Mechanistic Interpretability of Animacy Effects on Structure Choice in GPT-2*

Yue Li[1], Yan Cong[1,2], Elaine J. Francis[1]
[1]Department of Linguistics, Purdue University
[2]School of Languages and Cultures, Purdue University
{li4207, cong4, ejfranci}@purdue.edu

Abstract

Language models (LMs) exhibit human-like behavior across linguistic tasks, yet behavioral similarity does not establish mechanistic correspondence. Animacy — whether an entity is alive and sentient — is a well-documented semantic feature shaping linguistic behavior in humans. Although LMs show animacy sensitivity behaviorally, the mechanistic basis remains unexplored. In this study, we probe GPT-2 Small's internal circuitry to test whether animacy representations *causally* drive syntactic structure choice. Activation patching confirms causality: swapping animacy representations in the model shifts its downstream output. Critically, bidirectional patching reveals that animacy conditions differ in how strongly they commit to a structure: some animacy configurations resist perturbation and exert strong causal influence, while others remain flexible. We identify 22 attention heads mediating these effects, split between *passive-promoting* and *passive-suppressing* populations, suggesting GPT-2 Small's structure choice likely emerges from internal competition between opposing heads. These findings provide mechanistic grounding for animacy effects documented in extensive psycholinguistics research and demonstrate how interpretability methods can enrich and test psycholinguistic theory.

1 Introduction

A growing body of research has found that language models (LMs) exhibit human-like linguistic behavior. In surprisal-based analyses, models show sensitivity to syntactic structure that parallels human judgments, including subject-verb agreement across intervening material (Linzen et al., 2016; Gulordava et al., 2018), filler-gap dependencies (Wilcox et al., 2020), and island constraints (Wilcox et al., 2023). Model surprisal reliably predicts human reading times, with transformer-based models consistently outperforming n-gram baselines (Goodkind and Bicknell, 2018; Wilcox et al., 2020; Huang et al., 2024; Oh and Schuler, 2023). Direct prompting studies have found human-like performance on reasoning tasks (Cong, 2024; Hu and Levy, 2023; Fang et al., 2025a, 2026), semantic judgments (Ettinger, 2020), and semantic-syntax interface (Li et al., 2025; Hanna et al., 2023; Fang et al., 2025b). Neural alignment studies report that model representations correlate with brain activity during language comprehension (Schrimpf et al., 2021; Goldstein et al., 2022; Caucheteux and King, 2022). Moreover, LM-derived surprisal also indexes language development and recovery in non-native speakers and clinical populations, tracking second-language proficiency in degree expressions (Cong, 2025), detecting and subtyping aphasia from discourse (Cong et al., 2024), and tracking priming-induced syntactic recovery in aphasia (Cong and Lee, 2025). These convergent findings raise an intriguing possibility: LMs might serve as computational testbeds for investigating cognitive mechanisms found in humans. Yet behavioral alignment does *not* establish causal equivalence. Using mechanistic interpretability methods, we ask whether the model arrives at human-like behavior by relying on the same information humans use, or whether the correspondence arises from shallow heuristics that produce the correct outputs for the wrong reasons (McCoy et al., 2019).

Animacy, whether an entity is alive and sentient, is a semantic feature with well-documented effects on syntactic processing in psycholinguistics research. Specifically, object relative clauses (ORCs) with two animate nouns were found to be harder to process than those with mixed animacy, an effect attributed to *accessibility* and *interference* mechanisms (Gordon et al., 2001; Mak et al., 2002; Traxler et al., 2002; Bock and Warren, 1985; Gennari et al., 2012). Recent work has begun testing whether LMs show sensitivity to ani-

*All code is available at https://github.com/Yue00831/Animacy_LLMech_CoNLL2026.

Proceedings of the 30th Conference on Computational Natural Language Learning, pages 626–640

macy like humans. Hanna et al. (2023) found that LMs distinguish animate from inanimate entities and adapt to contextually atypical animacy, though not as flexibly as humans. The BLiMP benchmark includes animacy-based selectional restriction tests on which models achieve high accuracy (Warstadt et al., 2020). Li et al. (2025) found that DistilGPT-2's passive rates in ORCs vary systematically across animacy configurations in a pattern that mirrors human production data.

These findings establish behavioral alignment, but they leave open the mechanistic question: does animacy *causally* drive model behavior, or does the correlation arise through a different pathway?

The present study takes GPT-2 Small and asks whether animacy causally drives the model's structure choice, and if so, through what computational architecture. We chose GPT-2 Small because Li et al. (2025) already showed initial behavioral evidence of animacy sensitivity. As an extension, we investigate whether this alignment reflects shared computational principles. Also, GPT-2 is one of the most widely studied and validated models in recent mechanistic interpretability studies (Wang et al., 2023; Conmy et al., 2023; Vig et al., 2020). Its moderate size allows us to perform controlled interventions that might not be accessible or feasible in larger-scale, closed models such as GPT-5 (Mueller et al., 2025).

Building on this line of work, we focus on structure choice in English ORCs, where speakers choose between active (e.g., *the journalist that the newspaper contacted*) and passive (e.g., *the journalist that was contacted by the newspaper*) in four animacy configurations defined by the head noun and agent noun animacy: **AA** (both animate), **IA** (inanimate head, animate agent), **AI** (animate head, inanimate agent), and **II** (both inanimate). Using Li et al. (2025) controlled stimulus materials, we probe the computational mechanisms underlying this alignment through a series of studies: We first ask whether GPT-2 encodes animacy as a generalizable categorical feature (Study 1) and whether within-sentence nouns' similarity varies across animacy configurations (Study 2). Then we use activation patching (Zhang and Nanda, 2024; Heimersheim and Nanda, 2024) to test whether these representations *causally* drive structure choice (Study 3) and identify the specific attention heads mediating this effect (Study 4). Our main contributions are:

1. Using activation patching, we provide causal evidence that animacy drives ORC structure choice in GPT-2, not just behavioral alignment.

2. Through similarity analysis and bidirectional patching, we show that *interference* and *accessibility* operate jointly, and their combined strength determines how strongly a given animacy configuration commits to passive.

3. Via targeted ablation, we localize 22 attention heads with opposing *passive-promoting* and *passive-suppressing* populations, revealing that GPT-2's structure choice involves a more complex internal process than a single unified bias.

