Mechanisms vs. Outcomes: Probing for Syntax Fails to Explain Performance on Targeted Syntactic Evaluations

Ananth Agarwal, Jasper Jian, Christopher D. Manning, Shikhar Murty
Stanford University
{ananthag, manning, smurty}@cs.stanford.edu, jjian@stanford.edu

Abstract

Large Language Models (LLMs) exhibit a robust mastery of syntax when processing and generating text. While this suggests internalized understanding of hierarchical syntax and dependency relations, the precise mechanism by which they represent syntactic structure is an open area within interpretability research. Probing provides one way to identify syntactic mechanisms linearly encoded in activations; however, no comprehensive study has yet established whether a model's probing accuracy reliably predicts its downstream syntactic performance. Adopting a "mechanisms vs. outcomes" framework, we evaluate 32 open-weight transformer models and find that syntactic features extracted via probing fail to predict outcomes of targeted syntax evaluations across English linguistic phenomena. Our results highlight a substantial disconnect between latent syntactic representations found via probing and observable syntactic behaviors in downstream tasks.

1 Introduction

The remarkable ability of LLMs to process and generate text that respects a rich diversity of syntactic constraints strongly suggests that models have sophisticated syntactic knowledge. Yet our understanding of internal representations of these syntactic facts is lacking, which has motivated research into interpretability paradigms searching for syntactic *mechanisms*, such as mechanistic interpretability and causal abstractions (Geiger et al., 2022; Murty et al., 2022). We study probing, a prominent approach in this domain which assumes that model *activations* are the mechanism through which syntactic knowledge is encoded and that this knowledge is linearly recoverable using small supervised models called *probes* (Conneau et al., 2018; Hewitt and Manning, 2019; Tenney et al., 2019, among others). However, no comprehensive study has evaluated whether probing accuracy is indicative of targeted syntactic *outcomes*.

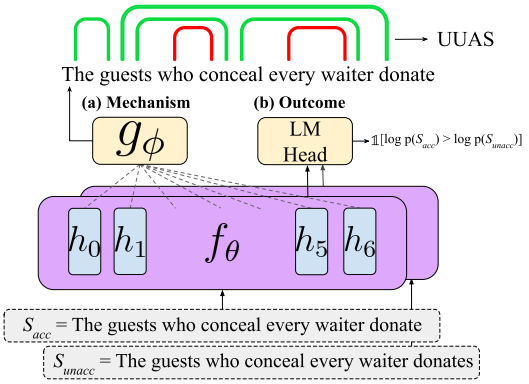

Figure 1: **Mechanisms vs. outcomes setup**. We find no convincing predictive power of syntax probing accuracy on downstream syntactic evaluation accuracy. (a) Mechanism: probe g_ϕ extracts a dependency parse tree from S_{acc} word-level hidden states h_i; UUAS = 4/6 edges in this toy example. (b) Outcome: evaluate minimal pair accuracy.

We deepen understanding of the relationship between syntactic mechanisms and outcomes by asking: *Do probing accuracies, measured by dependency tree attachment scores of syntactic probes, effectively predict a model's downstream performance on targeted syntactic evaluations?* We evaluate syntactic outcomes using BLiMP (Warstadt et al., 2020), a benchmark of minimal pairs spanning well-established English grammatical contrasts. Assigning higher probability to the acceptable sentence in an acceptable-unacceptable pair has been established as a desired linguistic outcome (Marvin and Linzen, 2018; Warstadt et al., 2020).

We train three syntax probes on 32 open-weight transformer models. We then fit Ordinary Least Squares (OLS) regressions modeling BLiMP minimal pairs accuracy as a function of probe attachment scores on the grammatically acceptable sentences (Fig 1). We demonstrate that across all levels of analysis—amalgamation of all BLiMP linguistic phenomena, individual phenomena, and targeted

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 33737–33757
November 4-9, 2025 ©2025 Association for Computational Linguistics

suites of difficult minimal pairs—syntax probing accuracy shows no clear predictive power for downstream outcomes.[1]

Concretely, we find that probing performance—measured via directed and undirected unlabeled attachment scores (UAS/UUAS)—fails to predict model accuracy on BLiMP minimal pairs. Across 32 models, none of the three probes yields a statistically significant regression fit on the full overall dataset. At the per-phenomenon level, only anaphor agreement shows a significant correlation, but the control experiment reveals that even a non-syntactic probe can produce a similar fit, suggesting the result is likely spurious. In finer-grained evaluations targeting difficult syntactic paradigms, we test whether probes can recover critical tree edges that are plausibly relevant for resolving minimal pairs, such as the subject-verb edge in subject–verb agreement pairs. However, probe recovery of these edges only weakly aligns with minimal pair outcome (40–60% match rate across models). Finally, within each BLiMP paradigm we test whether UUAS score distributions differ between correct and incorrect minimal pair outcomes. In over 85% of paradigms, the distributions overlap substantially for most models.

Taken together, our results suggest that while syntactic probing is a dominant interpretability paradigm, it fails to robustly reveal a model's latent syntactic knowledge. We establish a disconnect between the syntactic information revealed through conventional probing for dependency trees and the latent syntactic knowledge that far more often than not assigns grammatical text a higher likelihood than ungrammatical text. Our results argue for the use of external targeted evaluations as the gold standard for establishing model syntactic competence, and support the continuing development of BLiMP-style resources beyond English (Jumelet et al., 2025; Başar et al., 2025; Taktasheva et al., 2024, among others).

2 Setup

Given an input sentence $S \triangleq (w_1, \ldots, w_N)$, a pretrained transformer f_θ produces contextual vectors $h_i \in \mathbb{R}^n$ for each word w_i at every layer. Our probes, denoted $g_\phi : \mathbb{R}^n \to \mathbb{R}^k$ with parameters ϕ, are learned linear functions that project contextual hidden states into a k-dimensional subspace, where $k \leq n$. Linear and bilinear probes have been shown to achieve high selectivity; they effectively capture linguistic properties of the probed representation (Hewitt and Liang, 2019).

2.1 Mechanisms

We train three syntax probes with different optimization objectives for comprehensive coverage. We evaluate the correctness with which each probe extracts the dependency parse of a sentence.

Structural Probe. Proposed in Hewitt and Manning (2019), the structural distance probe g_ϕ^{struct} is a linear transformation $g_\phi^{\text{struct}}(h) = B^{\text{struct}}h$ with parameters $\phi = \{B^{\text{struct}} \in \mathbb{R}^{k \times n}\}$ that learns to encode the dependency tree distance d_{ij} between word pair w_i and w_j in the probe's projected space by minimizing

$$\min_\phi \sum_S \frac{1}{|S|^2} \sum_{i,j} \left| d_{ij} - \|g_\phi^{\text{struct}}(h_i - h_j)\|_2^2 \right|. \tag{1}$$

The evaluation metric is undirected unlabeled attachment score (UUAS): the fraction of gold tree edges $E_{\text{gold}} = \{\{w_i, w_j\} : (w_i, w_j) \in \mathcal{G}\}$ included in the minimum spanning tree $E_{\text{pred}} = \{\{w_i, w_j\} : (w_i, w_j) \in \hat{\mathcal{G}}\}$ calculated from the probe's predicted tree distances. Punctuation and the root relation are excluded.

$$\text{UUAS} = (|E_{\text{pred}} \cap E_{\text{gold}}|)/|E_{\text{gold}}|.$$

Orthogonal Structural Probe. Limisiewicz and Mareček (2021) replace the original structural probe linear transformation with an orthogonal transformation and scaling vector. The orthogonal probe g_ϕ^{ortho} is the transformation $g_\phi^{\text{ortho}}(h) = \bar{d} \odot Vh$ with parameters $\phi = \{V \in \mathbb{R}^{n \times n}, \bar{d} \in \mathbb{R}^n\}$ where V is orthogonal. The authors maintain orthogonality during training using Double Soft Orthogonality Regularization (DSO; Bansal et al., 2018) with regularization $\lambda_o = 0.05$. The overall probe training objective is

$$\min_\phi \sum_S \frac{1}{|S|^2} \sum_{i,j} \left| d_{ij} - \|g_\phi^{\text{ortho}}(h_i - h_j)\|_2^2 \right| + \lambda_o \text{DSO}(V). \tag{2}$$

Limisiewicz and Mareček (2021) report g_ϕ^{ortho} performs on par with g_ϕ^{struct} at predicting dependency trees while being less prone to memorizing

[1] Code is at https://github.com/agananth/SyntaxMechanismsOutcomes

training trees. Moreover, the scaling vector $\bar{d}$ enables interpretation of the relative importance of each dimension in V. The evaluation metric is UUAS, as with g_ϕ^{struct}.

