Beyond Binary Animacy: A Multi-Method Investigation of LMs' Sensitivity in English Object Relative Clauses

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

Animacy is a well-documented factor affecting language production, but its influence on Language Models (LMs) in complex structures like Object Relative Clauses (ORCs) remains underexplored. This study examines LMs' sensitivity to animacy in English ORC structure choice (passive vs. active) using surprisal-based and prompting-based analyses, alongside human baselines. In surprisal-based analysis, DistilGPT-2 best mirrored human preferences, while GPT-Neo and BERT-base showed rigid biases, diverging from human patterns. Prompting-based analysis expanded testing to GPT-4o-mini, Gemini models, and DeepSeek-R1, revealing GPT-4o-mini's stronger human alignment but limited animacy sensitivity in Gemini models and DeepSeek-R1. Some LMs exhibited inconsistencies between analyses, reinforcing that prompting alone is unreliable for assessing linguistic competence. Corpus analysis confirmed that training data alone cannot fully explain animacy sensitivity, suggesting emergent animacy-aware representations. These findings underscore the interaction between training data, model architecture, and linguistic generalization, highlighting the need for integrating structured linguistic knowledge into LMs to enhance their alignment with human sentence processing mechanisms.

1 Introduction

Animacy belongs to a set of semantic factors known to affect language production due to its centrality in human communication (Cooper and Ross, 1975). Previous studies have found that the animacy status of nouns affects how structures are formed. Specifically, one commonly investigated structure is object relative clauses (ORC). Many studies found that ORCs with animate head nouns are more likely to be produced in the passive structure instead of the active structure (e.g., Gennari et al., 2012; Humphreys et al., 2016; Wu et al., 2022). For example, English speakers overwhelmingly prefer passive structures like (1-a) in Table 1 over their active counterparts (1-b), whereas in describing an inanimate target, (1-c) and (1-d) are equally probable.

The concept of animacy—distinguishing between living and non-living entities—is rooted in human beings' perceptual, cognitive, and linguistic development (Gelman, 1981; Leslie, 1994; Rakison and Poulin-Dubois, 2001). However, transformer-based pre-trained language models (LMs) learn solely from text, raising the question of whether they exhibit human-like sensitivity to animacy when processing complex syntactic structures like ORCs, or if their behavior differs due to their text-based learning paradigm.

Recent studies have begun exploring this question, finding that while some LMs demonstrate sensitivity to animacy constraints, this varies across models (Hanna et al., 2023; Kauf et al., 2023; Yun et al., 2023), leaving open the question of whether LMs encode animacy as an abstract linguistic feature or simply reflect statistical patterns in text. Particularly relevant to the current study, Yun et al. (2023) reported ChatGPT-3.5's higher probability of generating active ORCs when the head noun was inanimate and the agent noun was animate than when both were animate. However, their study did not fully control for all possible animacy configurations of head and agent nouns. And the rapid advancement of LMs highlights the need for continued research to refine our understanding of their linguistic processing.

To address the gap, we use psycholinguistically guided minimal pairs to systematically test how animacy influences active vs. passive ORC structure choice across a full list of animacy conditions: AA (animate head noun + animate agent), IA (inanimate head noun + animate agent), AI (animate head noun + inanimate agent), and II (inanimate head noun + inanimate agent). This approach allows

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 184–196
May 3, 2025 ©2025 Association for Computational Linguistics

No.	Cond.	Head Noun	Agent Noun	Structure	Example
(1-a)	AA	animate	animate	passive	*the man who's being punched by the woman*
(1-b)	AA	animate	animate	active	*the man that the woman is punching*
(1-c)	IA	inanimate	animate	passive	*the sandbag that's being punched by the woman*
(1-d)	IA	inanimate	animate	active	*the sandbag that the woman is punching*

Table 1: Sample ORCs varied by head noun animacy.

us to determine whether LMs replicate humanlike animacy effects or diverge from human processing, providing insight into the role of animacy in LMs' ORC structure selection.

2 Related Works

2.1 Animacy in object relative clauses

One of the widely studied structures affected by animacy in psycholinguistics is the object relative clause (ORC): the animacy status of nouns involved in the ORC was found to affect whether the ORC is produced in passive or active structures in many languages (e.g., Gennari et al., 2012; Hsiao and MacDonald, 2016; Rodrigo et al., 2018; Wu et al., 2022). Specifically, the passive ORC is strongly preferred when both the head noun and the agent noun are animate (Condition AA), but this preference diminishes when the head noun is inanimate and the agent noun remains animate (Condition IA).

One explanation for the preference for passive ORCs is the animacy-based accessibility mechanism (Gordon et al., 2001), which suggests that animate nouns are conceptually salient and more likely to take the subject role in ORCs, leading to a passive preference (J. K. Bock and Warren, 1985). Alternatively, the similarity-based competition mechanism (K. Bock et al., 1992; McDonald et al., 1993) argues that two animate nouns (e.g., *man* and *woman* in Table 1) create higher cognitive load than inanimate-animate pair (e.g., *sandbag* and *woman*) due to conceptual competition in working memory. To ease this load, speakers prefer passives, which postpone the agent noun (Gennari et al., 2012). While both mechanisms predict animacy effects on ORC structure choice, they differ in their explanations for the passive preference in animate-head ORCs. However, past studies have only tested two (AA, IA: Gennari et al., 2012; Hsiao and MacDonald, 2016; Humphreys et al., 2016; Wu et al., 2022) or three (AA, IA, AI: Rodrigo et al., 2018) conditions, leaving gaps in understanding the full scope of animacy effect.

