From LLM–Human Behavioral Alignment to Mechanistic Interpretability

Animacy and Similarity-Based Interference in Object Relative Clauses

Yue Li

PhD Candidate in Linguistics · Purdue University
Exling Lab & CALM Lab

slides

The SEAL Lab

experimental meaning at northwestern

May 2026 · SEAL Lab & Experimental Meaning Group · Northwestern University

How Production Planning Shapes Structure Choice in Humans

THEORETICAL FRAMEWORK

When speakers have grammatical alternatives, what determines which structure they choose?

Production-Distribution-Comprehension (PDC)

MacDonald (2013): production biases shape syntactic choices, give rise to distributional regularities, and guide comprehension.

Three production biases

Easy First -- start with the most accessible element

Reduce Interference -- space out similar, competing elements

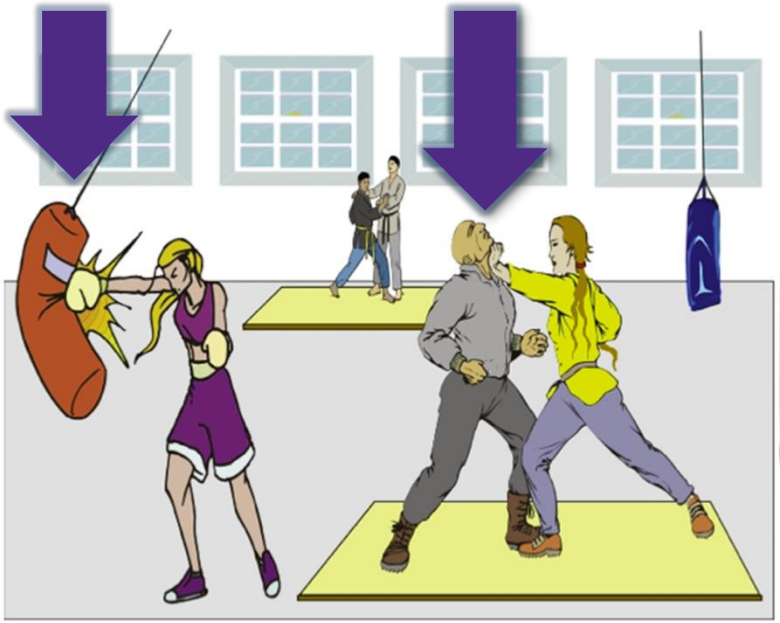
Plan Reuse -- reuse well-practiced syntactic plans

EMPIRICAL WINDOW

Animacy in object relative clause (ORC)

MacDonald (2013) · Li & Francis (under review, Cognitive Linguistics) · Francis, Li et al. (in prep, 2026)

Animacy → Object Relative Clauses



Gennari et al. (2012)

Animate

the man that's punched by the woman

the man that the woman punches

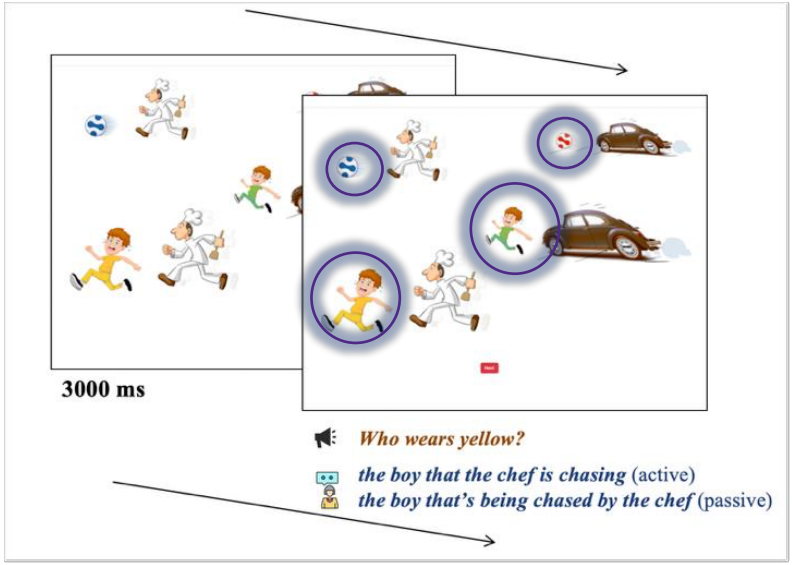
Inanimate

the sandbag that's punched by the woman

the sandbag that the woman punches

Gennari et al. (2012); Humphrey et al. (2016)

Experimental Method



Example scene depicting four animacy configurations (AA, AI, IA, II)

Participants & Design

39 native English-speaking adults (M = 19.85, SD = 2.17)