Head Word Probe. Inspired by Clark et al. (2019), we define the head word probe g_ϕ^{head} as a linear transformation $g_\phi^{\text{head}}(h) = B^{\text{head}}h$ with parameters $\phi = \{B^{\text{head}} \in \mathbb{R}^{k\times n}\}$ trained using cross-entropy loss. For each dependent w_i (including punctuation), the probe predicts its head in $\{\text{ROOT}, w_1, \ldots, w_N\} \setminus \{w_i\}$. Let H be cross entropy loss and $\text{head}(i)$ be the head of w_i:

$$\hat{d}_{ij} = \|g_\phi^{\text{head}}(h_i - h_j)\|_2$$
$$\min_\phi \sum_S \frac{1}{|S|} \sum_i H(\{\hat{d}_{ij} \mid i \neq j\}, \text{head}(i)). \quad (3)$$

Since the edge prediction is directed, the evaluation metric is unlabeled attachment score (UAS). Here $E_{\text{gold}} = \{(i, \text{head}(i)) : 1 \leq i \leq N\}$ and $E_{\text{pred}} = \{(i, \hat{\text{head}}(i)) : 1 \leq i \leq N\}$ such that

$$\text{UAS} = (|E_{\text{pred}} \cap E_{\text{gold}}|)/|E_{\text{gold}}|.$$

Control Probe. Importantly, the performance of syntax probes and downstream syntactic performance may partly reflect non-syntactic factors such as the quality of lexical semantic representations in a model. The overall quality of model hidden states is thus a confounding variable affecting both probe and syntactic evaluation performance. To isolate this confounder, we propose a simple control probing task where performance depends only on hidden state quality and not latent syntactic structure. Concretely, for each sentence, we form all word pairs where the treebank-specific part-of-speech (XPOS) of the two words is the same and train a probe g_ϕ^{ctrl} to recover the L_2 distance between the two words' GloVe vectors (Pennington et al., 2014). g_ϕ^{ctrl} is a linear transformation $g_\phi^{\text{ctrl}}(h) = B^{\text{control}}h$ with parameters $\phi = \{B^{\text{control}} \in \mathbb{R}^{k\times n}\}$. Let v_i denote the GloVe vector of w_i. The training objective for the control probe is

$$\tilde{d}_{ij} = \|g_\phi^{\text{ctrl}}(h_i - h_j)\|_2$$
$$\min_\phi \sum_S \frac{1}{|S|} \sum_{\substack{(i,j) \\ \text{XPOS}(w_i)=\text{XPOS}(w_j)}} \text{Huber}(\tilde{d}_{ij}, \|v_i - v_j\|_2). \quad (4)$$

We use Huber loss for outlier robustness. The evaluation metric for the probe is the Spearman correlation ρ_s between predicted and actual GloVe distances.

This probe is trained to recover non-syntactic information as modeled by an uncontextualized word embedding model, GloVe, whose vector geometries have been shown to encode linguistic information like lexical semantics (Brown et al., 2023). We further ensure that our control probe is not indirectly capturing syntactic part-of-speech information by only training the probe to recover within word-pair distances where the words share the same XPOS. Crucially, any predictive success of the control probe on syntactic evaluations warrants a more cautious interpretation of the results of syntactic probes, as it suggests that the latter's explanatory power may not be attributable solely to the recoverability of syntactic representations.

2.2 Outcomes

Our benchmark for linguistic knowledge outcomes is BLiMP accuracy. BLiMP has 67 template-generated sub-datasets called paradigms grouped into 13 linguistic phenomena across English morphology, syntax, and semantics (Warstadt et al., 2020).[2] Each paradigm has 1000 acceptable-unacceptable minimal pairs (S_{acc}, S_{unacc}) that exhibit a single grammatical contrast. The example pair from Fig 1 is from the *distractor agreement relative clause* paradigm within subject–verb agreement. The verb agreement contrast point in the pair between the acceptable (top) and unacceptable (bottom) sentences is shown below in bold:

The guests (Subject (plural)) who conceal every waiter (Distractor (singular)) (Relative Clause) **donate**.

The guests (Subject (plural)) who conceal every waiter (Distractor (singular)) (Relative Clause) **donates**.

Model f_θ accuracy on a set of pairs $\mathcal{D}$ is computed as

$$\frac{1}{|\mathcal{D}|} \sum_{\mathcal{D}} \mathbb{1}\{\log P_{f_\theta}(S_{acc}) > \log P_{f_\theta}(S_{unacc})\}. \quad (5)$$

For decoder models, $\log P_{f_\theta}(S)$ is the sum of token level log-probabilities. For encoder and encoder-decoder models, we instead use pseudo-log-likelihood scores (PLL; Salazar et al., 2020).

[2]The two semantic phenomena, quantifiers and NPI licensing, still require knowledge of hierarchical syntactic structure, e.g., to determine scope for NPI licensing (Ladusaw, 1979).

2.3 Models

We test 32 open-weight pretrained decoder, encoder, and encoder-decoder models up to 8B parameters. Appendix A has the full list.

3 Experimental Setup

Probe Training. We train each of our four probe "families" on Stanza (Qi et al., 2020) Universal Dependencies (de Marneffe et al., 2021) dependency parses of the Penn Treebank (PTB; Marcus et al., 1993) using its standard train and validation splits. We report training parameters and compute in Appendix B. g_ϕ^{struct}, g_ϕ^{head}, and g_ϕ^{ctrl} have fixed probe dimension $k = 256$, while g_ϕ^{ortho} has $k = n$ for each model to ensure orthogonality is feasible. Let L be the number of layers in model f_θ. To reduce the influence of a word's semantic properties, we subtract word embeddings from each layer's contextual states

$$h_i^\ell = h_i^\ell - h_i^0 \quad \forall \ell \in \{1, \ldots, L\}.$$

For each of g_ϕ^{struct} and g_ϕ^{head}, we train L probes per model—one per layer—and select the probe with the best PTB test metric. For g_ϕ^{ortho}, given limited compute resources and the fact that it is derived from g_ϕ^{struct}, we train a single probe per model on the layer with the best g_ϕ^{struct} PTB test metric. Since the best layer can differ between g_ϕ^{struct} and g_ϕ^{head}, and the objective of g_ϕ^{ctrl} is to control for hidden state quality, we train a separate g_ϕ^{ctrl} instance for each on the corresponding best layer. Control results are reported with the associated probe.

Control Probe Validation. We construct a simple evaluation dataset to test how well our control probe abstracts away contextual linguistic information to verify its design. For 500 random words chosen from SimLex999 (Hill et al., 2015), we prompt GPT-4o (OpenAI et al., 2024) to generate 10 sentences that contain the word in substantially different contexts. Refer to Appendix C for the prompt and sample sentences.

We first measure the variance of each word's *contextual* hidden states (at the layer g_ϕ^{ctrl} is trained at) across the 10 samples, and report the average across all 500 words. We then apply g_ϕ^{ctrl} on these hidden states and recompute this quantity. If the probe successfully abstracts away contextually-determined linguistic information like syntax and retains non-contextual information like lexical semantics, we expect the mean variance of a given word hidden state to significantly decrease, since lexical semantics of a given word should remain the same regardless of context.

Mechanisms vs. Outcomes Regression. Our objective is model-level mechanisms vs. outcomes regression because we want to study how effectively the method of syntactic probing predicts an arbitrary model's latent syntactic knowledge. We compute regressions separately at two BLiMP granularities: (1) overall (averaging predictor and response values across all paradigms in the full dataset), and (2) each of the 13 phenomena separately with Holm-Bonferroni correction. Holm-Bonferroni is an effective method for adjusting significance tests to account for multiple comparisons (13 in our case).

For each linear regression at a given granularity, the predictor variables are probe scores: UUAS (g_ϕ^{struct}, g_ϕ^{ortho}), UAS (g_ϕ^{head}), or Spearman correlation ρ_s (g_ϕ^{ctrl}). We run regressions separately for each probe family, with one data point per model. Fig 1 illustrates the approach: the selected probe for each model is applied to every S_{acc} sentence at the given granularity to compute an average probe score (*mechanism*), and each S_{acc} and S_{unacc} feed into minimal pair evaluation (*outcome*).

We fit simple (eqn 6) and multiple (eqn 7) OLS regressions for each probe family at each granularity:

$$y = \beta_0 + \beta_1 x_1 + \epsilon \tag{6}$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \tag{7}$$

where y is minimal pairs accuracy (eqn 5) and x_1 is the probe score. For multiple regression, x_1 is UUAS or UAS and x_2 is ρ_s. Comparing β_1 and its p-value for the syntax probe when it is the sole predictor versus when paired with the control reveals the extent to which it predicts grammaticality outcomes. Since the two regression models are nested, we also include log-likelihood ratio tests (LRT) for robust statistical comparisons.

Critical Edge Study in Difficult Paradigms. Drilling down to the paradigm level, we select the two paradigms in each of the subject–verb agreement (*distractor agreement relational noun*, *distractor agreement relative clause*) and filler–gap (*wh. vs. that with gap*, *wh. vs. that with gap long distance*) phenomena that scored lowest in model accuracy in Warstadt et al. (2020). Rather than

tree UUAS/UAS, we determine if the induced syntactic tree for each S_{acc} contains a *critical edge* corresponding to the phenomenon targeted by the minimal pairs—an `nsubj` edge for subject–verb agreement, and an `obj` or `obl` edge for filler–gap. Appendix D discusses critical edge criteria in-depth. We hypothesize that the binary outcome of the probe's ability to extract the critical edge on S_{acc} is linked to the minimal pair binary outcome.

4 Results & Discussion

Attachment score typically peaks in the middle layers. Fig 2 plots PTB test g_ϕ^{struct} UUAS across model layers. Best layers for GPT-2 (Radford et al., 2019) match with previous results from the literature (Eisape et al., 2022), validating our probe training procedure. In Appendix E, Fig 13 plots PTB test g_ϕ^{head} UAS, and we further show that our syntax probes' abilities to recover encoded parse tree information in BLiMP sentences correlate well across training objective (g_ϕ^{struct} vs. g_ϕ^{head}, Fig 14) and architecture (g_ϕ^{struct} vs. g_ϕ^{ortho}, Fig 15). Furthermore, the comparable range of UUAS/UAS scores between the PTB test set and BLiMP shows the probes generalize from human-written PTB text to template-generated BLiMP.

