Language and Science · **MAY 13th 2026** · **University of Chinese Academy of Sciences**

Does a language model know what's alive?

Mechanistic interpretability of animacy representation in GPT-2

animate · *inanimate*

Yue Li
PhD Candidate, Linguistics · Purdue University
ExLing Lab & CALM Lab

Animacy is everywhere, but mostly outside language

IN COGNITION

Animacy is a cognitive priority.

- **Memory.** Animate words are remembered better.
- **Attention.** We spot animals faster than artifacts.
- **Development.** Babies tell living from non-living.
- **Brain.** Distinct brain regions for each.

IN LANGUAGE

...a trace

Selectional restrictions

✓ The **woman** thinks about the trip.

✗ The **rock** thinks about the trip.

Only animate things can think, walk, want...

Rich in the mind. Quite restricted in the text.

Language gives only an ***indirect*** signal of a much bigger cognitive distinction.

Nairne et al. (2013) · New et al. (2007) · Rakison & Poulin-Dubois (2001) · Caramazza & Shelton (1998)

So… can an LLM pick up the trace?

*Humans build animacy from the **whole world**. LLMs only get the **text**.*

HUMANS

Multi-modal experience

- See animals move on their own
- Touch fur, feathers, stone, metal
- Hear that some things talk back
- And then, read about it in words

→ ***Animacy is everywhere in the input.***

LANGUAGE MODELS

Mostly text.

- Billions of tokens of written text
- ✗ No touch, no world
- ✗ No body, no developmental story
- → Animacy must come from text patterns alone

→ ***Can they still pick it up?***

Today's question: Do LLMs respond to animacy in language the way humans do?

One angle: do LLMs adapt when animacy gets atypical?

Hanna, Belinkov & Pezzelle (EMNLP 2023) ***"When Language Models Fall in Love"***

TYPICAL ANIMACY

A shoe stays a shoe.

✓ Naomi had cleaned a fork.

✗ That book had cleaned a fork.

Word ↔ usual animacy. Easy match.

Already known: LLMs handle this.

ATYPICAL ANIMACY

A peanut falls in love.

*"A lucky **peanut** had a big smile.*
*The **peanut** was **elated**..."*

Context flips the entity's animacy.

New question: can LLMs do this?

The question. LLMs only see text. Can they still detect animacy?

LLMs adapt, even from a single word of context

Three convergent findings across GPT-2, OPT, and LLaMA (up to 30B params).

1 **Typical animacy: matches humans.**

On BLiMP minimal pairs, models reach ~80–90% accuracy, close to the human ceiling. Baseline competence confirmed.

2 **Atypical animacy: models adapt over a story.**

Replicating Nieuwland & van Berkum's N400 study with surprisal: at first mention, the model is shocked by "the peanut." By the 3rd–5th mention, animate and inanimate surprisal converge, like the human N400. Stronger models adapt more fully.

Takeaway. Despite text-only training, LLMs pick up animacy as a functional feature.

?

*Does the model know "animacy" internally? If so, **how**?*

A controlled testbed

*To go from **does it respond** to **how does it do so**, we need a phenomenon we can **reverse-engineer**.*

THREE THINGS A GOOD MECHANISTIC TESTBED NEEDS

① A robust behavioral effect

Animacy-driven syntactic choice is one of the most replicated effects in human studies.

② A single, quantitative outcome

Passive rate per condition → a continuous, comparable dependent variable across humans and LMs.

③ Minimal-pair manipulation

A 2×2 design lets us flip one variable (animacy) and hold everything else constant → exactly what activation patching needs.

THE EFFECT

Animacy affects relative clause structure choice.

Robust across studies.

