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. In-context learning of representations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=pXlmOmlHJZ.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. Dissecting contextual word embeddings: Architecture and representation. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1499–1509, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1179. URL https://aclanthology.org/D18-1179/.

Tiago Pimentel and Clara Meister. How to compute the probability of a word. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18358–18375, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1020. URL https://aclanthology.org/2024.emnlp-main.1020/.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4609–4622, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.420. URL https://aclanthology.org/2020.acl-main.420/.

Keith Rayner. Eye movements in reading and information processing. *Psychological Bulletin*, 85(3): 618–660, 1978. ISSN 0033-2909. doi: 10.1037/0033-2909.85.3.618. URL http://dx.doi.org/10.1037/0033-2909.85.3.618.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. Visualizing and measuring the geometry of bert. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/159c1ffe5b61b41b3c4d8f4c2150f6c4-Paper.pdf.

Soo Hyun Ryu and Richard Lewis. Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. In Emmanuele Chersoni, Nora Hollenstein, Cassandra Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus (eds.), *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pp. 61–71, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.cmcl-1.6. URL https://aclanthology.org/2021.cmcl-1.6/.

Andrea Santi and Yosef Grodzinsky. Working memory and syntax interact in broca's area. *NeuroImage*, 37(1):8–17, August 2007. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2007.04.047. URL http://dx.doi.org/10.1016/j.neuroimage.2007.04.047.

Adam Shai, Paul M. Riechers, Lucas Teixeira, Alexander Gietelink Oldenziel, and Sarah Marzen. Transformers represent belief state geometry in their residual stream. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=YIB7REL8UC.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27 (3):379–423, July 1948. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, 2014.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SJzSgnRcKX.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Matthew W. Wagers, Ellen F. Lau, and Colin Phillips. Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, 61(2):206–237, August 2009. ISSN 0749-596X. doi: 10.1016/j.jml.2009.04.002. URL http://dx.doi.org/10.1016/j.jml.2009.04.002.

Jennifer C. White, Tiago Pimentel, Naomi Saphra, and Ryan Cotterell. A non-linear structural probe. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 132–138, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.12. URL https://aclanthology.org/2021.naacl-main.12/.

A Appendix

A.1 Polar Probe implementation

Each edge (dependency) $e \in E_W$ in the graph (dependency tree) is associated with two functions: U and C, which determine its directionality and label (dependency type) respectively.

$$\begin{aligned} U &: E_W \to \{-1, 1\}, \quad e \mapsto u(e), \\ C &: E_W \to C, \quad e \mapsto c(e). \end{aligned} \tag{10}$$

Similarily to the Structural Probe, the Polar Probe Diego-Simon et al. (2024) defined as $\mathbf{B}_P \in \mathbb{R}^{m \times d}$ linearly transforms the LLM's embedding space $\mathcal{H}$ into a subspace $\mathcal{P}$ following eqs. (5) to (7).

$$\begin{aligned} \mathbf{p}^l(w_i, w_j) &= \mathbf{B}_P \delta^l(w_i, w_j), \\ \hat{\mathbf{M}}_{ij} &= ||\mathbf{p}^l(w_i, w_j)||_2^2 \end{aligned} \tag{11}$$

Additionally to the structural objective (eq. (7)), the Polar Probe introduces an angular objective $\mathcal{L}_A$ so that $\mathbf{p}^l(w_i, w_j)$ where $\{w_i, w_j\} \in E_W$ encodes information about U and C. For that, a set of edges Ω_e are extracted across sentences from dataset $\mathcal{D}$. Then, the following objective is minimized:

$$\begin{aligned} \mathcal{L}_A(\mathbf{B}_P) = \frac{1}{|\Omega_e|} \sum_{e, e' \in \Omega_e} \Big(& \measuredangle(\mathbf{p}^l_e u(e), \mathbf{p}^l_{e'} u(e')) \\ & - \mathbb{1}[c(e) = c(e')] \Big)^2. \end{aligned} \tag{12}$$

where $\measuredangle(\cdot, \cdot)$ denotes the cosine similarity, and we write $\mathbf{p}^l_e$ as a shorthand for $\mathbf{p}^l(w_i, w_j)$ for an edge $e = \{w_i, w_j\} \in E_W \subset \Omega_e$, by slight abuse of notation.