AA	AI	II	IA
boy-chef	boy-car	Football-car	Football-chef

Tasks

Production: picture-based elicitation

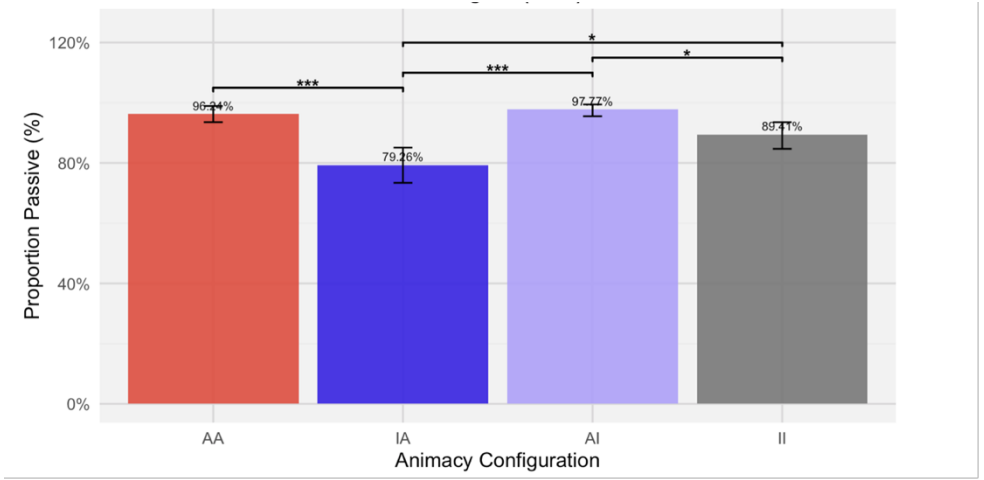
Comprehension: decision task

Experimental Results (Human Data)

Mechanism that promotes *passive*	AA boy-chef	IA football-chef	AI boy-car	II football-car

Production: animacy effect on ORC structure choice

- Animacy → main effect
- **Head animacy** drives passive choice → accessibility
- **II > IA** isolates similarity-based competition



Passive ORC production rates across the four animacy conditions

From human production to LLM expectations

1. HUMAN PRODUCTION

Structure choice distribution across AA, AI, IA, II

aggregated over speakers

2. NATURALISTIC TEXT

Production biases imprint on corpus distributions

training

3. LLM SURPRISAL

Should mirror the same animacy effect?

Production-Distribution-Comprehension (MacDonald, 2013)

Why look at LLMs?

Under the **PDC framework**, production biases shape the distributional regularities of language input.

LLMs trained on naturalistic text serve as a **distributional probe**: do their next-word expectations track the same effect observed in human production?

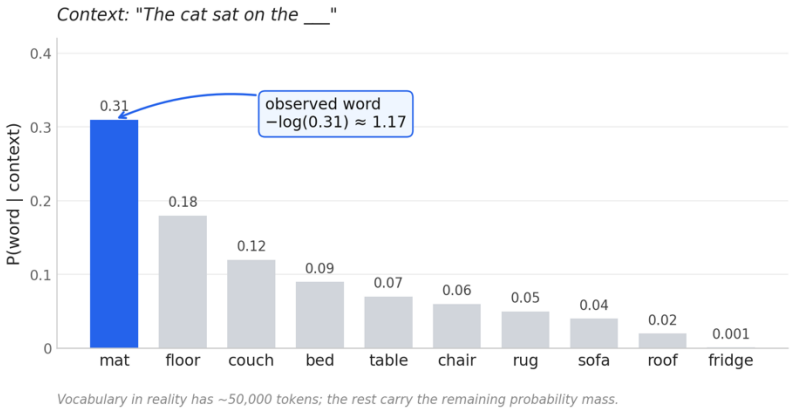
LLM evaluation work 1

Li, Cong & Francis (2025): Does the **animacy effect** we just saw in humans appear in LMs too?

What is surprisal?

A behavioral measure read directly off the model's probability distribution.

- **–log P(word | context)**
  - **Tokens** get mapped to vectors. The model crunches them through transformer layers.
  - **Output:** a probability distribution over the entire vocabulary (~50K tokens).
  - **Surprisal** = **–log** of the probability the model assigned to the word that actually appeared
  - High prob → **low surprisal** (the model finds it expected).
  - Low prob → **high surprisal**.



If the actual word is "mat" with P = 0.31,
surprisal = −log(0.31) ≈ 1.17

How we tested animacy effect with the LMs

For each context, we score a passive ORC and its active counterpart: whichever has lower surprisal is the model's choice.

1. How a minimal pair works

Context story

There are two babies, a mother, and a father in the scene. The father holds the crying baby. The mother holds the smiling baby.

Which baby is crying?

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

Whichever has **lower mean surprisal** = the model's "choice"

Repeat × 384 pairs (96 per animacy condition: AA, IA, AI, II)

2. Four LMs: 2 × 2 design

	Smaller	Larger
Causal *(left context only)*	**DistilGPT-2** 82M	**GPT-Neo** 1.3B
Masked *(left + right context)*	**BERT-base** 110M	**BERT-large** 340M

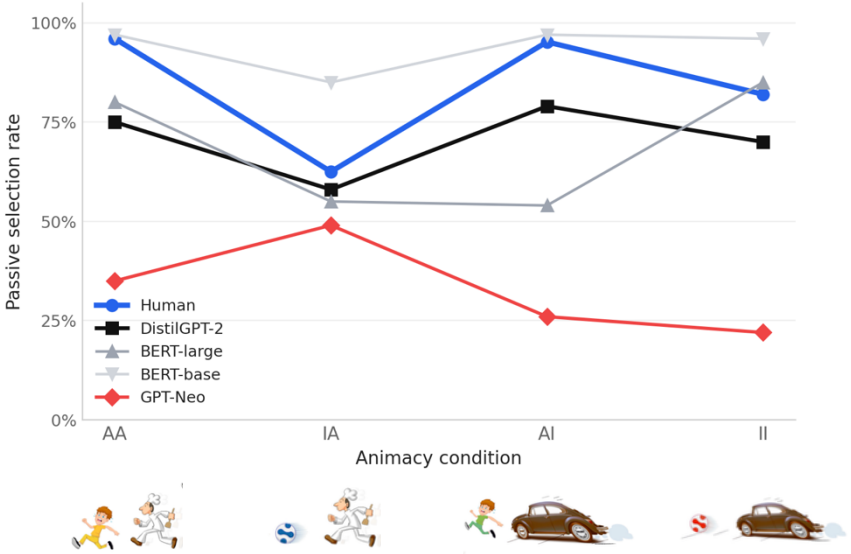
Causal LMs match how humans process incrementally; masked LMs see full context. Two sizes per family let us check whether scale drives any effect.