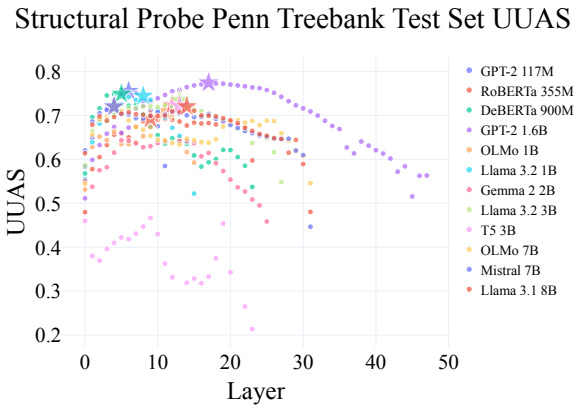

Figure 2: **Penn Treebank test set g_ϕ^{struct} UUAS for each layer of a sample of our models**. The star icon for a model indicates the layer with the best test set accuracy that is used for BLiMP evaluation. For most models, this occurs in the first half. Our results for GPT-2 124M and GPT-2 1.6B closely match those of Eisape et al. (2022). For T5 3B, although most layers yield low UUAS, the best layer (13) still exceeds 0.7.

Control probe erases contextual information. Fig 3 shows that for all models, regardless of the scale of the variance in the average word's contextual hidden states across sentences, the variance in the control probe's projected representation space is near-zero. This shows the probe nullifies context and uncovers GloVe-like linguistic information within hidden states.

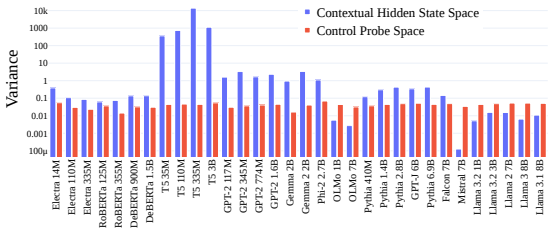

Figure 3: **Control probes consistently erase contextual information in hidden states, as evidenced by near-zero variance of word contextual hidden states in the projected representation space**. Control probes shown here are trained on structural probe best layers.

No significant fit on overall BLiMP. The first column of Fig 4 shows OLS simple regression result plots at the full BLiMP dataset granularity. Each of the three syntax probes demonstrates a minimal effect size with none reaching statistical significance. g_ϕ^{head} has no measurable effect at all. Its UAS values generally exceed the UUAS values of other probes, reflecting successful learning of relation directionality, but lack of y-axis stratification prevents predictive power.

Table 1 compares simple and multiple regression statistics. For g_ϕ^{head}, the significant p-value of the LRT conveys that considering non-syntactic signal from activations provides better fit on full BLiMP than relation direction-oriented signal alone. Even so, the adjusted R^2 is still negligible, showing that neither regression meaningfully explains BLiMP accuracy. For g_ϕ^{struct} and g_ϕ^{ortho}, adding the control probe offers no meaningful improvement.

Only anaphor agreement has statistical significance at the per-phenomenon granularity. Out of the three syntax probes and *all* phenomena, the only statistically significant correlations are g_ϕ^{struct} and g_ϕ^{ortho} on anaphor agreement. Strikingly, probes show no predictive power even for phenomena like subject–verb agreement and filler–gap dependencies whose behavior we might expect to be reliant on syntactic representations. Refer to the second, third, and fourth columns of panels in Fig 4 for simple regression plots for these select phenomena.

Table 1 shows that aside from anaphor agreement, there is no consistent result of the simpler

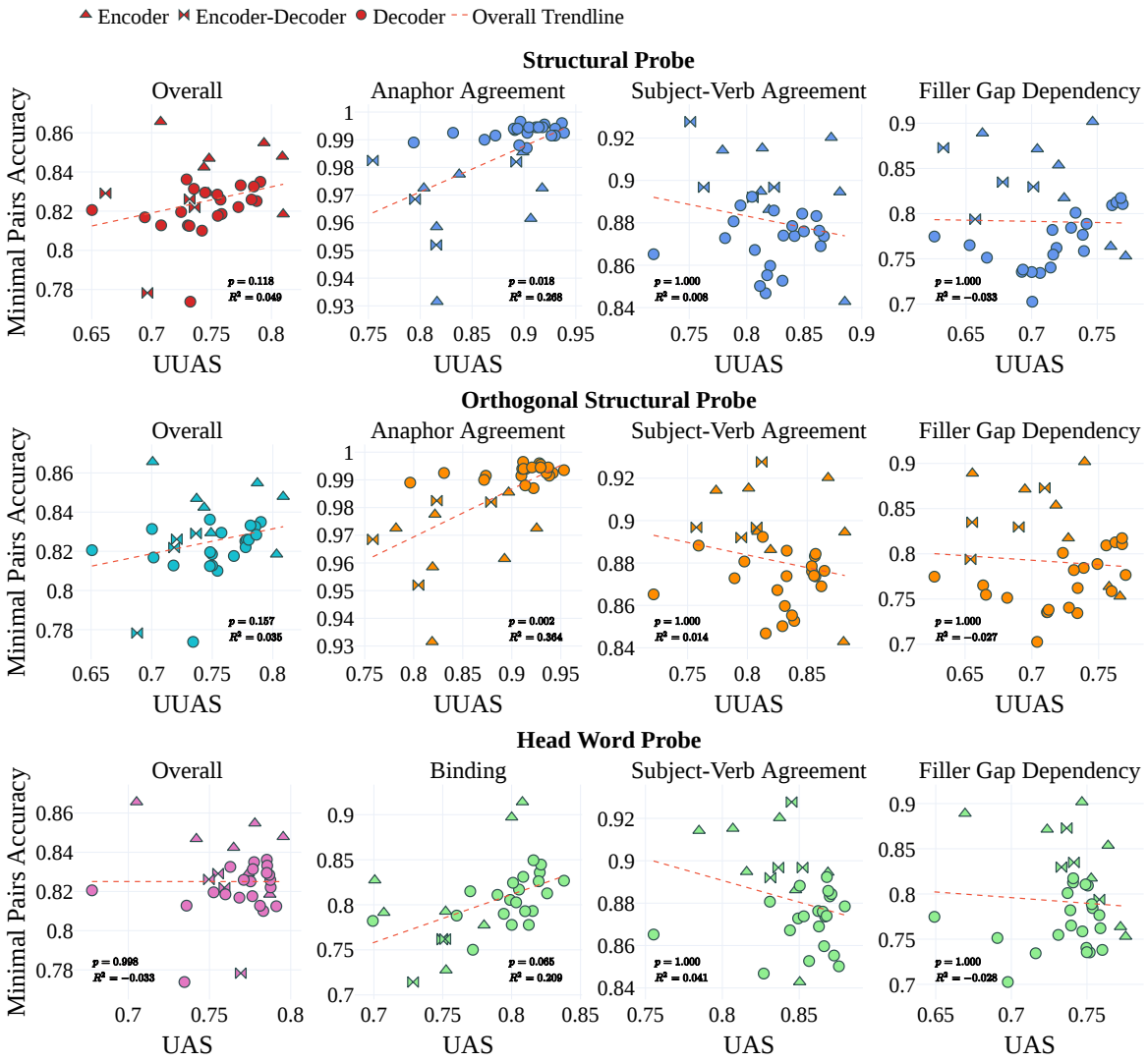

Figure 4: **Simple regression plots for g_ϕ^{struct} (top row), g_ϕ^{ortho} (middle row), and g_ϕ^{head} (bottom row)**. Each panel is annotated with adjusted R^2 and the p-value of β_1. Per-phenomenon results have Holm-Bonferroni correction. The first column of panels shows that at the full dataset granularity, no probe explains the spread in minimal pairs accuracy with any statistical significance. At per-phenomenon granularity, the second column contains the phenomenon with the lowest p-value per probe. We additionally highlight subject–verb agreement (third column) and filler–gap (fourth column) as strongly syntactic tasks with critical edges that we identify in Appendix D.

regression model providing a statistically significant better fit than the regression that includes the control. As a further example, g_ϕ^{head} encodes hierarchical dominance of the head word. We would expect this to influence Negative Polarity Item (NPI) licensing, as NPIs must be in the structural scope of their licensor. However, our results do not support this expectation; Table 3c reports a corrected p-value of 1.0 and an adjusted R^2 of zero.

The anaphor agreement result warrants caution. Fig 5 shows the surprising result that the control probe, which is trained on a non-morphosyntactic task, can be highly correlated to the BLiMP irregular forms morphosyntactic task. Non-syntactic probes *can* predict syntactic benchmark performance. Without broad systematic correlations, isolated syntax probe–task fits are not evidence that the syntactic information in hidden states that is recoverable by these probes is a primary explanatory variable for downstream behavior; even probes with no syntactic basis can one-off align with syntactic tasks. The anaphor agreement finding could be an artifact of minimal pair accuracy saturation.

Critical edge predication does not align with minimal pair evaluation. Fig 6 shows that for 3 out of the 4 challenging subject–verb agreement and filler–gap paradigms, the match rate between

Dataset	Simple Regression ($x_1 = g_\phi^{\text{struct}}$)			Multiple Regression ($x_1, x_2 = g_\phi^{\text{ctrl}}$)			LRT	
	β_1	p-value	Adj.R^2	β_1	p-value	Adj.R^2	Stat	p-value
Overall	0.133	0.117	0.049	0.109	0.214	0.051	1.161	0.281
Anaphor Agreement	0.165	**0.019**	0.268	0.168	**0.040**	0.243	0.027	1.0
Subject–Verb Agr	-0.108	1.0	0.007	-0.047	1.0	0.128	5.236	0.243
Filler–Gap Dependency	-0.028	1.0	-0.033	0.033	1.0	-0.05	0.546	1.0

(a) $x_1 = g_\phi^{\text{struct}}$ UUAS.