Therefore, the Polar Probe is the result of the following objective function:

$$\mathbf{B}^*_P = \arg\min_{\mathbf{B}_P} \mathcal{L}_S(\mathbf{B}_P) + \lambda \mathcal{L}_A(\mathbf{B}_P) \tag{13}$$

A.2 Additional baselines:

- **Linearly informed random prediction**

  In this baseline, for two words at positions i and j in a sentence, we define their distance to be the absolute difference in their positions, plus a small random perturbation ϵ_{ij}:

  $$\hat{\mathbf{M}}_{ij} \;=\; |i - j| \;+\; \epsilon_{ij}.$$

  Here, ϵ_{ij} is drawn i.i.d. ϵ_{ij} is sampled i.i.d. for $i < j$ from a noise distribution and we set $\epsilon_{ij} = \epsilon_{ji}$ and $\epsilon_{ii} = 0$ to keep $\hat{\mathbf{M}}$ symmetric and zero-diagonal.

- **Random prediction**

  To establish a lower bound, we assign completely random distances:

  $$\hat{\mathbf{M}}_{ij} \;=\; \epsilon_{ij},$$

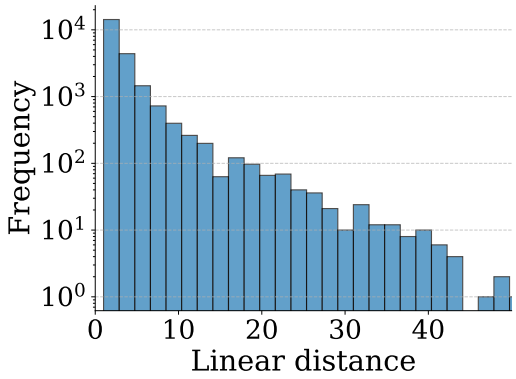


Figure 5: Histogram of linear distances for dependencies in the naturalistic sentences

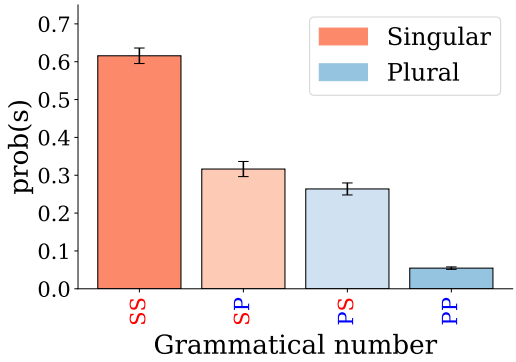


Figure 6: Probability for predicting the token ‘s’ for different grammatical number combinations of the subject and the attractor.

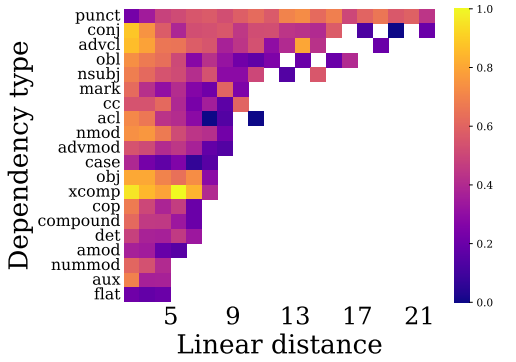


Figure 7: Polar probe accuracy for dependencies by their linear distance and dependency type.