Which structure would speakers prefer when producing ORCs with inanimate head nouns and inanimate agents (condition II)? Would they equally choose passive or active because there are no animate head nouns urgently in need of a subject role? Or would they still strongly favor passives because due to the cognitive load imposed by competition between two similar inanimate nouns? Due to the lack of studies incorporating all four animacy conditions, the relationship between animacy status and ORC structure preference is not clear. This gap extends beyond psycholinguistics to LMs, as investigating animacy-driven structure choices in LMs can provide insights into whether they reflect human-like processing or rely on different underlying mechanisms. Conversely, exploring these patterns in LMs may also offer predictions about what to expect in the underexplored conditions, guiding future psycholinguistic research. To bridge these gaps, the current study first exposes human participants to all four animacy conditions to establish a baseline. This not only fills a critical gap in psycholinguistics but also lays the groundwork for evaluating LMs' animacy-sensitivity in making syntactic decisions in the following steps.

2.2 Animacy in LMs

The role of animacy in language modeling has been a topic of interest in computational linguistics. Early work by Elman (1990) showed that a simple recurrent network trained on synthetic language data formed distinct clusters for animate and inanimate entities, suggesting that basic LMs developed animacy-sensitive representations.

More recent studies have examined how animacy is integrated into broader linguistic behavior in LMs. Kauf et al. (2023) found that LMs exhibit sensitivity to animacy as it relates to selectional constraints, indicating that animacy is integrated into their broader event knowledge. Hanna et al. (2023) found that LMs can infer animacy from contextual cues and adjust their processing accordingly, though not always to the same extent as humans.

Several studies have also explored how animacy affects syntactic structure choice. Futrell and Levy (2018) found that recurrent neural network language models (RNN LMs) learn animacy as an abstract feature that influences word order, though its effect was weaker and less consistent than other factors like constituent length. In a more targeted investigation, Yun et al. (2023) prompted GPT-3.5 with sentence fragments and observed significantly more active ORCs when the head noun was inanimate than when it was animate, suggesting that animacy influences structural choices in LMs. Papadimitriou (2024) found that animacy is a strong predictor of subjecthood in mBERT's embedding space: animate nouns were more likely to be classified as agents, even when controlling for syntactic role. This finding supports the idea that LMs encode subjectivity in gradient and functionally-driven ways, with animacy as a core dimension.

Building on this line of research, our study goes beyond the typical binary manipulation of head noun animacy in ORC configurations. We introduce a four-way animacy design that systematically varies both head noun and agent animacy across conditions (AA, IA, AI, II). Our investigation consists of three complementary experiments: (1) surprisal-based analysis, (2) training corpus examination, and (3) direct prompting-based analysis. Our goal is to determine whether LMs show animacy sensitivity in ORC processing, and if so, whether their animacy effects reflect an emergent linguistic pattern or are merely artifacts of training data biases. We hypothesize that (1) LMs will exhibit systematic surprisal-based animacy effects, but with model variations, (2) corpus distributions alone will not fully account for LMs' structure choices, and (3) prompting analysis will reveal animacy-driven patterns in ORC selection for some LMs, if not all.

3 Psycholinguistic Data

Design Fruitful previous studies, including Gennari and MacDonald (2009) with 82 native English speakers, Montag and MacDonald (2015) with 30, and Humphreys et al. (2016) with 16, have consistently found that animacy affects the choice between passive and active ORCs, particularly in Conditions AA and IA, using similar picture-based elicitation tasks. In the current study, we used 20 illustrated scenes created with Procreate and supplemented with licensed clip art (See Appendix A for an example). Each scene depicted four distinct events, all involving the same action (e.g., hitting, pulling, pushing, chasing, lifting), varying by the animacy of the agent and patient: AA: Animate Agent – Animate Patient (e.g., a woman lifting a boy); IA: Animate Agent – Inanimate Patient (e.g., a woman lifting a box); AI: Inanimate Agent – Animate Patient (e.g., balloons lifting a boy); II: Inanimate Agent – Inanimate Patient (e.g., balloons lifting a box). We also included 50 filler scenes depicting unrelated events (e.g., riding bikes, playing cards), designed to elicit a range of structures including simple and subject relative clauses. Participants viewed the images and responded to questions. Their choice of active or passive relative clause structure was analyzed. As a proof-of-concept psycholinguistic study, five adult native English speakers each produced twenty responses. Their structure choices were coded accordingly. This preliminary study establishes a human baseline for evaluating LM behavior, as no prior work has systematically investigated all four animacy configurations of ORCs.

Result Our preliminary results align with previous research in two key ways: (1) a general preference for passives overall (Gennari et al., 2012; Montag et al., 2017), and (2) higher passive usage in AA and AI compared to IA (e.g., Humphreys et al., 2016; Rodrigo et al., 2018). As shown in Figure 1, passive structures were strongly preferred in AA (96%) over IA (63%), with a significant difference ($B = 2.69$, $p = 0.02*$) confirmed by binomial mixed-effects logistic regression. AI also showed a high passive rate (95%), comparable to AA, consistent with findings by Rodrigo et al. (2018) in Spanish and Japanese. Our study further provides new insights into the II condition. While II did not differ significantly from other conditions ($p > 0.5$), its passive rate (82%) was noticeably higher than IA (63%), suggesting that even without an animate noun, similarity-based competition between two inanimates may still promote passive use.

These results reinforce the complex role of animacy in ORC structure choice. The strong passive preference in AA and AI aligns with the expectation that animate head nouns favor the subject position, making passivization the preferred structure (Gennari et al., 2012; Rodrigo et al., 2018). The IA condition, which lacks both an animate

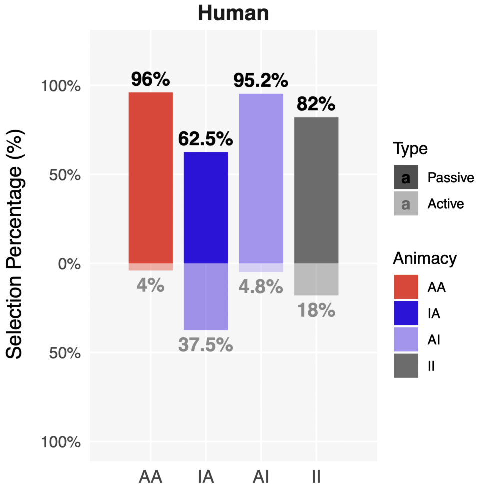

Figure 1: Human responses: ORC structure choice by animacy condition. AA represents ORCs with animate head noun and animate agent; IA: inanimate head noun, animate agent; AI: animate head noun, inanimate agent; II: inanimate head noun and agent.

head noun and animacy congruence, showed the lowest passive preference, suggesting that the absence of these factors results in weaker motivation for passivization. The II condition, despite the absence of an animate noun, exhibited a higher passive rate than IA, suggesting that *similarity-based competition* may still influence structure choice even among inanimate referents. While we acknowledge the limitations of our sample size[1], the clear alignment of our findings with prior research and the observed significant effects suggest that animacy effects in ORC processing extend beyond a simple binary contrast and involve a more complex interaction between competition and accessibility mechanisms.

