Causal Drawbridges: Characterizing Gradient Blocking of Syntactic Islands in Transformer LMs

Sasha Boguraev **Kyle Mahowald**
The University of Texas at Austin
{sasha.boguraev,kyle}@utexas.edu

Abstract

We show how causal interventions in Transformer models provide insights into English syntax by focusing on a long-standing challenge for syntactic theory: syntactic islands. Extraction from coordinated verb phrases is often degraded, yet acceptability varies gradiently with lexical content (e.g., "I know what he hates art and loves" vs. "I know what he looked down and saw"). We show that modern Transformer language models replicate human judgments across this gradient. Using causal interventions that isolate functionally relevant subspaces in Transformer blocks, attention modules, and MLPs, we demonstrate that extraction from coordination islands engages the same filler–gap mechanisms as canonical *wh*-dependencies, but that these mechanisms are selectively blocked to varying degrees. By projecting a large corpus of unrelated text onto these causally identified subspaces, we derive a novel linguistic hypothesis: the conjunction "and" is represented differently in extractable versus non-extractable constructions, corresponding to expressions encoding relational dependencies versus purely conjunctive uses. These results illustrate how mechanistic interpretability can inform syntax, generating new hypotheses about linguistic representation and processing.

1 Introduction

Our aim in this work is to study how, mechanistically, Language Models (LMs) handle *syntactic islands*, as a paradigmatic case of a complex and gradient linguistic phenomenon that is nonetheless learned by these models. In addition to giving empirical evidence for fine-grained similarity in how LMs and humans process the construction, our main goal is to uncover the mechanism by which these phenomena are handled in LMs. We do so using causal intervention techniques to tell the story of how these phenomena are handled.

An enduring question in human language research is how human learners come to master long-distance filler-gap constructions like (1). Contrast this with the gapless example in (2). (1) is considered a filler-gap construction since there is a gap (represented by the blank) that is licensed by (and in a sense "filled by") the *wh*- word "what".

(1) I know [what]$_t$ she smiled at $__{}_t$. (2) I know [that] she smiled at it.

Sometimes, though, this filler-gap mechanism is blocked, leaving the sentence unacceptable (or at least questionably acceptable). Consider:

(3) **?** I know [what]$_t$ she hates art and loves $__{}_t$.

This has a semantic interpretation, but syntactically sounds off. This construction is called a syntactic island since the gap is "stranded" (Ross, 1967). The linguist Cedric Boeckx writes: "Most modern grammarians would agree with me if I said that island effects are perhaps the most important empirical finding in modern theoretical linguistics" (Boeckx, 2012). There are several reasons for all the interest. We posit that the reasons that make islands linguistically interesting also make them interesting for interpretability more generally.

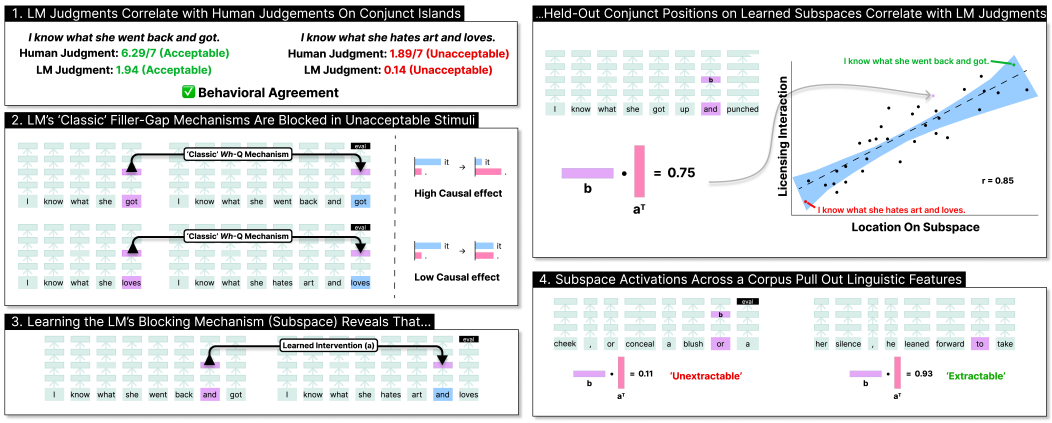


Figure 1: **1.** LM judgments of gradiently acceptable conjuncts correlate with human judgments. **2.** These constructions rely on a more general-purpose filler-gap mechanism, blocked in the island case. **3.** We identify the relevant blocking subspaces, showing they correlate with model acceptability, and **4.** pick out linguistically meaningful structures in a corpus.

First, linguistic theory suggests that filler-gap dependencies require rich abstract structure, to keep track of long-distance dependencies. To know how to end (1), you have to keep track that there was a gap-licensing *wh*-element ("what"). Second, the *blocking* of the filler-gap mechanism, as in the island case, requires learning what constructions block it. Here, it's the verb-phrase (VP) conjunct "hates art and loves ___" that forms the island.

Third, depending on the content of the VP conjunct, the filler-gap mechanism is not actually blocked. People typically accept (4), suggesting that the strength of the island is dependent on the lexical content of the conjunct – making these constructions a case where high-level syntactic generalization depends on lower-level lexical information.

(4) I know [what]$_t$ she looked down and saw $__{}_t$.

Fourth, filler-gaps are rare, and it's claimed children receive little positive evidence that syntactic islands are disallowed. Yet all adult English speakers learn as much. Filler-gaps are thus taken as examples *par excellence* of the Poverty of Stimulus problem (Pearl, 2022).

As a result of this state of affairs, Boeckx writes: "The sophisticated analyses [...] put forth in order to capture island effects offer perhaps the strongest case for a rich, abstract, domain-specific mental module for language" (Boeckx, 2012).

Given that, it would certainly be interesting if we could study how a fairly general-purpose learner trained on a finite amount of data processed these constructions. Our results suggest that, for the specific case of coordination islands illustrated here, modern LMs provide such a system. First, we show that LMs are gradiently sensitive, reflecting human judgments, to coordination islands both in the general case and in ways that vary with the lexical items. Second, we use mechanistic interpretability techniques, specifically causal interventions, to show extraction from VP conjuncts is handled by classical filler-gap mechanisms which get blocked in the island case. Third, we show we can identify the "drawbridge" mechanism necessary for stranding and unstranding syntactic islands. Finally, we use the identified subspaces to find other similar activations in a larger corpus of text. We use this method to motivate the hypothesis that the conjunction in extractable and non-extractable conjuncts have subtly lexically distinct representations—a claim we argue could be tested in humans.

2 Background

2.1 Syntactic Island Constraints

Our focus is on islands that arise from the Coordinate-Structure-Constraint (CSC), i.e., 'do not extract from a single conjunct of a coordinated structure' (Ross, 1967), as in (3).

While Ross admits violations to various island constraints, the broad class of island effects are often described as universal constraints on language (Ross, 1967; Chomsky, 1977). From this observation and analyses of the linguistic input, a Poverty of the Stimulus argument follows, suggesting that if a learner only encounters positive evidence ('acceptable' sentences) throughout development, they would be unable to learn these constraints without an innate linguistic faculty (Chomsky et al., 1973; Phillips, 2013; Pearl, 2022).

Yet there is also a large body of evidence arguing for gradience in these effects and/or sensitivity to non-syntactic factors (see, e.g., Kluender, 1991; Ambridge and Goldberg, 2008; Abeillé et al., 2020; Liu et al., 2022a;b; Namboodiripad et al., 2022; Cuneo and Goldberg, 2023; Fergus et al., 2025, i.a.), such as cognitive processing (Kluender and Kutas, 1993; Hofmeister et al., 2013) or discourse effects (Kehler, 2002; Goldberg, 2006; 2011; Abeillé et al., 2020).

Our work builds upon that of Fergus et al. (2025), whose materials and human ratings we use. They showed that while the CSC contests that one cannot extract the argument of a single conjunct within a verb phrase (correctly predicting (3) as ungrammatical), across a set of 46 such coordinated verb phrases, human acceptability judgments are gradient. They suggest the gradience is explained by the degree to which the coordinated VP represents a single complex state that matters, but also note the potential effect of the first verb's transitivity – many of the more acceptable variants have intransitive verbs as the first verb.

2.2 LMs and the Filler-Gap Task

As interest grew in whether LMs can process complex syntax, a cottage industry emerged assessing whether they could handle long-distance constructions (Linzen et al., 2016) and, in particular, filler-gaps and island constraints (e.g., Wilcox et al., 2018; Futrell et al., 2019; Warstadt et al., 2020; Kobzeva et al., 2023; Wilcox et al., 2023). While success was mixed on early LMs, today's LMs successfully handle these constructions (cf. Lan et al., 2024).

As a result, they have also been the locus for work on neural interpretability, with a focus on *how* they do so. In particular, there has been a focus on whether the filler-gap mechanism is shared across linguistically related long-distance constructions (Prasad et al., 2019; Bhattacharya and van Schijndel, 2020; Howitt et al., 2024; Lan et al., 2024; Boguraev et al., 2025).

Unlike this prior work, our goal is not to study how coordination islands relate to other constructions, but to zoom in on the construction with as much granularity as we can. We want to understand how the filler-gap mechanism operates in these constructions, how it is blocked in the island setting, and how it is unblocked in sentences that look superficially like islands but are not, as in Fergus et al. (2025). To do that, we separately consider the Output of Transformer Blocks and Attention Modules, as well as MLP Activations – further examining how these hidden features correspond to linguistic features. Our goal in doing so is in part to understand how the particulars of the Transformer architecture are relevant in this task (as in Elhage et al., 2021; Wang et al., 2023, *i.a.*).

3 General Methods

3.1 Data

We focus on verb-phrase conjuncts – and corresponding human-ratings – collected by Fergus et al. (2025). Their stimuli span 46 conjuncts, with Likert scale ratings from 1-7. Ratings are sourced from 442 participants with about twenty individuals judging each conjunct.

These stimuli are in the form of matrix *wh*-questions, making corresponding minimal pairs difficult to generate. As such, we transform each conjunct into the corresponding embedded *wh*-questions. We adapt the resulting data to sentential templates à la Arora et al. (2024) and Boguraev et al. (2025) for ease of sampling large numbers of unique minimally-paired data (see §A for details). We filter out all such stimuli in which the extracted *wh*-item is not *what* for consistency and those with continuations after the extracted object due to inherent limitations of autoregressive LMs. This leaves us with 33 minimally-paired templates.

3.2 Behavioral Analysis

To measure how robustly an LM represents extraction for given stimuli, we follow the methodology of Wilcox et al. (2018). Specifically, we get minimal sentence pairs: *wh*, which contains a *wh*-licensor, and *th*, which does not. They have corresponding labels l_{wh} (gap) and l_{th} (no gap). We then compute the model's surprisal, S, for each label given each sentence.

This 2×2 setup allows calculation of the gap-less interaction: the surprisal difference of a gap-less continuation for a sentence with a *wh*-licensor, $S(l_{th}|wh)$, and without a licensor, $S(l_{th}|th)$. It also allows us to calculate the gap interaction: the surprisal difference for a gap continuation when comparing sentences with a *wh*-licensor, $S(l_{wh}|wh)$, to those without one, $S(l_{wh}|th)$. If an LM robustly represents a long-distance dependency, the gap-less interaction, $S(l_{th}|wh) - S(l_{th}|th)$, should be positive, and the gap interaction, $S(l_{wh}|wh) - S(l_{wh}|th)$, negative. The overall *wh*-licensing term is then the difference between these effects: $(S(l_{th}|wh) - S(l_{th}|th)) - (S(l_{wh}|wh) - S(l_{wh}|th))$. Higher values indicate more salient representations of the long-distance dependency. A schematic of this method is in Figure 2.[1]

Island Violation (Extractable)

	it	.
I know what she looked down and saw...	High	Low
I know that she looked down and saw...	Low	High
Surprisal Difference	Positive	Negative
Wh-Licensing	Strongly Positive	

Conjunct Island (Unextractable)

	it	.
I know what she drank whiskey and played...	High	High
I know that she drank whiskey and played...	Low	High
Surprisal Difference	Positive	Neutral
Wh-Licensing	Weakly Positive	

Figure 2: Exemplar *wh*-licensing calculations for extractable (top) and unextractable (bottom) conjuncts respectively.