2 Related Work

2.1 Animacy effects on sentence structure choice

Two mechanisms have been proposed in psycholinguistics research to explain how animacy influences structure choice in ORCs. Gennari et al. (2012) found that ORCs with two animate nouns (AA) are more likely to be produced in passive voice than those with an inanimate head noun (IA). To explain this finding, they argue that two animate nouns (both potential agents), trigger similarity-based competition during processing, and that passive voice helps resolve this competition (***interference mechanism***, Gordon et al. (2001, 2004)). An alternative explanation is the ***accessibility mechanism***, according to which animate nouns are more salient and thus more accessible to the subject position within the relative clause, promoting passive use when the head noun is animate (Bock and Warren, 1985; Prat-Sala and Branigan, 2000; Rodrigo et al., 2018).

Li et al. (2025) tested all four animacy configurations in both humans and DistilGPT-2 and found that AA and AI showed the highest passive rates, IA the lowest, and II intermediate between these. They interpreted this through joint activation of the two mechanisms: AA activates both *interference* and *accessibility*, IA activates neither, AI activates *accessibility* only, and II activates *interference* only. However, they were unable to explain why the accessibility mechanism seemed to have a stronger effect than the interference mechanism, resulting in higher passive rates for AI compared to II. Behavioral methods alone are inadequate to resolve these questions, because they observe the joint outcome

and relative weightings of the mechanisms without specifying how they are computationally integrated. The present study aims to address this gap.

2.2 LMs as testbeds of human language processing

The question of whether LMs can inform theories of human cognition has generated substantial debate. On the one hand, LMs achieve remarkable behavioral alignment with humans across diverse linguistic phenomena, including syntactic dependencies (Linzen et al., 2016; Gulordava et al., 2018), filler-gap processing (Wilcox et al., 2018), reading time prediction (Huang et al., 2024; Oh and Schuler, 2023; Goodkind and Bicknell, 2018), and good-enough processing (Cong and Rayz, 2025). On the other hand, critics argue that behavioral similarity does not establish mechanistic correspondence: models may achieve human-like outputs through entirely different internal processes (Bowers et al., 2023; Guest and Martin, 2023). Ivanova (2023) articulates this concern directly: behavioral and neural alignment between LMs and humans may reflect superficial similarities rather than shared underlying algorithms.

Recent studies have examined how animacy is represented by LMs. Hanna et al. (2023) found that models like GPT-2 and LLaMA distinguish animate from inanimate entities and show human-like behavior in adapting to atypical animacy. They argue that despite limited signal, LMs acquire sensitivity to relevant lexical semantic distinctions. Li et al. (2025) found that GPT-2's passive rates in ORCs across the four animacy configurations mirror human production data: both show the highest passive rates for AA and AI, moderate rates for II, and the lowest for IA, and this pattern cannot be fully explained by biases in its training data (Li et al., 2025). These findings suggest an emergent animacy-sensitivity of LMs, but fail to pinpoint its cause.

Does animacy actually drive structure choice in LMs? If so, then manipulating animacy representations inside the model should change its behavior; otherwise, manipulation should have little effect (Heimersheim and Nanda, 2024). Answering this question matters beyond model evaluation, because a model that not only behaves like humans but processes information in a similar way would offer a far more productive testbed for investigating theories of human sentence processing.

2.3 Mechanistic interpretability and causal methods

Mechanistic interpretability aims to reverse-engineer neural networks into human-understandable algorithms (Zhang and Nanda, 2024; Heimersheim and Nanda, 2024). A central technique is *activation patching*, which isolates the causal contribution of specific representations by substituting activations from one forward pass into another (Vig et al., 2020; Meng et al., 2022). Heimersheim and Nanda (2024) provide methodological guidance for this approach, distinguishing between denoising and noising interventions and discussing how different circuit architectures manifest in patching results. Zhang and Nanda (2024) offer best practices for metrics and corruption methods.

Recent work has begun applying these methods to linguistic phenomena. Finlayson et al. (2021) examined causal effects of syntactic agreement representations, and Lasri et al. (2022) introduced usage-based probing with causal interventions to test whether BERT actually relies on its encoding of grammatical number. However, to our knowledge, no prior work has used causal intervention methods to investigate how semantic features like animacy affect sentence structure. The present study addresses this gap, using activation patching not merely to localize where information is stored but to test competing theoretical accounts of how multiple constraints are integrated during structure selection.

3 Study 1: Probing animacy representations

We first ask whether animacy is encoded in GPT-2 Small. Study 1 tests whether animacy information generalizes across lexical items. If a classifier learns a boundary between animate and inanimate nouns from a subset of items and successfully applies it to held-out nouns, this indicates the representational space is structured along an animacy-relevant dimension.

Method We compiled 38 animate nouns (e.g., *baby*, *farmer*) and 57 inanimate nouns (e.g., *trophy*, *table*) using Li et al. (2025)'s psycholinguistics-driven stimulus set, split into training (70%) and test (30%) sets stratified by animacy. For each noun at each layer, we placed it in a minimal context ("The noun is") and extracted the hidden state

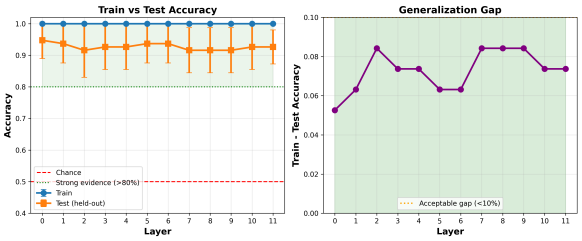

Figure 1: Train and test accuracy for animacy classification (5-fold stratified cross-validation). Left: High test accuracy on held-out nouns. Right: Generalization gap remains below the 10% threshold at all layers.

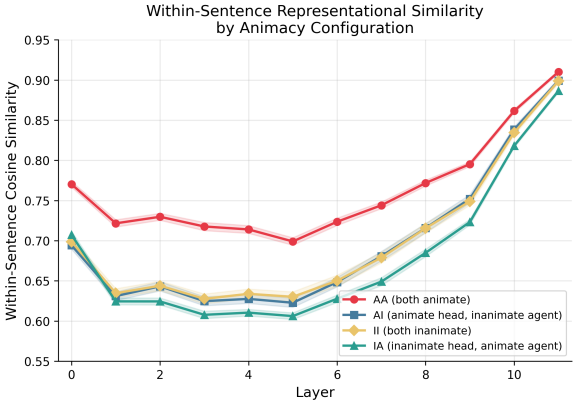

Figure 2: Within-sentence cosine similarity between head and agent noun representations, averaged across active and passive constructions (active and passive produced virtually identical patterns, r=.994).

from the residual stream. We trained logistic regression classifiers to predict animacy, then evaluated on held-out nouns. To address the potential confound that the probe exploits subword tokenization rather than animacy itself (Belinkov, 2022), we verified that the animate and inanimate noun sets were matched by *subword tokenization*, with comparable mean token counts per noun and consistent with the stable sentence-level OOV (out-of-vocabulary) rates reported in Li et al. (2025).