Outcome variable: structure choice per condition, per LM; compared to human baseline.

Method: minimal-pair surprisal scoring (Cong 2022; Misra 2022) · Li, Cong & Francis (CMCL 2025)

LMs Behavior Evaluation

Results



LLM-Human alignment found but varied by model

- **DistilGPT-2 mirrors humans the most:** r = 0.98, RMSE = 0.14

Model size doesn't predict alignment

Is DistilGPT-2's alignment just a frequency echo?

From the last slide: DistilGPT-2 mirrored humans nicely (r = 0.98). But did it *learn* animacy, or just memorize the corpus distribution?

We mined ORCs from DistilGPT-2's training corpus

- **Source:** OpenWebText (open-source reproduction of GPT-2's training data)
- **Sample:** ~8,000 sentences randomly drawn
- **Pipeline:** custom SpaCy parser detects ORCs (head noun + embedded VP + agent), then **manual validation**
- **Coding:** each ORC labeled for **structure** (active/passive) and **animacy** of head + agent (AA / IA / AI / II)
- **Test:** Pearson correlation: do corpus frequencies predict (a) human responses, (b) DistilGPT-2's surprisal pattern?

Look into the training data

If the model's alignment **is** a frequency echo:

→ *corpus distribution should predict the surprisal pattern (and the human pattern)*

If it's **not**:

→ *corpus shouldn't predict it well, suggesting an emergent representation*

OpenWebText corpus (Gokaslan & Cohen 2019) · Li, Cong & Francis (CMCL 2025)

Corpus is skewed; alignment *isn't* a frequency echo

- **Corpus analysis: results**

The corpus is heavily skewed

- Among ORCs: **71% active, 29% passive**; the **opposite** of human production
- **Active-IA dominates:** 53% of all ORCs; the corpus heavily includes actives in IA

ORCs in OpenWebText sample (% of total)

	AA	IA	AI	II	Total
Passive	0.93	13.08	0.93	14.02	28.97
Active	3.74	**53.27**	0.93	13.08	71.03
Total	**4.67**	**66.36**	**1.87**	**27.10**	**100**

Active-IA alone = 53% of all ORCs

Corpus distribution *doesn't* predict the patterns surprisal showed.

- Corpus → humans: $R^2 = 0.12$, $p = .66$ | Corpus → DistilGPT-2: $R^2 = 0.26$, $p = .49$
- DistilGPT-2 alone → humans: $R^2 = 0.96$, $p = .02$ | Adding corpus barely changes it ($R^2 = 0.99$, n.s.)

→ DistilGPT-2's alignment isn't a frequency echo: it looks emergent.

Prompting LLMs

- **Prompting-based analysis: method**

Surprisal (last 2 slides)

Reads the model's probability distribution **directly:** what the model *represents*.

- Needs token-level access
- Limits us to open-weights LMs (BERT, DistilGPT-2, GPT-Neo)

Prompting

Asks the model to **judge** which sentence is more natural, what the model can *report*.

- Works on more advanced models (GPT-4o-mini, Gemini, DeepSeek)
- Tests metalinguistic awareness, not just representation

Hu & Levy (2023): the two methods can disagree within the same model: running both gives convergent evidence and tests a wider model set.

The prompt

`Read the following context carefully … Two possible answers are provided. Choose the answer that sounds most natural to a native English speaker. Respond with "1" for Passive or "2" for Active.`

Same 384 minimal pairs (96/condition) as the surprisal analysis.

Models tested

The 4 from before:

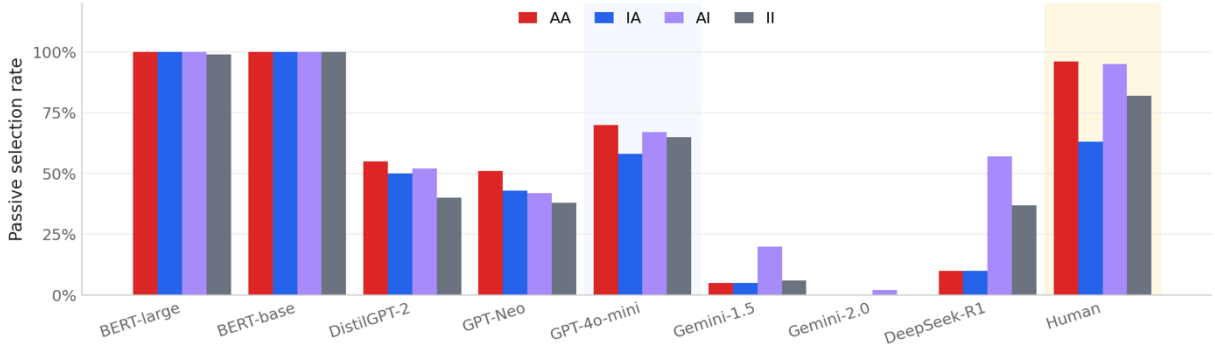
BERT-base, BERT-large, DistilGPT-2, GPT-Neo