Animate head → ***passive***

"the man that was punched by the woman"

Inanimate head → ***active***

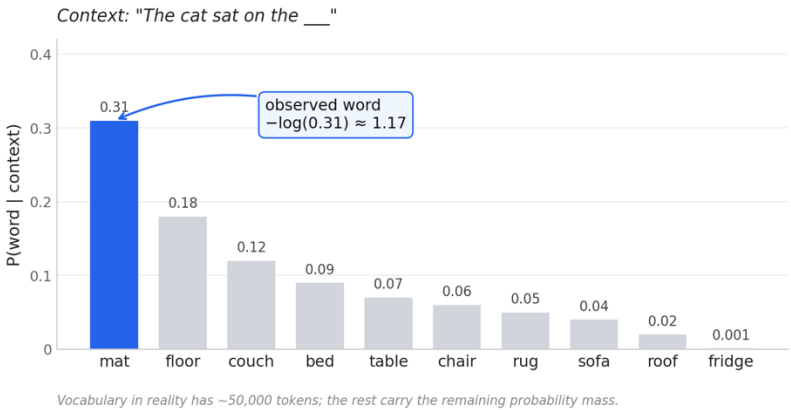
"the sandbag that the woman punched"

Gennari et al. (2012) · Bock & Warren (1985) · Mak et al. (2002) · Humphrey et al. (2016) — among many others

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"

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

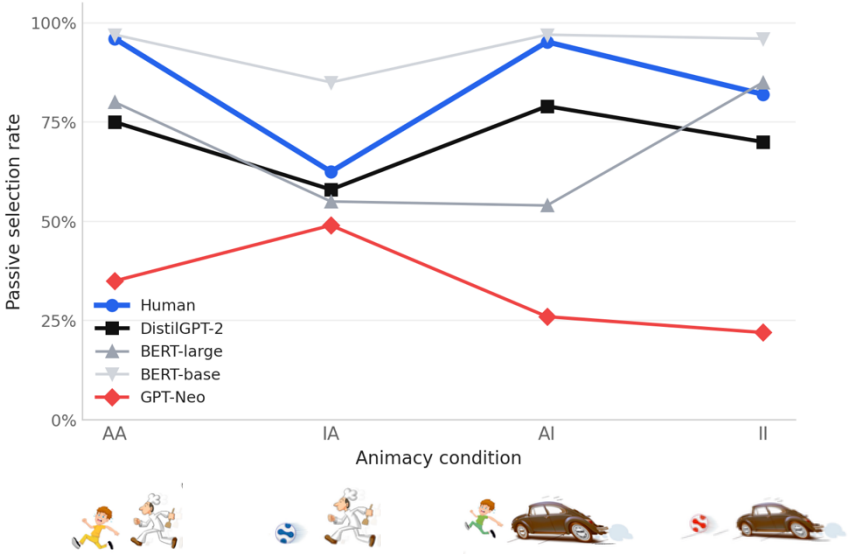
Repeat × 384 pairs (96 per animacy condition) (4 animacy configurations)

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)

Study 1 -- LMs Behavior Evaluation

Results



LLM-Human alignment found but varied by model

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

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)

Study 2 – Corpus Analysis

Corpus analysis: results

The corpus is heavily skewed

- the **opposite** of human behavior
- ***Not*** aligned with model surprisal-behavior

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 that LM-surprisal showed.

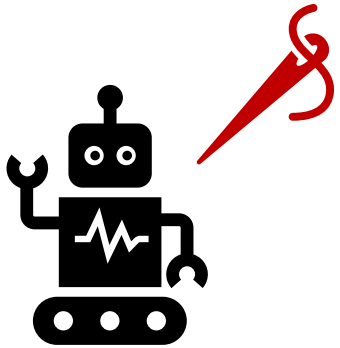
- Corpus → humans: $R^2 = 0.12$, $p = .66$ | Corpus → DistilGPT-2: $R^2 = 0.26$, $p = .49$
- DistilGPT-2 alone → humans: $R^2 = 0.96$, $p = .02$ | Adding corpus barely changes it ($R^2 = 0.99$, n.s.)

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

The question

So GPT2 behaviorally aligned with humans, but that human-alignment did not come from its training corpus.

Then, how and where did it show the animacy-sensitivity?



Study 3: 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."