3.3 Interpretability Analysis

Distributed Alignment Search To localize the internal mechanisms utilized by LMs to process these conjuncts, we use Distributed Alignment Search (DAS; Geiger et al., 2024), consistent with prior work on syntactic interpretability (Arora et al., 2024; Boguraev et al., 2025). DAS is a supervised interpretability method which finds d-dimensional subspaces in an LM causally responsible for specified behavior. In particular, DAS acts on a base value $b \in \mathbb{R}^n$ – the tensor at a given internal site when the model processes an input (e.g., Example 1) – and corresponding source value $s \in \mathbb{R}^n$ – the tensor at the same site when the model processes a minimally-paired input (e.g., Example 2). The intervention performed by DAS is

$$\mathbf{b} + (\mathbf{s}\mathbf{a}^\top - \mathbf{b}\mathbf{a}^\top)\mathbf{a}$$

where $a \in \mathbb{R}^{n \times d}$ is a rotation learned through gradient-descent on the cross-entropy loss of the LM's prediction under intervention. Intuitively, DAS aims to find subspaces on which intervening from s to b maximizes the corresponding prediction, l_s. We learn 1D interventions, localizing features in three sites: (1) the output of Transformer blocks, (2) the output of attention-heads before head-mixing (i.e., an LM's OV-circuit outputs; Elhage et al., 2021) and (3) MLP activations. We choose these sites to probe how LMs' different sub-components process these syntactic phenomena in addition to their joint processing.

Training We train DAS at each position of our minimal pairs and layer of our LMs. For every intervention, we also learn a control intervention trained on the same minimal pairs but flipped labels. By comparing efficacies, we can evaluate the degree to which our interventions represent the LMs true mechanisms, as opposed to merely the separability of the labels. We train 5 seeds for each intervention. For more training details see §B.

Evaluation We evaluate interventions with the **ODDS** metric introduced by Arora et al. (2024). This metric measures the increase in probability of the intervened label after an intervention, relative to the decrease in probability of the expected label. Higher **ODDS** indicate higher causal efficacy. We report the average Δ**ODDS**, that is, the difference between our critical condition **ODDS** and the corresponding control condition, averaged across seeds.

[1]Code for replicating experiments is at `https://github.com/SashaBoguraev/causal-drawbridges`.

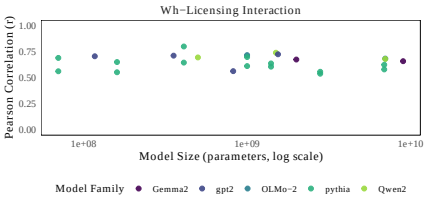
4 Exp. 1: LMs Agree with Human Judgments of Coordination Islands

We first examine whether LMs share human gradient acceptability judgments across filler-gap sentences with conjoined verb phrases.

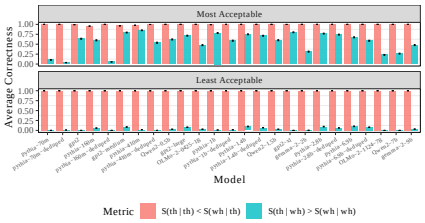
Methods We sample 400 minimal pairs of each conjunct and calculate the mean *wh*-interaction across pairs as described in §3.2. We then measure the correlation (Pearson *r*-value) between human judgments and an LM's mean *wh*-interaction. We test 25 LMs across 5 model families: `OLMo2` (OLMo et al., 2025), `Qwen2` (Yang et al., 2024), `Gemma 2` (Team et al., 2024), `gpt2` (Radford et al., 2019) and `pythia` (Biderman et al., 2023), testing a host of different sizes from 70 million to 9 billion parameters.

Results Figure 3a shows strong correlation between LM *wh*-licensing and human acceptability judgments (between 0.54 and 0.80). Figure 3b further plots the proportion that an LM's surprisal for a given stimuli's correct label is less than the mismatched label, averaged across the four most and least acceptable conjuncts (as judged by people).

Our *wh*-licensing results suggest LM's judgments broadly agree with human judgments. That is, generally, the more acceptable an extraction is deemed by humans, the more saliently LMs represent the long-distance dependency. This is reinforced by our binary comparison results. LMs consistently judge that for a *th* stimuli, l_{th} is more probable than l_{wh}, whether or not the minimally-paired *wh* stimuli is acceptable or not (i.e., the red bars are high in both facets). Further, most LMs correctly display a preference for extraction in the extractable *wh* case, and much less so in the non-extractable case.



(a) Pearson *r* correlations between LM *wh*-licensing interaction and human acceptability judgments, shown as a function of model size. Full scatter plots are in §F.



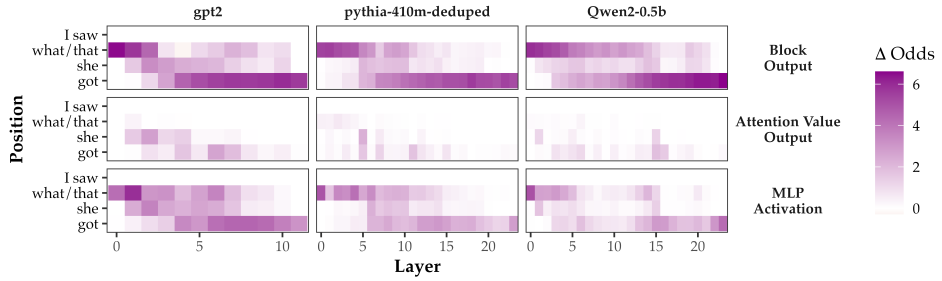
(b) LMs' predicting of the correct label given a stimulus for the four most- and least-extractable conjuncts (measured by human acceptability), ordered left-to-right by size. All models perform well in the gapless case (red), while for the gap cases, performance is gradiently related to extractability in the expected direction (green).

Figure 3: LM behavioral metrics.

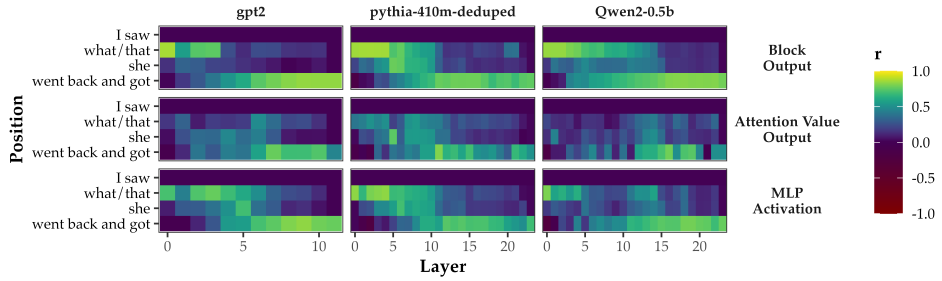
Discussion Our results provide evidence that LMs process our coordination island stimuli in a manner that agrees behaviorally with human judgments. Interestingly, this correlation with human judgments does not increase monotonically with scale. Correlations range narrowly between roughly 0.55 and 0.80 across model sizes, with as much variation across model families at a given size as across sizes. This suggests that factors other than raw parameter count may better predict how closely a model's behavior tracks human gradient judgments, though we leave a systematic investigation of this to future work. In the next sections, we turn our eye towards developing a mechanistic understanding of exactly how LMs come to these judgments. We focus our analysis on three LMs with strong behavioral correlation: `gpt2`, `pythia-410m-deduped`, and `Qwen2-0.5b`.

5 Exp. 2: LMs Process Conjuncts with 'Classic' Filler-Gap Mechanisms

To find the mechanisms LMs use to process extraction (or lack thereof) in these conjuncts, we use the methodology of Boguraev et al. (2025) to probe generalization between constructions. That is, we train DAS on non-conjunct embedded *wh*-question sentences (e.g., Example 5) to find interventions causally responsible for mediating extraction. We then measure the causal efficacy of these interventions on our conjunct stimuli (e.g, Example 6). The more efficacious these interventions, the stronger the evidence for shared processing mechanisms.



(a) **ODDS** for learned embedded *wh*-question interventions evaluated on a held-out test set. Strong causal efficacy suggests DAS has robustly recovered the filler-gap mechanism.



(b) Pearson *r* correlation between **ODDS** for 'classic' filler-gap interventions performed on conjunct stimuli and mean licensing-score. We see high correlation, suggesting greater generalization of mechanisms to 'extractable' conjuncts.

Figure 4: We use DAS to find the mechanisms to process embedded *wh*-questions before measuring the generalization of said mechanisms to our conjunct stimuli.

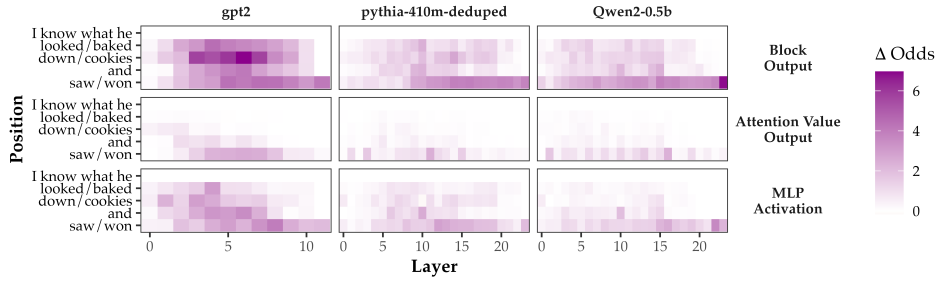
(5) a. I saw **what** she got ⇒ **.**
b. I saw **that** she got ⇒ **the**

(6) a. I saw **what** she went back and got ⇒ **.**
b. I saw **that** she went back and got ⇒ **the**

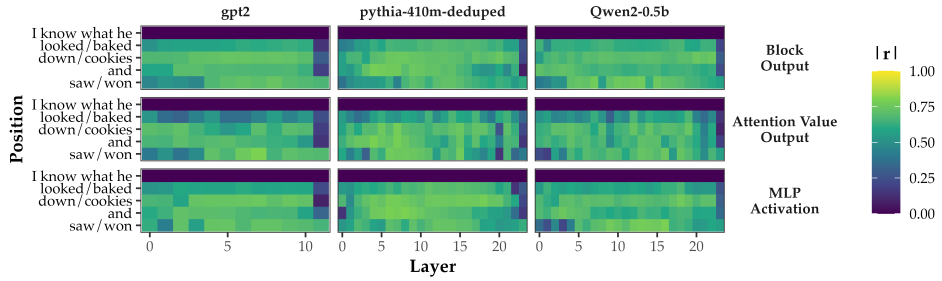
Methods We sample training sets of embedded *wh*-questions without coordinated verb-phrases and train DAS as described in §3.3. We first validate these alignments on a held-out test set of 100 minimal pairs sampled from the same distribution as the training set before evaluating them on 100 minimal pairs for each of our 33 conjuncts to see how much the mechanisms generalize. We then compute the mean correlation across seeds between the resulting **ODDS** for each conjunct and the corresponding *wh*-licensing.

Results Figure 4a illustrates the **ODDS** of the learned 'classic' filler-gap interventions when evaluated on a held-out test set. Across all three intervention sites, we see strong causal efficacy – suggesting our interventions robustly capture the mechanisms of 'classic' filler-gap extraction. We further see that MLPs are involved more ubiquitously in processing, with the attention modules showing strong causal effect at sites preceding MLP efficacy. This suggests that the attention modules are generally responsible for moving information (consistent with Elhage et al., 2021; Meng et al., 2022), with the MLPs responsible for linguistic processing of that moved information. Figure 4b further shows high correlation between *wh*-interaction score and **ODDS** across all positions of the conjuncts in which we see high **ODDS**. We further note that, in all such positions, correlations are significant. And there is low variance across seeds (exact details in §D).

Discussion Taken together, our results demonstrate that the more saliently extraction is represented by an LM, the more 'classic' filler-gap mechanisms generalize to it. That is, when an LM predicts extraction from a VP conjunct, it makes that prediction through the same pathways it uses to process extraction from 'classic' embedded *wh*-questions. However, in cases in which the LM does not predict extraction from a coordinated verb-phrase, this 'classic' mechanism gets blocked, only permitting extraction to a gradient extent.



