Language models' processing of animacy parallels human processing

When Language Models Fall in Love: Animacy Processing in Transformer LMs

Michael Hanna[1], Yonatan Belinkov[2], and Sandro Pezzelle[1]
m.w.hanna@uva.nl
[1]ILLC, University of Amsterdam
[2]The Technion—Israel Institute of Technology

Animacy in Language (Models)

Animate entities can think, act, feel of their own will.

Inanimate entities cannot.

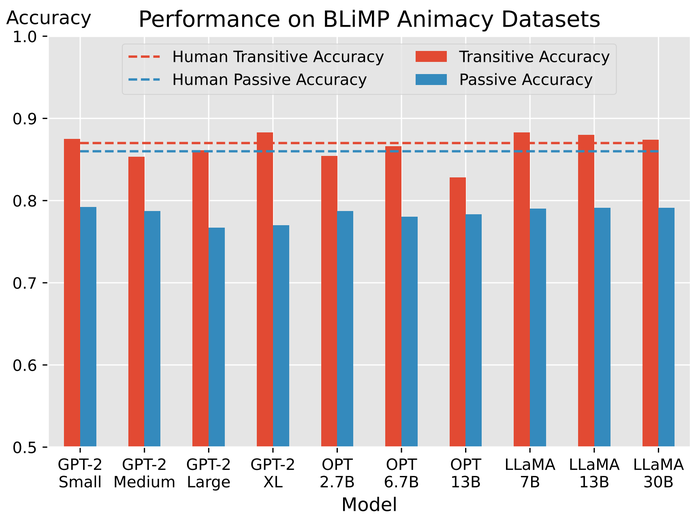
In English, animacy appears as indirect constraints; only animate entities can *be happy*, or *walk*. So can LMs, exposed only indirectly to animacy, capture this phenomenon?

Typical Animacy

We test LMs' animacy responses via BLiMP[1]:

	Acc?	Sentence
T	✓	**Naomi** had cleaned a fork.
T	✗	**That book** had cleaned a fork.
P	✓	Lisa was kissed by the **boys**.
P	✗	Lisa was kissed by the **blouses**.

Models prefer the acceptable sentence!

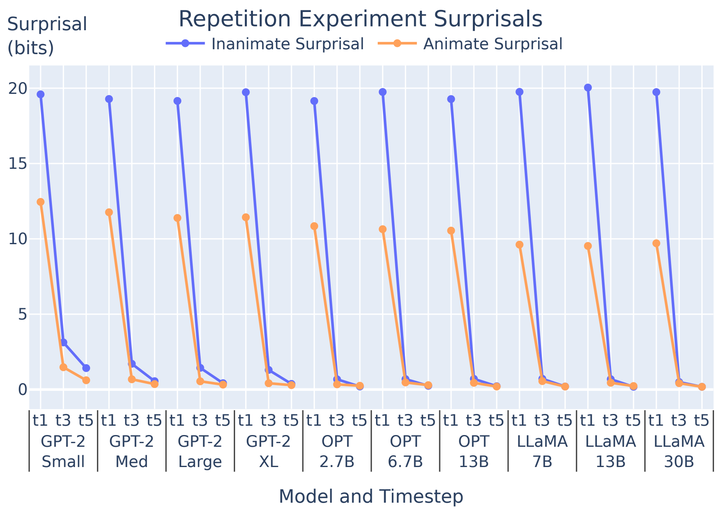


Atypical Animacy

Experiment 1: We replicate Nieuwland and van Berkum's (2006) study[2], which showed humans are initially surprised by (T1), but quickly adapt to (T3, T5) atypical animacy.

> A nurse was talking to the **sailor/oar [1]** who'd been in a violent boating accident. The sailor/oar cried for a long time over the storm that had raged over the lake for hours. The nurse consoled the **sailor/oar [3]**, saying that he'd soon be well again. The sailor/oar complained of a bad headache that wouldn't go away. The nurse gave the **sailor/oar [5]** a large dose of aspirin.