4 LMs and Experiments

4.1 Surprisal-based analysis

Dataset Following our psycholinguistic paradigm and previous research (e.g., Gennari and MacDonald, 2009; Humphreys et al., 2016), we designed experimental English prompt minimal pairs (Cong, 2022), each consisting of a written *context story* and a pair of *target sentences*. Expanding on our psycholinguistics experiment, we developed a set of 384 prompt pairs, with 96 pairs per animacy condition. Each pair includes both a passive ORC target sentence and its active counterpart. For instance, in Table 2, the target sentences (passive: "*The baby that is held by the father is crying*"; active: "*The baby that the father holds is crying*") both serve as grammatically valid answers to the final question in the *context story*. We hypothesize that LMs will select different target sentences depending on the animacy condition, despite potential variations due to model differences.

Experimental Design and LMs We evaluated the performance of various LMs on ORCs' structure choice: DistilGPT-2 (Sanh et al., 2019), GPT-Neo (Black et al., 2021; Gao et al., 2020), BERT-large-uncased (Devlin et al., 2018), and the BERT-base-uncased (Devlin et al., 2018). See Table 3 for summary. These differences, including size, architecture (masked vs. causal), and training data diversity, are likely to influence how each LM processes syntactic structure and animacy-sensitive patterns, and thus are important for interpreting model–human comparisons.

In the current analysis, the preference for a particular answer is measured by the *surprisal score* of each target sentence given by LMs (Cong et al., 2023; Hale, 2001; Michaelov and Bergen, 2022). For GPT-type LMs, surprisal was calculated as the negative log probability of the word given left context (Levy, 2013). We computed the *surprisal score* at the sentence level. When the LMs tokenizer splits the target in more than one token, we take the average of the *surprisal score* of its subtokens (See Appendix B for out-of-vocabulary (OOV) ratios by animacy condition for each LM). For BERT-type models, which are bidirectional and trained via a masked language modeling objective, *surprisal* was calculated as the sum of the negative log probabilities of each word, conditioned on both its preceding and following context—normalized by the total number of tokens in the sentence. This sentence-level *surprisal* aligns with BERT's bidirectional training: unlike autoregressive models that rely solely on left context, denoising autoencoding models like BERT and RoBERTa are explicitly trained to make word predictions based on both left and right contexts. Our *surprisal* calculation for BERT therefore mirrors its underlying architecture and learning objective, supporting a more principled comparison with GPT-style models. To keep consistency in operation, we used minicons (Misra, 2022) for both BERT and GPT-type LMs, specifically the scorer module for the masked language models such as BERT (i.e., the scorer.MaskedLMScorer class), and standard au-

[1] Our ongoing psycholinguistic research with 35 participants replicates these findings. Details will be published in a forthcoming paper.

Context Story	Structure	Target Sentence
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.

Table 2: Example prompt pair for surprisal analysis.

Model	Arch.	Size	Training Data
BERT-base	masked	110M	BooksCorpus, Wikipedia
BERT-large	masked	340M	BooksCorpus, Wikipedia
DistilGPT-2	causal	82M	OpenWebText
GPT-Neo	causal	1.3B	The Pile

Table 3: Summary of LMs used in the surprisal-based analysis.

toregressive language models such as DistilGPT-2 (i.e., the scorer.IncrementalLMScorer class). When the passive structure in a prompt pair receives a lower mean surprisal score than its active counterpart, we coded the outcome variable *choose-psv* as 1, otherwise as 0.

For statistical analysis, binomial logistic mixed-effects model was fitted for each LM with *choose-psv* as the dependent variable, Animacy as the main predictor (categorical). The random-effects structure included only Items. The LME4 package in R (Bates, 2014) was used for statistics modeling. Post-hoc comparisons were conducted with the *emmeans* package (Lenth, 2019), applying *Tukey* adjustments for pairwise comparisons. Our implementation is available on our Github page.

Results Figure 2 shows structure selection rates by animacy for each LM, with darker bars indicating passive selection and lighter bars representing active selection. Several key patterns emerged. First, different LMs exhibited distinct structural biases: BERT-large ($B = 0.79$, $p < 0.001$), BERT-base ($B = 2.71$, $p < 0.001$), and DistilGPT-2 ($B = 0.92$, $p < 0.001$) showed overall strong passive preference, whereas GPT-Neo significantly favored actives across conditions, shown by its significant negative intercept ($B = -0.78$, $p < 0.01$).

Next, for each LM, the results (see Figure 2) showed significant animacy effects for BERT-large ($p < 0.001$), DistilGPT-2 ($p < 0.01$), and GPT-Neo ($p < 0.001$), while BERT-base did not reach significance ($p = 0.06$). BERT-large showed significantly lower passive selection rates in IA and AI conditions compared to AA and II, suggesting that BERT-large is less likely to choose passives when the head noun and agent differ in animacy features.

DistilGPT-2 chose significantly fewer passives in IA, indicating an increased selection for actives when the head noun is animate and the agent is inanimate. GPT-Neo, unlike other models, showed a stronger passive preference in IA compared to AI and II. BERT-base, due to its exceptionally high passive selection rates across all conditions, did not exhibit significant effect of animacy.