+ 4 new state-of-the-art LMs (2025):

GPT-4o-mini, Gemini-1.5-flash, Gemini-2.0-flash, DeepSeek-R1

GPT-4o-mini wins; the rest collapse

- **Prompting-based analysis: results**



- **BERT models:** floor pegged at ~100% passive across all conditions; no animacy effect
- **Gemini-1.5 / 2.0:** reverse: ~100% active; Gemini-2.0 explicitly justified actives as "more direct"
- **DeepSeek-R1:** varies, but in a theoretically ungrounded way (no AA > IA contrast)
- **GPT-4o-mini:** mirrors human decision the best; $r \approx 0.98$, $R^2 = 96\%$, RMSE = 0.21

Surprisal vs. prompting

Some models give **different answers under each method:**

- DistilGPT-2 won on surprisal but only explains 12% of human variance via prompting
- BERT-large was sensitive on surprisal, not on prompting

→ Confirms Hu & Levy: prompting alone is unreliable for assessing linguistic competence.

From behavioral to mechanistic

Behavioral alignment doesn't establish causal correspondence.

WHAT WE'VE SHOWN

LLM-Human Behavioral alignment

DistilGPT-2's surprisal mirrored human passive rates across all four animacy conditions.

r = 0.98

Not reducible *to corpus frequencies (R^2 = 0.12, n.s.).*

BUT...

Right answer, right reason?

A model can match human behavior using **completely different internal pathways**.

Does animacy ***causally drive*** structure choice, or just correlate with it?

(McCoy et al. 2019; Ivanova 2023)

WHAT'S NEXT

Mechanistic test

If we **swap animacy representations** inside the model and structure choice *shifts* → animacy is causal.

If nothing changes → the alignment was *incidental*.

Method: activation patching (Vig et al. 2020; Heimersheim & Nanda 2024)

Setup. Same stimuli (Li et al. 2025). | **Model: GPT-2 Small** (12 layers, 144 attention heads: tractable for mechanistic analysis).
4 studies: probing → representational similarity → activation patching → circuit identification

Study 1: Does GPT-2 even encode animacy?

Before asking if animacy drives behavior, we check that the model represents it as a generalizable feature.

- **What is a probe?**

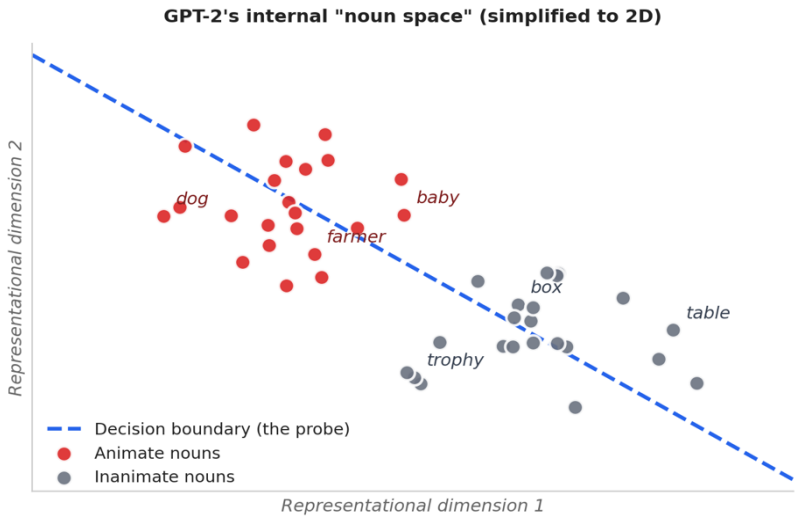
When GPT-2 reads a noun, it converts it into a **vector:** a list of 768 numbers that lives in a high-dimensional "noun space."

A **probe** is a simple decision rule (logistic regression) that asks:

> *"Can I draw a line through this space that puts animate nouns on one side and inanimate nouns on the other?"*

If yes → the model has organized its noun space along an **animacy dimension**.

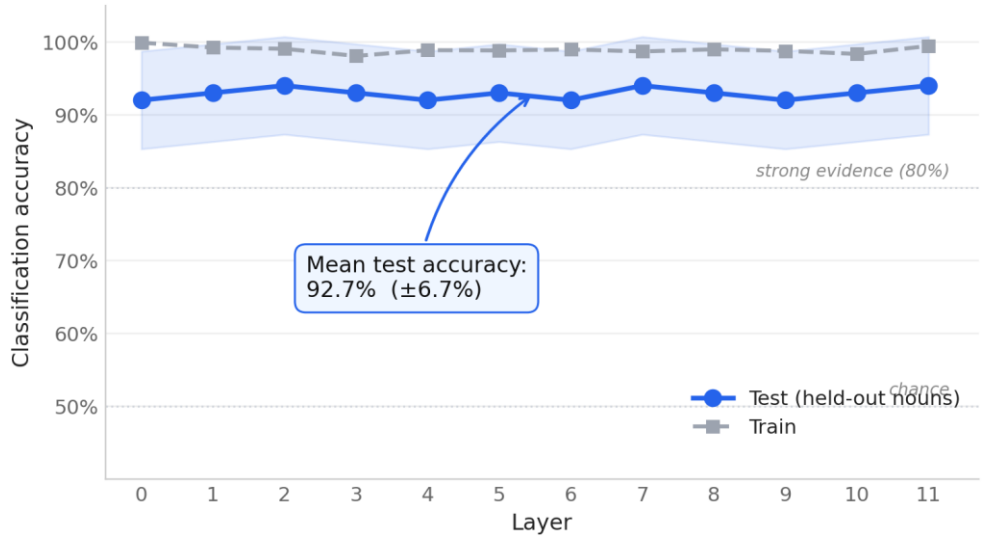
Test: train the probe on **some** nouns, test on **new** ones.