(a) **ODDS** for learned interventions at each position of the conjuncts and layer of every LM when evaluated on a held-out test set. We see strong causal efficacy suggesting DAS has robustly discovered the associated blocking mechanism.



(b) Absolute correlation between each conjunct's average position along the learned subspaces and *wh*-licensing. We see strong generalization of these mechanisms to all conjuncts along the gradients of behavior, despite being trained on a subset of conjuncts.

Figure 5: We use DAS to find the causal drawbridges responsible for stranding and unstranding coordination islands.

6 Exp. 3: LM Judgments are Modulated by Causal Drawbridges

To find the mechanisms responsible for LMs' gradient behavior, we again use DAS. Specifically, we learn interventions that can causally manipulate an LM to predict extraction from unextractable stimuli. In doing so, we aim to find the causal drawbridges which strand and unstrand these syntactic islands. Recognizing that faithful mechanisms must explain the LMs' full behavior, we evaluate these learned interventions on held-out, gradiently acceptable conjuncts. The degree to which these mechanisms reflect gradient behavior serves as a measure for faithfulness.

Methods We utilize a set of 16 conjuncts: 8 deemed highly acceptable by both humans and LMs and 8 similarly deemed highly unacceptable. A full list of conjuncts, and acceptability ratings, are in §E. For a given experiment, we randomly partition these into 6 conjuncts to train on and 2 for held-out evaluation, yielding different training and validation sets for each seed. We then sample 400 minimal pairs of our train conjuncts, differing in the coordinated VP from which extraction is occurring, and train DAS at each position. Example 7 shows exemplar minimal pairs. To evaluate, we sample 100 minimal pairs from the held-out conjuncts on which to validate the efficacy of our interventions.

(7) a. I saw what she **looked down and saw** ⇒ **.**
b. *I saw what she **baked cookies and won** ⇒ **the**

DAS minimal pairs are typically sampled with each class of minimal pair as base and source during training. However, due to one of our minimal pairs being necessarily ungrammatical (i.e., Example 7b has no grammatical continuation), we fix our 'unextractable' stimuli as our base and 'extractable' variants as our source. In doing so, DAS learns subspaces which cause models to predict 'unacceptable' stimuli to be 'acceptable'.

To measure whether these subspaces faithfully represent an LM's gradient behavior, we sample evaluation sets of 100 stimuli for all conjuncts, including those not seen during training, and measure their position on these subspaces. If our interventions faithfully represent the LM's 'blocking' mechanisms, not only would 'acceptable' and 'unacceptable' stimuli fall on the subspace's extremes, but the gradient items would fall correspondingly in the middle. To this end, we measure the absolute correlation between a conjunct's subspace position and its *wh*-licensing interaction. A high correlation constitutes evidence that our interventions faithfully represent the mechanisms responsible for an LM's gradience. We measure the absolute correlation because DAS only optimizes the subspace such that swapping activations swaps outputs, meaning the direction it learns can be oriented either way (acceptable more positive or more negative). Absolute correlation duly captures strong alignment regardless of sign.

Our hypothesis is that the mechanism for blocking will be primarily localized to the second verb, since that is the point at which the extractability becomes clear.

Results Our results suggest these interventions capture LMs' gradient behavior. That is, Figure 5a shows the interventions capture causally efficacious information flow through the LMs and Figure 5b suggests these interventions generalize across conjuncts. Interestingly we observe causal effects (to varying extents) starting at the *first* verb position, suggesting relevant information is available as soon as the first verb. Post-hoc, we believe this is because the first verb in extractable conjuncts tend to be qualitatively different (e.g., less transitive, semantically lighter) than in unextractable conjuncts. To provide empirical support for these differences, we train a logistic regression on GLoVe embeddings (Pennington et al., 2014) of the first verb and the first VP (averaged) to predict extractability. Using leave-one-out validation across the 16 train conjuncts, we find accuracy of 75% and 100% respectively, suggesting the verbs (and complements) have meaningfully different character.

We also see qualitatively different mechanisms across LMs. That is, while `pythia-410m-deduped` and `Qwen2-0.5b` show weak signal until the final position, `gpt2` shows strong efficacy in all positions from the first verb's complementizer all the way through the last verb. This suggests that `gpt2` is more heavily reliant on cues from the first VP. Further, that we see these mechanisms early in the conjuncts and that they are strongly explanatory of the gradient behavior (high absolute r across positions) suggests that the LMs may employ incremental processing from early on in the conjunct, with the different relative efficacies suggesting that this is employed to different degrees. This mirrors psycholinguistic accounts of incremental processing, where verbs constrain downstream reference resolution (Altmann and Kamide, 1999) and could be tested in humans using eye-tracking experiments testing how the lexical semantics of the first VP constrain predictions on extraction.

Discussion These results show evidence of causal drawbridges: mechanisms that modulate the gradient acceptability of filler-gap extractions. In the next section, we show these causal drawbridges can help identify related constructions, providing linguistic insights.[2]

7 Exp. 4: Learned Subspaces Represent Linguistic Features

So far, we have found the abstract mechanisms (i.e., subspaces in an LM's latent space) responsible for gradient extraction across conjuncts. However, this has revealed few linguistic insights. Our last experiment explores the linguistic features encoded in such subspaces.

Methods We focus on qualitative analyses of the subspaces learned in Section 6, which constitute the 'blocking mechanism'. To do so, we chunk the Gutenberg Corpus (an open-source English corpus; Project Gutenberg, n.d.) into segments matching the length of our stimuli (8 tokens). We then randomly select 100,000 chunks, collect the activations of an LM processing those chunks, and project those activations onto the relevant subspaces. By analyzing the chunks' positions within these subspaces, we aim to gain insight into what properties they encode and why they are causally relevant for LM behavior.

[2]For similar results and discussion correlating subspace location with human acceptability see §F.

Position	Layer	ΔODDS	$\lvert r \rvert$	'Extractable' Examples	'Unextractable' Examples
AND	6	5.03	0.67	...his life he had crept up **and** stolen...	...you men, we offer you meat **and** bread...
				...valleys, so that men look up **and** see...	...the groom's books **and** papers...
				...I think we're in **for** a...	...offers on the best terms for cash **or** approved...
				...it in motion, she started up **and** said...	...chief took a spoonful of meat **and** liquor...
				...ieved in and preached what he had **just** heard...	...as pigs, fowl, dogs**,** cattle...
				...ourd Wood, we sat down **to** make...	...a. I hab no wife **nor** child...
				...dozen willing hands reached down **to** catch...	...A broken chair, a stool **or** two...

Table 1: Exemplar chunks activated near 'extractable' and 'unextractable' conjuncts in the Transformer Block Output of `gpt2`. Full lists of top 75 such chunks at this site are in §H.

Our method is akin to work associating features with SAEs and Transcoders through qualitative analysis of a corpus' top activating examples (Cunningham et al., 2023; Bricken et al., 2023; Dunefsky et al., 2024). The difference is that other methods find features which take on a value in a range from "activated" to "not activated", which have been trained in an unsupervised manner to represent singular features. In contrast, our method identifies gradable features defining causally-relevant subspaces found through supervised training.

We focus on `gpt2`, specifically on a highly efficacious intervention in the Transformer Block Output at the AND position. While the second verb is not yet available at this site, the results in §6 show that information relevant for extraction is already available. This experiment is an attempt to understand which linguistic features the LM is tracking. We do not claim to have reverse-engineered the full blocking mechanism, leaving this to future work. To that end, we release all chunks and corresponding values at every site of every model.[3] We stress that this analysis is qualitative and noisy: the subspace was learned to discriminate extractable from unextractable conjuncts, but we cannot directly observe what linguistic features drive that discrimination, nor rule out surface correlates such as transitivity.

Results We see many chunks with *and*, likely given the lexical nature of our conjuncts. However, the *and* tokens near the 'extractable' and 'unextractable' items possess different senses. In particular, the *and* near extractable items seems akin to an intra-event connective, conjoining verb-phrases (such as *crept up and stolen*, *look up and see*...). Conversely, in the 'unextractable' examples, the *and* functions as a logical conjunction, conjoining nominal entities or conceptual categories (e.g., *meat and bread*, *books and papers*...).

However, it is particularly instructive to examine activated "non-and" tokens. We see a trend in which tokens fall where on the subspaces, with *to*, acting as a purpose marker, *for*, expressing a causal consequent, and *just*, marking immediate precedence, activated in the 'extractable' case. And in contrast, in the 'unextractable' case, logical disjunctions like *or*, *nor*, and conjunctive punctuation are activated.

Discussion These results highlight a novel hypothesis: in unextractable cases the conjunction behaves akin to a logical conjunction, but in extractable cases it takes on the character of function words encoding relational dependencies. This hypothesis is well situated in the literature. First, LMs are known to be highly sensitive to polysemous usages (Chronis and Erk, 2020), grounding the plausibility of this distinction. Secondly, this distinction tracks with prior work in theoretical linguistics aimed at explaining CSC violations. Ross himself noticed the asynchrony between extractable and unextractable conjuncts and proposed that those permitting extraction, such as *went to the store and bought*, were derived from the underlying structure *went to the store to buy* (Ross, 1967), and are not in fact conjunctions at all, merely resembling them on the surface. While this analysis has been convincingly argued against (Schmerling, 1975), it captures similar intuitions to ours, namely that the conjunction in these cases functions not as purely logical conjunction, but as a purposive or sequential marker. Ross, however, commits strongly to this distinction being a categorical, syntactic one, whereas our analysis suggests this is a gradable property of the conjunction's lexical semantics. Separately, Kehler (2002) classifies VP coordination into Parallel relations, where the two coordinated VPs are unrelated, Occasion relations, in which the two VPs are in natural sequence and Cause-Effect relations in which the first VP causes the second VP. He posits that extraction is not permitted from a Parallel relation but is permitted from

[3]Chunks and accompanying visualizer can be found at this link.

Occasion and Cause-Effect relations. And accordingly, we see the 'extractable' chunks are these such permissible relations, whereas the 'unextractable' chunks are not.

We encourage future psycholinguistic work testing this hypothesis, e.g, syntactic priming paradigms measuring whether people process phrases like *went out and got* more akin to activated phrases like *riding up to him* than classical conjuncts like *drank whiskey and ate*. We are optimistic our results point to further work on the lexical semantics of conjunctions.

8 Conclusion

There have been recent claims that LMs aid in developing insights into linguistic structure and representations (e.g., Futrell and Mahowald, 2025). Here, we hope to have delivered some possible insights of this nature. We showed that a variety of LMs show human-like sensitivity to the gradient acceptability of supposed syntactic islands, and that these constructions are handled by a more general filler-gap mechanism. We then show that we can use causal interventions to understand the drawbridge mechanism that controls the gradience of acceptability. We end by providing evidence that these interventions can help generate novel linguistic insights.

9 Limitations

We claim that we generate linguistic insights: to see if those insights generalize to humans, human experiments would be helpful. We suggest possible directions for human experiments, but such experiments were outside of the scope of this project.

We focus on English constructions here but note that some of the importance of syntactic islands in linguistic history depends on their generalizability across languages – particularly in the case of the Coordinate Structure Constraint which is claimed to be remarkably robust multi-lingually. Thus, extending this work to non-English languages would enhance its generalizability, particularly given differential performance of models on filler-gaps in non-English languages (Kobzeva et al., 2023).

We recognize that the models we tested see more data than human learners see, and should thus moderate any claims about Poverty of Stimulus arguments. At the same time, we also note that the models we test are smaller and weaker than frontier models and so these results do not necessarily reflect either the performance or internal analyses of state-of-the-art LMs, which could differ both in behavioral performance and mechanism.

10 LLM Usage Statement

We used Claude to provide feedback on the clarity of our writing during the editing process. We further used Claude to aid in the development of the code used to perform Experiment 4, and to do post-hoc optimization of the code written for Experiments 2 and 3 which allowed for faster experimentation.

11 Acknowledgments

We thank Abigail Fergus and Adele Goldberg for their LSA presentation that introduced us to their data set and for making their data available. We also thank Andy Kehler for his insights and discussion on an initial version of this work. We acknowledge funding from NSF CAREER grant 2339729 to Kyle Mahowald.