Dataset	Simple Regression ($x_1 = g_\phi^{\text{ortho}}$)			Multiple Regression ($x_1, x_2 = g_\phi^{\text{ctrl}}$)			LRT	
	β_1	p-value	Adj.R^2	β_1	p-value	Adj.R^2	Stat	p-value
Overall	0.127	0.157	0.035	0.102	0.266	0.041	1.299	0.254
Anaphor Agreement	0.173	**0.002**	0.364	0.173	**0.005**	0.342	0.004	1.0
Subject–Verb Agr	-0.119	1.0	0.014	-0.059	1.0	0.132	5.183	0.251
Filler–Gap Dependency	-0.1	1.0	-0.027	-0.058	1.0	-0.049	0.402	1.0

(b) $x_1 = g_\phi^{\text{ortho}}$ UUAS.

Dataset	Simple Regression ($x_1 = g_\phi^{\text{head}}$)			Multiple Regression ($x_1, x_2 = g_\phi^{\text{ctrl}}$)			LRT	
	β_1	p-value	Adj.R^2	β_1	p-value	Adj.R^2	Stat	p-value
Overall	0.0	0.998	-0.033	-0.048	0.700	0.091	5.194	**0.023**
Binding	0.534	0.064	0.209	0.522	0.108	0.184	0.085	1.0
Subject–Verb Agr	-0.205	1.0	0.041	-0.221	1.0	0.025	0.56	1.0
Filler–Gap Dependency	-0.125	1.0	-0.028	-0.15	1.0	-0.055	0.256	1.0

(c) $x_1 = g_\phi^{\text{head}}$ UAS.

Table 1: **Comparison of simple and multiple regression statistics**. Each subtable corresponds to a different syntax probe as x_1, and $x_2 = g_\phi^{\text{ctrl}}$ ρ_s. Per-phenomenon p-values are Holm-Bonferroni corrected and bold text indicates statistical significance ($p < 0.05$). The datasets shown correspond to those in Fig 4. For both g_ϕ^{struct} and g_ϕ^{ortho}, the simple regression model fits anaphor agreement better than the multiple regression model does, but no consistent signal emerges for other phenomena. The higher capacity of g_ϕ^{ortho} relative to g_ϕ^{struct} does not translate to improved regression fit. For g_ϕ^{head}, despite the statistical significance of the LRT at the full dataset granularity, the adjusted R^2 of the multiple regression model is very weak. Table 3 in Appendix F has regression results for all phenomena.

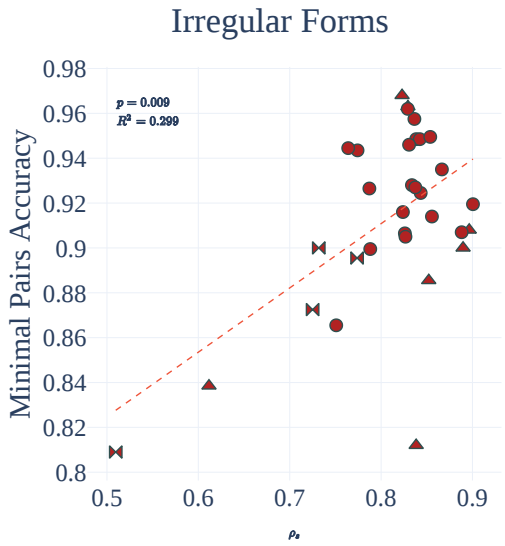

Figure 5: **Control g_ϕ^{ctrl} (trained on g_ϕ^{struct} best layers) unexpectedly achieves statistical significance for predicting the BLiMP irregular forms phenomenon, which is a morphosyntactic task**. Tables 4 and 5 in Appendix F have full control simple regression results.

g_ϕ^{struct} correctly predicting the critical edge and the model resolving the minimal pair correctly is between 40–60% for the majority of models. The wide spread of minimal pair accuracies along the y-axis of the scatter plots underscores the relative difficulty of these paradigms; by contrast, in Fig 4 plots most accuracies cluster well above 0.7. However, high Hamming distances between edge prediction and minimal pair evaluation mean that errors do not align as much as expected. Extracting critical edges, despite relevance to deciding between the words that are different between the minimal pairs (Appendix D), does not in fact predict minimal pair outcome. Fig 16 in Appendix G yields the same conclusion from g_ϕ^{head} results.

UUAS score distributions largely overlap for minimal pair outcomes. While cross-model correlations between probing and downstream performance have proved elusive, a natural follow-up investigation for each model is the sentence-level

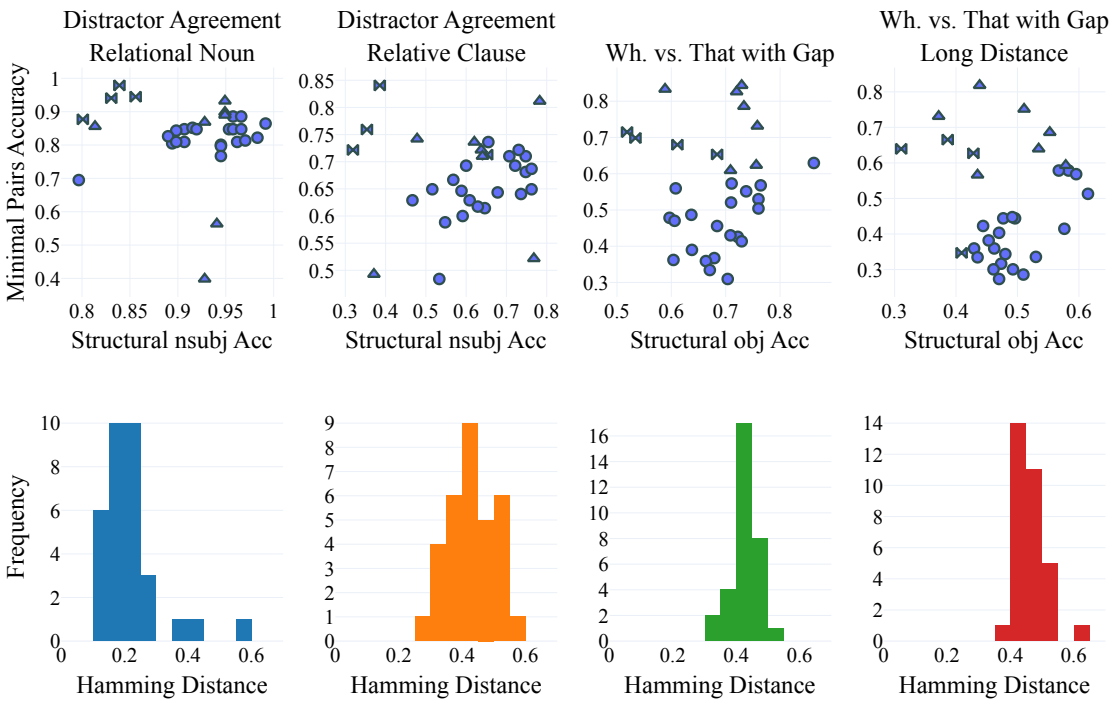

Figure 6: **Challenging subject–verb agreement and filler–gap phenomena critical edge prediction results for** $g_\phi^{\textbf{struct}}$. We compute the Hamming distance between the probe's binary outcome of predicting the critical edge in S_{acc} correctly and the model's binary outcome of resolving the minimal pair correctly. Successful critical edge prediction does not translate to successful minimal pair resolution for 3 out of 4 paradigms. The majority of models show mismatches occurring in approximately 40% to 60% of cases (Hamming distance buckets 0.4 to 0.6). Harder paradigms, indicated by lower minimal pair accuracies in the scatter plots, have higher Hamming distance.

comparison of if there is a statistically significant difference in the UUAS score distribution between the pool of sentences where the minimal pair is evaluated correctly and the pool where it is evaluated incorrectly. For each BLiMP paradigm, we run a two-sample one-sided t-test for each model on these pools of UUAS scores and apply Holm-Bonferroni correction on the test p-values. The results in Fig 7 show that only 8 out of 67 BLiMP paradigms have statistical evidence for the mean of UUAS scores for correctly predicted minimal pairs being higher than that of incorrectly predicted minimal pairs for at least 5 models. Yet again, we conclude that for most models and paradigms, UUAS score of the probe at the sentence level (mechanism) does not predict whether the model will succeed at the corresponding minimal pair evaluation (outcome).

5 Related Work

Our work builds on the extensive literature examining the syntactic capabilities of language models through probing (Hewitt and Manning, 2019; Clark et al., 2019; Müller-Eberstein et al., 2022; Limisiewicz and Mareček, 2021; Diego-Simón et al., 2024, among others).

Causal Analysis. Recent probing work aims to make *causal* claims between probed representations and observed model downstream behavior. Causal analysis intervenes in model representations to understand their effects. Tucker et al. (2021) demonstrate that counterfactual interventions on contextual embeddings—designed to erase information identified by a probe—can predictably affect downstream behavior. However, they find that such effects are not robust across models or probes. Similarly, Eisape et al. (2022) perform causal intervention on transformer states during autoregressive generation from GPT-2. Arora et al. (2024) introduce a causal interpretability method benchmark, CausalGym, to evaluate interventions sourced from minimal pairs on next-token predictions. Rather than evaluating causality, we focus on probing's statistical predictive power.