Results Figure 1 shows classifier performance across layers. We performed 5-fold stratified cross-validation to assess generalization. Mean test accuracy across folds was 92.7% (±6.7%), and performance was flat across depth, indicating that linear separability of animacy is consistent throughout the representational hierarchy and present from the embedding layer.

Discussion GPT-2 organizes animate and inanimate nouns into distinct regions of representational space, and this organization generalizes to unseen nouns. But probe generalization does not establish that GPT-2 uses animacy functionally (Belinkov, 2022). What Study 1 shows is that the *representational prerequisites* for animacy-based processing are in place. Whether this information causally influences structure choice is the question we address in Studies 3 and 4.

4 Study 2: Representational similarity across animacy configurations

Motivation Study 1 established that animacy is encoded in GPT-2. We next ask whether the joint animacy configuration is reflected in how similarly the model represents the two nouns within a sentence. *The interference account* predicts that representationally similar nouns should compete more during processing and thus will promote passive use more (Gordon et al., 2001; Gennari et al., 2012).

Method With Li et al. (2025)'s stimuli set, we have 96 items per animacy condition (AA, AI, IA, II). We extracted hidden states for both noun positions at each layer and computed the cosine similarity between head noun and agent noun representations within each sentence.

Results Within-sentence similarity varied systematically by animacy configuration (Figure 2; all ANOVAs $p < .001$ across layer ranges). AA showed the highest similarity ($M = 0.763$), significantly exceeding all other conditions (all $p < 10^{-63}$, Cohen's $d = 0.73$–0.97) (e.g., *rabbit/fox*: 0.817, *girl/man*: 0.811). IA showed the lowest similarity ($M = 0.681$) (e.g., *vase/gardener*: 0.577, *box/woman*: 0.635), significantly below AI and II ($p < .001$), though with small effect sizes ($d \approx 0.18$–0.19). Interestingly, AI ($M = 0.698$) and II ($M = 0.700$) were statistically indistinguishable ($p = .73$, $d = -0.01$). All conditions showed a U-shaped trajectory across layers, with between-condition differences largest in mid-layers.

We additionally verified that the similarity pattern is not affected by the presence of the explicit relativizer *that*. The AA > AI ≈ II > IA ordering held with virtually identical effect sizes for both full relative clauses (e.g., *the baby that the father holds*) and reduced relative clauses (*the baby the father holds*), with no relativizer main effect or interaction (details reported in Appendix D).

Discussion Results align with predictions of the *interference account* at the extremes (Gordon et al.,

2001; Gennari et al., 2012): AA shows the highest similarity, as two animate nouns both possess agentive properties that form competition, while IA shows the lowest, as the inanimate head noun creates no competition for agenthood with the animate agent (thematic role is clear). II shows substantially lower similarity than AA despite both being matched-animacy conditions, suggesting that animacy congruency alone does not guarantee high representational overlap. Animate nouns (AA) showed a coherent cluster of properties (agenthood, volition, sentience) that produce tight representational clustering. In contrast, II pairs range from functionally related objects with strong similarity (*bus/car*) to unrelated artifacts with much lower overlap (*vase/shelf*), showing the greater heterogeneity. Although Study 1 confirmed the animate–inanimate categories are distinct for GPT-2, inanimate nouns seem to show more diversity within their category.

The results of Study 2 point to an interesting interaction: when the agent is animate, changing the head noun from animate to inanimate produces a large similarity difference but when the agent is inanimate, the same change has virtually no effect. This suggests that head noun and agent noun animacy contribute to the representational similarity interactively. In Studies 3–4, we will clarify whether and how these representational patterns channel into causal influence on structure choice.

5 Study 3: Causal effects of animacy on structure choice

Motivation Studies 1 and 2 established that GPT-2 encodes animacy and that representations reflect joint animacy configuration. However, probe accuracy cannot establish whether animacy *causally* affects sentence structure (Belinkov, 2022). Study 3 addresses this using activation patching.

Method We patched the residual stream activations at the noun token positions, in order to tease apart and isolate the causal contribution of head noun animacy and agent noun animacy separately. Each transition holds one noun's animacy constant while flipping the other:

- Transitions **AA↔IA** and **AI↔II** hold agent animacy constant and vary head animacy, isolating the causal effect of *head* animacy.
- Transitions **AA↔AI** and **IA↔II** hold head animacy constant and vary agent animacy, isolating the causal effect of *agent* animacy.

Specifically, for each layer l, we ran the source sentence (condition A) through the model, extracted its post-block residual stream vectors at the two noun token positions, and substituted them into the corresponding positions during a forward pass of the target sentence (condition B) (Heimersheim and Nanda, 2024; Zhang and Nanda, 2024). Crucially, this intervention does not replace the input tokens, rather, it replaces the model's internal encoding at noun positions, making the model process the target sentence as if its nouns carried the animacy properties of the source sentence. Structure preference was measured as the difference in length-normalized log-probabilities between passive ($\bar{\ell}_p$) and active ($\bar{\ell}_a$) variants:

$$\Delta = \bar{\ell}_p - \bar{\ell}_a \tag{1}$$

$$\bar{\ell}_x = \frac{1}{n_x} \sum_{i=1}^{n_x} \log P(t_i^x \mid t_{<i}^x) \tag{2}$$

where probabilities are computed autoregressively over all BPE tokens in each sentence. Length normalization accounts for the different token counts between active and passive. GPT-2's BPE tokenizer handles leading whitespace internally, so no additional whitespace correction was applied.

We tested four forward transitions (AA→AI, AA→IA, AI→II, IA→II) and their reverses to assess bidirectional asymmetry. Two null controls validated the procedure: *self-patching* (patching sentences to themselves, expected effect: exactly zero) to confirm that the operation itself introduces no artifacts; *within-animacy patching* (swapping sentences sharing the same animacy configuration) established a lexical variation noise floor. Experimental effects were later corrected by subtracting within-animacy noise floor. All analyses applied FDR correction.

Results *Null control validations* Self-patching yielded effects of exactly 0.000 ($M = 0.000$, $SD = 0.000$), as expected, so patching itself introduces no artifacts. Within-animacy patching produced small effects ($M = 0.033$, Cohen's $d = 0.16$). Experimental patching's effect ($M = 0.177$) was 5.3× larger than noise floor, indicating predominant animacy-driven causation (Figure 3).