References

Anne Abeillé, Barbara Hemforth, Elodie Winckel, and Edward Gibson. Extraction from subjects: Differences in acceptability depend on the discourse function of the construction.

Cognition, 204:104293, 2020.

Gerry TM Altmann and Yuki Kamide. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3):247–264, 1999.

Ben Ambridge and Adele E Goldberg. The island status of clausal complements: Evidence in favor of an information structure explanation. *Cognitive Linguistics*, 19(3):357–389, 2008.

Aryaman Arora, Dan Jurafsky, and Christopher Potts. 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, 2024.

Debasmita Bhattacharya and Marten van Schijndel. Filler-gaps that neural networks fail to generalize. In Raquel Fernández and Tal Linzen, editors, *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 486–495, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.conll-1.39. URL https://aclanthology.org/2020.conll-1.39/.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023. URL https://arxiv.org/abs/2304.01373.

Cedric Boeckx. *Syntactic Islands*. Cambridge University Press, 2012.

Sasha Boguraev, Christopher Potts, and Kyle Mahowald. Causal interventions reveal shared structure across English filler–gap constructions. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25032–25053, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1271. URL https://aclanthology.org/2025.emnlp-main.1271/.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features.

Noam Chomsky. On wh-movement. In Peter Culicover, Thomas Wasow, and Adrian Akmajian, editors, *Formal Syntax*, pages 71–132. Academic Press, New York, 1977.

Noam Chomsky, Stephen Anderson, and Paul Kiparsky. Conditions on transformations. *1973*, 232:286, 1973.

Gabriella Chronis and Katrin Erk. When is a bishop not like a rook? when it's like a rabbi! multi-prototype BERT embeddings for estimating semantic relationships. In Raquel Fernández and Tal Linzen, editors, *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 227–244, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.conll-1.17. URL https://aclanthology.org/2020.conll-1.17/.

Nicole Cuneo and Adele E Goldberg. The discourse functions of grammatical constructions explain an enduring syntactic puzzle. *Cognition*, 240:105563, 2023.

Hoagy Cunningham, Aidan Ewart, Logan Riggs Smith, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *ArXiv*, abs/2309.08600, 2023. URL https://api.semanticscholar.org/CorpusID:261934663.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable llm feature circuits. *Advances in Neural Information Processing Systems*, 37:24375–24410, 2024.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Abigail Fergus, Arielle Belluck, Nicole Cuneo, and Adele Goldberg. Islands result from clash of functions: Single-conjunct wh-qs. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47, 2025.

Richard Futrell and Kyle Mahowald. How linguistics learned to stop worrying and love the language models. *Behavioral and Brain Sciences*, page 1–98, 2025. doi: 10.1017/S0140525X2510112X.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. Neural language models as psycholinguistic subjects: Representations of syntactic state. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1004. URL https://aclanthology.org/N19-1004.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pages 160–187. PMLR, 2024.

Adele E Goldberg. *Constructions at work: The nature of generalization in language*. OUP Oxford, 2006.

Adele E Goldberg. Backgrounded constituents cannot be "extracted". In *Experimental syntax and island effects*, pages 221–238. Cambridge University Press, 2011.

Philip Hofmeister, Laura Staum Casasanto, and Ivan A. Sag. Islands in the grammar? standards of evidence. In Jon Sprouse and Norbert Hornstein, editors, *Experimental Syntax and Island Effects*, pages 42–63. Cambridge University Press, Cambridge, 2013.

Katherine Howitt, Sathvik Nair, Allison Dods, and Robert Melvin Hopkins. Generalizations across filler-gap dependencies in neural language models. In Libby Barak and Malihe Alikhani, editors, *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 269–279, Miami, FL, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.conll-1.21. URL https://aclanthology.org/2024.conll-1.21/.

Andrew Kehler. *Coherence, reference, and the theory of grammar*, volume 380. CSLI publications Stanford, CA, 2002.

Robert Kluender and Marta Kutas. Subjacency as a processing phenomenon. *Language and Cognitive Processes*, 8(4):573–633, 1993.

Robert E Kluender. *Cognitive constraints on variables in syntax*. University of California, San Diego, 1991.

Anastasia Kobzeva, Suhas Arehalli, Tal Linzen, and Dave Kush. Neural networks can learn patterns of island-insensitivity in norwegian. In *Proceedings of the Society for Computation in Linguistics 2023*, pages 175–185, 2023.

Tatsuki Kuribayashi, Yohei Oseki, Souhaib Ben Taieb, Kentaro Inui, and Timothy Baldwin. Large language models are human-like internally. *Transactions of the Association for Computational Linguistics*, 13:1743–1766, 2025.

Nur Lan, Emmanuel Chemla, and Roni Katzir. Bridging the empirical-theoretical gap in neural network formal language learning using minimum description length. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13198–13210, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.713.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016.

Yingtong Liu, Rachel Ryskin, Richard Futrell, and Edward Gibson. A verb-frame frequency account of constraints on long-distance dependencies in english. *Cognition*, 222:104902, 2022a.

Yingtong Liu, Elodie Winckel, Anne Abeillé, Barbara Hemforth, and Edward Gibson. Structural, functional, and processing perspectives on linguistic island effects. *Annual Review of Linguistics*, 8:495–525, 2022b.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Kanishka Misra. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*, 2022.

Savithry Namboodiripad, Nicole Cuneo, Mathew A Kramer, Yourdanis Sedarous, Felicia Bisnath, Yushi Sugimoto, and Adele Goldberg. Backgroundedness predicts island status of non-finite adjuncts in english. In *Proceedings of the annual meeting of the cognitive science society*, volume 44, 2022.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2025. URL https://arxiv.org/abs/2501.00656.

Lisa Pearl. Poverty of the stimulus without tears. *Language Learning and Development*, 18(4): 415–454, 2022.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162/.

Colin Phillips. On the nature of island constraints ii: Language learning and innateness. *Experimental syntax and island effects*, pages 132–157, 2013.

Grusha Prasad, Marten van Schijndel, and Tal Linzen. Using priming to uncover the organization of syntactic representations in neural language models. In Mohit Bansal and Aline Villavicencio, editors, *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 66–76, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/K19-1007. URL https://aclanthology.org/K19-1007/.

Project Gutenberg. Project gutenberg. https://www.gutenberg.org, n.d. Retrieved December 3, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

John Robert Ross. *Constraints on Variables in Syntax*. PhD thesis, MIT, Cambridge, MA, 1967.

Susan F Schmerling. Asymmetric conjunction and rules of conversation. In *Speech acts*, pages 211–231. Brill, 1975.

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, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

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

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. What do RNN language models learn about filler–gap dependencies? In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi, editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing*

and Interpreting Neural Networks for NLP, pages 211–221, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5423. URL https://aclanthology.org/W18-5423/.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. Using computational models to test syntactic learnability. *Linguistic Inquiry*, pages 1–44, 2023.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020. URL https://arxiv.org/abs/1910.03771.

Zhengxuan Wu, Atticus Geiger, Aryaman Arora, Jing Huang, Zheng Wang, Noah Goodman, Christopher Manning, and Christopher Potts. pyvene: A library for understanding and improving PyTorch models via interventions. In Kai-Wei Chang, Annie Lee, and Nazneen Rajani, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 158–165, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-demo.16. URL https://aclanthology.org/2024.naacl-demo.16/.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

A Stimuli Conversion

The stimuli we source from Fergus et al. (2025) are in the form of matrix wh-questions. This makes minimal pair templates difficult to form. As such, we convert them into embedded wh-questions. Below, we detail this conversion process.

(8) What did she bake cookies and win?

Given the stimuli in (8), we convert it into the corresponding embedded question, as in (9).

(9) I know what she baked cookies and won → .

To facilitate this, we make the stimuli past-tense as well. The example in (9) then serves as one of our minimal pairs (in particular the one with a gap), with the corresponding minimal pair shown in (10).

(10) I know that she baked cookies and won → the

This is just one such minimal pair sample from our templates, with the prefix (in this pair '*I know*') and subject ('*she*') sampled from a large set of potential items to allow for diverse minimal pairs.

We note that such changes have the potential to induce different argument structures, affecting the licensing. However, we believe these effects to be minimal, and expect the acceptability judgments collected by Fergus et al. (2025) to serve as reliable judgments for the embedded *wh*-questions.

B Training and Evaluation Details

We access all LMs used in this study through the `transformers` python package (Wolf et al., 2020). In order to calculate LM surprisals for our stimuli, we utilize the `minicons` library (Misra, 2022). To train DAS and leverage the resulting interventions, we use the `pyvene` library (Wu et al., 2024).

When training DAS, we train for one epoch on 400 unique minimal pairs, utilizing a batch size of 4, and gradients are accumulated over 4 batches per step. We use the Adam optimizer with a learning rate of $5e-3$ and apply 100 warm-up steps. In the case of token length mismatch of aligned stimuli, we train interventions on the last token of a given span.

All experiments were run on 1 NVIDIA A40 GPU. Behavioral experiments take approximately 8 hours for all models.

When training DAS on `gpt2` each intervention takes approximately 5 seconds per intervention, totaling approximately 20 minutes and 25 minutes per seed for Experiment Two and Three respectively. Training DAS on `pythia-410m-deduped` takes approximately 12 seconds per intervention, totaling approximately 55 minutes and 75 minutes per seed for Experiment Two and Three respectively. Finally, for `Qwen2-0.5b` each intervention takes approximately 17 seconds, totaling approximately 100 minutes per seed.

Evaluation for Experiment Two takes approximately 30 seconds per intervention for `gpt2`, totaling approximately 45 minutes per seed, 40 seconds per intervention for `pythia-410m-deduped` totaling approximately 150 minutes per seed, and 60 seconds per intervention for `Qwen2-0.5b` totaling approximately 210 minutes per seed.

Evaluation for Experiment Three takes approximately 20 seconds per intervention for `gpt2`, totaling approximately 45 minutes per seed, 20 seconds per intervention for `pythia-410m-deduped` totaling approximately 90 minutes per seed, and 40 seconds per intervention for `Qwen2-0.5b` totaling approximately 180 minutes per seed.

Evaluating the chunks in Experiment 4 takes approximately 20 seconds per intervention, totaling approximately 45 minutes per seed for `gpt2` and 90 minutes per seed for `pythia-410m-deduped` and `Qwen2-0.5b`.

C Human and LM Behavioral Correlation

In Figure 6 we provide a full set of scatter plots for each LM displaying the given LM mean licensing interaction plotted as a function of human acceptability rating for each conjunct in our stimuli. We further include the reported Pearson r value for each conjunct.

D Descriptive Statistics for Correlation Heatmaps

In §5 and §6, we report the raw correlations between **ODDS** and *wh*-licensing and the position on the learned DAS subspace and *wh*-licensing respectively. In this section we provide further descriptive statistics. In particular for each set of correlations, we provide tables with explicit β and significance values in Tables 4 and 5 and variance across seeds in Tables 6 and 7.

E DAS Blocking Mechanism Training Data

A full list of conjuncts used to train DAS to find the blocking mechanism, along with their respective human acceptability ratings and LM *wh*-licensing scores, can be seen in Table 2.

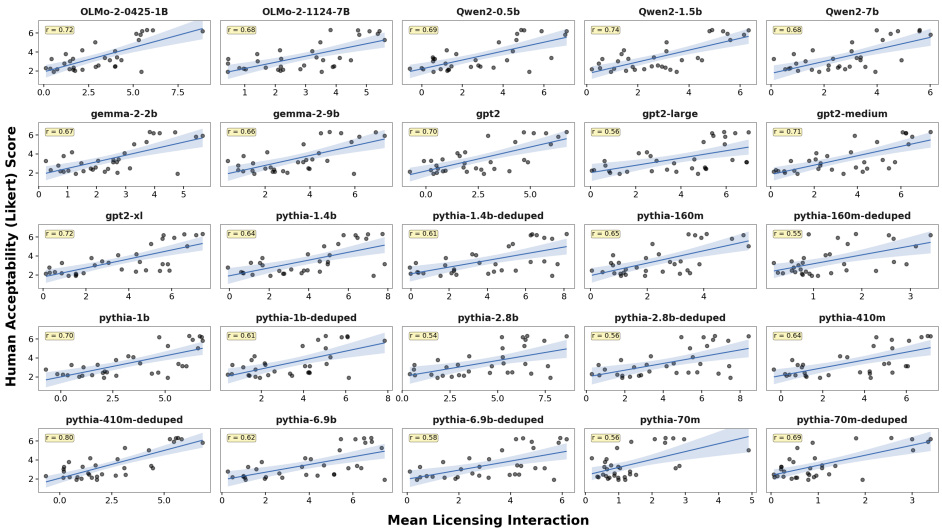


Figure 6: Correlation between LM mean licensing interaction and human acceptability judgments across embedded *wh*-questions with verb-phrase conjuncts. Across a broad class of LM families and sizes, we largely find agreement (high Pearson r correlation coefficient) between the two measures.