To evaluate the alignment between LMs and human responses, we conducted Pearson correlation analyses and RMSE (Root Mean Square Error) calculations between each LM's passive selection rates and human data. The results reported DistilGPT-2's highest Pearson correlation ($r = 0.98$) and lowest RMSE (0.14), suggesting closer alignment with human patterns. GPT-Neo showed the lowest Pearson correlation ($r = -0.66$) and highest RMSE (0.55), indicating its substantial divergence from human patterns. Figure 3 visualizes passive selection rates across animacy conditions for each LM, with the red line representing human response patterns from psycholinguistic data. The figure further highlights DistilGPT-2's closer alignment to human behavior (navy blue line), while GPT-Neo exhibits the greatest divergence (light blue line).

4.2 OpenWeb corpus analysis

While surprisal-based experiment found varying degrees of animacy sensitivity in LMs, an open question remains: Is this sensitivity an emergent linguistic property or merely a reflection of the distribution in the training data? Specifically, do LMs assign surprisal scores based on inherent animacy effects, or are these scores simply mirroring the animacy-driven distribution of ORCs in the training data?

Method To address this question, we exam-

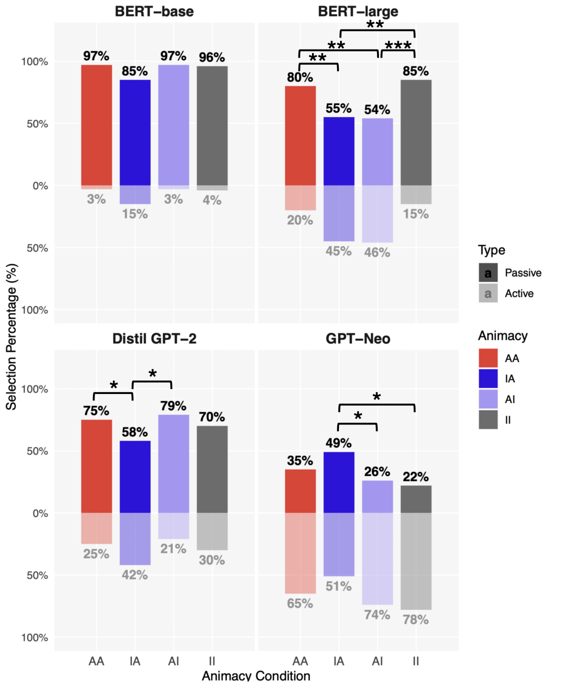

Figure 2: Structure selection rate by animacy and LM according to surprisal-based analysis.

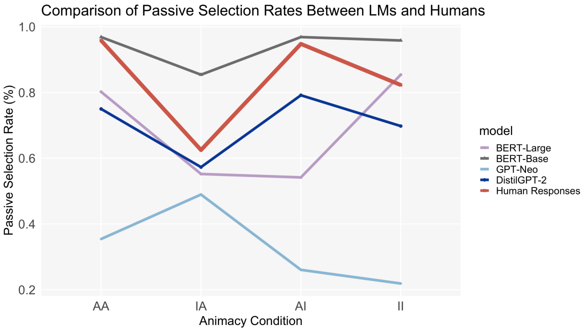

Figure 3: Passive selection pattern comparison between LMs and human data.

ined the structural distribution of active and passive ORCs across animacy conditions in the OpenWeb corpus (Gokaslan and Cohen, 2019), an open-source reproduction of OpenAI's WebText dataset, which was used to train GPT-2. We randomly selected over 8,000 sentences from the corpus and used a custom syntactic parsing pipeline using SpaCy to automatically extract sentences containing ORCs. The extraction procedure identified ORCs based on the presence of a head noun, an embedded verb phrase, and an agent noun—accounting for both overt and omitted relative pronouns. Manual examination of the automatically identified ORCs was conducted and only those with correct annotations were retained. Each validated ORC was then categorized as active or passive, and the animacy status of both the head noun and agent noun was annotated. This allowed us to quantify the frequency of active and passive ORC structures across different animacy conditions.

Then, we conducted Pearson correlation tests to assess whether the ORC distribution in OpenWeb alone could account for the animacy-driven ORC patterns found in human responses, to identify which LMs' surprisal scores best aligned with human responses, and to determine whether incorporating corpus data could enhance the explanatory power of LMs in modeling human behavior.

Results Only 1.34% of our examined sentences were found to contain a complete ORC. As shown in Table 4, these ORCs are skewed toward actives (71.03%) over passives (28.97%) and are imbalanced by animacy, with IA conditions dominating (66.36%). Particularly, active IA ORCs alone account for 53.27% of all ORCs, suggesting a pronounced structural bias in the training corpus. In contrast, AA and AI conditions are rare, comprising only 4.67% and 1.87% of the total ORCs, respectively.

Struct.	**AA**	**IA**	**AI**	**II**	**Sum**
	(%)	**(%)**	**(%)**	**(%)**	**(%)**
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

Table 4: ORCs found in OpenWeb sample grouped by animacy and structure.

Pearson correlation tests (Table 5) indicate that corpus data alone has low predictive power for both human responses ($R^2 = 0.12$, $p = 0.66$) and DistilGPT-2's surprisal values ($R^2 = 0.26$, $p = 0.49$). While the corpus shows moderate predictive power for GPT-Neo ($R^2 = 0.78$, $p = 0.12$), the negative correlation ($r = -0.89$) suggests that GPT-Neo follows an opposite trend from corpus-based distributions.

Among the four tested LMs (DistilGPT2, GPT-Neo, BERT-large, and BERT-base), DistilGPT2 accounted for the highest variance in human responses ($R^2 = 0.96$, $p = 0.02$), explaining 95.6% of the variance with statistical significance. Adding OpenWeb further increased the explained variance to 98.7%, but the lack of significance suggests that the combined model did not outperform DistilGPT-2 alone. The other three LLMs showed weaker alignment with human behavior. GPT-Neo exhibited strong divergence, as indicated by its negative estimates and low R^2 values, suggesting an opposite structure preference. BERT-large explained only 9% of the variance (not significant), indicating it is a weak predictor of human responses. BERT-base captured 88.4% of the variance but was not significant, and incorporating OpenWeb did not improve its predictive power.