Advancing Probing Research. Belinkov (2022) surveys the promises and shortcomings of probing research. Our paper addresses several core issues raised and advances the discussion with new empirical evidence. First, Belinkov emphasizes the im-

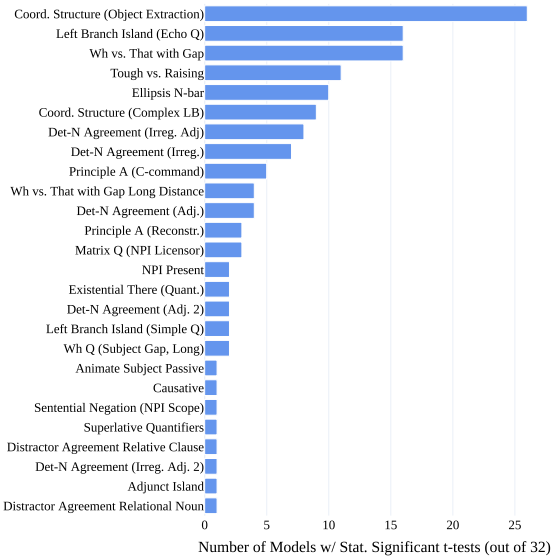

Figure 7: **Sentence-level UUAS score distribution t-test results**. Out of 67 BLiMP paradigms, only the 26 shown here have at least one model that has a statistically significant difference in the UUAS score means.

portance of controls for confounders. We introduce a control probe trained to recover non-syntactic information and demonstrate that it can show correlations with syntactic benchmarks. Second, he notes probe complexity affects interpretation, with simpler probes (e.g., linear) preferred to avoid the probe independently learning the probing task. Our work uses three linear probes with different objective functions and metrics for robustness. Finally, Belinkov warns that probing experiments may conflate properties of the model with properties of the datasets used for training or probing, making it difficult to disentangle the two. We use different datasets for training probes (Penn Treebank) and evaluation (BLiMP), which helps ensure our findings are not artifacts of a single dataset.

While some shortcomings of probing are already known, systematic studies like ours are necessary to thoroughly test hypotheses. Notably, our scope spans a broader range of models and linguistic properties than prior work, including Ravichander et al. (2021), who focus on verb tense, subject number, and object number in Natural Language Inference with pre-transformer architectures, and Elazar et al. (2021), who apply amnesic probing to sequence tagging tasks with BERT (Devlin et al., 2019).

Future Multilingual Study. Just within the last few months, we have seen new work substantially advancing BLiMP for multilingual use cases. The largest cross-lingual development is MultiBLiMP 1.0 (Jumelet et al., 2025), which has over 128,000 minimal pairs across 101 languages specifically for subject–verb agreement. Furthermore, Başar et al. (2025) contribute TurBLiMP, a Turkish minimal pairs benchmark with 16 linguistic phenomena. We emphasize in our work that a model's ability to solve minimal pairs correctly is a desired linguistic property and critical evaluation tool, which Başar et al. (2025) also echo. Much of our evaluation framework is applicable to these new benchmarks: MultiBLiMP uses Universal Dependencies and TurBLiMP includes BLiMP phenomena. As momentum builds towards targeted syntactic analysis for more and lower-resource languages, we believe that future work on our framework towards multilinguality would help identify language-specific probing and downstream performance artifacts. While probing has little predictive power in English, replicating the study could yield novel insights from languages with for instance, richer subject–verb agreement systems than English or greater word order flexibility as seen in Turkish.

6 Conclusion

Targeted evaluations like BLiMP are widely used to benchmark model syntactic knowledge. Given the importance of minimal pairs, our key contribution is to show systematically that probing fails to reliably extract the latent syntactic knowledge within models that correlates with minimal pair accuracy. The high BLiMP scores observed across the open-weight models studied suggest effective learning of syntactic phenomena, and probing task performance is likewise strong. However, across models, probing performance does not align with BLiMP performance, even in fine-grained settings where such alignment might reasonably be expected.

Returning to our "mechanisms vs. outcomes" framework, the mechanism of recovering syntactic knowledge from model activations through linear probing *does not* at all predict targeted outcomes. A common question in probing research is whether the probed representations are actually used to produce model outputs, motivating the line of work on causality. However, our negative finding regarding predictive power preempts this causal question. Extensive research on probing and interpretability more broadly has advanced our knowledge, but a robust understanding of how syntax is represented within models remains to be developed.

7 Limitations

Our evaluation only uses English grammatical benchmarks and could usefully be extended to other languages. Extending our work to multilingual benchmarks would strengthen the study. This would require formalization of a unified evaluation framework to handle differing linguistic phenomena across languages. Moreover, while we design our syntax probes to be lightweight and to learn to extract syntax with different training objectives, they do not encompass all possible syntactic probing tasks or architectural variants. Other probe architectures could yield different results. Due to compute limitations, we also do not train probes on models larger than 8B parameters.

Acknowledgements

The authors would like to thank the anonymous reviewers for constructive feedback during the review period. CM is a fellow in the CIFAR Learning in Machines and Brains program. We thank members of the Stanford NLP Group for their insightful comments and support throughout the project.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The Falcon series of open language models. *Preprint*, arXiv:2311.16867.

Aryaman Arora, Dan Jurafsky, and Christopher Potts. 2024. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14638–14663, Bangkok, Thailand. Association for Computational Linguistics.

Nitin Bansal, Xiaohan Chen, and Zhangyang Wang. 2018. Can we gain more from orthogonality regularizations in training deep CNNs? *Preprint*, arXiv:1810.09102.

Ezgi Başar, Francesca Padovani, Jaap Jumelet, and Arianna Bisazza. 2025. TurBLiMP: A Turkish benchmark of linguistic minimal pairs. *Preprint*, arXiv:2506.13487.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Kevin S. Brown, Eiling Yee, Gitte Joergensen, Melissa Troyer, Elliot Saltzman, Jay Rueckl, James S. Magnuson, and Ken McRae. 2023. Investigating the extent to which distributional semantic models capture a broad range of semantic relations. *Cognitive Science*, 47(5):e13291.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *ICLR 2020*.

Alexis Conneau, German 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 of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Pablo Diego-Simón, Stéphane D'Ascoli, Emmanuel Chemla, Yair Lakretz, and Jean-Rémi King. 2024. A polar coordinate system represents syntax in large language models. *Preprint*, arXiv:2412.05571.

Tiwalayo Eisape, Vineet Gangireddy, Roger Levy, and Yoon Kim. 2022. Probing for incremental parse states in autoregressive language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2801–2813, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. 2022. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7324–7338. PMLR.

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.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, and 24 others. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *Preprint*, arXiv:2306.11644.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Jasper Jian and Siva Reddy. 2023. Syntactic substitutability as unsupervised dependency syntax. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2341–2360, Singapore. Association for Computational Linguistics.

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.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Preprint*, arXiv:2504.02768.

William A. Ladusaw. 1979. *Polarity Sensitivity as Inherent Scope Relations*. Doctoral dissertation, University of Texas, Austin.

Tomasz Limisiewicz and David Mareček. 2021. Introducing orthogonal constraint in structural probes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 428–442, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. 2022. Probing for labeled dependency trees. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7711–7726, Dublin, Ireland. Association for Computational Linguistics.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher D. Manning. 2022. Characterizing intrinsic compositionality in transformers with tree projections. *Preprint*, arXiv:2211.01288.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. 2024. RuBLiMP: Russian benchmark of linguistic minimal pairs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9268–9299, Miami, Florida, USA. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024a. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Gemma Team, 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, and 179 others. 2024b. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

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, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Mycal Tucker, Peng Qian, and Roger Levy. 2021. What if this modified that? Syntactic interventions with counterfactual embeddings. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 862–875, Online. Association for Computational Linguistics.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

A Models

Model	HuggingFace ID
Architecture: Decoder	
GPT-2 124M (Radford et al., 2019)	`gpt2`
GPT-2 345M	`gpt2-medium`
GPT-2 774M	`gpt2-large`
GPT-2 1.6B	`gpt2-xl`
Pythia 410M (Biderman et al., 2023)	`EleutherAI/pythia-410m`
Pythia 1.4B	`EleutherAI/pythia-1.4b`
Pythia 2.8B	`EleutherAI/pythia-2.8b`
Pythia 6.9B	`EleutherAI/pythia-6.9b`
GPT-J 6B (Wang and Komatsuzaki, 2021)	`EleutherAI/gpt-j-6b`
Falcon 7B (Almazrouei et al., 2023)	`tiiuae/falcon-7b`
Llama 3.2 1B (Grattafiori et al., 2024)	`meta-llama/Llama-3.2-1B`
Llama 3.2 3B	`meta-llama/Llama-3.2-3B`
Llama 2 7B (Touvron et al., 2023)	`meta-llama/Llama-2-7b-hf`
Llama 3 8B	`meta-llama/Meta-Llama-3-8B`
Llama 3.1 8B	`meta-llama/Meta-Llama-3.1-8B`
Mistral 7B (Jiang et al., 2023)	`mistralai/Mistral-7B-v0.1`
Gemma 2B (Team et al., 2024a)	`google/gemma-2b`
Gemma 2 2B (Team et al., 2024b)	`google/gemma-2-2b`
OLMo 1B (Groeneveld et al., 2024)	`allenai/OLMo-1B-hf`
OLMo 7B (Groeneveld et al., 2024)	`allenai/OLMo-1.7-7B-hf`
Phi 2 2.7B (Gunasekar et al., 2023)	`microsoft/phi-2`
Architecture: Encoder	
ELECTRA-Small 14M (Clark et al., 2020)	`google/electra-small-discriminator`
ELECTRA-Base 110M	`google/electra-base-discriminator`
ELECTRA-Large 335M	`google/electra-large-discriminator`
RoBERTa 125M (Liu et al., 2019)	`FacebookAI/roberta-base`
RoBERTa 355M	`FacebookAI/roberta-large`
DeBERTa v2 900M (He et al., 2021)	`microsoft/deberta-v2-xlarge`
DeBERTa v2 1.5B	`microsoft/deberta-v2-xxlarge`
Architecture: Encoder-Decoder	
T5 35M (Raffel et al., 2020)	`google-t5/t5-small`
T5 110M	`google-t5/t5-base`
T5 335M	`google-t5/t5-large`
T5 3B	`google-t5/t5-3b`

Table 2: All 32 models used for every experiment.