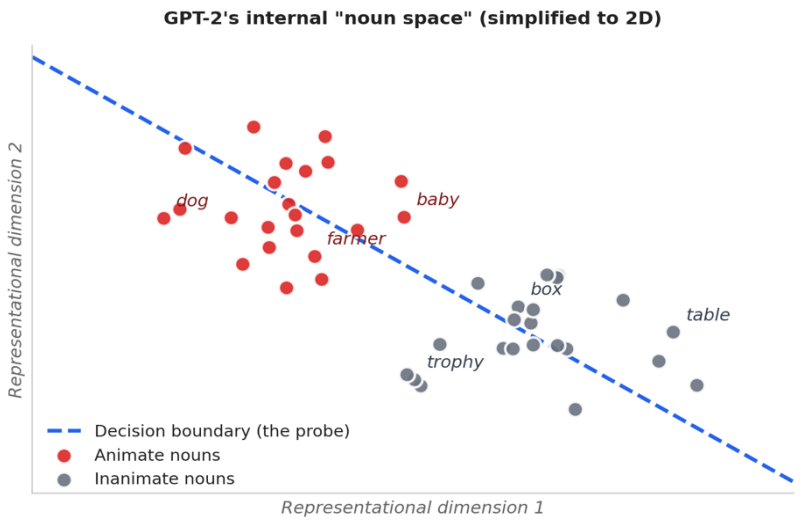
A **probe** is a simple decision rule (logistic regression) that asks:

> *"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**.

Machine Learning Paradigm:

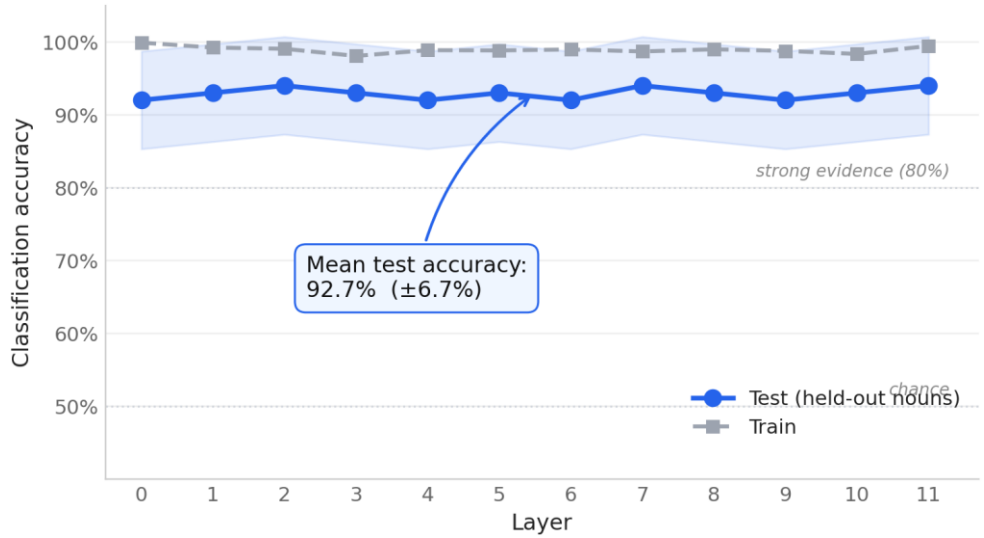
Test: train the probe on **some** nouns, test on **new** ones.



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

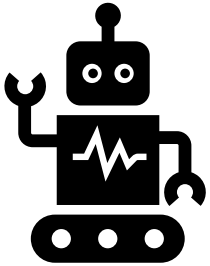
- Animacy **generalizes:** not memorized lexical knowledge
- The **representational prerequisites** for animacy-based processing are in place

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 the **next Studies**. *(Belinkov 2022)*

Study 4: Does animacy *cause* structure choice variation?