The real space is 768-D, but the same logic applies.

Yes, and it generalizes

- **Study 1: result & what it does and doesn't show**



What this tells us

- Animacy **generalizes:** not memorized lexical knowledge
- The **representational prerequisites** for animacy-based processing are in place

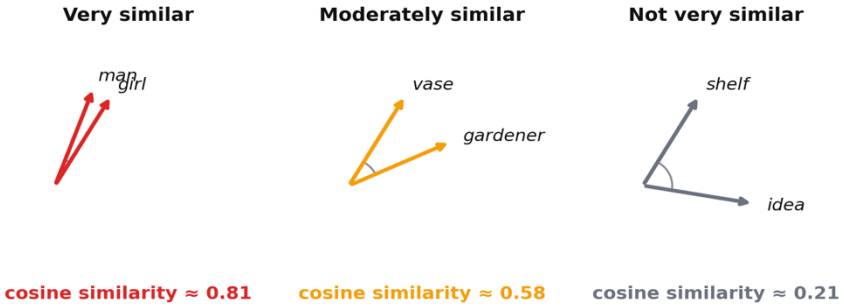
What it does NOT show

Encoding ≠ functional use. The model *has* the information, but whether it *uses* it for structure choice is the question for **Studies 3–4**. *(Belinkov 2022)*

Study 2: How similar are the two nouns to each other?

Similarity-based Interference theory *(Gennari et al., 2012, MacDonald, 2013) predicts: similar nouns compete for the subject role, so we need a way to measure how similar two noun representations are.*

- **What is cosine similarity?**



the man **that** ***the woman*** **hugs**

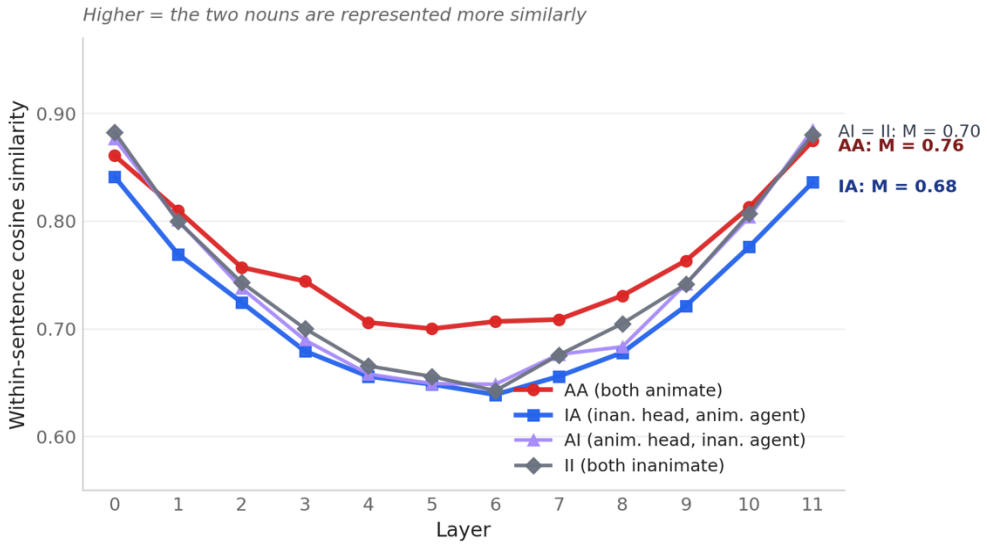
What we calculate:

Cos_Similarity between head noun and agent noun

Prediction: if interference is indeed playing a role, AA should show the highest similarity (two competing animate agents).