B Probe Training Details

All training and inference occurs on a single NVIDIA RTX 6000 Ada Generation or NVIDIA RTX A6000 GPU depending on cluster availability. Training probes on the largest models could take up to 4 days to complete. In the Penn Treebank, the training set has 39,831 sentences, validation set 1,700 sentences, and test set 2,416 sentences. In the following sections, we specify training hyperparameters for each probe.

B.1 Structural Probe (g_ϕ^{struct})

- Batch size: 32
- Max epochs: 300
- Early stopping criteria: validation loss; patience of 50 epochs; eval interval of 1 epoch
- Learning rate: $1e^{-4}$; AdamW optimizer
- Learning rate schedule: Linear decay with warmup for 10% of max epochs

B.2 Orthogonal Structural Probe (g_ϕ^{ortho})

- Batch size: 32, 64
- Max epochs: 50
- Early stopping criteria: validation loss; patience of 5 epochs; eval interval of 1 epoch
- Learning rate: $1e^{-4}$; AdamW optimizer

B.3 Head Word Probe (g_ϕ^{head})

- Batch size: 32
- Max epochs: 300
- Early stopping criteria: validation UAS; patience of 50 epochs; eval interval of 1 epoch
- Learning rate: $1e^{-4}$; AdamW optimizer
- Learning rate schedule: Linear decay with warmup for 10% of max epochs

B.4 Control Probe (g_ϕ^{ctrl})

- Batch size: 128
- Max epochs: 300
- Early stopping criteria: validation loss; patience of 50 epochs; eval interval of 1 epoch
- Learning rate: $1e^{-4}$; AdamW optimizer
- Learning rate schedule: Linear decay with warmup for 10% of max epochs

C Control Probe Validation

We prompt GPT-4o to elicit sentences with varied contextual usage for a given word:

> I will give you a word. Generate 10 sentences that use the word in substantially different contexts. Do not capitalize or add any punctuation to the word, but you can add punctuation elsewhere in the sentences. Do not make the word the first word of the sentence; instead it should be near the end! Return only the list of sentences in the format of a Python list without Markdown. Word: {{word}}

Outputs for the words "window" and "flour":

```
"window": [
    "The cat loves to sit by the window
        and watch the birds.",
    "After the update, the software has
        a new window that is more user-
        friendly.",
    "She opened the window to let in
        some fresh air.",
    "There was a brief window of
        opportunity to complete the task
        .",
    "The display showed a colorful
        window of stained glass.",
    "He peered through the window to see
         what was happening inside.",
    "The startup process will initiate a
         pop-up window for further
        instructions.",
    "In the story, the hero jumps out of
         a high window to escape.",
    "The interview was scheduled within
        a tight window of time.",
    "She painted a beautiful scene of a
        sunrise in the window of her art
         studio."
],
"flour": [
    "To bake the cake, you'll need to
        sift the flour thoroughly.",
    "She dusted the countertop with
        flour before kneading the dough
        .",
    "The recipe calls for almond flour
        instead of regular wheat.",
    "A fine layer of flour covered the
        baker's apron after a long day
        .",
    "If you want a crispy crust, coat
        the chicken with flour before
        frying.",
    "They decided to try a gluten-free
        flour for their cookies this
        time.",
    "The bread's texture depends largely
         on the quality of the flour
        used.",
    "After spilling the bag, a cloud of
        flour filled the air in the
        kitchen.",
```

```
    "He couldn't believe how much more
        expensive organic flour was at
        the store.",
    "The children made a mess while
        trying to measure the flour for
        the pancakes."
]
```

D Critical Tree Edges for BLiMP Phenomena

We investigate if the syntax probe's accuracy in predicting a *critical edge* we define on a per-paradigm basis in the acceptable sentence's dependency parse tree is consistent with the model's outcome of assigning the acceptable sentence a higher log probability than the unacceptable one. The critical edge involves the minimal difference in the pair's sentences. In the following sections, we justify the critical edges for the subject–verb agreement and filler–gap paradigms that we experiment with:

- Subject–verb agreement
  - Distractor agreement relational noun
  - Distractor agreement relative clause
- Filler–gap dependencies
  - *Wh.* vs. *that* with gap
  - *Wh.* vs. *that* with gap long distance

D.1 Subject–Verb Agreement

The critical edge is one that connects the relevant subject and the verb targeted by the minimal pair. Consider the following test instance from the *distractor agreement relational noun* suite:

The prints of every vase aggravate Nina. (acc.)
The prints of every vase aggravates Nina. (unacc.)

Agreement on the verb 'aggravate' is determined by its subject, the plural noun 'prints'. This is represented in parse trees as an edge connecting 'prints' and 'aggravate,' labeled `nsubj`. Fig 8 shows the parse tree for the acceptable sentence. If a model's syntactic representations are related to its success on subject–verb agreement, `nsubj` is the critical dependency to evaluate – a model successfully representing this edge means that it is minimally aware of what noun is its subject, and thus, which noun it should use to determine whether 'aggravate' or 'aggravates' should be assigned a higher log probability (see Clark et al., 2019; Jian and Reddy, 2023 for similar discussions of syntactic attention heads).

Under the UD formalism whose parses Stanza produces, the verb which agrees with the subject is not always the head of the `nsubj` relation in a sentence. For example, in Fig 10 the verb which agrees with the subject is the auxiliary 'have', but the UD parse assigns 'alarmed,' which does not agree, as the head. For these critical edge experiments only, we filter out all incongruent cases where UD assigns no edge between the subject and critical verb, which leaves **236** valid minimal pairs for *distractor agreement relational noun* and **345** for *distractor agreement relative clause*.

D.2 Filler-gap

We establish the critical edge for *wh. vs. that with gap* and *wh. vs. that with gap long distance* in the same way. Fig 11 and Fig 12 show sample acceptable parse trees for *wh. vs. that with gap* and *wh. vs. that with gap long distance*. The critical edge is the edge between the *wh*-word and the verb it is the direct (`obj`) or oblique object (`obl`) of; any other case is filtered out. After filtering, there are **972** valid minimal pairs for *wh. vs. that with gap* and **885** for *wh. vs. that with gap long distance*.

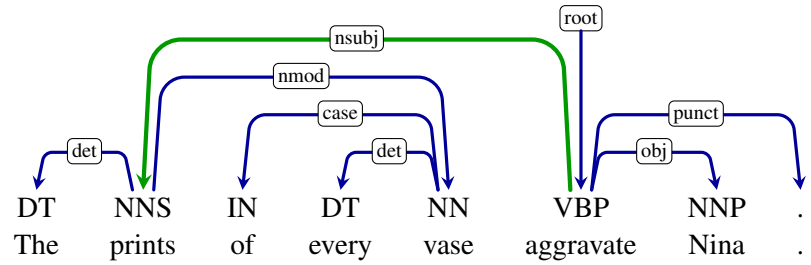

Figure 8: ***Distractor agreement relational noun* sample well-formed sentence Stanza dependency parse tree**. The `nsubj` edge in green is the critical edge.

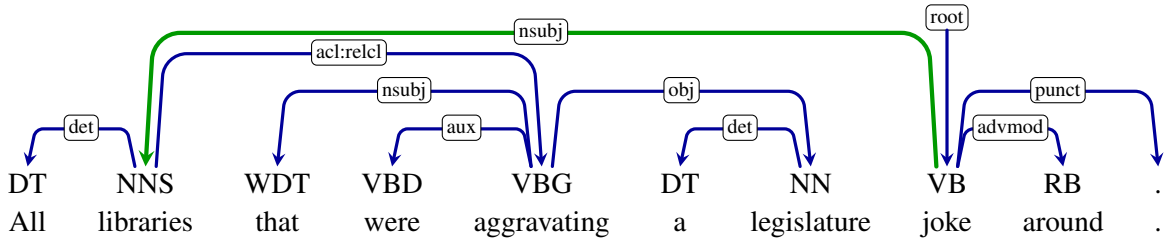

Figure 9: ***Distractor agreement relative clause* sample well-formed sentence Stanza dependency parse tree**. The `nsubj` edge in green is the critical edge.

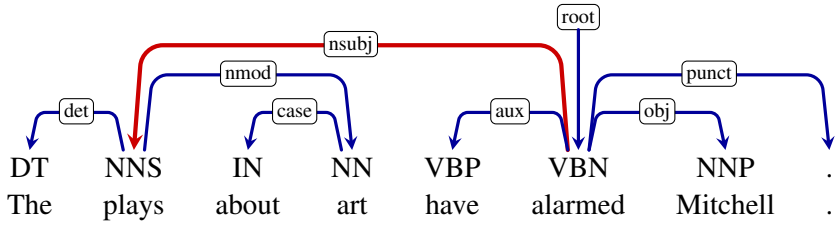

Figure 10: ***Distractor agreement relative clause* sample that we filter out since the critical word is the auxiliary "have", which is not connected to the subject "plays"**.

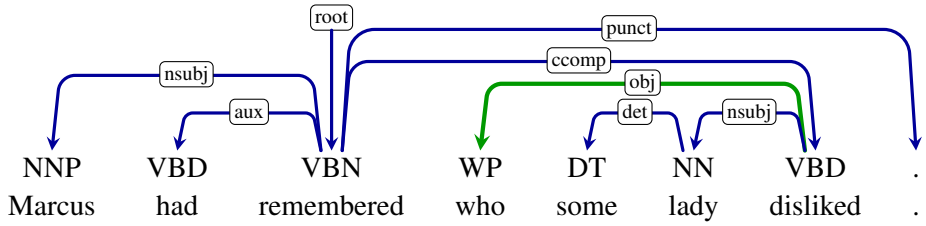

Figure 11: ***Wh. vs. that with gap* sample well-formed sentence Stanza dependency parse tree**. The `obj` edge in green is the critical edge.