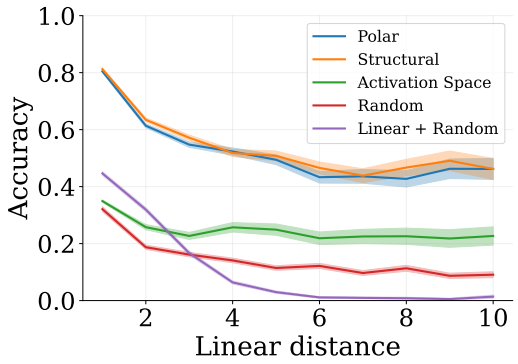


Figure 8: Accuracy as a function of linear distance in UD-EWT

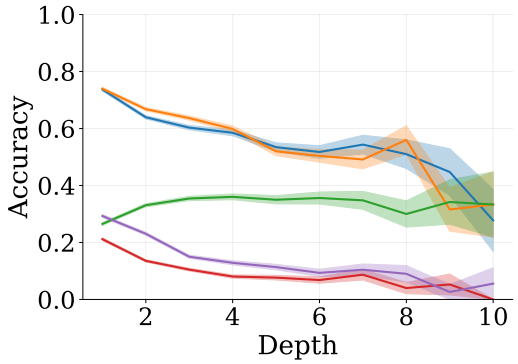


Figure 9: Accuracy as a function of syntactic depth in UD-EWT

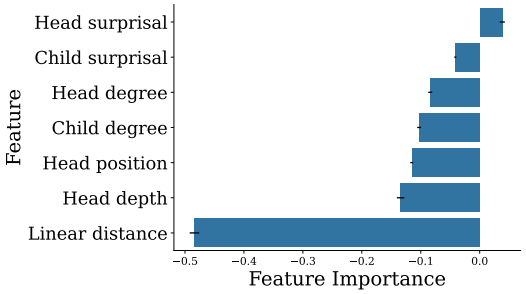


Figure 10: Feature importance values in Random Forest model with additional features.