Causal effects After subtracting within-animacy baselines (noise floor), all four forward transitions showed substantial corrected effects: AA→IA

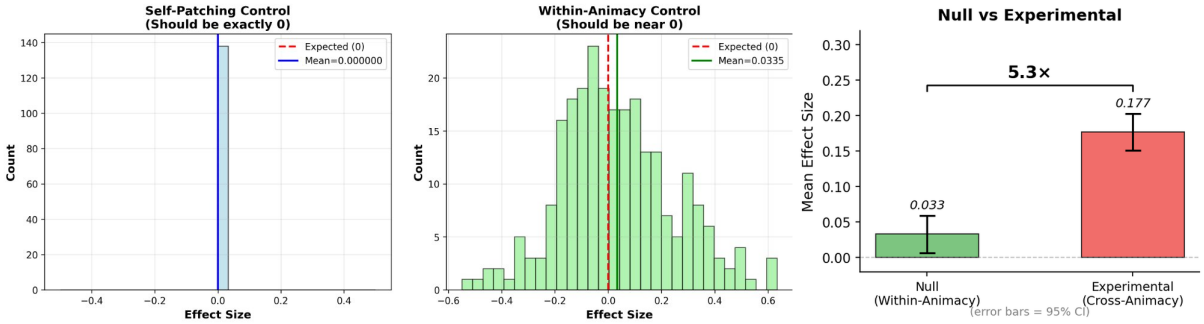

Figure 3: (Left) Self-patching yields exactly zero effect. (Middle) Within-animacy patching produces small effects ($M = 0.033$). (Right) Experimental cross-animacy effects are 5.3× larger than the within-animacy control.

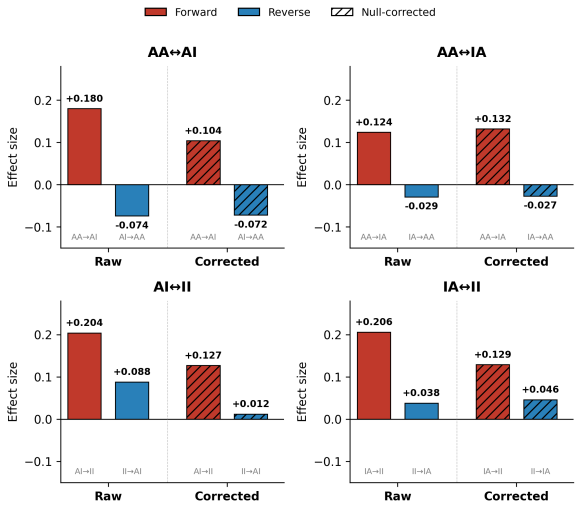

Figure 4: Study 3 bidirectional patching with null correction. Positive values indicate increased passive preference; negative values indicate increased active preference. Solid bars: raw effects; hatched bars: corrected effects (subtracting target condition's within-animacy null baseline).

(+0.132), IA→II (+0.129), AI→II (+0.127), and AA→AI (+0.104), all with 95% CIs excluding zero (Appendix Table 2). Bidirectional patching revealed consistent *asymmetry*: AA→IA promoted passive by +0.132, while IA→AA promoted actives only by 0.027. II representations exerted particularly weak causal influence on others after correction (II→AI +0.012; II→IA +0.046), which were minimal compared to AA's strong effects (AA→AI: +0.104; AA→IA: +0.132) (Figure 4). Seven of eight transitions showed effects directionally consistent with behavioral results in Li et al. (2025), but IA→II presented an anomaly: despite IA having the lowest behavioral passive rate (63%) compared to II (82%), patching IA into II unexpectedly increased passive likelihood. We report layer-wise results in Appendix Figure 6.

Discussion Manipulating animacy representations causally affected structure choice, indicating that the animacy sensitivity observed behaviorally in prior work (Hanna et al., 2023; Warstadt et al., 2020; Li et al., 2025) is mechanistically grounded. A caveat is that residual stream patching swaps entire representations, which is a known interpretive challenge for activation patching (Heimersheim and Nanda, 2024). Our within-animacy null control addresses this: swapping representations between sentences that share the same animacy configuration but differ in lexical content produced effects 5.3× smaller than cross-animacy patching. While this control cannot eliminate every possible confound, it demonstrates that most of the observed effect is attributable to the animacy differences between conditions rather than to incidental lexical properties.

The bidirectional asymmetry shows that AA representations seem to be more committed to passives and exert strong causal influence on other conditions. In contrast, II representations are causally weak despite involving matched-animacy nouns like AA. This AA–II divergence cannot be explained by similarity alone. We return to this contrast in the General Discussion.

The IA→II anomaly points to a limitation of residual stream patching that requires further analysis. We suggest feature-level patching as the next step. Specifically, by isolating animacy within a low-dimensional subspace of the model's representations and intervening only along that subspace, we could potentially determine whether this anomaly reflects genuine animacy dynamics or contributions from co-varying features. Relevant techniques include nullspace projection on representations (Ravfogel et al., 2020) and distributed alignment search for locating interpretable causal vari-

ables in distributed representations (Geiger et al., 2024), though subspace interventions themselves require careful validation to avoid interpretability illusions (Makelov et al., 2024). We leave this for future work, but flag it as the natural next step for this line of inquiry.

6 Study 4: Identifying critical attention heads

Building upon Study 3's finding, Study 4 now addresses *which specific components* of GPT-2 mediate the animacy effect.

Method We applied a three-stage filter to GPT-2's 144 attention heads: 1) *structural criterion*: we retained heads where the verb token allocated $\geq$10% attention to both noun positions in $\geq$50% of sentences; 2) *sensitivity criterion*: we tested whether attention patterns differed across animacy conditions using Kruskal-Wallis tests with FDR correction; 3) *causal necessity criterion*: ablating each head by zeroing its per-head output vector at `hook_z`, removing the head's contribution to the residual stream, and measured resulting changes in structure preference (Conmy et al., 2023; Wang et al., 2023).

To verify that ablation was targeted rather than destructive, we also measured the change in mean sentence log-likelihood under ablation. Zero-ablation can in principle disrupt model performance beyond the removal of the targeted component's function (Li and Janson, 2024; Heimersheim and Nanda, 2024). This diagnostic examines whether any observed effects on structure preference reflect targeted intervention rather than collateral damage.