4.3 Prompting-based analysis

LMs To further strengthen our investigation, we conducted a supplementary analysis using prompt engineering. In addition to the four previously examined LMs, we included four recent state-of-the-art models: GPT-4o-mini (Achiam et al., 2023), Gemini-1.5-flash (Team et al., 2023), Gemini-2.0-flash, and DeepSeek-R1 (Guo et al., 2025). This analysis used the same dataset as the surprisal analysis, which consists of 384 context stories paired with sentences containing passive and active ORCs.

Method Each LM was prompted to select the more appropriate syntactic structure based on the given context. The structured prompt explicitly instructed the model as follows: *"Read the following context carefully, which includes a short story and a question at the end. Two possible answers are provided. Your task is to choose the answer that sounds most natural to a native English speaker. Please respond with either "1" for the first option (Passive) or "2" for the second option (Active)".*

Same as surprisal-based analysis, LMs' choice was recorded as 1 for *passive* and 0 for *active* in the variable *choose-psv* for each trial. The passive selection rate was calculated as the proportion of trials in which the model selected *passive* within each animacy condition.

For model comparison, we computed Pearson correlation, MSE and RMSE. Pearson correlation evaluates the linear relationship, while RMSE quantifies the average deviation of model predictions from human responses, with lower values indicating better fit. Together, these measures provide a comprehensive evaluation of how closely or differently each LM perform compared to human.

Design Considerations While our human experiment used picture-based elicitation (See Appendix A for an example), we opted for a controlled, text-based prompting design in this analysis. This choice was made to avoid confounds introduced by image recognition and visual reasoning, which current LMs may not reliably handle in a standardized way. Instead, we used context stories that mirrored the structural and referential properties of the original visual stimuli, allowing us to isolate syntactic preference.

That said, a potential alternative design could involve describing the visual scene and posing a direct question (e.g., “Who is wearing red?”), then analyzing the model's free-text response. Such a design could more closely simulate the referential pressure that led to ORC production in humans and may be explored in future work.

Results As shown in Figure 4, the structure choices made by different LMs in the prompt engineering experiment show great variation. Several noticeable patterns emerged. First, BERT models (BERT-large, BERT-base) exhibit limited variation in response, overwhelmingly favoring passive ORCs (near 100%) across all conditions. Gemini models (Gemini-1.5-flash, Gemini-2.0-flash), on the other hand, strongly prefer actives, with Gemini-2.0-flash selecting active ORCs in nearly 100% of all conditions. Both model families seem to lack human-like variation in structure choice. GPT models (DistilGPT2, GPT-Neo, GPT-4o-mini) and DeepSeek-R1 show more variation. ANOVA analysis confirms significant differences among LMs compared to human responses ($df = 8$, $p < 0.001$). Post-hoc tests indicate that while all LMs deviate from human responses to some extent, GPT-4o-mini exhibits the smallest difference ($diff. = 0.19$, $p = 0.01$).

Model evaluation (see Figure 5) showed GPT-4o-mini as the top performer, with the highest correlation to human data ($estimate = 0.98$), highest explained variance ($R^2 = 96.4\%$), and lowest RMSE ($estimate = 0.21$). BERT models (especially BERT-base) performed the worst, as they explained almost no variance in human data and had weak correlations. DistilGPT2 and GPT-Neo showed moderate alignment, indicating they capture some trends but weren't very strong predictors. Gemini models and DeepSeekR1 performed inconsistently, they had low variance explained and high RMSE, suggesting they aren't reliable in matching human responses.

5 Discussion

5.1 LMs show animacy sensitivity with model-specific variations

Our surprisal-based and prompting-based analyses revealed LMs' varying sensitivity to animacy in

	R^2	Adjusted R^2	F-statistic	p-value
How Corpus Explains Human Responses				
OpenWeb (corpus)	0.12	-0.33	0.26	0.66
How Corpus Explains GPT Models				
DistilGPT-2	0.26	-0.12	0.69	0.49
GPT-Neo	0.78	0.67	7.13	0.12
How LMs Explain Human Responses				
DistilGPT-2	0.96	0.93	42.93	0.02*
DistilGPT-2 + OpenWeb	0.99	0.96	37.11	0.12
GPT-Neo	0.44	0.16	1.58	0.34
GPT-Neo + OpenWeb	0.72	0.15	1.26	0.53
BERT(large)	0.09	-0.36	0.20	0.70
BERT(large) + OpenWeb	0.18	-1.43	0.12	0.90
BERT(base)	0.88	0.83	15.29	0.06
BERT(base) + OpenWeb	0.93	0.80	6.95	0.26

Table 5: Regression results: corpus vs. LMs and human ORC structure choice.

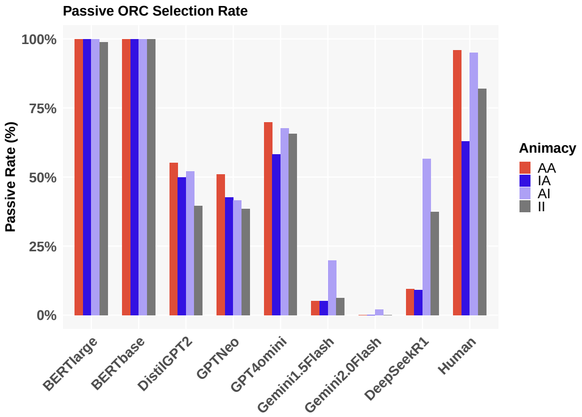

Figure 4: Passive selection rate by animacy and LM in prompting-based analysis

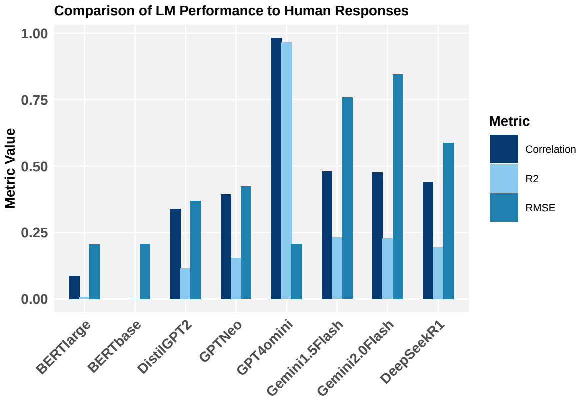

Figure 5: Evaluation of LMs' performance by human responses in prompting-based analysis