Verb$_1$	Complement	And	Verb$_2$	Human Acceptability	gpt2 Mean *Wh*-Licensing	pythia-410m-ded. Mean *Wh*-Licensing	Qwen2-0.5b Mean *Wh*-Licensing	Class
looked	down	and	saw	6.29	6.75	5.68	4.90	Acceptable
went	home	and	got	6.26	4.70	5.58	5.04	Acceptable
woke	up	and	smelled	6.16	5.76	5.85	7.13	Acceptable
came	back	and	talked about	6.17	5.17	5.29	5.89	Acceptable
ran	out	and	got	5.89	2.99	5.48	4.69	Acceptable
woke	up	and	thought of	5.79	6.35	6.86	7.06	Acceptable
drove	to the market	and	bought	5.24	4.29	3.13	4.66	Acceptable
sat	back	and	enjoyed	5.00	5.99	4.28	3.80	Acceptable
picked	strawberries	and	wrote	1.89	0.37	0.82	0.59	Unacceptable
loved	painting	and	enrolled in	2.00	0.53	1.05	1.27	Unacceptable
baked	cookies	and	won	2.16	0.83	0.86	0.08	Unacceptable
ate	candy	and	developed	2.11	-0.76	0.19	1.34	Unacceptable
read	a magazine	and	enrolled in	2.11	1.97	1.55	1.31	Unacceptable
played	soccer	and	planted	2.22	0.95	1.80	-0.57	Unacceptable
called	his boyfriend	and	painted	2.28	0.25	-0.65	0.01	Unacceptable
cooked	dinner	and	read	2.76	0.51	0.21	1.17	Unacceptable

Table 2: **Left Block:** The coordinated clause of the coordination-islands used to train DAS on the blocking mechanism. The top items are 'acceptable' coordination-islands (strong preference for extraction), and the bottom four 'unacceptable' (strong lack of preference for extraction). **Right Block**: Human Acceptability scores along a Likert scale and mean licensing-scores for each LM we mechanistically evaluate. We use conjuncts which **a)** broadly demonstrate calibration between human and LM judgments and **b)** have length-matched tokens.

F Subspaces in LM Hidden Spaces Correlate with Human Judgments

We replicate the analysis of the blocking mechanism as in Section 6, but instead correlate conjunct position on the subspace with human acceptability judgments rather than LM *wh*-licensing score.

Our results can be seen in Figure 7. Interestingly, we find extremely strong correlations across the majority of positions – notably higher than the correlation with LM *wh*-licensing score shown in Figure 4b. While we do not investigate these results further, we find them

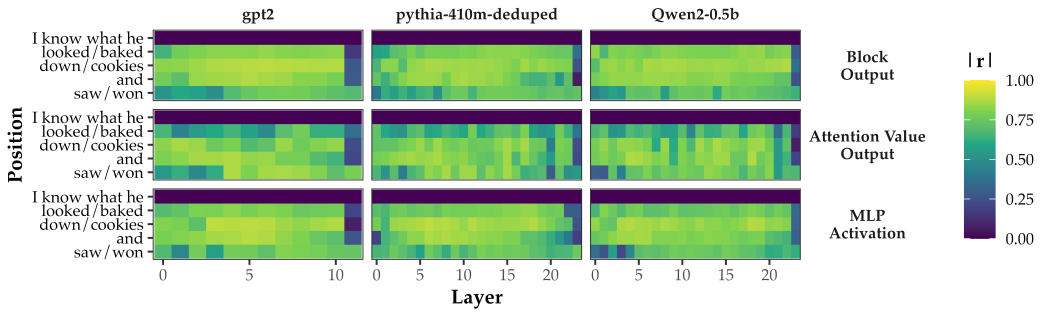


Figure 7: Absolute correlation between each conjunct's average position along the learned subspaces (on which an intervention is performed) and human judgments. We see strong correlation between position on subspace and human judgments.

extremely exciting, and in line with prior work which has shown that LLM internals align strongly with human behavior (Kuribayashi et al., 2025). We see these as preliminary results suggesting that there are subspaces, discoverable through methods like DAS, which strongly correlate LM internals with human behavior. Future work can build upon these results, discovering further behavior correlating with LMs internal subspaces and localizing where specifically in the models this correlation occurs.

G Chunk Extraction Protocol

As discussed in the main text for Experiment 3 report the absolute Pearson-r in our results due to the fact that the position on a subspace can correlate either positively or negatively with *wh*-licensing. This is because the interventions are merely optimized such that intervening on the subspace changes the model's output accordingly, and this is agnostic to which side of a subspace each class of a minimal pair gets seperated to. Due to this fact, it is non-trivial to take the top-activating chunks averaged over the five seeds. Our solution to this problem is once we have calculated the correlation with model *wh*-licensing, we multiply all activations by the sign of that correlation, before normalizing by z-score. We then average across the resulting normalized activations.

H Top-Activating Chunks

We provide the top 75 most activated chunks at the given position we study, averaged across the five trained sets of interventions for `gpt2` at the AND positions in Table 3 . Note that we filter out the chunks for which the token activated at a given position was whitespace for heightened clarity of results.

	Top 'Extractable' Examples	Top 'Unextractable' Examples
1	\nhis life he had crept up **and** stolen	you men, we offer you meat **and** bread
2	orsyth and Ronald were walking back **and** forth	partly to discover the places where gold **and**\r
3	her room she saw him pacing back **and** forth	the eating of animal food, wine **and** marriage
4	quilt. Her fingers went in **and** out	\nyoung men how he threw spears **and** har
5	\nUneasy wanders to **and** fro	the groom's books **and** papers
6	\r\n\r\nSo far, **so** clear	the most nutritive of all meat **and**\r
7	! Churn!\r\nUp **and** down	a lot of old muskets **and** ten
8	think, one has to be up **and**\r	ast assortment of duodec bombs **and** other
9	valleys, so that men look up **and** see	representing tomatoes, radishes, apples, p
10	cher and his lanky friend entered **and** draped	and Mère Marie found some pots **and**\r
11	. Perhaps some one will come down **and** show	\n sort of grammar **and** sense
12	\n\r\nbeer the world has **ever** known	\r\n\r\ndesirable dogs **and** his
13	it in motion, she started up **and** said	baking of the squaw corn cakes **and** the
14	When the village was passed I stood **and** looked	an egg, a bit of bacon, ordered
15	\r\n\r\nreached up **and** tried	whatever about the different kinds of trees **and** he
16	she had hoped. He got up **and** carried	not tell you how glad my wife **and** sons
17	!"\r\nThe Lieutenant staggered out **and** almost	of heroes **and** law
18	At last, at last we turned **and** stood	me, he clutched both rifle **and** belt
19	\nMother Tiger, then he went **and** hid	of them left his Bible, cane **and** hat
20	to wander\r\n\r\nup **and** down	which I put some wool **and** moss
21	O and the T up-and-down	\nmountain. They used spells **and** magic
22	ishing their hot stew! Far down **and** nearer	arkan, built of whalebone **and** rubber
23	village was passed I stood and looked **back**.	\r\n\r\n\r\nFood **and** the
24	if he wakes at midnight he rises **and** begins	vapor of iodine, bromine, or
25	Only."\r\nI went to **bed** at	\r\n\r\nWisdom literature **and** Phil
26	again next moment, and I got **hold** of	are constantly spoken of as heretics **and** false
27	I think we're in **for** a	ers on the best terms for cash **or** approved
28	\r\n\r\ncame forth **and** stood	!' And she brought wood **and**\r
29	thought\r\n\r\nwould cease **and** determine	soja beans, beggarweed,\r
30	\r\nto youth, leading through **and** beyond	\ntwice as much real education **and** mental
31	\n\r\nthey all sat down **and** waited	my breast **and** yelled
32	\r\nMr. McCool goes **and** gives	\r\nshould be given nearby oak **and** h
33	The green posters were distributed far **and** wide	handles, broom-sticks, **and** s
34	and shivering, they stumbled outside **and** by	the roar of its hundred heavy guns **and** mort
35	ieved in and preached what he had **just** heard	\ngrain, and introducing clover **and** new
36	corn for you.\r\nCome **and** eat	17.A great deal of rain **and** very
37	\r\n\r\nTO THE IND**IF**FER	four-lobed calyx **and** sometimes
38	please.\r\n"There **you** are	ur Gilson. Take the food **and**\r
39	. Antwerp will fall to-morrow	\n\r\nbloated flesh **and** redd
40	\r\nThis was refreshing doctrine to **come** from	ponderous, helpless knowledge of books, for
41	for the thought.\r\nNow **and** then	spoonful of beef-tea,\r
42	'm getting paid for." He paused **and** removed	\n\r\nthe poor Princes **and** put
43	\ndispatch, the Press got **hold** of	chief took a spoonful of meat **and** liquor
44	list in her hand, which she **proceeded** to	\r\n\r\nmade admirable sailors **and** excellent
45	in despair.\r\nShe turned **round**.	as pigs, fowl, dogs, cattle
46	essential."\r\n"You have,**"** agreed	-dress of dark-blue cloth **and** a
47	turned his brain that he went home **and** determined	subsistence. Turtles, crabs, p
48	.\r\n"What are you **going** to	.\r\n\r\nBirds **and** flowers
49	\r\nattention. There is, first	mean an inordinate desire for drink, a
50	unselfish!' She turned, **and** caught	a. I hab no wife **nor** child
51	nd ed. © 15Sep50; A	A broken chair, a stool **or** two
52	place."\r\nHe looked up, and	, a gold watch-chain, **and** a
53	\nwhen his followers came riding up **to** him	\n\r\n"Only more hills **and** more
54	to England–Letters of End**ym**ion	deep in grass, with ivy **and** b
55	own concerns, and leave him und**ist**urbed	otton and linen rags, hemp, woods
56	\nWhen the men and women came **home** that	Lyons, the country of silks **and**\r
57	\r\n\r\nI have little **to** say	\r\nby smooth flint stones **and** gains
58	saw the dark bodies whirled round **and** round	\n\r\nsuffered from hunger **and** thirst
59	aham. © 20Mar47; A	\n grain **and** fed
60	. I had to go to **bed** for	\nevery day. They were carrots, of
61	end of the still figure and turned **back** the	might have been worse, the linen **and** plate
62	ourd Wood, we sat down **to** make	\r\ndrops of bromine, the
63	ed cadence art could ne'**er** attain	greater importance in education, the classics **or**\r
64	—I thought when I came down **to**\r	ism, anticipating the spirit of Christianity **and** the
65	\nAnd the next day he came **back**,	\nmeteorite-screens **and** wall
66	.\r\nSimon said unent**hus**i	geons, apothecaries, **and** physicians
67	Autry. © 30Aug50; AA	\r\nwith Indians, Kurds, **and** a
68	\nThe Miss Sahiba paused, **and** all	out on prejudice or anti-British **or**\r
69	. She opened the gate, and **came** up	bromine, iodus, **and** i
70	Wellard. © 26Jan50; A	of\r\n\r\nsoldiers **and** of
71	-sacrifice. There came **to** men	indulge hope for the church **and** the
72	of the time. I didn't	body there depended a decomposition blaster **and** an
73	my wife, Fritz had gone down **to** his	of hairline, width of cheeks **and** height
74	by Pa-khar or Ne**hes**i	\n\r\nthe hum of insects **and** the
75	\r\ndozen willing hands reached down **to** catch	\r\nMothers and Wives **and** House

Table 3: Top 75 Activating 'Extractable' and 'Unextractable' Chunks for GPT2 on the Subspace Learned at the Transformer Block Output of Layer 6. The relevant token is highlighted.