Results The filtering procedure found 22 critical heads (Table 1, Figure 5). These heads split into two functionally distinct types: 12 *passive-promoting* heads (ablating them decreased passives) and 10 *passive-suppressing* heads (ablating them increased passives). L0H9 emerged as the dominant passive-suppressing head, with an ablation effect (+0.338) approximately 4× larger than the next strongest head. The strongest passive-promoting heads were L11H0 (-0.069) and L0H7 (-0.045). Ablation changed mean sentence log-likelihood by only 0.65%, confirming minimal collateral disruption.

Discussion The split between passive-promoting and passive-suppressing heads suggests structure choice emerges from competition between opposing populations. One possibility is that the differences observed in Study 3 reflect how decisively one population dominates the other: configurations like AA may strongly activate passive-promoting heads, while open configurations like IA may produce more balanced activation across the two populations. However, the present single-head ablation approach cannot directly test this hypothesis. Characterizing how the balance between these populations shifts across animacy conditions is an important direction for future circuit-level analysis.

In the current study, we focused on attention heads because the animacy–structure computation requires integrating information across noun positions; prior circuit-level work on natural language tasks in GPT-2 has likewise localized comparable computations at this level (Wang et al., 2023). Head-level localization is a necessary first step toward a fuller mechanistic account, which would require decomposing each head's query–key and output–value circuits. Complementary analyses at

Head	Type	η^2	Ablation	p_{FDR}
L0H9	passive-suppressing	0.191	+0.338	***
L11H0	passive-promoting	0.052	−0.069	***
L11H8	passive-suppressing	0.037	+0.050	***
L0H7	passive-promoting	0.033	−0.045	***
L1H1	passive-suppressing	0.135	+0.023	***
L8H5	passive-suppressing	0.058	+0.022	***
L6H7	passive-promoting	0.095	−0.021	***
L3H7	passive-promoting	0.235	−0.018	***
	... (14 additional heads omitted for space)			

Table 1: Top 8 critical attention heads by ablation effect magnitude. All p-values are FDR-corrected and significant at $p < .001$ (***). Full results in Appendix Table 3.

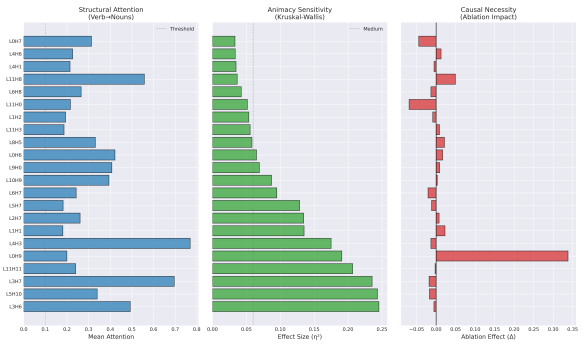

Figure 5: Critical attention heads and their causal effects. Left: Filtering funnel (144 → 74 → 27 → 22). Center: Layer distribution of critical heads. Right: Ablation effects showing passive-suppressing (positive) and passive-promoting (negative) heads.

finer-grained levels such as activation subspaces (Geiger et al., 2024) or individual neurons remain for future discussion.

7 General Discussion

7.1 Animacy causally drives structure choice

While prior causal interpretability work has focused on morphosyntactic phenomena (Finlayson et al., 2021; Vig et al., 2020), our findings extend this to the syntax-semantics interface, showing that animacy *causally* drives syntactic structure selection. By manipulating animacy representations and observing downstream structure choice shift, we found that the animacy sensitivity is causal, not just behavioral alignment (Hanna et al., 2023; Warstadt et al., 2020; Li et al., 2025). GPT-2 not only organizes animate and inanimate nouns into distinct representational regions that generalize to unseen items, but also encodes the joint animacy configuration of noun pairs at different similarity levels, as found in psycholinguistics work (Gordon et al., 2001; Gennari et al., 2012).

Critically, patching these representations between animacy conditions shifted structure choice with effects $5.3\times$ larger than the within-animacy noise floor, and this influence was mediated by a localized circuit of 22 attention heads whose *passive-promoting* and *passive-suppressing* populations suggest that structure choice likely emerges from competition between opposing forces. Together, these findings move beyond the question of *whether* LMs are sensitive to animacy to *how* animacy information is encoded, propagated, and used to determine structural preferences.

7.2 *Interference* and *accessibility* mechanism jointly create a structural commitment gradient

Psycholinguistic accounts propose two mechanisms: *interference*, where representationally similar nouns compete during processing (Gennari et al., 2012), and *accessibility*, where animate head nouns favor subject position (Bock and Warren, 1985). Li et al. (2025) interpreted behavioral patterns through these mechanisms but did not reveal how they operate together.

Our mechanistic evidence suggests the four animacy conditions fall along a *structural commitment gradient*: a continuum reflecting how strongly an animacy configuration commits to passives and how resistant that commitment is to perturbation.

AA–Committed: highest noun similarity, consistent with strong competition between two animate nouns for the subject role (Gennari et al., 2012). With both mechanisms strongly activated, the model commits decisively to passives, resisting perturbation while exerting strong influence to other configurations.

AI–Leaning: prior work attributed AI's high passive rate to *accessibility* alone (Rodrigo et al., 2018; Li et al., 2025). However, Study 2 revealed moderate similarity in AI—some inanimate agents (e.g., *car*, *drone*) possess quasi-agentive properties, suggesting AI activates both mechanisms, though less strongly than AA. Study 3 confirmed this: AI exerted causal influence on II (pushing it toward passive) but shifted toward active when patched onto AA.

II–Uncommitted: Under the *interference* account, II should activate competition due to matched animacy (Gennari et al., 2012). However, II similarity is lower than AA, and masks substantial heterogeneity: some pairs like *bus/car* approach AA levels, while other pairs like *vase/shelf* fall well below. Behaviorally, II shows moderate passive rates (Li et al., 2025), consistent with this intermediate similarity. Study 3 further showed that II representations exerted almost no causal influence on other conditions. Unlike AA, where two animate nouns both possess agentive properties, two inanimate nouns may create some competition but not strong enough to commit to passives as AA.

IA–Open: IA shows the lowest similarity and the lowest behavioral passive rate in LMs and humans (Li et al., 2025; Gennari et al., 2012). The inanimate head noun does not compete for the subject role, and low similarity produces low interference. Compared to the other three, for IA, active is more acceptable because no constraint pushes against it. Study 3 confirmed this malleability: patching other configurations onto IA increased passive preference.

Therefore, we propose that *interference* and *accessibility* operate jointly and their combined strength determines how decisively a configuration commits to passive. This extends constraint-based models of sentence production (MacDonald, 2013; Bock and Warren, 1985) by providing mechanistic evidence with GPT-2.