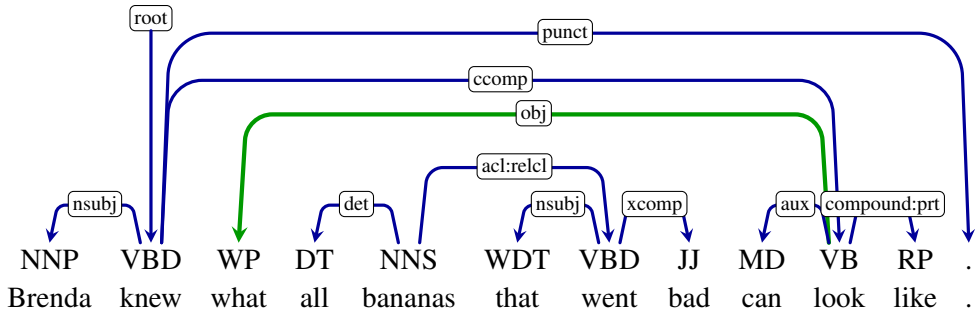

Figure 12: ***Wh. vs. that with gap long distance* sample well-formed sentence Stanza dependency parse tree**. The `obj` edge in green is the critical edge.

E Probe Training and Comparison

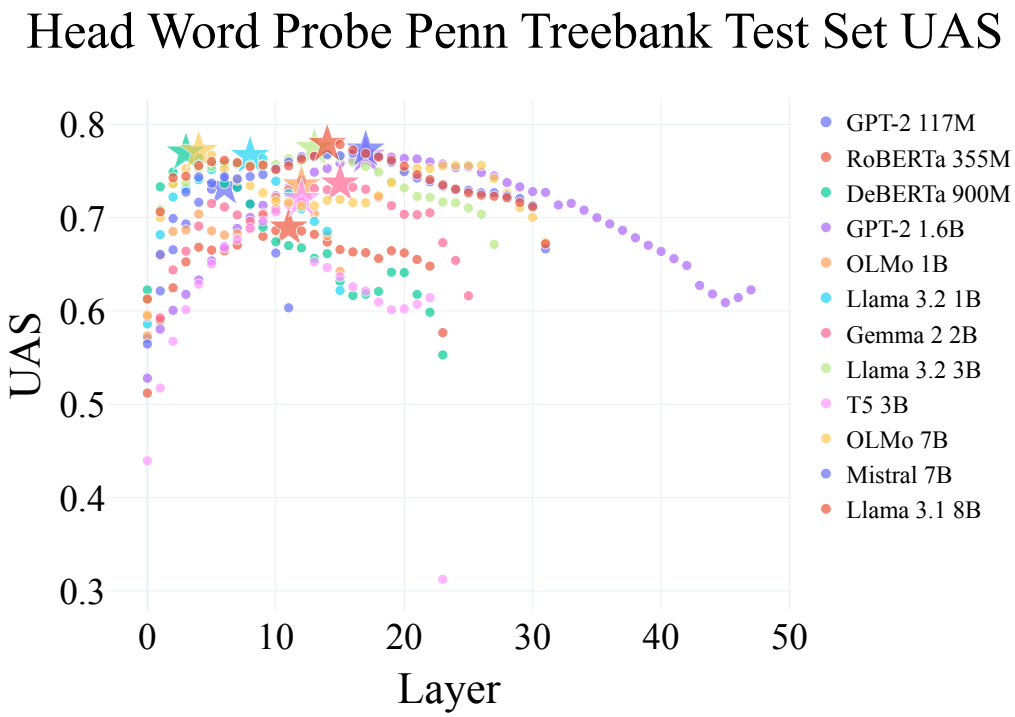

Figure 13: **Penn Treebank test set g_ϕ^{head} UAS for each layer of a sample of our models**. We train a probe on each layer of the model using the train and validation splits and select the probe for the layer with the best test set accuracy – indicated in the plot with a star icon – for BLiMP evaluation. For most models, this occurs in the first half.

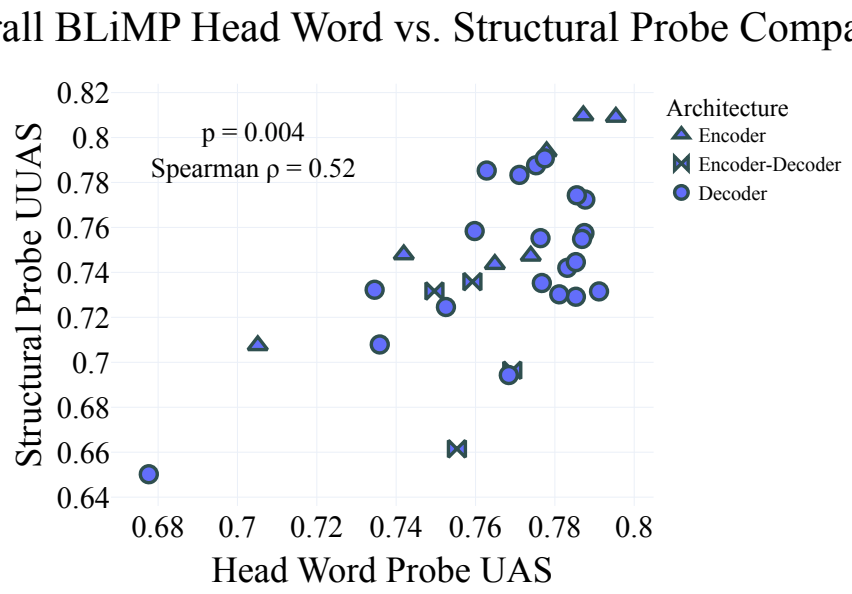

Figure 14: **Correlating g_ϕ^{head} and g_ϕ^{struct} BLiMP attachment scores.** Across all models, g_ϕ^{head} and g_ϕ^{struct} syntax probes show a significant, moderately positive correlation in accuracy averaged across BLiMP paradigms. Despite having an additional directional aspect, g_ϕ^{head} trains similarly to g_ϕ^{struct}.

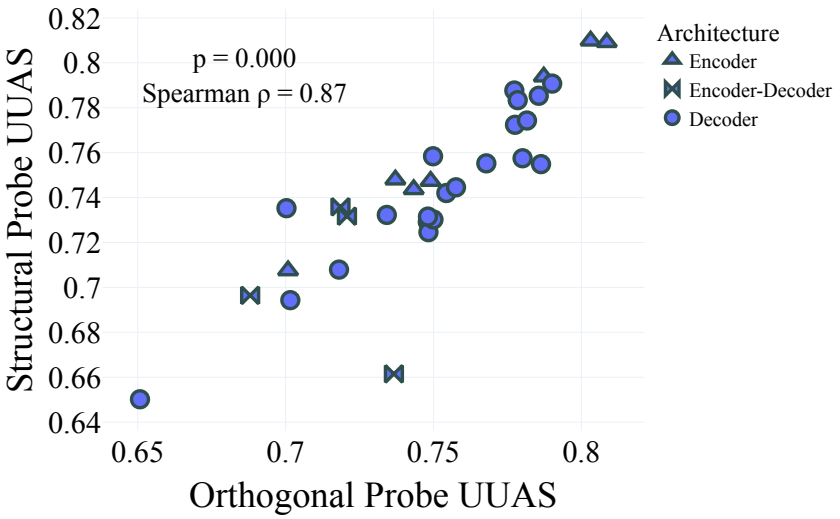

Figure 15: **Correlating g_ϕ^{ortho} and g_ϕ^{struct} BLiMP attachment scores.** Across all models, g_ϕ^{ortho} and g_ϕ^{struct} syntax probes show a strongly positive correlation in accuracy averaged across BLiMP paradigms. Despite the larger capacity of the orthogonal probe, performance largely mirrors its simpler variant, suggesting that the extra dimensions are not critical to learning the dependency tree probing task.

F Full BLiMP and per-Phenomenon OLS Regressions

Dataset	Simple Regression ($x_1 = g_\phi^{\text{struct}}$)			Multiple Regression ($x_1, x_2 = g_\phi^{\text{ctrl}}$)			LRT	
	β_1	p-value	Adj.R^2	β_1	p-value	Adj.R^2	Stat	p-value
Overall	0.133	0.117	0.049	0.109	0.214	0.051	1.161	0.281
Anaphor Agreement	0.165	**0.019**	0.268	0.168	**0.040**	0.243	0.027	1.0
Binding	0.398	0.106	0.178	0.386	0.155	0.155	0.18	1.0
Ellipsis	-0.421	0.106	0.181	-0.426	0.152	0.153	0.011	1.0
NPI Licensing	0.408	0.121	0.165	0.404	0.155	0.143	0.266	1.0
S-Selection	-0.249	0.242	0.125	-0.213	0.374	0.255	6.254	0.149
Island Effects	0.483	0.45	0.087	0.435	0.862	0.065	0.326	1.0
Filler–Gap Dependency	-0.028	1.0	-0.033	0.033	1.0	-0.05	0.546	1.0
Quantifiers	0.267	1.0	0.008	0.253	1.0	-0.019	0.232	1.0
Determiner Noun Agr	-0.063	1.0	-0.013	-0.062	1.0	-0.046	0.067	1.0
Argument Structure	0.198	1.0	0.006	0.164	1.0	0.043	2.295	1.0
Control/Raising	0.118	1.0	0.003	0.101	1.0	0.009	1.295	1.0
Subject–Verb Agr	-0.108	1.0	0.007	-0.047	1.0	0.128	5.236	0.243
Irregular Forms	-0.046	1.0	-0.029	-0.027	1.0	0.277	12.376	**0.006**

(a) $x_1 = g_\phi^{\text{struct}}$ UUAS.