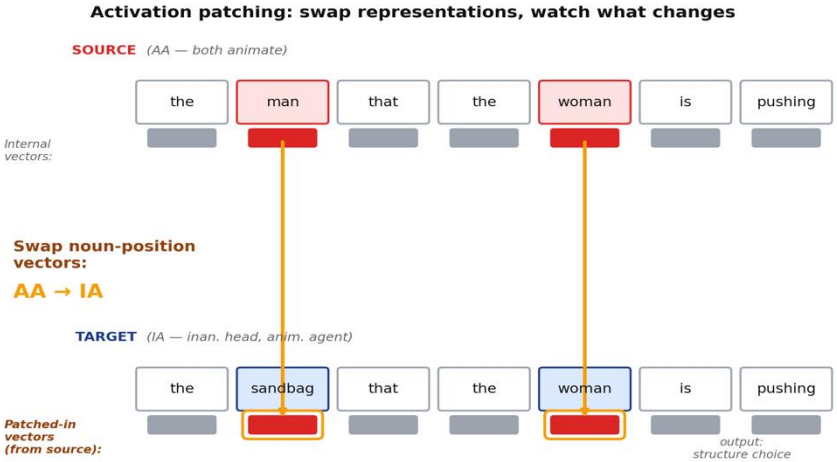
the	[inanimate]	that	the	woman	is	pushing

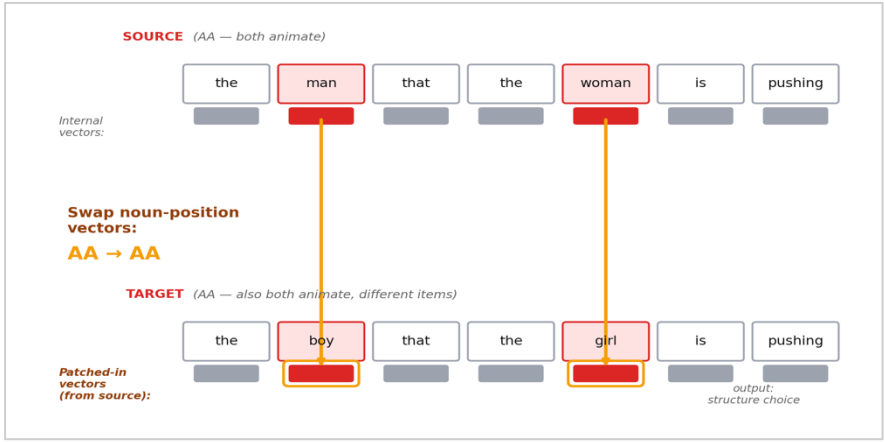


?

Study 4: 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.

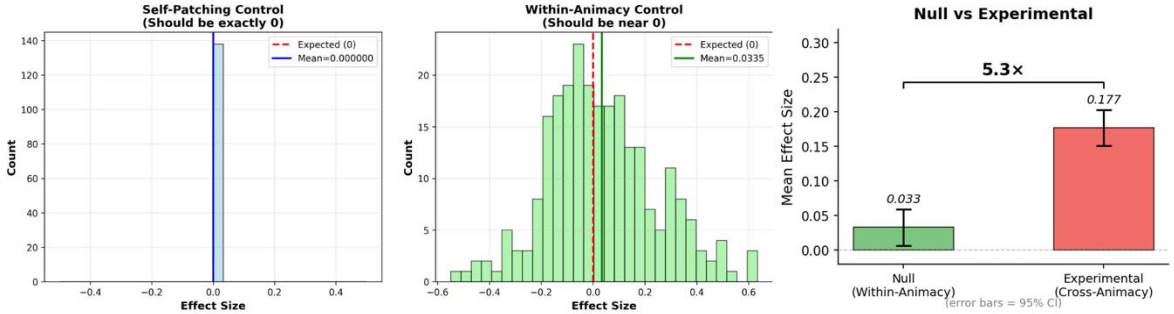
Null control:

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

Study 4: 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

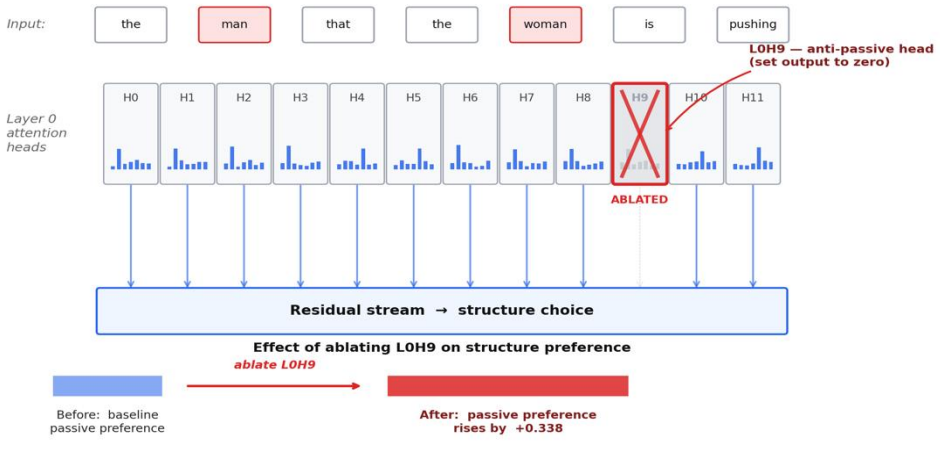
What we think this means:

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 5: 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**



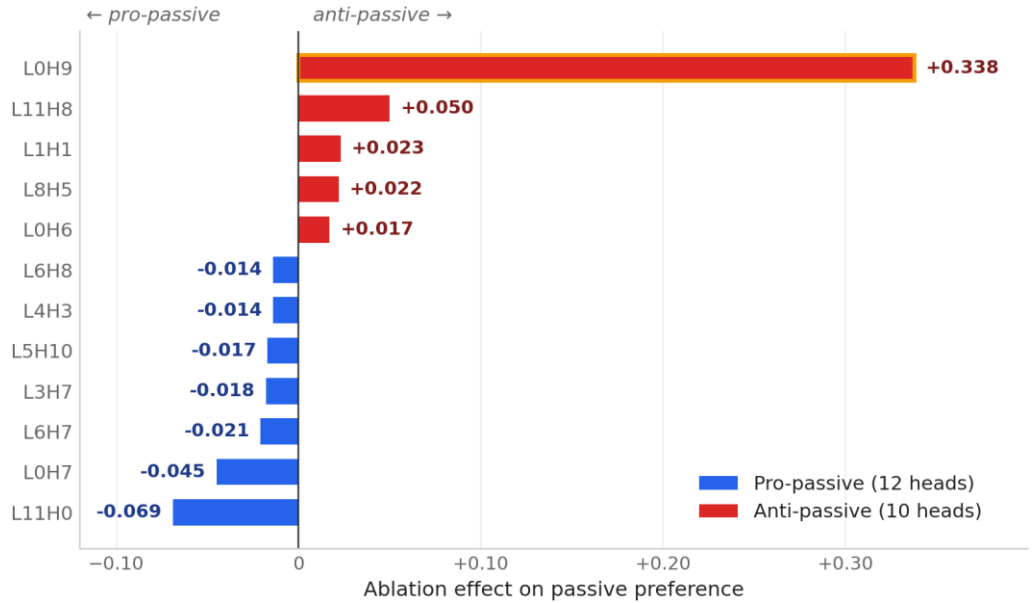
Removing the head reveals its causal contribution: it was suppressing passives.

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 5: 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 5: 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
LM-Human Behavior Alignment
DistilGPT2
Most human-like behavior

STUDY 2
Not distributional bias
Corpus ***failed*** to predict what model does

STUDY 3
Probing
Animacy is **encoded** as a generalizable feature.
92.7%
held-out accuracy

STUDY 4
Patching
Animacy **causally** changed the downstream task
5.3×
experimental vs. noise floor

STUDY 5
Circuit
Attention heads split into **opposing populations**.
12 pro · 10 anti
22 critical attention heads

Contribution. Mechanistic grounding for animacy sensitivity in LM.

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