7.3 LMs as mechanistic testbeds for psycholinguistic theory

This work also showcases how mechanistic interpretability can help us understand psycholinguistic theories in ways that behavioral methods alone cannot. Human experiments have established animacy-driven structure choice (MacDonald, 2013; Gennari et al., 2012; Bock and Warren, 1985; Gordon et al., 2001) but cannot manipulate internal representations. Neuroimaging provides correlational windows but lacks precision to isolate specific computations (Poeppel and Embick, 2005). ERP (event-related potential) studies have identified neural signatures such as N400 modulations for animacy violations (Nieuwland and Van Berkum, 2006), but these signals reflect aggregate neural responses rather than isolable computational steps. LMs offer a complementary approach: we intervene on representations and identify components responsible for behavioral effects (Heimersheim and Nanda, 2024).

The activation patching methodology exemplifies this (Vig et al., 2020; Meng et al., 2022). By swapping animacy representations between conditions and measuring downstream effects, we established causality. Crucially, within-animacy null controls allowed us to approximate the animacy-specific contribution despite animacy being inherently confounded with lexical identity, a methodological control we encourage future computational psycholinguistics work to consider (Heimersheim and Nanda, 2024; Zhang and Nanda, 2024).

A natural follow-up question is whether our findings generalize beyond GPT-2 Small, such as the more modern open-weights models Pythia and Llama-3-8B. Mechanistically, recent work indicates that circuits identified in small models often replicate at larger scales: Tigges et al. (2024) showed that circuits found in models with different scales of parameters, have the same algorithms implemented, even when specific attention heads differ. For next step, exploring whether the structural commitment gradient and passive-promoting/passive-suppressing head split we identify here replicate across model scale and architecture properties can be a productive direction for follow-up work.

We do not claim GPT-2's mechanisms are identical to human sentence production. The model lacks embodiment, communicative intent, and developmental trajectory (Bender and Koller, 2020), and behavioral similarity alone does not establish mechanistic correspondence (Bowers et al., 2023). However, the parallels between model and human behavior suggest some computational principles may be shared. At minimum, this study attempted to examine the glass box to see how behavioral patterns *can* emerge from a model's internal representations.

8 Conclusion

This study demonstrates that animacy representations in GPT-2 causally drive syntactic structure choice, moving beyond behavioral alignment. Studies 1–2 established that animacy is encoded categorically and that joint animacy configurations produce different representational similarities. Studies 3–4 showed these representations causally influence structure preference through a circuit of 22 attention heads with opposing *passive-promoting* and *passive-suppressing* populations. Activation patching revealed a *structural commitment gradient*: AA commits decisively to passive and resists perturbation, AI leans toward passive but remains relatively flexible, II shows weak commitment despite matched animacy, and IA remains open to either structure. Our evidence suggests *interference* and *accessibility* jointly function and their combined strength determines how strongly a configuration commits to passive. These findings provide mechanistic grounding for animacy effects on syntactic choice and demonstrate how interpretability methods can enrich and test the predictions of psycholinguistic theories.

9 Limitations

We acknowledge several limitations. First, our stimuli used restricted sentence frames based on psycholinguistics-driven experiment (Li et al., 2025) to control for confounds, however, this approach might limit generalization to naturalistic production. Study 1's probing analysis used a relatively small noun set, and the held-out test nouns, while disjoint from the probe's training set, are indeed common English nouns that GPT-2 has in its pretraining data. Our generalization test therefore might not establish true out-of-distribution generalization to novel or rare nouns. Stronger tests of representational generalization would extend the probe to low-frequency items, nonce words, or contextually atypical animacy. Second, we chose GPT-2 Small for the extensive prior interpretability work

on it (Conmy et al., 2023; Wang et al., 2023; Vig et al., 2020), but it is a relatively small model by current standards. Replicating these findings in more modern open-weight models such as Pythia (Biderman et al., 2023) or Llama-3 (Grattafiori et al., 2024) would establish whether the structural commitment patterns we observe generalize across architectures and scales, or are particular to GPT-2. Third, our activation patching operates on entire residual stream representations. While within-animacy null corrections control for item-level lexical variation, they cannot fully rule out category-level properties that might covary with animacy (e.g., concreteness). Future work should employ feature-level patching (Ravfogel et al., 2020; Geiger et al., 2024) along probe-defined animacy directions to isolate the animacy-specific causal contribution with greater precision. Fourth, our circuit analysis focused exclusively on attention heads; MLP layers may also contribute to animacy-driven structure selection. Moreover, while we identified 22 critical heads, the specific computations they perform, whether analogous to retrieval interference, accessibility checking, or constraint integration, remain to be characterized through detailed circuit analysis.

Future directions should also consider extending this approach to other animacy-sensitive alternations (e.g., dative, genitive) and integrating with human neuroimaging data to test whether the layer-wise dynamics observed here correspond to temporal dynamics in human sentence production.

Code and Data Availability

All code is available at `https://github.com/Yue00831/Animacy_LLMech_CoNLL2026`.

Acknowledgments

We thank the anonymous reviewers for their insightful and constructive feedback. We gratefully acknowledge funding support from Department of English at Purdue University. We acknowledge the Computation and Linguistic Meaning (CALM) Lab and Experimental Linguistics Lab (ExLing) at Purdue for additional support.

References

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

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.

J. Kathryn Bock and Richard K. Warren. 1985. Conceptual accessibility and syntactic structure in sentence formulation. *Cognition*, 21(1):47–67.

Jeffrey S Bowers, Gaurav Malhotra, Marin Dujmović, Milton Llera Montero, Christian Tsvetkov, Valerio Biscione, Guillermo Puebla, Federico Adolfi, John E Hummel, Rachel F Heaton, and 1 others. 2023. Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 46:e385.

Charlotte Caucheteux and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134.

Yan Cong. 2024. Manner implicatures in large language models. *Scientific Reports*, 14(1):29113.

Yan Cong. 2025. Second language learning of degree expressions: A computational approach. *Natural Language Processing*, 31(5):1187–1209.

Yan Cong, Arianna N LaCroix, and Jiyeon Lee. 2024. Clinical efficacy of pre-trained large language models through the lens of aphasia. *Scientific Reports*, 14(1):15573.

Yan Cong and Jiyeon Lee. 2025. Tracking priming-induced language recovery in aphasia with pre-trained language models. *Frontiers in Artificial Intelligence*, 8:1668399.