	Block Output			Attention Value Output			MLP Activation		
Layer	WH/THAT	SHE	VP_2	WH/THAT	SHE	VP_2	WH/THAT	SHE	VP_2
GPT2									
0	**0.87*****	**0.15*****	0.02	**0.16*****	0.04	**−0.07***	**0.71*****	**0.09****	**−0.05***
1	**0.56*****	**0.32*****	**0.17*****	**0.32*****	**0.25*****	**0.12*****	**0.43*****	**0.29*****	−0.05
2	**0.74*****	**0.30*****	**0.28*****	**0.08*****	**0.35*****	**0.39*****	**0.73*****	**0.29*****	**0.11***
3	**0.75*****	**0.21*****	**0.40*****	**0.18*****	**0.34*****	**0.28*****	**0.77*****	**0.38*****	**0.27*****
4	**0.25*****	**0.20*****	**0.52*****	**0.16*****	**0.47*****	**0.43*****	**0.69*****	**0.51*****	**0.54*****
5	**0.13*****	**0.15*****	**0.61*****	**0.18*****	**0.36*****	**0.39*****	**0.60*****	**0.70*****	**0.56*****
6	**0.24*****	**0.11*****	**0.74*****	**0.48*****	**0.43*****	**0.58*****	**0.49*****	**0.37*****	**0.71*****
7	**0.23*****	**0.07*****	**0.79*****	**0.36*****	**0.29*****	**0.77*****	**0.32*****	**0.29*****	**0.75*****
8	**0.15***	0.01	**0.82*****	**0.22*****	**0.08*****	**0.69*****	**0.24*****	**0.12*****	**0.81*****
9	**0.14****	0.03	**0.84*****	**0.24*****	**0.04***	**0.69*****	**0.22*****	**0.14*****	**0.83*****
10	**0.15*****	**0.07*****	**0.84*****	**0.14*****	**0.05****	**0.73*****	**0.13*****	**0.06****	**0.78*****
11	−0.01	−0.01	**0.84*****	−0.01	−0.01	**0.45*****	−0.01	−0.01	**0.77*****
pythia-410m-deduped									
0	**0.86*****	**0.21*****	**−0.16*****	**0.40*****	0.03	**−0.11*****	**0.83*****	0.02	0.02
1	**0.86*****	**0.36*****	0.03	**0.43*****	**0.25*****	−0.01	**0.66*****	**0.24*****	**−0.11****
2	**0.87*****	**0.32*****	0.05	**0.51*****	**0.21*****	**−0.13*****	**0.80*****	**0.26*****	**−0.21*****
3	**0.86*****	**0.49*****	**0.26*****	**0.45*****	**0.36*****	**0.34*****	**0.82*****	**0.47*****	**0.12*****
4	**0.86*****	**0.53*****	**0.24*****	**0.50*****	**0.40*****	**0.14*****	**0.83*****	**0.47*****	0.07
5	**0.72*****	**0.77*****	**0.39*****	**0.19*****	**0.75*****	**0.39*****	**0.68*****	**0.56*****	**0.20*****
6	**0.76*****	**0.76*****	**0.44*****	**0.28*****	**0.30*****	**0.34*****	**0.69*****	**0.70*****	**0.33*****
7	**0.66*****	**0.65*****	**0.53*****	**0.51*****	**0.55*****	**0.45*****	**0.64*****	**0.57*****	**0.32*****
8	**0.61*****	**0.64*****	**0.59*****	**0.54*****	**0.55*****	**0.41*****	**0.62*****	**0.61*****	**0.48*****
9	**0.56*****	**0.62*****	**0.65*****	**0.46*****	**0.54*****	**0.57*****	**0.61*****	**0.59*****	**0.62*****
10	**0.55*****	**0.55*****	**0.68*****	**0.43*****	**0.53*****	**0.57*****	**0.55*****	**0.57*****	**0.62*****
11	**0.33*****	**0.31*****	**0.77*****	**0.42*****	**0.37*****	**0.81*****	**0.41*****	**0.35*****	**0.57*****
12	**0.15*****	**0.18*****	**0.76*****	**0.32*****	**0.37*****	**0.64*****	**0.29*****	**0.26*****	**0.74*****
13	**0.17*****	**0.21*****	**0.78*****	**0.21*****	**0.21*****	**0.68*****	**0.28*****	**0.29*****	**0.72*****
14	**0.18*****	**0.25*****	**0.78*****	**0.28*****	**0.22*****	**0.59*****	**0.27*****	**0.31*****	**0.74*****
15	**0.14*****	**0.15*****	**0.78*****	**0.24*****	**0.16*****	**0.70*****	**0.17*****	**0.17*****	**0.71*****
16	**0.21*****	**0.12*****	**0.78*****	**0.25*****	**0.12*****	**0.51*****	**0.25*****	**0.17*****	**0.72*****
17	**0.19*****	**0.14*****	**0.75*****	**0.27*****	**0.15*****	**0.48*****	**0.23*****	**0.17*****	**0.72*****
18	**0.18*****	**0.14*****	**0.68*****	**0.23*****	**0.12*****	**0.53*****	**0.22*****	**0.17*****	**0.67*****
19	**0.15*****	**0.12*****	**0.73*****	**0.21*****	**0.14*****	**0.41*****	**0.21*****	**0.15*****	**0.66*****
20	**0.37*****	**0.15*****	**0.73*****	**0.23*****	**0.12*****	**0.29*****	**0.16*****	**0.12*****	**0.57*****
21	**0.37*****	**0.14*****	**0.68*****	**0.17*****	**0.10*****	**0.67*****	**0.19*****	**0.11*****	**0.53*****
22	**0.06****	**0.11***	**0.74*****	**0.07*****	0.00	**0.61*****	0.03	**0.20*****	**0.49*****
23	0.02	0.02	**0.76*****	0.02	0.02	**0.51*****	0.02	0.02	**0.69*****
Qwen2-0.5b									
0	**0.85*****	**0.16*****	−0.04	**0.22*****	**0.14*****	**−0.18*****	**0.80*****	**0.17*****	**−0.26*****
1	**0.84*****	**0.39*****	0.04	**0.14*****	**0.44*****	0.08	**0.58*****	**0.29*****	**−0.08****
2	**0.84*****	**0.32*****	**0.09*****	**0.13*****	**0.14*****	0.08	**0.63*****	**0.25*****	**0.22*****
3	**0.80*****	**0.30*****	**0.44*****	**0.15*****	**0.11*****	**0.31*****	**0.54*****	**0.18*****	**0.13*****
4	**0.78*****	**0.41*****	**0.50*****	0.06	**0.13*****	**0.21*****	**0.63*****	**0.15*****	**0.44*****
5	**0.71*****	**0.48*****	**0.48*****	**0.15*****	**0.27*****	**0.33*****	**0.39*****	**0.38*****	**0.16*****
6	**0.69*****	**0.40*****	**0.57*****	**0.15*****	**0.22*****	**0.47*****	**0.28*****	**0.28*****	**0.43*****
7	**0.70*****	**0.35*****	**0.58*****	**0.14*****	**0.15*****	**0.20*****	**0.29*****	**0.25*****	**0.48*****
8	**0.63*****	**0.35*****	**0.61*****	0.04	**0.24*****	**0.41*****	**0.21*****	**0.17*****	**0.42*****
9	**0.61*****	**0.36*****	**0.65*****	**0.06****	**0.15*****	**0.10*****	**0.23*****	**0.10*****	**0.48*****
10	**0.63*****	**0.36*****	**0.67*****	**0.20*****	**0.27*****	**0.37*****	**0.24*****	**0.18*****	**0.48*****
11	**0.57*****	**0.43*****	**0.70*****	**0.25*****	**0.20*****	**0.08*****	**0.40*****	**0.30*****	**0.65*****
12	**0.52*****	**0.41*****	**0.73*****	**0.34*****	**0.32*****	**0.58*****	**0.51*****	**0.34*****	**0.64*****
13	**0.46*****	**0.41*****	**0.74*****	**0.38*****	**0.46*****	**0.63*****	**0.50*****	**0.39*****	**0.68*****
14	**0.40*****	**0.39*****	**0.76*****	**0.42*****	**0.49*****	**0.63*****	**0.58*****	**0.43*****	**0.72*****
15	**0.19*****	**0.33*****	**0.78*****	**0.29*****	**0.36*****	**0.71*****	**0.23*****	**0.36*****	**0.74*****
16	**0.14*****	**0.15*****	**0.81*****	**0.20*****	**0.16*****	**0.79*****	**0.16*****	**0.20*****	**0.73*****
17	**0.14*****	**0.16*****	**0.82*****	**0.35*****	**0.23*****	**0.50*****	**0.25*****	**0.21*****	**0.73*****
18	**0.16*****	**0.15*****	**0.82*****	**0.29*****	**0.26*****	**0.72*****	**0.35*****	**0.35*****	**0.79*****
19	**0.13*****	**0.14*****	**0.81*****	**0.12*****	**0.14*****	**0.74*****	**0.16*****	**0.18*****	**0.79*****
20	**0.11*****	**0.12*****	**0.82*****	**0.20*****	**0.13*****	**0.46*****	**0.10*****	**0.11*****	**0.77*****
21	**0.14*****	**0.14*****	**0.81*****	**0.22*****	**0.12*****	−0.05	**0.11*****	**0.11*****	**0.76*****
22	**0.15*****	**0.15*****	**0.82*****	**0.34*****	**0.19*****	**0.50*****	**0.17*****	**0.17*****	**0.69*****
23	0.02	0.02	**0.79*****	0.02	0.02	**0.50*****	0.02	0.02	**0.77*****

Table 4: Pearson r (bold, starred by significance: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$; p-values averaged across seeds) between **ODDS** of the learned 'classic' filler-gap interventions and mean *wh*-licensing interaction, for every layer, stream, position, and model, corresponding to Figure 4b. Positions abbreviated as WH/THAT = 'what'/'that', VP_2 = 'went back and got'.