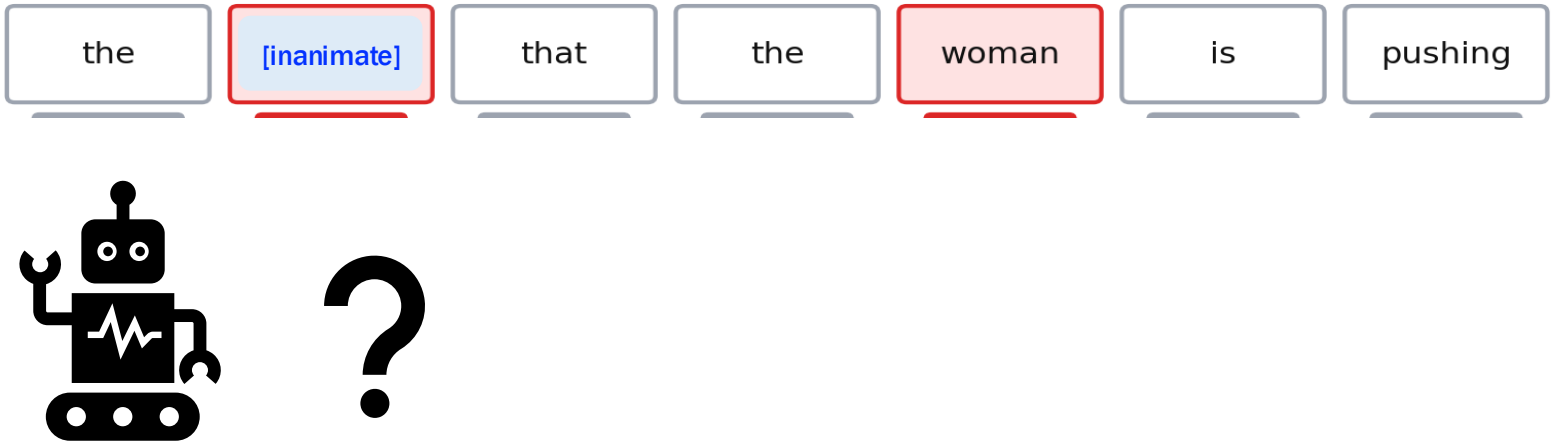
Partial match to interference theory

- **Study 2: result**



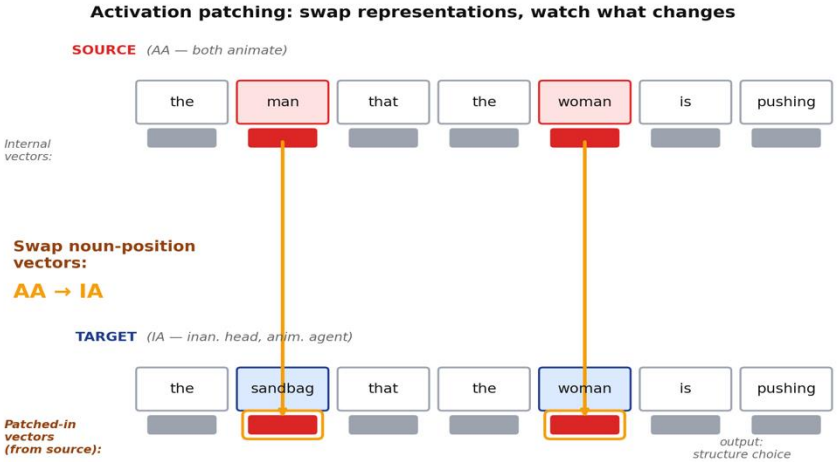
- **AA highest cos_sim**: two animate nouns share agentive properties; they compete for subject role → Highest passive rate (Li et al., 2025)
- **IA lowest cos_sim**: inanimate head can't compete for agenthood; thematic role is clear → lowest passive rate (Li et al., 2025)

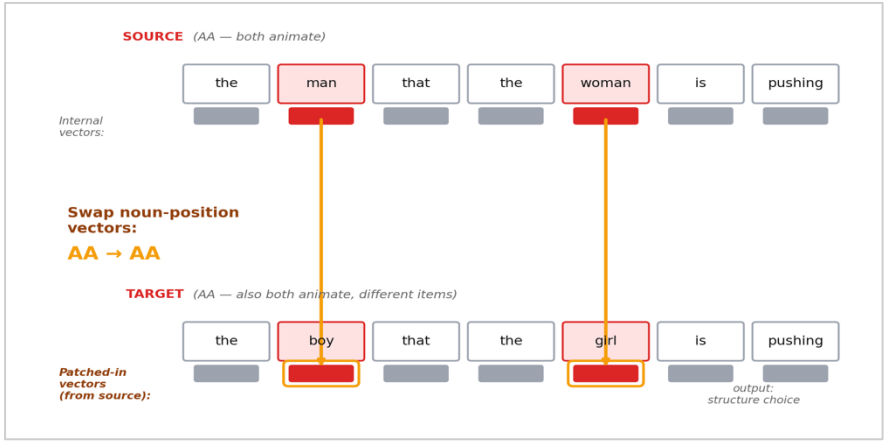
Study 3: Does animacy *cause* structure choice variation?



Study 3: Does animacy *cause* structure choice variation?

- **Activation patching: the method** *(Heimersheim and Nanda, 2024; Zhang and Nanda, 2023).*





The logic

- If the swap **shifts** structure choice → animacy is causal.
- If nothing happens → the alignment was incidental.

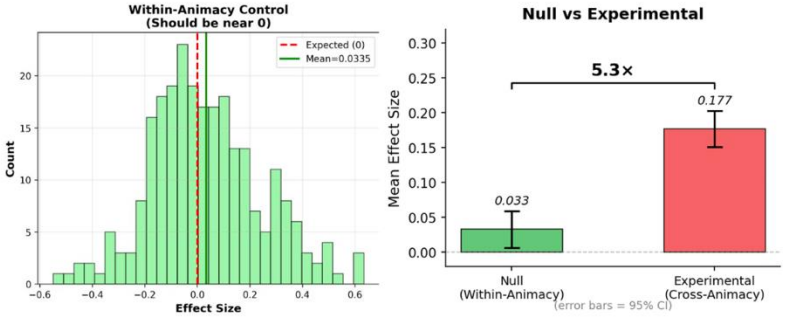
Null control:

- **Within-animacy patching**: noise floor for lexical variation

Study 3: Does animacy *cause* structure choice variation?

Animacy is causally driving structure choice, not a frequency echo, not a lexical confound.