Yan Cong and Julia Rayz. 2025. Language models demonstrate the good-enough processing seen in humans. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pages 16318–16352. Curran Associates, Inc.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Shaohua Fang, Yue Li, and Yan Cong. 2025a. Quantifier scope interpretation in language learners and LLMs. *arXiv preprint arXiv:2509.10860*.

Shaohua Fang, Yue Li, and Yan Cong. 2025b. Understanding quantifier scope with large language models: How many children climbed trees? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.

Shaohua Fang, Yue Li, and Yan Cong. 2026. Semantic capacity in language learners and LLMs: A case study of quantifier scope. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 9602–9617. ELRA Language Resources Association (ELRA).

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1828–1843, Online. Association for Computational Linguistics.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. 2024. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 160–187. PMLR.

Silvia P. Gennari, Jelena Mirković, and Maryellen C. MacDonald. 2012. Animacy and competition in relative clause production: A cross-linguistic investigation. *Cognitive Psychology*, 65(2):141–176.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, and 1 others. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*, pages 10–18.

Peter C Gordon, Randall Hendrick, and Marcus Johnson. 2001. Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(6):1411–1423.

Peter C Gordon, Randall Hendrick, and Marcus Johnson. 2004. Effects of noun phrase type on sentence complexity. *Journal of Memory and Language*, 51(1):97–114.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Olivia Guest and Andrea E Martin. 2023. On logical inference over brains, behaviour, and artificial neural networks. *Computational Brain & Behavior*, 6(2):213–227.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Michael Hanna, Yonatan Belinkov, and Sandro Pezzelle. 2023. When language models fall in love: Animacy processing in transformer language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12120–12135, Singapore. Association for Computational Linguistics.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

Anna A Ivanova. 2023. Running cognitive evaluations on large language models: The do's and the don'ts. *arXiv:2312.01276v1*.

Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau, and Ryan Cotterell. 2022. Probing for the usage of grammatical number. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8818–8831, Dublin, Ireland. Association for Computational Linguistics.

Maximilian Li and Lucas Janson. 2024. Optimal ablation for interpretability. In *Advances in Neural Information Processing Systems*, volume 37, pages 109233–109282. Curran Associates, Inc.

Yue Li, Yan Cong, and Elaine J. Francis. 2025. Beyond binary animacy: A multi-method investigation

of LMs' sensitivity in English object relative clauses. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 184–196, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

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

Maryellen C MacDonald. 2013. How language production shapes language form and comprehension. *Frontiers in Psychology*, 4:226.

Willem M. Mak, Wietske Vonk, and Herbert Schriefers. 2002. The influence of animacy on relative clause processing. *Journal of Memory and Language*, 47(1):50–68.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. 2024. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *The Twelfth International Conference on Learning Representations*.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, and 4 others. 2025. MIB: A mechanistic interpretability benchmark. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 45069–45108. PMLR.

Mante S. Nieuwland and Jos J. A. Van Berkum. 2006. When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, 18(7):1098–1111.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

David Poeppel and David Embick. 2005. Defining the relation between linguistics and neuroscience. In Anne Cutler, editor, *Twenty-first century psycholinguistics: Four cornerstones*, pages 103–120. Lawrence Erlbaum Associates, Mahwah, NJ.

Mercè Prat-Sala and Holly P Branigan. 2000. Discourse constraints on syntactic processing in language production: A cross-linguistic study in english and spanish. *Journal of Memory and Language*, 42(2):168–182.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Laura Rodrigo, José M Igoa, and Hiromu Sakai. 2018. The interplay of relational and non-relational processes in sentence production: The case of relative clause planning in japanese and spanish. *Frontiers in Psychology*, 9:1573.

Martin Schrimpf, Idan A. Blank, Greta Tuckute, Carina Kauf, Elham A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. 2024. Llm circuit analyses are consistent across training and scale. In *Advances in Neural Information Processing Systems*, volume 37, pages 40699–40731. Curran Associates, Inc.

Matthew J Traxler, Robin K Morris, and Rachel E Seely. 2002. Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language*, 47(1):69–90.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

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

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of*

the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Ethan G Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society (CogSci 2020)*.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Fred Zhang and Neel Nanda. 2024. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*.

A Appendix

A.1 Appendix A: Layerwise patching results

Figure 6 shows the layer-wise activation patching effects for all eight bidirectional transitions in Study 3.

A.2 Appendix B: Full statistical reporting of study 3, bidirectional patching

Table 2 reports the full statistical details for all eight bidirectional transitions in Study 3. Raw effects reflect the mean change in structure preference (Δ) from activation patching. The null column shows the condition-matched within-animacy baseline for each target condition. Corrected effects subtract this baseline to isolate animacy-specific causation. Confidence intervals are computed on the corrected effects.

A.3 Appendix C: Full Study 4 critical heads

Table 3 lists all 22 attention heads meeting the structural, sensitivity, and causal criteria described in Study 4.

A.4 Appendix D: Relativizer absence

To verify that the within-sentence similarity pattern is not an artifact of the explicit relativizer *that*, we recomputed all analyses separately for stimuli containing *that* (full relative clauses; $n = 192$) and for reduced relative clauses without it ($n = 192$).

- Active *The baby (that) the father holds is crying.*
- Passive *The baby (that is) held by the father is crying.*

Overall means by condition were nearly identical across relativizer conditions (Table 4). All pairwise contrasts that were significant in the pooled analysis remained significant in both subsets, with effect sizes within 0.03 of each other (Table 4). A two-way ANOVA (animacy $\times$ relativizer) on within-sentence similarity confirmed a robust main effect of animacy ($F(3, 4600) = 174.78$, $p < 10^{-100}$), no main effect of relativizer ($F(1, 4600) = 0.08$, $p = .77$), and no interaction ($F(3, 4600) = 0.03$, $p = .99$). The animacy-driven similarity pattern is therefore not an artifact of the relativizer cue.