	Block Output				Attention Value Output				MLP Activation			
Layer	VP_1	NP	AND	VP_2	VP_1	NP	AND	VP_2	VP_1	NP	AND	VP_2
GPT2												
0	**0.51****	**0.68*****	**0.59*****	0.46	**0.52****	**0.70*****	**0.66*****	0.40	0.52	**0.66*****	**0.60*****	**0.62*****
1	**0.57*****	**0.67*****	**0.59***	**0.52***	0.36	**0.69*****	**0.66*****	**0.51****	**0.53****	**0.66****	**0.64*****	**0.43***
2	**0.58*****	**0.68*****	**0.70*****	0.50	0.32	**0.72*****	**0.71*****	0.39	**0.60*****	0.58	**0.72*****	**0.72*****
3	**0.58*****	**0.72*****	**0.70*****	0.46	**0.44***	**0.65****	**0.68*****	**0.48***	**0.59*****	**0.73*****	**0.70*****	0.56
4	**0.58*****	**0.72*****	**0.67*****	**0.68*****	**0.53****	**0.61*****	**0.69*****	**0.72*****	**0.58*****	**0.74*****	**0.70*****	**0.73*****
5	**0.59*****	**0.74*****	**0.68*****	**0.71*****	**0.36***	**0.53****	**0.70*****	**0.67*****	**0.58*****	**0.74*****	**0.70*****	**0.74*****
6	**0.62*****	**0.73*****	**0.67*****	**0.72*****	**0.37***	**0.59*****	**0.67*****	**0.76*****	**0.61*****	**0.74*****	**0.68*****	**0.68****
7	**0.61*****	**0.73*****	**0.67*****	**0.71*****	**0.45****	**0.70*****	**0.66*****	**0.80*****	**0.58*****	**0.75*****	**0.66*****	**0.73*****
8	**0.62*****	**0.73*****	**0.67*****	**0.74*****	**0.56*****	**0.63*****	**0.63*****	**0.65*****	**0.59*****	**0.71*****	**0.67*****	**0.73*****
9	**0.63*****	**0.73*****	**0.68*****	**0.69*****	**0.45****	**0.70*****	**0.63*****	**0.71*****	**0.60*****	**0.74*****	**0.66*****	**0.70*****
10	**0.62*****	**0.70*****	**0.64*****	**0.68*****	**0.53****	**0.68*****	**0.53****	**0.69*****	**0.59*****	**0.72*****	**0.58*****	**0.69*****
11	0.14	0.29	0.18	**0.67*****	0.24	0.18	0.14	**0.68*****	0.23	0.08	0.33	**0.66*****
pythia-410m-deduped												
0	0.43	0.57	**0.57*****	0.38	**0.55****	**0.58****	**0.64*****	**0.60****	**0.54***	**0.67*****	0.14	**0.53****
1	0.45	**0.66*****	**0.54****	0.49	**0.45***	**0.67*****	**0.71*****	0.43	**0.52****	**0.48***	**0.53****	**0.58*****
2	**0.55****	**0.66*****	**0.62*****	**0.53***	0.36	**0.60****	**0.74*****	0.55	**0.60*****	**0.66*****	**0.63*****	**0.68*****
3	**0.56***	0.59	**0.72*****	**0.54****	**0.42***	**0.69*****	**0.74*****	0.30	**0.61*****	**0.72*****	**0.68*****	**0.65*****
4	**0.63*****	**0.70*****	**0.78*****	0.54	**0.56*****	**0.69*****	**0.79*****	**0.68****	**0.60*****	**0.75*****	**0.76*****	**0.58****
5	**0.59****	**0.71*****	**0.77*****	0.56	**0.62*****	**0.76*****	**0.75*****	**0.63***	**0.62*****	**0.73*****	**0.76*****	0.55
6	**0.66*****	**0.75*****	**0.77*****	**0.64*****	**0.54****	**0.69*****	**0.73*****	**0.76*****	**0.66*****	**0.78*****	**0.79*****	**0.71*****
7	**0.65*****	**0.75*****	**0.77*****	0.44	**0.51****	**0.78*****	**0.78*****	**0.75*****	**0.62*****	**0.77*****	**0.78*****	**0.69*****
8	**0.64*****	**0.75*****	**0.70****	**0.66****	**0.60****	**0.70*****	**0.75*****	**0.76*****	**0.67*****	**0.76*****	**0.80*****	**0.61***
9	**0.64*****	**0.74*****	**0.75*****	**0.71*****	**0.59*****	**0.69*****	**0.71*****	**0.75*****	**0.66*****	**0.74*****	**0.77*****	**0.71*****
10	**0.67*****	**0.73*****	**0.74*****	**0.69*****	**0.52****	**0.68*****	**0.69*****	**0.73*****	**0.68*****	**0.71*****	**0.75*****	**0.66****
11	**0.67*****	**0.72*****	**0.72*****	**0.63***	**0.65*****	**0.64*****	**0.63*****	**0.71*****	**0.72*****	**0.75*****	**0.72*****	**0.70*****
12	**0.69*****	**0.72*****	**0.73*****	**0.70*****	**0.49****	**0.69*****	**0.63*****	**0.56****	**0.69*****	**0.75*****	**0.72*****	**0.74*****
13	**0.69*****	**0.72*****	**0.70*****	**0.72*****	**0.57*****	**0.69*****	**0.56****	**0.69*****	**0.70*****	**0.75*****	**0.71*****	**0.73*****
14	**0.68*****	**0.71*****	**0.72*****	**0.72*****	**0.67*****	**0.72*****	**0.67*****	**0.66*****	**0.69*****	**0.75*****	**0.69*****	**0.73*****
15	**0.67*****	**0.70*****	**0.67*****	**0.70*****	**0.60*****	**0.75*****	**0.74*****	**0.76*****	**0.70*****	**0.76*****	**0.68*****	**0.71*****
16	**0.68*****	**0.67*****	**0.65*****	**0.67*****	**0.58*****	**0.67*****	**0.61*****	**0.57*****	**0.69*****	**0.76*****	**0.68*****	**0.71*****
17	**0.68*****	**0.72*****	**0.56****	**0.64*****	**0.66*****	**0.72*****	**0.71*****	**0.53****	**0.68*****	**0.76*****	**0.61*****	**0.67*****
18	**0.66*****	**0.70*****	**0.52****	**0.62*****	**0.49****	**0.72*****	**0.59*****	**0.46****	**0.67*****	**0.78*****	**0.62*****	**0.58*****
19	**0.63*****	**0.72*****	**0.52****	**0.60*****	**0.40***	**0.65*****	**0.63*****	**0.58****	**0.65*****	**0.74*****	**0.57*****	**0.59*****
20	**0.65*****	**0.75*****	**0.60*****	**0.63*****	**0.61*****	0.26	**0.42***	0.14	**0.59*****	**0.74*****	**0.57*****	**0.53****
21	**0.57*****	**0.69*****	**0.48****	**0.62*****	**0.65*****	**0.71*****	**0.67*****	**0.67*****	**0.63*****	**0.66*****	**0.53****	**0.55****
22	**0.57*****	**0.68*****	**0.63*****	**0.55****	**0.50****	**0.63*****	**0.65*****	0.34	0.14	**0.60****	**0.36***	**0.57****
23	0.25	0.21	0.07	**0.62*****	0.33	0.22	0.15	**0.41***	0.28	0.21	0.16	**0.63*****
Qwen2-0.5b												
0	**0.65*****	**0.55***	**0.71*****	0.41	**0.52****	**0.62*****	**0.64*****	0.23	**0.54***	**0.67*****	**0.68*****	0.35
1	0.48	**0.65*****	**0.69*****	0.44	**0.57***	**0.64*****	**0.67*****	0.18	**0.64****	**0.62****	**0.64*****	0.24
2	**0.62*****	**0.68*****	**0.70*****	0.52	**0.46***	**0.71*****	**0.62****	**0.57***	**0.59****	**0.59***	**0.71*****	**0.47***
3	**0.62*****	**0.65*****	**0.72*****	0.51	**0.47***	**0.70*****	**0.65*****	0.42	**0.57****	**0.70*****	**0.75*****	0.21
4	**0.63*****	**0.69*****	**0.71*****	**0.72*****	**0.52****	**0.72*****	**0.72*****	**0.67*****	**0.60*****	**0.71*****	**0.72*****	0.42
5	**0.61*****	**0.65*****	**0.68*****	**0.73*****	**0.46***	**0.69*****	**0.67*****	**0.61****	**0.58****	**0.68*****	**0.70*****	**0.69*****
6	**0.61*****	**0.70*****	**0.66*****	**0.72*****	**0.50***	**0.67*****	**0.69*****	**0.69*****	**0.55****	**0.71*****	**0.69*****	**0.63*****
7	**0.63*****	**0.69*****	**0.64*****	**0.77*****	**0.63*****	0.59	**0.64*****	**0.67*****	**0.58****	**0.70*****	**0.61*****	**0.70*****
8	**0.65*****	**0.71*****	**0.66*****	0.55	**0.52****	**0.70*****	**0.67*****	**0.65*****	**0.61*****	**0.70*****	**0.62*****	**0.71*****
9	**0.65*****	**0.71*****	**0.66*****	**0.72*****	0.31	**0.53****	**0.49****	**0.70*****	**0.61*****	**0.70*****	**0.60*****	**0.77*****
10	**0.62*****	**0.70*****	**0.65*****	**0.76*****	**0.53***	**0.64*****	**0.62*****	**0.68*****	**0.58*****	**0.67*****	**0.62*****	**0.74*****
11	**0.61*****	**0.69*****	**0.63*****	**0.75*****	**0.47****	**0.66*****	**0.66*****	**0.60*****	**0.58*****	**0.69*****	**0.59*****	**0.69*****
12	**0.61*****	**0.70*****	**0.63*****	**0.76*****	**0.67*****	**0.73*****	**0.63*****	**0.68*****	**0.61*****	**0.71*****	**0.61*****	**0.76*****
13	**0.61*****	**0.70*****	**0.65*****	**0.77*****	**0.53***	**0.66*****	**0.72*****	**0.74*****	**0.59*****	**0.70*****	**0.66*****	**0.78*****
14	**0.61*****	**0.71*****	**0.65*****	0.63	**0.62*****	**0.73*****	**0.71*****	**0.72*****	**0.55****	**0.69*****	**0.63*****	**0.77*****
15	**0.62*****	**0.70*****	**0.66*****	**0.75*****	**0.58*****	**0.67*****	**0.71*****	**0.69*****	**0.60*****	**0.68*****	**0.64*****	**0.76*****
16	**0.63*****	**0.68*****	**0.62*****	**0.72*****	**0.57*****	**0.70*****	**0.57*****	**0.75*****	**0.63*****	**0.68*****	**0.60*****	**0.80*****
17	**0.66*****	**0.68*****	**0.64*****	**0.71*****	**0.55****	**0.70*****	**0.65*****	**0.54****	**0.63*****	**0.72*****	**0.64*****	**0.64*****
18	**0.66*****	**0.70*****	**0.64*****	**0.69*****	**0.59*****	**0.68*****	**0.70*****	**0.73*****	**0.67*****	**0.76*****	**0.66*****	**0.66*****
19	**0.60*****	**0.70*****	**0.65*****	**0.66*****	**0.67*****	**0.74*****	**0.65*****	**0.68*****	**0.56****	**0.72*****	**0.66*****	**0.63*****
20	**0.59*****	**0.75*****	**0.63*****	**0.65*****	**0.44***	**0.62*****	**0.60*****	**0.62*****	**0.52****	**0.66*****	**0.59*****	**0.54****
21	**0.64*****	**0.73*****	**0.61*****	**0.63*****	0.35	**0.43***	**0.46****	**0.49****	**0.54****	**0.71*****	**0.60*****	**0.55****
22	**0.64*****	**0.75*****	**0.63*****	**0.61*****	**0.47****	**0.60*****	**0.59*****	**0.66*****	**0.54****	**0.71*****	**0.59*****	**0.59*****
23	0.20	0.30	0.23	**0.58****	0.14	0.09	0.18	**0.56****	0.26	0.31	0.35	**0.56****

Table 5: Absolute Pearson r (bold, starred by significance: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$; p-values averaged across seeds) between each conjunct's position on the learned DAS subspace and mean *wh*-licensing interaction, for every layer, stream, position, and model, corresponding to Figure 5b. Positions abbreviated as VP_1 = 'looked'/'baked', NP = 'down'/'cookies', VP_2 = 'saw'/'won'.

