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

Michael Hanna¹, Yonatan Belinkov², and Sandro Pezzelle¹

¹ILLC, University of Amsterdam
²The Technion—Israel Institute of Technology

Animacy in Language (Models)

Animacy in Language (Models)

An animate can think. Inanimate entities cannot. Act, feel of their own will.

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

Typical Animacy

We test LMs’ animacy responses via BLiMP¹:

<table>
<thead>
<tr>
<th>Acc</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>T ×</td>
<td>Naomi had cleaned a fork.</td>
</tr>
<tr>
<td>P ×</td>
<td>Lisa was kissed by the boys.</td>
</tr>
</tbody>
</table>

Models prefer the acceptable sentence!

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

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

Models also grow less surprised over time!

Atypical Animacy

Atypical Animacy

Experiment 1: We replicate Nieuwland and van Berkum’s (2006) study², which showed humans are initially surprised by (T1), but quickly adapt to (T3, T5) atypical animacy.

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, steep, 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) data² into English. Our paper replicates Boudewyn et al.’s (2019) animacy N400 study³, originally in English.

References

1. Alex Warstadt, Alicja Parish, Hanuk Liu, Anhui Mohanogey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. TACL