- **Activation patching: Result**



Finding:

- experimental effect: **5.3× larger** than the noise floor

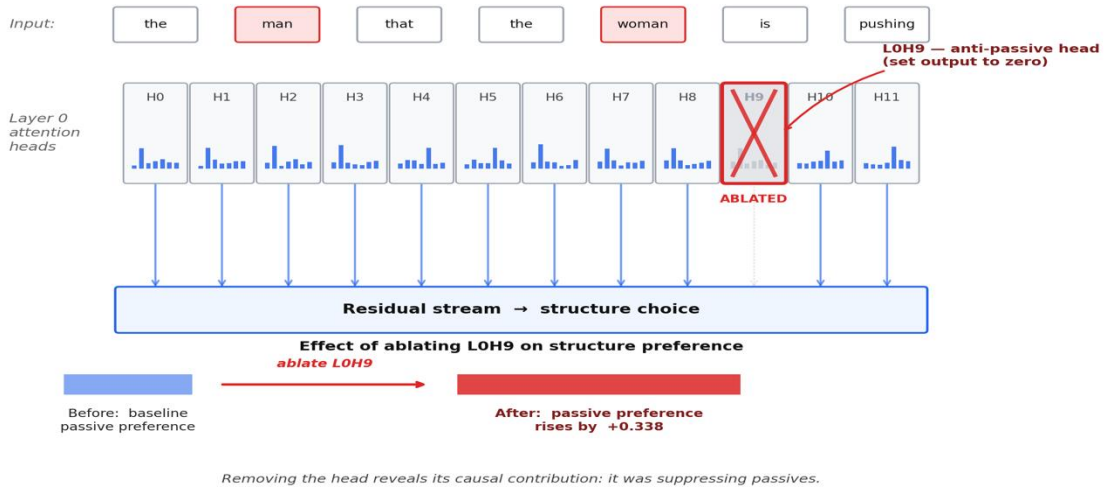
What we think this means:

The bulk of the effect might not merely be lexical confound (within-animacy controls for that). Animacy itself is what's driving structure choice.

Study 4: Where in the model is animacy doing its work?

Zero out one head's output, leave everything else intact, and watch how structure choice shifts.

- **Ablation = causal subtraction**



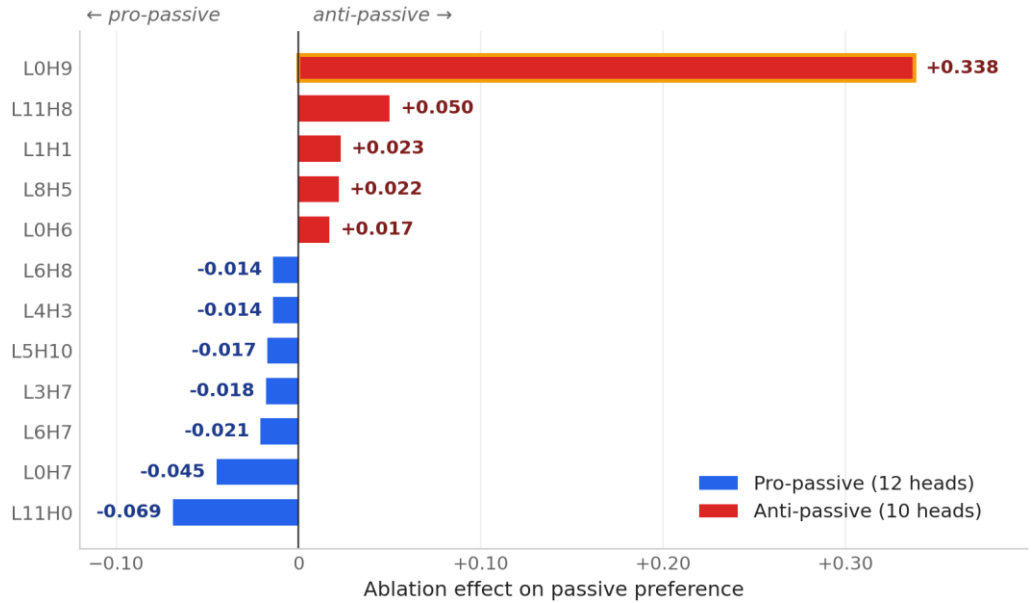
Removing the head reveals its causal contribution: it was suppressing passives.

The logic. If removing a head **changes the model's output**, that head was **causally necessary**. The size of the change tells us how much, and the direction tells us what kind of work the head was doing.

Study 4: Where in the model is animacy doing its work?

Structure choice emerges from competition between pro-passive and anti-passive heads, not a single unified bias.

- **Study 4: result**



- **Two functional populations:** 12 pro-passive (ablating decreases passives) and 10 anti-passive (ablating increases passives)
- **L0H9 dominates:** ablation effect of +0.338, ~4× the next strongest head

The bigger picture

Structure choice isn't a single unified mechanism, it's the **net result of competition** between opposing forces inside the model.

Animacy causally drives structure choice in GPT-2

Mechanistic interpretability moves us from behavioral alignment to causal grounding.

Summary & contribution

STUDY 1

Probing

Animacy is **encoded** as a generalizable feature.

92.7%

held-out accuracy

STUDY 2

Similarity

AA highest, IA lowest. Inanimates are **heterogeneous**.

AA > AI ≈ II > IA

cosine similarity rank

STUDY 3

Patching

Animacy **causally** shifts structure choice.

5.3×

experimental vs. noise floor

STUDY 4

Circuit

Heads split into **opposing populations**.

12 pro · 10 anti

22 critical attention heads

Contribution. Mechanistic grounding for animacy effects in psycholinguistic theory

Future direction 1: Teasing apart accessibility vs. competition

So far, patching swaps both nouns at once, conflating the two mechanisms. Targeted patches can dissociate them.