	Block Output			Attention Value Output			MLP Activation		
Layer	WH/THAT	SHE	VP_2	WH/THAT	SHE	VP_2	WH/THAT	SHE	VP_2
GPT2									
0	0.007	0.035	0.015	0.028	0.035	0.030	0.081	0.042	0.013
1	0.048	0.025	0.021	0.039	0.020	0.021	0.114	0.038	0.068
2	0.082	0.014	0.025	0.019	0.015	0.035	0.020	0.015	0.060
3	0.061	0.014	0.023	0.053	0.019	0.032	0.018	0.026	0.030
4	0.142	0.014	0.028	0.021	0.020	0.023	0.046	0.029	0.012
5	0.037	0.006	0.015	0.051	0.011	0.025	0.060	0.022	0.026
6	0.080	0.005	0.012	0.017	0.025	0.022	0.080	0.031	0.017
7	0.020	0.008	0.002	0.019	0.018	0.009	0.018	0.024	0.007
8	0.071	0.004	0.008	0.009	0.008	0.007	0.078	0.015	0.006
9	0.059	0.003	0.007	0.034	0.002	0.008	0.109	0.006	0.003
10	0.171	0.002	0.005	0.016	0.007	0.007	0.007	0.011	0.004
11	0.028	0.028	0.005	0.028	0.028	0.023	0.028	0.028	0.007
pythia-410m-deduped									
0	0.018	0.016	0.049	0.027	0.025	0.011	0.009	0.015	0.015
1	0.011	0.022	0.024	0.026	0.024	0.033	0.016	0.031	0.051
2	0.006	0.007	0.038	0.027	0.033	0.026	0.009	0.038	0.022
3	0.007	0.027	0.018	0.025	0.036	0.015	0.010	0.026	0.026
4	0.017	0.023	0.033	0.032	0.025	0.034	0.018	0.025	0.032
5	0.024	0.010	0.011	0.045	0.010	0.017	0.023	0.032	0.039
6	0.016	0.019	0.022	0.051	0.019	0.034	0.027	0.011	0.028
7	0.009	0.026	0.029	0.017	0.018	0.035	0.018	0.026	0.020
8	0.020	0.029	0.018	0.017	0.025	0.038	0.015	0.020	0.013
9	0.037	0.027	0.020	0.026	0.018	0.027	0.015	0.020	0.012
10	0.028	0.028	0.016	0.024	0.019	0.040	0.015	0.027	0.035
11	0.022	0.024	0.020	0.034	0.030	0.015	0.036	0.022	0.032
12	0.013	0.017	0.007	0.022	0.016	0.020	0.025	0.012	0.008
13	0.017	0.013	0.009	0.017	0.015	0.011	0.018	0.018	0.009
14	0.021	0.020	0.013	0.017	0.010	0.018	0.022	0.018	0.014
15	0.011	0.014	0.010	0.020	0.013	0.012	0.013	0.016	0.011
16	0.016	0.010	0.013	0.018	0.012	0.019	0.023	0.013	0.017
17	0.019	0.012	0.013	0.019	0.015	0.024	0.021	0.015	0.014
18	0.015	0.011	0.032	0.019	0.011	0.028	0.019	0.016	0.017
19	0.010	0.012	0.018	0.014	0.014	0.025	0.018	0.016	0.016
20	0.097	0.019	0.017	0.019	0.030	0.018	0.013	0.012	0.025
21	0.027	0.009	0.016	0.005	0.032	0.017	0.010	0.026	0.027
22	0.014	0.075	0.013	0.005	0.015	0.012	0.003	0.051	0.028
23	0.015	0.015	0.012	0.015	0.015	0.008	0.015	0.015	0.015
Qwen2-0.5b									
0	0.011	0.009	0.054	0.033	0.028	0.046	0.014	0.021	0.040
1	0.009	0.063	0.030	0.074	0.021	0.088	0.032	0.056	0.034
2	0.004	0.034	0.016	0.057	0.029	0.070	0.046	0.077	0.038
3	0.007	0.021	0.010	0.037	0.027	0.018	0.028	0.026	0.034
4	0.019	0.022	0.012	0.049	0.032	0.066	0.011	0.017	0.024
5	0.021	0.017	0.018	0.031	0.019	0.017	0.043	0.025	0.063
6	0.025	0.040	0.012	0.026	0.024	0.020	0.034	0.036	0.026
7	0.019	0.033	0.013	0.018	0.031	0.045	0.016	0.013	0.018
8	0.024	0.026	0.009	0.024	0.017	0.024	0.028	0.025	0.024
9	0.020	0.031	0.008	0.023	0.015	0.009	0.020	0.029	0.020
10	0.010	0.012	0.018	0.013	0.008	0.022	0.020	0.018	0.028
11	0.018	0.009	0.013	0.083	0.019	0.026	0.049	0.030	0.012
12	0.017	0.016	0.014	0.044	0.011	0.028	0.017	0.023	0.020
13	0.022	0.018	0.014	0.008	0.018	0.017	0.031	0.052	0.024
14	0.024	0.017	0.012	0.015	0.013	0.022	0.014	0.009	0.011
15	0.012	0.023	0.012	0.029	0.014	0.010	0.024	0.022	0.009
16	0.013	0.009	0.009	0.015	0.016	0.009	0.071	0.021	0.009
17	0.013	0.010	0.009	0.028	0.013	0.029	0.080	0.026	0.008
18	0.060	0.013	0.008	0.017	0.016	0.008	0.016	0.019	0.007
19	0.013	0.015	0.009	0.020	0.017	0.012	0.019	0.017	0.010
20	0.010	0.012	0.010	0.020	0.016	0.011	0.011	0.007	0.009
21	0.013	0.013	0.011	0.023	0.023	0.022	0.009	0.012	0.011
22	0.016	0.015	0.007	0.029	0.024	0.023	0.020	0.006	0.017
23	0.015	0.015	0.012	0.015	0.015	0.029	0.015	0.015	0.007

Table 6: Standard deviation of Pearson r across the five training seeds at every site, position, and layer, for the **ODDS** vs. *wh*-licensing correlations shown in Figure 3a and Table 4. Standard deviations are generally low, indicating consistency across the five evaluated seeds.

	Block Output				Attention Value Output				MLP Activation			
Layer	VP_1	NP	AND	VP_2	VP_1	NP	AND	VP_2	VP_1	NP	AND	VP_2
GPT2												
0	0.073	0.015	0.022	0.155	0.041	0.011	0.024	0.140	0.179	0.025	0.025	0.043
1	0.036	0.027	0.150	0.186	0.161	0.015	0.021	0.084	0.093	0.105	0.017	0.118
2	0.035	0.019	0.012	0.247	0.070	0.019	0.014	0.179	0.046	0.266	0.017	0.020
3	0.064	0.014	0.012	0.284	0.043	0.113	0.015	0.100	0.047	0.021	0.011	0.261
4	0.044	0.019	0.017	0.059	0.045	0.004	0.012	0.013	0.034	0.022	0.013	0.019
5	0.044	0.017	0.014	0.017	0.065	0.060	0.012	0.028	0.052	0.017	0.016	0.036
6	0.048	0.017	0.014	0.013	0.062	0.039	0.016	0.016	0.063	0.021	0.015	0.119
7	0.041	0.013	0.023	0.044	0.009	0.020	0.013	0.008	0.051	0.022	0.028	0.022
8	0.018	0.015	0.017	0.034	0.026	0.013	0.024	0.019	0.034	0.009	0.026	0.036
9	0.020	0.013	0.023	0.031	0.025	0.033	0.040	0.017	0.037	0.010	0.037	0.021
10	0.040	0.010	0.029	0.043	0.012	0.035	0.024	0.035	0.031	0.012	0.062	0.032
11	0.145	0.135	0.136	0.041	0.184	0.155	0.092	0.025	0.144	0.087	0.211	0.033
pythia-410m-deduped												
0	0.167	0.231	0.030	0.276	0.056	0.078	0.056	0.105	0.132	0.011	0.000	0.042
1	0.170	0.012	0.018	0.190	0.116	0.080	0.036	0.169	0.027	0.119	0.015	0.024
2	0.082	0.013	0.020	0.191	0.104	0.113	0.039	0.249	0.034	0.030	0.017	0.068
3	0.164	0.275	0.019	0.067	0.052	0.030	0.019	0.182	0.019	0.027	0.015	0.082
4	0.044	0.019	0.011	0.211	0.018	0.021	0.012	0.137	0.043	0.014	0.008	0.133
5	0.115	0.028	0.011	0.217	0.017	0.024	0.014	0.189	0.025	0.030	0.013	0.262
6	0.032	0.015	0.008	0.092	0.050	0.040	0.017	0.070	0.040	0.007	0.003	0.052
7	0.030	0.011	0.010	0.328	0.016	0.008	0.012	0.035	0.007	0.010	0.004	0.052
8	0.036	0.008	0.170	0.112	0.069	0.012	0.006	0.039	0.038	0.018	0.007	0.188
9	0.029	0.014	0.005	0.018	0.058	0.022	0.019	0.061	0.025	0.008	0.011	0.030
10	0.035	0.013	0.008	0.033	0.032	0.024	0.014	0.019	0.039	0.026	0.007	0.155
11	0.015	0.021	0.008	0.182	0.012	0.011	0.018	0.024	0.039	0.015	0.012	0.027
12	0.021	0.015	0.009	0.042	0.013	0.019	0.043	0.047	0.019	0.017	0.007	0.039
13	0.024	0.016	0.010	0.059	0.015	0.028	0.044	0.045	0.031	0.012	0.009	0.049
14	0.017	0.012	0.004	0.042	0.012	0.017	0.017	0.046	0.019	0.014	0.010	0.047
15	0.019	0.010	0.009	0.048	0.015	0.019	0.018	0.017	0.017	0.014	0.012	0.042
16	0.016	0.041	0.021	0.043	0.034	0.039	0.052	0.065	0.024	0.017	0.019	0.049
17	0.013	0.007	0.037	0.039	0.021	0.006	0.023	0.054	0.017	0.019	0.034	0.045
18	0.022	0.009	0.028	0.068	0.018	0.005	0.016	0.046	0.022	0.013	0.042	0.057
19	0.030	0.011	0.031	0.049	0.104	0.052	0.032	0.061	0.031	0.016	0.027	0.049
20	0.052	0.013	0.068	0.053	0.024	0.116	0.039	0.122	0.020	0.015	0.025	0.075
21	0.036	0.043	0.066	0.050	0.017	0.038	0.026	0.012	0.022	0.014	0.047	0.053
22	0.015	0.076	0.030	0.053	0.070	0.032	0.066	0.088	0.029	0.109	0.049	0.054
23	0.183	0.146	0.029	0.054	0.210	0.121	0.102	0.081	0.173	0.168	0.158	0.050
Qwen2-0.5b												
0	0.081	0.192	0.005	0.202	0.080	0.055	0.069	0.215	0.187	0.009	0.030	0.234
1	0.275	0.024	0.020	0.223	0.152	0.045	0.018	0.151	0.109	0.123	0.038	0.208
2	0.091	0.020	0.014	0.278	0.078	0.018	0.118	0.155	0.090	0.182	0.011	0.131
3	0.068	0.059	0.017	0.274	0.079	0.025	0.036	0.209	0.097	0.012	0.009	0.192
4	0.077	0.017	0.016	0.052	0.041	0.017	0.007	0.124	0.081	0.012	0.018	0.366
5	0.077	0.068	0.033	0.048	0.072	0.032	0.008	0.092	0.091	0.011	0.006	0.065
6	0.064	0.013	0.015	0.086	0.113	0.043	0.025	0.070	0.111	0.020	0.019	0.086
7	0.079	0.028	0.009	0.035	0.049	0.294	0.036	0.071	0.128	0.018	0.032	0.118
8	0.063	0.023	0.011	0.331	0.073	0.024	0.014	0.057	0.070	0.022	0.025	0.069
9	0.061	0.009	0.015	0.050	0.057	0.023	0.014	0.019	0.056	0.011	0.019	0.073
10	0.068	0.010	0.008	0.036	0.123	0.024	0.016	0.046	0.062	0.007	0.018	0.067
11	0.073	0.012	0.012	0.037	0.020	0.033	0.018	0.031	0.027	0.019	0.023	0.073
12	0.067	0.014	0.020	0.044	0.019	0.008	0.016	0.033	0.020	0.016	0.023	0.054
13	0.064	0.017	0.017	0.037	0.183	0.025	0.015	0.033	0.012	0.009	0.015	0.062
14	0.064	0.010	0.009	0.308	0.052	0.009	0.010	0.023	0.030	0.009	0.022	0.069
15	0.057	0.009	0.013	0.027	0.034	0.018	0.015	0.046	0.026	0.015	0.022	0.058
16	0.049	0.030	0.019	0.043	0.013	0.019	0.044	0.041	0.052	0.016	0.038	0.049
17	0.056	0.025	0.017	0.046	0.063	0.022	0.022	0.099	0.042	0.014	0.025	0.083
18	0.046	0.013	0.018	0.053	0.032	0.036	0.013	0.033	0.047	0.012	0.023	0.089
19	0.053	0.013	0.013	0.054	0.032	0.060	0.029	0.063	0.066	0.020	0.015	0.081
20	0.060	0.035	0.010	0.065	0.022	0.011	0.010	0.031	0.044	0.028	0.016	0.099
21	0.061	0.026	0.016	0.070	0.096	0.024	0.021	0.026	0.033	0.011	0.014	0.100
22	0.051	0.023	0.010	0.075	0.056	0.058	0.062	0.060	0.032	0.018	0.011	0.066
23	0.130	0.206	0.124	0.072	0.076	0.054	0.136	0.085	0.189	0.149	0.124	0.060

Table 7: Standard deviation of absolute Pearson r across the five training seeds at every site, position, and layer, for the DAS-subspace vs. *wh*-licensing correlations shown in Figure 5b and Table 5. Standard deviations are generally low, indicating consistency across the five evaluated seeds.