B Llama2-7B

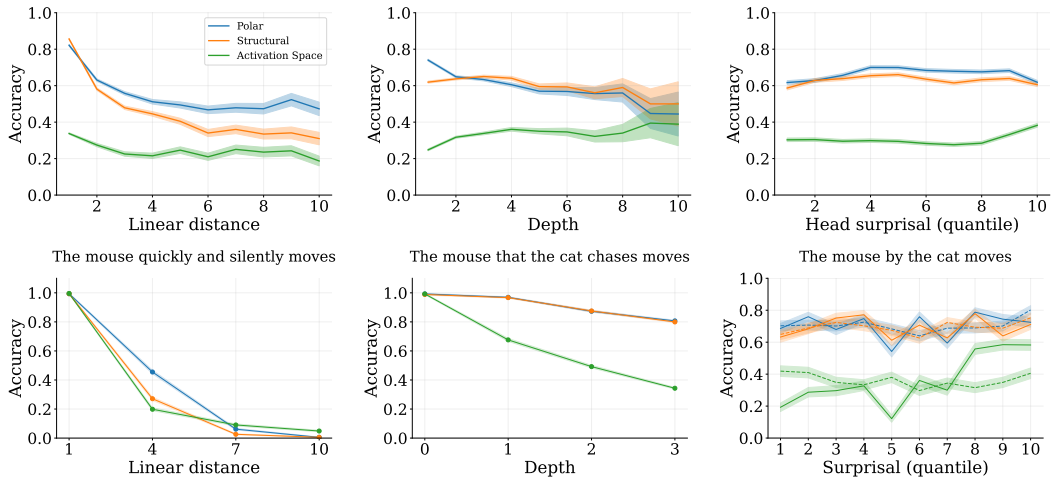


Figure 11

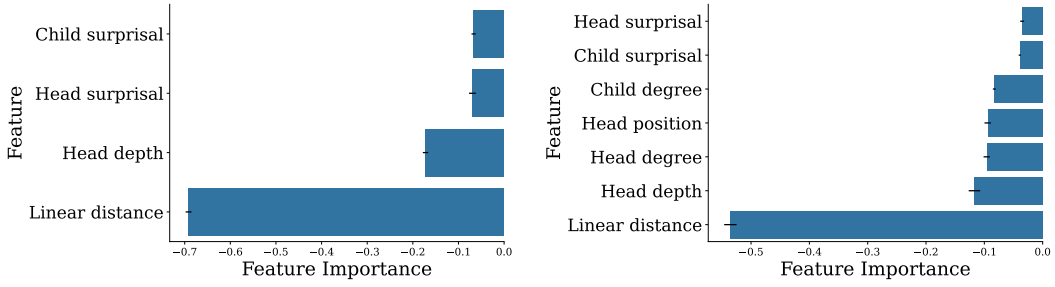


Figure 12

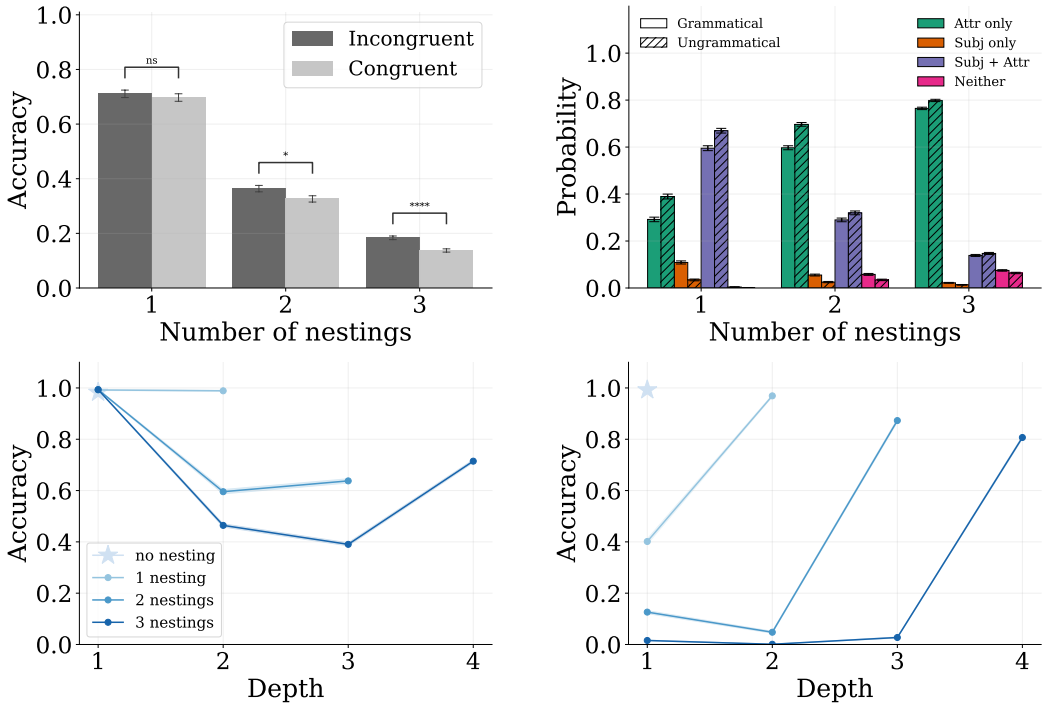

Incongruent
Congruent
ns
*
****
Accuracy
Number of nestings
Grammatical
Ungrammatical
Attr only
Subj only
Subj + Attr
Neither
Probability
Number of nestings
no nesting
1 nesting
2 nestings
3 nestings
Accuracy
Depth
Accuracy
Depth