ORC structure choice, aligning with Hanna et al. (2023). The surprisal-based analysis reveals that DistilGPT-2's lower passive selection rate in IA compared to AA and AI aligns well with human data and psycholinguistic predictions (Gennari et al., 2012; Hsiao and MacDonald, 2016). According to the similarity-based competition mechanism (Gennari et al., 2012), passives should be more frequent in animacy-congruent conditions (AA and II). Among the tested LMs, only BERT-large followed this expected pattern, while BERT-base consistently over-selected passives, diverging from human data. Similar to Ettinger (2020), we report greater sensitivity of BERT-large to linguistic constraints than BERT-base. GPT-Neo showed a general preference for actives but unexpectedly showed its highest passive rate in IA, contradicting human data and psycholinguistic theories.

Some LMs performed inconsistently across prompting- and surprisal-based analyses. DistilGPT-2 and BERT-large performed poorly in prompting, explaining only 11.55% and 0.75% of human variance, respectively, likely due to fundamental task differences. As Hu and Levy (2023) pointed out, prompting is not a substitute for direct probability measurements in LMs, and results may vary within the same LM.

Among the four newly tested LMs in prompting-based analysis, GPT-4o-mini best mirrored human patterns, despite an overall lower rate of passive selection. In contrast, Gemini models (Gemini-1.5-Flash and Gemini-2.0-Flash) showed minimal variation across animacy conditions, suggesting that their internal representations likely do not align with established linguistic theories (Cong, 2024). Gemini-2.0-Flash, in particular, overwhelmingly favored active structures (∼100%), justifying its choices by claiming actives sound more direct and natural in English, whereas passives feel overly formal. DeepSeek-R1 exhibited structural variation across animacy conditions but in a theoretically ungrounded way. While psycholinguistic studies consistently report higher passive rates in AA than IA (Gennari et al., 2012; Hsiao and MacDonald, 2016), DeepSeek-R1 showed little distinction between these conditions, deviating from both human behavior and psycholinguistic predictions.

5.2 Training data alone fails to explain animacy sensitivity in LMs

Our analysis of ORC distribution in OpenWeb suggests that training data alone is a weak predictor of LMs' structure choices, as reflected in surprisal results. While training data influences LM behavior (Chai et al., 2024), it fails to fully account for observed animacy effects, challenging the idea that these effects stem solely from training biases. Instead, our findings suggest that some LMs, particularly DistilGPT-2 (surprisal-based) and GPT-4o-mini (prompting-based), develop emergent animacy sensitivity beyond exposure, aligning with human data and psycholinguistic predictions (Gennari et al., 2012; Hsiao and MacDonald, 2016), despite training corpus' limited explanatory power. DistilGPT-2 alone explains 95.6% of the variance in human responses, indicating that its animacy sensitivity cannot be attributed to corpus distributions alone.

That said, the predominance of active ORCs in IA conditions in the corpus may still contribute to LMs' preference for active structures in these cases. This pattern is consistent with Roland et al. (2007), they also found higher percentage of active ORCs in IA conditions compared to AA in both the Brown corpus (IA: 53%, AA: 25%) and the Switchboard corpus (IA: 69%, AA: 9%). Our corpus analysis revealed an even stronger dominance of active ORCs in IA conditions, reinforcing the influence of corpus-based biases.

Ultimately, while corpus distributions shape structure choices to some extent, they fail to explain the deeper, human-like patterns observed in surprisal-based and prompting-based analyses. The strong alignment between certain LMs and human responses suggests that animacy sensitivity in LMs arises from more than just statistical learning—it may reflect deeper linguistic generalization.

5.3 Optimize LMs with psycholinguistic knowledge

Despite carefully controlled input pairs and explicit instructions, many LMs failed to capture human-like animacy effects, with only a few demonstrating satisfactory sensitivity. Gemini-1.5-Flash, Gemini-2.0-Flash, DeepSeek-R1, and GPT-Neo showed little alignment with human patterns.

It is likely that these LMs struggle with the syntactic-semantic interface required for ORC structure choice processing, particularly when two structures convey the same meaning. Their training on large, diverse datasets may not emphasize fine-grained semantic features that guide human sentence processing. Future LM development and optimization could benefit from explicit integration of semantic and syntactic knowledge and targeted training on animacy effects and structural dependencies. Moving beyond surface-level pattern recognition towards deeper linguistic representation would improve LMs' alignment with human-like reasoning and formal (psycho-)linguistic theories.

6 Conclusion

To conclude, we found that LMs exhibited animacy sensitivity, though the extent varied across models, as reflected in their ORC structure choices. While some models aligned closely with human data, others diverged significantly, highlighting variation in how LMs process animacy in syntactic structures. DistilGPT-2 and GPT-4o-mini showed the strongest alignment, while Gemini models, DeepSeek-R1, and GPT-Neo failed to capture animacy effects meaningfully.

While training data influences LM behavior to some extent, it does not fully explain their animacy sensitivity, suggesting that some models develop emergent linguistic generalizations beyond mere statistical learning. To improve LMs' alignment with human cognition, future development should integrate psycholinguistic insights, refine semantic-syntactic training, and move beyond surface-level pattern learning. Strengthening linguistic representations will inspire the development of psychologically plausible models.