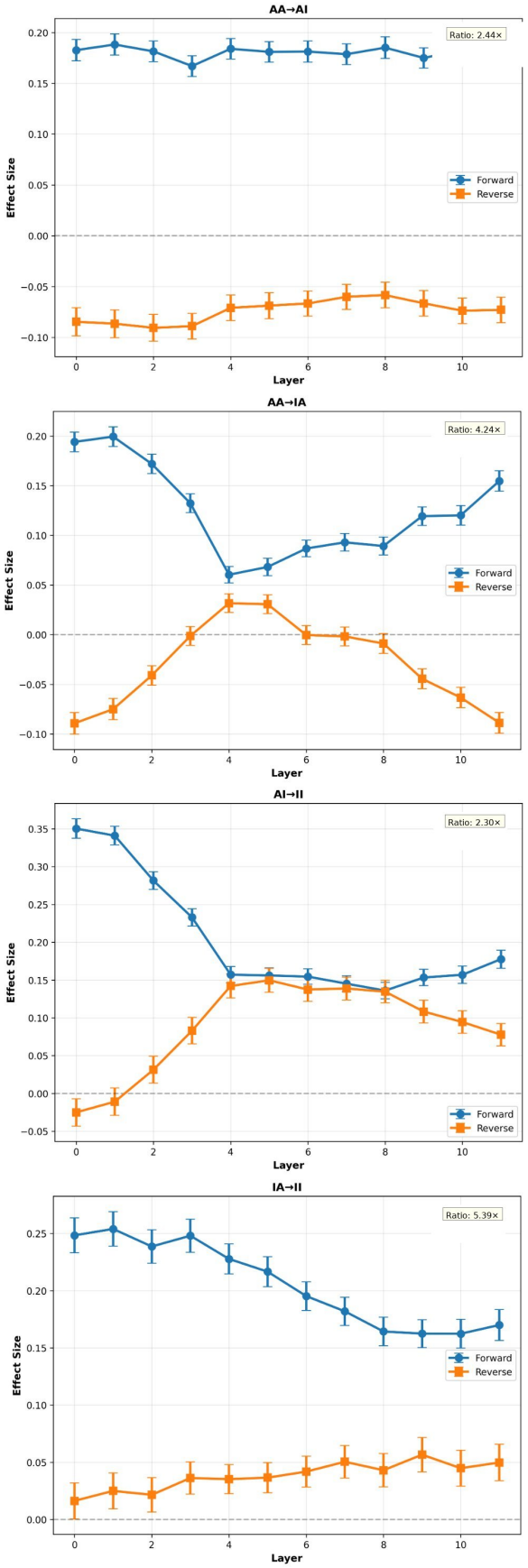

Figure 6: Study 3 activation patching: layer-wise results for all eight transitions. Forward (blue) and reverse (yellow) directions shown for each transition pair.

Transition	Direction	Raw Effect	Null Baseline	Corrected Effect	95% CI	n
AA→AI	Forward	+0.180	0.076	+0.104	[+0.099, +0.110]	3,840
AA→IA	Forward	+0.124	−0.008	+0.132	[+0.127, +0.138]	4,080
AI→II	Forward	+0.204	0.077	+0.127	[+0.120, +0.133]	3,648
IA→II	Forward	+0.206	0.077	+0.129	[+0.121, +0.136]	3,876
AI→AA	Reverse	−0.074	−0.002	−0.072	[−0.079, −0.064]	3,840
IA→AA	Reverse	−0.029	−0.002	−0.027	[−0.033, −0.021]	4,080
II→AI	Reverse	+0.088	0.076	+0.012	[+0.003, +0.022]	3,648
II→IA	Reverse	+0.038	−0.008	+0.046	[+0.038, +0.054]	3,876

Table 2: Study 3: Raw effects, condition-matched null baselines, null-corrected effects, and 95% confidence intervals for all eight bidirectional transitions. n = number of patching observations per transition (items × layers).

Head	Type	η^2	Ablation Effect	p_{FDR} (ablation)
L0H9	passive-suppressing	0.191	+0.338	$<10^{-57}$
L11H0	passive-promoting	0.052	−0.069	$<10^{-75}$
L11H8	passive-suppressing	0.037	+0.050	$<10^{-87}$
L0H7	passive-promoting	0.033	−0.045	$<10^{-4}$
L1H1	passive-suppressing	0.135	+0.023	$<10^{-14}$
L8H5	passive-suppressing	0.058	+0.022	$<10^{-37}$
L6H7	passive-promoting	0.095	−0.021	$<10^{-20}$
L3H7	passive-promoting	0.235	−0.018	$<10^{-15}$
L5H10	passive-promoting	0.243	−0.017	$<10^{-11}$
L0H6	passive-suppressing	0.065	+0.017	$<10^{-4}$
L4H3	passive-promoting	0.175	−0.014	$<10^{-4}$
L6H8	passive-promoting	0.043	−0.014	$<10^{-19}$
L4H6	passive-suppressing	0.034	+0.013	$<10^{-11}$
L5H7	passive-promoting	0.129	−0.012	$<10^{-9}$
L9H0	passive-suppressing	0.069	+0.009	$<10^{-9}$
L1H2	passive-promoting	0.054	−0.009	$<10^{-10}$
L11H3	passive-suppressing	0.056	+0.009	$<10^{-12}$
L2H7	passive-suppressing	0.135	+0.008	$<10^{-11}$
L3H6	passive-promoting	0.245	−0.006	0.009
L4H1	passive-promoting	0.035	−0.005	0.004
L10H9	passive-suppressing	0.087	+0.003	0.008
L11H11	passive-promoting	0.207	−0.003	0.013

Table 3: All 22 critical attention heads meeting structural, sensitivity, and causal criteria, sorted by ablation effect magnitude. η^2: effect size from Kruskal-Wallis test of animacy sensitivity. Ablation effect: change in passive preference (Δ) when the head is zeroed. Positive = passive-suppressing; negative = passive-promoting.

(a) Overall means				(b) Pairwise comparisons				
Condition	With *that*	Reduced	Pooled	Comparison	With *that*		Reduced	
					d	p	d	p
AA	0.762	0.764	0.763	AA vs AI	0.73	$< 10^{-32}$	0.76	$< 10^{-34}$
II	0.699	0.700	0.700	AA vs IA	0.96	$< 10^{-53}$	0.98	$< 10^{-55}$
AI	0.698	0.698	0.698	AA vs II	0.71	$< 10^{-31}$	0.74	$< 10^{-33}$
IA	0.680	0.682	0.681	AI vs IA	0.18	.002	0.17	.004
				AI vs II	−0.01	.84 (n.s.)	−0.02	.77 (n.s.)
				IA vs II	−0.20	$< .001$	−0.19	.001

Table 4: Robustness of the within-sentence similarity pattern to relativizer presence. **(a)** Within-sentence cosine similarity by animacy condition, separately for each relativizer condition and pooled across both ($n = 576$ layer×item observations per cell in each split; $n = 1{,}152$ pooled). **(b)** Pairwise condition comparisons by relativizer condition. Effect sizes (Cohen's d) and p-values from independent-samples t-tests. The same significant comparisons emerge in both relativizer conditions, with effect sizes differing by at most 0.03.