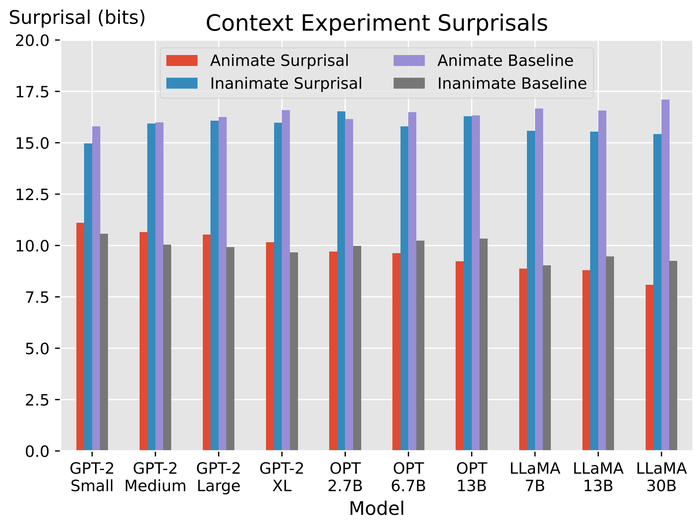
Models also grow less surprised over time!



Experiment 2: Could surprisal decrease be due to the repetition of the target word? We replicate another study without this flaw:

> A girl sat next to a diamond who was always doing strange things. The diamond told her that he liked to eat erasers. The girl ignored the diamond and his stories. Then the diamond said he also liked to sing songs. The diamond was quite **foolish/valuable** but secretly also very funny.

In animate-implying contexts, humans **and models** expect an animate adjective!



Conclusions

- Models respect animacy constraints, much like humans, in typical animacy scenarios.
- They also adapt to atypical animacy.
- Adaptation occurs even in cases without repetition, and in very short contexts.

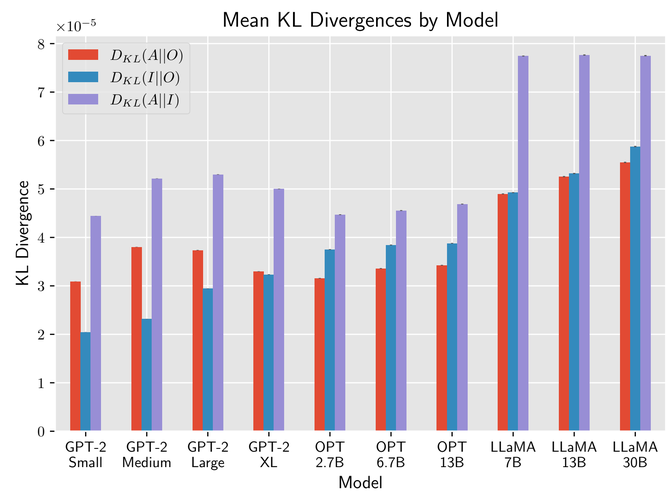
Low-Context Adaptation

In previous experiments, LMs had access to longer contexts, which they could have relied on to adapt. Can LMs adapt to atypical animacy even with little context?

We create a dataset for this, consisting of triplets of sentences (O, I, A) like:

- O: The [chair] spoke and began to"
- I: The [chair] began to"
- A: "The [woman] began to"

We compare distributions over atypically animate continuations ($p(w|O)$), typically inanimate continuations ($p(w|I)$), and typically animate continuations ($p(w|A)$).



$D_{KL}(A||O)$ is lower than $D_{KL}(A||I)$; the atypically animate context yields more animate continuations, suggesting models can adapt even with short contexts.

But adaptation is inconsistent; only some contexts yield animate continuations:

- The ion misunderstood and began to: *get*, *cry*, *run*, *walk*, *feel*
- The firewood replied and was very: *helpful*, *happy*, *friendly*, *good*, *pleased*
- The road gulped and became very: *narrow*, *stee*, *dark*, *wide*, *rough*
- The telephone waited and began to: *ring*, *be*, *d*, *vu*, *b*

Model and Dataset Details

We test autoregressive English LMs from the GPT-2, OPT, and LLaMA families.

We translate Nieuwland and van Berkum's (2006) data[2] into English. Our paper replicates Boudewyn et al.'s (2019) animacy N400 study[3], originally in English.

References
1. 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. TACL
2. Mante S. Nieuwland and Jos J. A. van Berkum. 2006. When peanuts fall in love: N400 evidence for the power of discourse. J. Cognitive Neuroscience
3. Megan Boudewyn, Adam Blalock, Debra Long, and Tamara Swaab. 2019. Adaptation to animacy violations during listening comprehension. Cognitive, Affective, & Behavioral Neuroscience

EMNLP 2023