Limitations

While this study offers valuable insights into LMs' sensitivity to animacy in English ORC structure choice, several limitations remain.

Our current surprisal analysis computed average surprisal across tokens at the sentence level. While this approach simplifies comparison across sentence types, future work could adopt additive surprisal values, which better reflect joint probabilities over token sequences. Moreover, exploring surprisal at more localized levels—such as word- or phrase-level surprisal given left context—may better align with psycholinguistic processing and production (for this, we thank our anonymous reviewer for the suggestion). In addition, analyzing surprisal using a binary outcome variable (*choose-psv*) was conducted to mirror human production, but using raw surprisal differences as the dependent measure could potentially yield additional insights. This is an alternative analysis that can be done in future work to identify more fine-grained distinctions in model preferences.

Our psycholinguistic proof-of-concept study involved a limited number of human participants. While our findings are consistent with prior literature documenting animacy effects—particularly in AA, IA, and AI conditions—a larger sample size would strengthen empirical comparisons with LMs. Furthermore, in this project, we did not collect separate animacy norming data for our stimuli, which could improve future experimental control and interpretation in the future.

Our corpus analysis used a representative sample of the OpenWeb corpus to approximate natural distributional patterns, but it does not reconstruct LMs' full pretraining data. Broader corpus comparisons and controlled datasets would offer a more robust estimate of the linguistic patterns LMs are exposed to.

Lastly, although not the focus of the current work, future studies could incorporate layer-wise probing to explore whether animacy effects arise during lexical encoding, syntactic composition, or higher-level integration processes.

Acknowledgments

This project was supported by Purdue-internal funding sources, including the Ross-Lynn Research Scholar Fund, the College of Liberal Arts PROMISE Award, the School of Interdisciplinary Studies, the Department of Linguistics, and the Department of English. We also gratefully acknowledge external support from the 2023 Gorilla and Prolific Grant. We thank the Gilbreth Cluster at the Rosen Center for Advanced Computing (RCAC), Purdue University, for providing GPU computational resources; Gorilla Experiment Builder for the implementation of the psycholinguistic experiment. We also thank our research assistants—Marley Mack, Golsa Khodadadi, and Yongjia Deng—for their help with data collection in psycholinguistic experiment; the members of Purdue Experimental Linguistics Lab—Shaohua Fang, Vanessa Sheu, Jingying Hu, and Kyle Swanson—for their valuable feedback throughout the development of this project. We thank Yue Li's dissertation committee members—Alejandro Cuza, John Sundquist, and Mariko Wei—for their guidance and support. Finally, we gratefully thank three anonymous reviewers for their detailed feedback on the psycholinguistic experiment as well as the three LM analyses, which improved the clarity and quality of the paper. All codes and materials for this project are available on our OSF page and GitHub repository.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bates, K. (2014). Empathy or entertainment? the form and function of violent crime narratives in early-nineteenth century broadsides. *Law, Crime Hist.*, *4*, 1.

Black, S., Gao, L., Wang, P., Leahy, C., & Biderman, S. (2021). Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow. *If you use this software, please cite it using these metadata*, *58*, 2.

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

Bock, K., Loebell, H., & Morey, R. (1992). From conceptual roles to structural relations: Bridging the syntactic cleft. *Psychological review*, *99*(1), 150.

Chai, Y., Liu, Q., Wang, S., Sun, Y., Peng, Q., & Wu, H. (2024). On training data influence of gpt models. https://arxiv.org/abs/2404.07840

Cong, Y. (2022). Psycholinguistic diagnosis of language models' commonsense reasoning. *Proceedings of the first workshop on com-*

monsense representation and reasoning (CSRR 2022), 17–22.

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

Cong, Y., Chersoni, E., Hsu, Y.-Y., & Lenci, A. (2023). Are language models sensitive to semantic attraction? a study on surprisal. *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (* SEM 2023)*, 141–148.

Cooper, W. E., & Ross, J. R. (1975). World order. *Papers from the parasession on functionalism*, *11*, 63–111.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211.

Ettinger, A. (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.

Futrell, R., & Levy, R. P. (2018). Do rnns learn human-like abstract word order preferences? *arXiv preprint arXiv:1811.01866*.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., & Nabeshima, N. (2020). The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Gelman, R. (1981). The development of thoughts about animate and inanimate objects: Implications for research in social cognition. *The development of social cognition in children*.

Gennari, S. P., & MacDonald, M. C. (2009). Linking production and comprehension processes: The case of relative clauses. *Cognition*, *111*(1), 1–23.

Gennari, S. P., Mirković, J., & MacDonald, M. C. (2012). Animacy and competition in relative clause production: A cross-linguistic investigation. *Cognitive psychology*, *65*(2), 141–176.

Gokaslan, A., & Cohen, V. (2019). Openwebtext corpus.

Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of experimental psychology: learning, memory, and cognition*, *27*(6), 1411.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *Second meeting of the north american chapter of the association for computational linguistics*.

Hanna, M., Belinkov, Y., & Pezzelle, S. (2023). When language models fall in love: Animacy processing in transformer language models. *arXiv preprint arXiv:2310.15004*.

Hsiao, Y., & MacDonald, M. C. (2016). Production predicts comprehension: Animacy effects in mandarin relative clause processing. *Journal of Memory and Language*, *89*, 87–109.

Hu, J., & Levy, R. (2023). Prompting is not a substitute for probability measurements in large language models. https://arxiv.org/abs/2305.13264

Humphreys, G. F., Mirković, J., & Gennari, S. P. (2016). Similarity-based competition in relative clause production and comprehension. *Journal of Memory and Language*, *89*, 200–221.

Kauf, C., Ivanova, A. A., Rambelli, G., Chersoni, E., She, J. S., Chowdhury, Z., Fedorenko, E., & Lenci, A. (2023). Event knowledge in large language models: The gap between the impossible and the unlikely. *Cognitive Science*, *47*(11), e13386.