- **Control one factor, vary the other**

INTERVENTION 1

Isolate accessibility

Patch **only the head noun**: replace its representation, leave the agent untouched.

e.g., "the box that the woman is pushing" → replace "box" with an animate-noun representation, but with similar cosine similarity. Same agent.

If accessibility → passive preference rises.

INTERVENTION 2

Isolate competition

Change the similarity level but hold animacy constant.

e.g., "the box that the woman is pushing" → replace "box" with another inanimate-noun representation with higher similarity to "woman".

If competition → passive preference rises.

Future direction 2: Attention metrics, not just surprisal

- **What can attention metrics offer?**

Surprisal → the **output** view

Tells us **that** a word is hard to predict.

Aggregates over the whole computation. Doesn't reveal which prior items the model accessed.

Attention metrics → the **process** view

Tell us **which prior items** the model is accessing, and how diffusely.

Direct analog to cue-based retrieval & encoding interference in memory-based theories.

Three attention metrics (Parker 2026, CMCL; Ryu & Lewis 2021, 2025)

Earth Mover's Distance

How different two attention distributions are. Higher = more representational reconfiguration between conditions.

Attention-to-target

How much attention the verb sends to the retrieval target. Lower = more interference at retrieval.

Attention entropy

How spread out the attention is. Higher = more uncertainty about which item to retrieve.

The bigger picture. Attention is a candidate **unified mechanism** for encoding and retrieval: with interference emerging from **representational geometry**, not feature loss.

Future direction 3: Beyond GPT-2

MODEL FAMILY 1

Pythia the developmental angle

A suite of open-source models (70M-12B) released with **~150 training checkpoints each**.

New question: When does the animacy effect *emerge* during training, and is the structural commitment gradient learned gradually or all at once?

MODEL FAMILY 2

Llama the scale angle

An open-weights frontier-class family (8B–405B), substantially more **capable** than GPT-2 Small.

New question: Do the same opposing pro/anti circuits *scale up*, or does competence change the architecture? Does AA–Committed survive?

Same pipeline: probing → representational similarity → activation patching → circuit identification

For the room: **Grounding well-documented psycholinguistic phenomena in the internal representations of neural networks is exciting. So much we can do and explore.**

References (Selected)

Berzak, Y., & Levy, R. (2023). Eye movement traces of linguistic knowledge in native and non-native reading. *Open Mind, 7*, 179–196.

Duan, X., Yao, Z., Zhang, Y., Wang, S., & Cai, Z. G. (2025). How syntax specialization emerges in language models. *arXiv:2505.19548*

Fang, S., Li, Y., & Cong, Y. (2025). Understanding quantifier scope with large language models: How many children climbed trees? *Proceedings of CogSci 2025.*

Fang, S.*, Li, Y.*, & Cong, Y. (2026). Semantic capacity in language learners and LLMs: A case study of quantifier scope. *LREC 2026.*

Fang, S.*, Li, Y.*, & Cong, Y. (under review). Processing of quantifier scope: Experimental and modeling evidence from English and Chinese. Under review at *Languages*.

Fang, S.*, Li, Y.*, & Cong, Y. (under review). Language models approximate human sensitivity to the syntax–pragmatics tension but show cross-linguistic and model variation. Under review at *Humanities and Social Sciences Communications.*

Fang, S., Li, Y., Lu, J., & Cong, Y. (in prep). Semantic representation generalization in BabyLMs.

Francis, E. (2022). *Gradient acceptability and linguistic theory.* Oxford University Press.

Francis, E., Li, Y., Khodadadi, G., Mack, M., Ok, S., Bahmanian, N., Fang, S., Michaelis, L. A., Sheu, V., & Weirick, J. (in prep). Effects of verb type and prior context on the production of relative clause extraposition in English.

Fu, Duan, & Cai (2026). SCALPEL: Selective capability ablation via low-rank parameter editing for large language model interpretability analysis. arXiv:2601.07411

Haller, P., Bolliger, L., & Jäger, L. (2024). Language models emulate certain cognitive profiles. *Findings of ACL 2024*, 7878–7892.

Huang, K. J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2024). Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language, 137*, 104510.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126–1177.

Li, Y., & Francis, E. (under review). Rethinking animacy in relative clauses: Accessibility and competition across agent–patient configurations. Under review at *Cognitive Linguistics.*

Li, Y., Cong, Y., & Francis, E. (2025). Beyond binary animacy: A multi-method investigation of LMs' sensitivity in English object relative clauses. *CMCL, NAACL 2025.*

Li, Y., Cong, Y., & Francis, E. (2026). Mechanistic interpretability of animacy effects on structure choice in GPT-2. *CoNLL 2026.*

Li, Y., Khodadadi, G., Sheu, V., & Fang, S. (in prep). Priming and lexical boost effects in the comprehension of English object relative clauses.

Nanda, N., & Bloom, J. (2022). TransformerLens. *GitHub repository.* https://github.com/neelnanda-io/TransformerLens

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2022). Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. *arXiv:2211.00593*

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S. F., & Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics, 8*, 377–392.

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics, 11*, 1451–1470.

Xu, W., Dillon, B., & Futrell, R. (2026). Memory efficiency and resource-rational encoding in human sentence processing.

Thank you

Questions/comments?

Yue Li

PhD Candidate in Linguistics · Purdue University

ExLing Lab & CALM Lab
Contact: li4207@purdue.edu

Dr. Elaine Francis

Dr. Yan Cong

Dr. Shaohua Fang