Figure 13

C Mistral-7B

C.1 Structural Probe

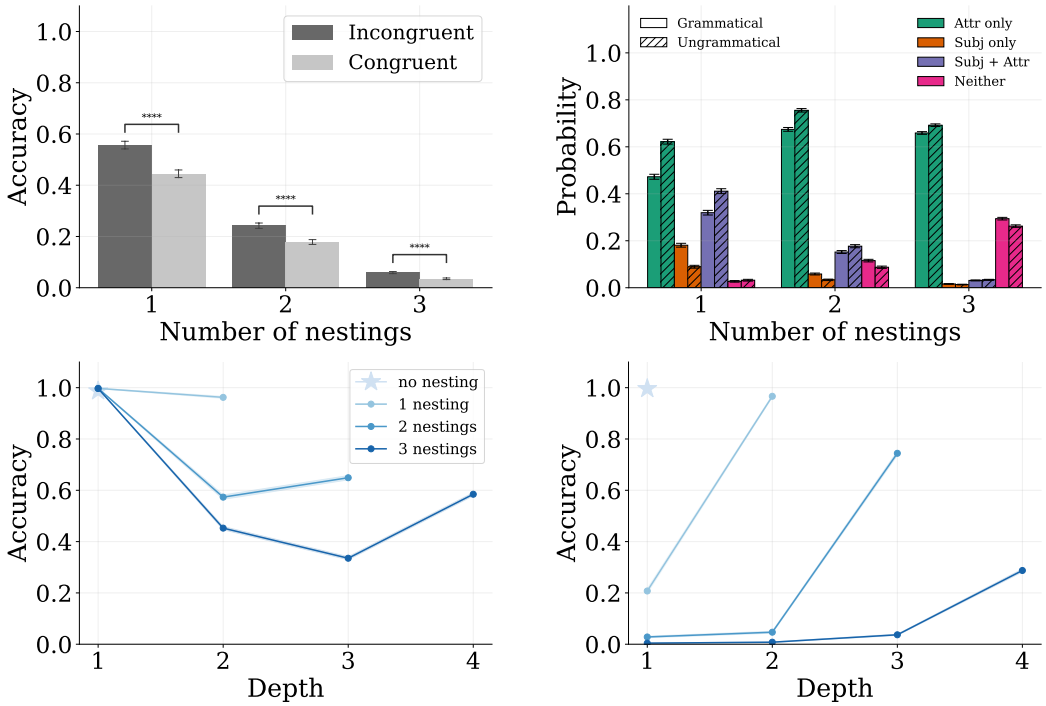


Figure 14

D BERT-large

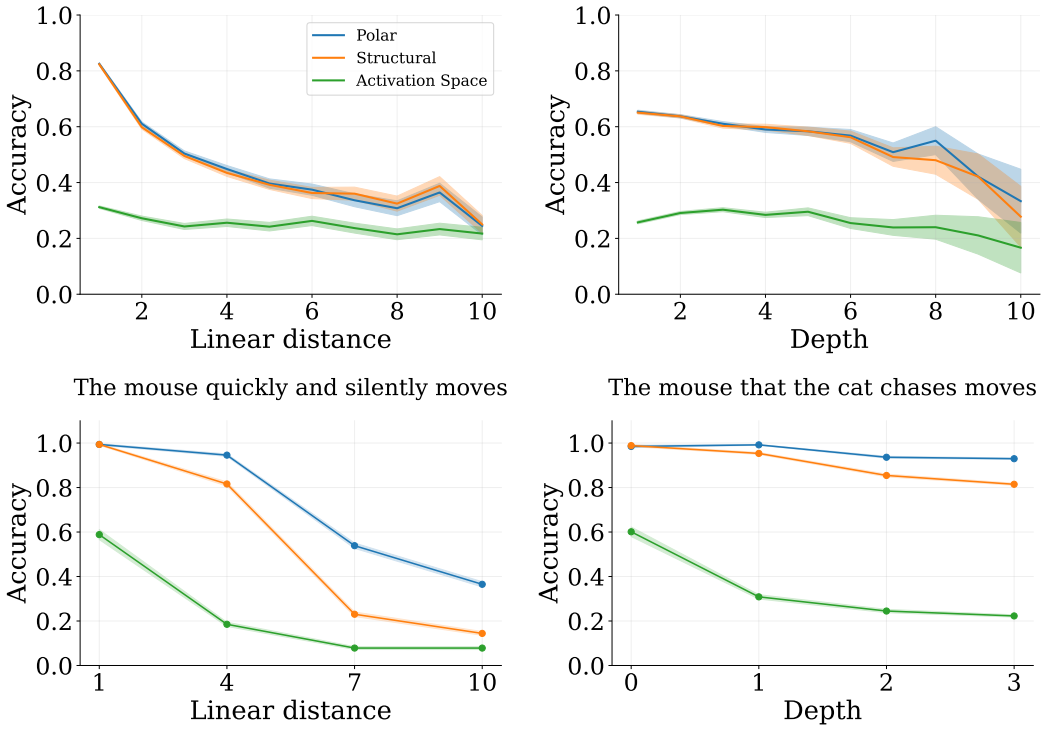


Figure 15

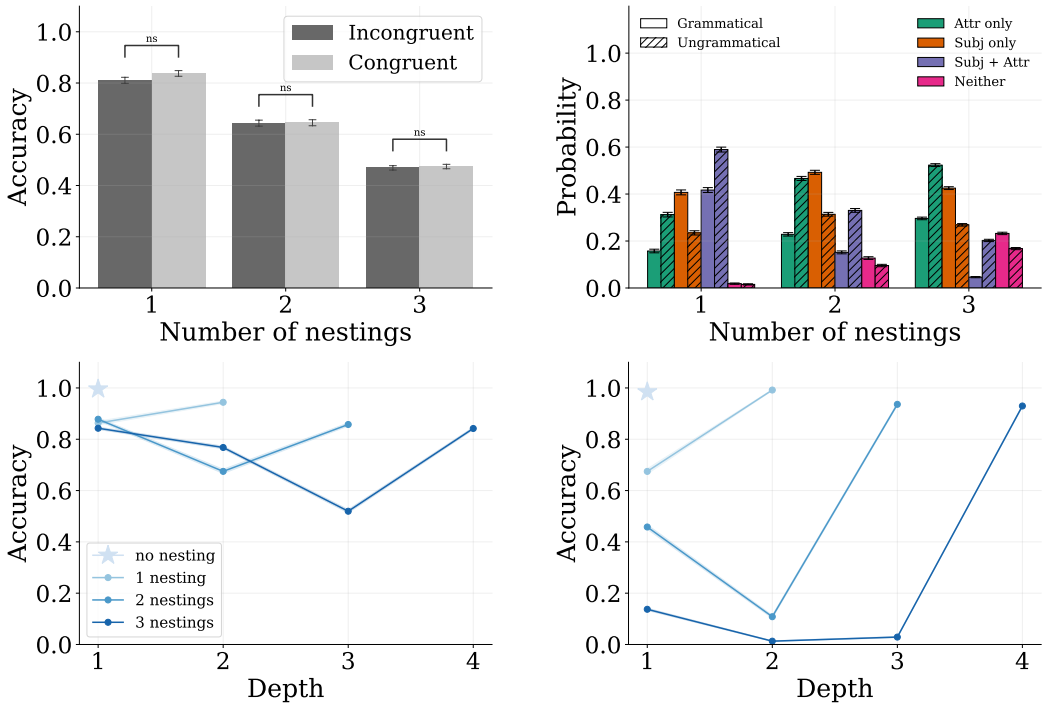


Figure 16

E Controlled Dataset

E.1 Prepositional Phrase

Sentence	Structure	Attractors	Fillers
some neighbours by some boys rest	pp	1	0
the teacher below some friends around some bakers plays	pp	2	0
a tailor above the boy near a teacher by the kids rests	pp	3	0
the girls past a boy quickly and carefully stand	pp	1	1
the grandma below a friend gently eagerly playfully happily and nervously plays	pp	1	2
a kid behind a soldier near the waiters playfully and silently sits	pp	2	1

Table 1: Examples of PP sentences in the controlled dataset.

E.2 Center Embedding

Sentence	Structure	Attractors	Fillers
some teachers that the tailor admires walk	ce	1	0
some players that the neighbors that a mom calls find rest	ce	2	0
the teacher that some moms that a tailor that the guests show finds lift stands	ce	3	0
the grandmas that some engineers loudly carefully and boldly help smile	ce	1	1
an artist that the player swiftly happily joyfully silently gently quietly and nervously hears sits	ce	1	2
the artist that the neighbor that the soldier joyfully boldly and quickly kicks meets walks	ce	2	1

Table 2: Examples of CE sentences in the controlled dataset.

E.3 Right Branching

Sentence	Structure	Attractors	Fillers
a friend assumes that the baker predicts	rb	1	0
a singer imagines that the guests feel that the neighbor realizes	rb	2	0
the girls realize that a writer predicts that some teachers think that an engineer imagines	rb	3	0
a boy feels that the sad shy and brave engineer expects	rb	1	1
a baker notices that the friendly young old angry funny lazy and wild dads believe	rb	1	2
some grandmas feel that some neighbors notice that the funny kind and shy boys know	rb	2	1

Table 3: Examples of RB sentences in the controlled dataset.