Lenth, R. (2019). Emmeans: Estimated marginal means, aka least-squares means. r package version 1.4. 3.01.

Leslie, A. M. (1994). Tomm, toby, and agency: Core architecture and domain specificity. *Mapping the mind: Domain specificity in cognition and culture*, *29*, 119–48.

Levy, R. (2013). Memory and surprisal in human sentence comprehension. In *Sentence processing* (pp. 78–114). Psychology Press.

McDonald, J. L., Bock, K., & Kelly, M. H. (1993). Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive psychology*, *25*(2), 188–230.

Michaelov, J. A., & Bergen, B. K. (2022). Rarely a problem? language models exhibit inverse scaling in their predictions following few-type quantifiers. *arXiv preprint arXiv:2212.08700*.

Misra, K. (2022). Minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Montag, J. L., & MacDonald, M. C. (2015). Text exposure predicts spoken production of complex sentences in 8-and 12-year-old children and adults. *Journal of Experimental Psychology: General*, *144*(2), 447.

Montag, J. L., Matsuki, K., Kim, J. Y., & MacDonald, M. C. (2017). Language specific and language general motivations of production choices: A multi-clause and multi-language investigation. *Collabra: Psychology*, *3*(1), 20.

Nair, S., & Resnik, P. (2023). Words, subwords, and morphemes: What really matters in the surprisal-reading time relationship? *arXiv preprint arXiv:2310.17774*.

Papadimitriou, I. V. (2024). *Jointly studying linguistic structure and language models: Methods for a bilateral science* [Doctoral dissertation, Stanford University].

Rakison, D. H., & Poulin-Dubois, D. (2001). Developmental origin of the animate–inanimate distinction. *Psychological bulletin*, *127*(2), 209.

Rodrigo, L., Igoa, J. M., & Sakai, H. (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*, 325103.

Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic english grammatical structures: A corpus analysis. *Journal of memory and language*, *57*(3), 348–379.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. (2023). Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Wu, S. H., Henderson, L.-M., & Gennari, S. P. (2022). Animacy-induced conflict in sentence production and comprehension from late childhood to adolescence. *Journal of experimental child psychology*, *217*, 105350.

Yun, H., Yi, E., & Song, S. (2023). Exploring ai-generated english relative clauses in comparison to human production. *Journal of Cognitive Science*, *24*(4).

A Appendix: Psycholinguistic Experiment Procedure

Figure 6 presents an example trail from the elicitation task used in the psycholinguistic experiment. Participants viewed the image for 3 seconds before hearing a prompt question (e.g., "*Who wears red?*"). They then responded based on their observation of the scene. To encourage the production of ORCs without directly instructing participants, we told them their responses would help another participant identify characters or objects in the images. To prevent reliance on surface-level features like color (e.g., "red") or position (e.g., "on the left"), participants were informed that these features would change for the next group, while the actions would remain constant. This setup subtly prompted the use of ORCs by emphasizing actions as the most stable and reliable descriptors.

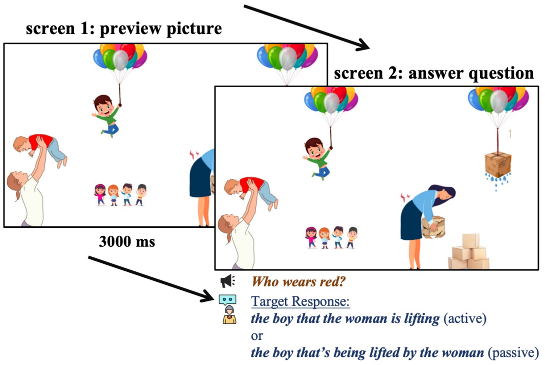
screen 1: preview picture
screen 2: answer question
3000 ms
Who wears red?
Target Response:
the boy that the woman is lifting (active)
or
the boy that's being lifted by the woman (passive)

Figure 6: Sample stimulus image illustrating an ORC elicitation scenario

B Appendix: Surprisal-based analysis: out-of-vocabulary ratio

To ensure that surprisal differences across animacy conditions were not artifacts of tokenization, we examined the out-of-vocabulary (OOV) rates for each

LM by animacy condition. We acknowledge that word-level splits in subword tokenization may reduce the psycholinguistic validity of surprisal at the individual item level. However, Nair and Resnik (2023) found that BPE surprisal retains predictive power when comparing condition-level means—a pattern directly relevant to our study design, and that BPE-based models like GPT-2 still yield reliable surprisal–reading time correlations at the aggregate level.

Figure 7 shows the OOV percentage across animacy condition within each LM in our surprisal-based analysis. We see that OOV rates within each LM were quite consistent across animacy conditions. For example, BERT models ranged from 19.6% (AA) to 21.5% (II), while GPT models ranged from 33.9% (IA) to 36.5% (AI/II) (Distil GPT2 and GPT neo, BERT-base and BERT-large were combined due to the same OOV score). This stability across conditions suggests that differences in surprisal are unlikely to be driven by variability in tokenization. Thus, while GPT-based models naturally exhibit higher OOV due to their subword vocabularies, the uniformity of these rates across animacy conditions allows for meaningful interpretation of surprisal trends in line with the broader goals of our study.

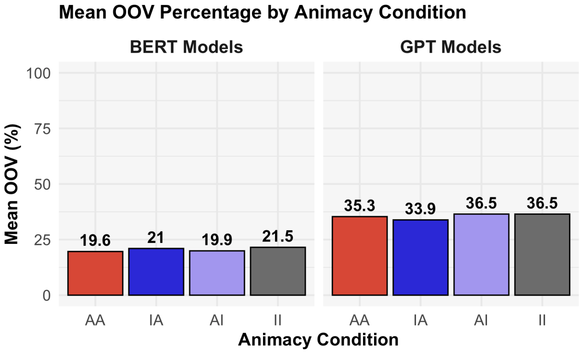

Figure 7: OOV by animacy by LM