Dataset	Simple Regression ($x_1 = g_\phi^{\text{ortho}}$)			Multiple Regression ($x_1, x_2 = g_\phi^{\text{ctrl}}$)			LRT	
	β_1	p-value	Adj.R^2	β_1	p-value	Adj.R^2	Stat	p-value
Overall	0.127	0.157	0.035	0.102	0.266	0.041	1.299	0.254
Anaphor Agreement	0.173	**0.002**	0.364	0.173	**0.005**	0.342	0.004	1.0
Binding	0.431	0.08	0.195	0.419	0.12	0.173	0.222	1.0
Ellipsis	-0.381	0.297	0.124	-0.376	0.414	0.095	0.01	1.0
NPI Licensing	0.326	0.618	0.082	0.32	0.715	0.057	0.215	1.0
S-Selection	-0.216	0.727	0.068	-0.175	1.0	0.205	6.164	0.156
Filler–Gap Dependency	-0.1	1.0	-0.027	-0.058	1.0	-0.049	0.402	1.0
Quantifiers	0.297	1.0	0.015	0.266	1.0	-0.012	0.21	1.0
Determiner Noun Agr	-0.089	1.0	-0.004	-0.089	1.0	-0.036	0.077	1.0
Argument Structure	0.174	1.0	-0.003	0.132	1.0	0.032	2.24	1.0
Control/Raising	0.165	1.0	0.041	0.153	1.0	0.048	1.304	1.0
Subject–Verb Agr	-0.119	1.0	0.014	-0.059	1.0	0.132	5.183	0.251
Island Effects	0.42	1.0	0.045	0.356	1.0	0.028	0.493	1.0
Irregular Forms	-0.093	1.0	-0.016	-0.078	1.0	0.287	12.45	**0.005**

(b) $x_1 = g_\phi^{\text{ortho}}$ UUAS.

Dataset	Simple Regression ($x_1 = g_\phi^{\text{head}}$)			Multiple Regression ($x_1, x_2 = g_\phi^{\text{ctrl}}$)			LRT	
	β_1	p-value	Adj.R^2	β_1	p-value	Adj.R^2	Stat	p-value
Overall	0.0	0.998	-0.033	-0.048	0.700	0.091	5.194	**0.023**
Binding	0.534	0.064	0.209	0.522	0.108	0.184	0.085	1.0
S-Selection	-0.2	1.0	-0.007	-0.255	1.0	0.068	3.553	0.654
NPI Licensing	0.25	1.0	0.006	0.233	1.0	-0.001	0.882	1.0
Filler–Gap Dependency	-0.125	1.0	-0.028	-0.15	1.0	-0.055	0.256	1.0
Quantifiers	0.434	1.0	0.045	0.385	1.0	0.1	2.98	0.843
Determiner Noun Agr	-0.096	1.0	-0.013	-0.098	1.0	-0.046	0.037	1.0
Ellipsis	-0.109	1.0	-0.02	-0.1	1.0	-0.053	0.087	1.0
Argument Structure	0.025	1.0	-0.033	-0.037	1.0	0.076	4.658	0.402
Control/Raising	0.241	1.0	0.047	0.201	1.0	0.076	2.083	1.0
Subject–Verb Agr	-0.205	1.0	0.041	-0.221	1.0	0.025	0.56	1.0
Island Effects	0.392	1.0	-0.001	0.397	1.0	-0.035	0.025	1.0
Anaphor Agreement	0.102	1.0	0.005	0.096	1.0	-0.027	0.069	1.0
Irregular Forms	-0.120	1.0	-0.026	-0.099	1.0	0.069	4.217	0.48

(c) $x_1 = g_\phi^{\text{head}}$ UAS. That 12 out of 13 phenomena have p-value 1.0 is a striking result.

Table 3: **Comparison of simple and multiple regression statistics for all BLiMP linguistic phenomena.** Per-phenomenon rows are sorted in order of increasing simple regression corrected p-value. Phenomena show substantial variation in strength of association. Zeroing into Subtable 3a, we note the adjusted R^2 jumps for s-selection and irregular forms after adding the control. The significant p-value of the irregular forms LRT concurs with Fig 5.

Dataset	Simple Regression ($x_1 = g_\phi^{\text{ctrl}}$)		
	β_1	p-value	Adj. R^2
Overall	0.127	0.165	0.032
Irregular Forms	0.287	**0.009**	0.299
S-Selection	-0.445	0.141	0.167
Subject–Verb Agr	-0.343	0.178	0.151
Binding	0.118	1.0	-0.009
NPI Licensing	0.086	1.0	-0.022
Filler–Gap Dependency	-0.193	1.0	-0.016
Quantifiers	0.232	1.0	-0.015
Determiner Noun Agr	-0.036	1.0	-0.031
Ellipsis	-0.134	1.0	-0.018
Argument Structure	0.208	1.0	0.048
Control/Raising	0.161	1.0	0.016
Island Effects	0.202	1.0	0.01
Anaphor Agreement	0.025	1.0	0.005

Table 4: **Simple linear regression results for the structural probe control (trained on the best layers for g_ϕ^{struct})** $x_1 = g_\phi^{\text{ctrl}}$ ρ_s. Per-phenomenon rows are sorted in order of increasing corrected p-value. Irregular forms has statistical significance and substantially higher R^2 than most of the syntax probe simple regressions for any phenomena (Table 3a and Table 3c). This critical result shows that even a non-syntactic control can achieve a significant positive association with one syntactic benchmark, which thus urges caution when interpreting the statistical signifiance of anaphor agreement for g_ϕ^{struct}.

Dataset	Simple Regression ($x_1 = g_\phi^{\text{ctrl}}$)		
	β_1	p-value	Adj. R^2
Overall	0.043	**0.031**	0.117
Argument Structure	0.085	0.5	0.106
Irregular Forms	0.078	0.568	0.096
Quantifiers	0.143	0.865	0.069
S-Selection	0.043	1.0	0.057
Binding	0.038	1.0	-0.007
NPI Licensing	0.038	1.0	-0.001
Filler–Gap Dependency	0.027	1.0	-0.027
Determiner Noun Agr	0.002	1.0	-0.033
Ellipsis	-0.023	1.0	-0.028
Control/Raising	0.049	1.0	0.053
Subject–Verb Agr	0.008	1.0	-0.026
Island Effects	-0.004	1.0	-0.033
Anaphor Agreement	0.004	1.0	-0.025

Table 5: **Simple linear regression results for the head word probe control (trained on the best layers for g_ϕ^{head})** $x_1 = g_\phi^{\text{ctrl}}$ ρ_s. Per-phenomenon rows are sorted in order of increasing simple regression corrected p-value. While all phenomena have extremely low or near-zero explanatory power that does not achieve statistical significance, the fit on the full dataset is significant but has near-zero effect size.

G Subject–Verb Agreement & Filler–Gap Experiments

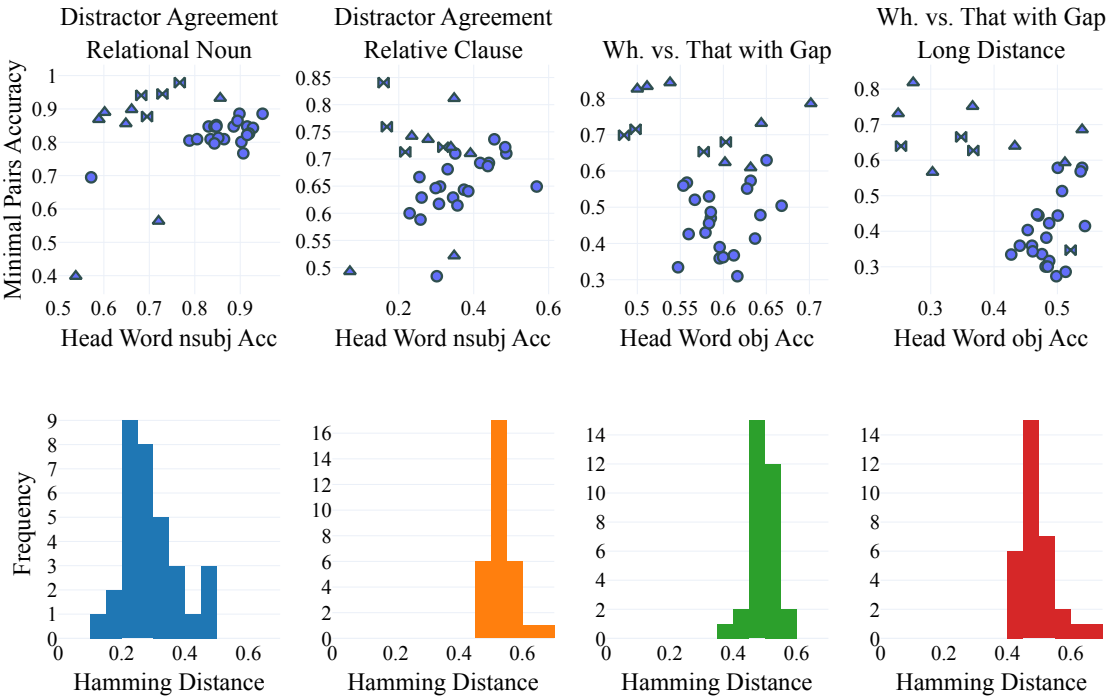

Figure 16: Analogous to Fig 6 but for the head word probe. Conclusion is similar.