Encoder-only and decoder-only models exhibit distinct patterns of linguistic bias

Anonymous ACL submission

Abstract

Large language models (LLMs) based on encoder and decoder architectures exhibit strong performance across a range of tasks, yet systematic comparisons of how training objectives shape their behavior remain limited. Building on prior work establishing systematic comparison between decoder-encoder differences (Weller et al., 2026), we ask whether this difference shapes different linguistic bias using two production-related tasks: semantic role assignment and syntactic structure choice (actives vs. passives). Across architectures, LLMs exhibit a robust animacy-first bias, preferring to place animate entities in subject position, diverging from human bias. However, when animacy cues are removed, encoders generalize better than decoders and show a stronger preference for passive constructions. We argue that these differences reflect training objectives: encoders' sensitivity to contextual compatibility supports flexible mapping from semantics to structure, whereas decoders' reliance on surface frequency favors dominant patterns and disfavors infrequent structures such as passives.

1 Introduction

Encoder and decoder architectures differ fundamentally in their training objectives. Encoder-only models (henceforth encoders), such as BERT (Devlin et al., 2019), are typically trained with masked language modeling, predicting tokens from bidirectional context and encouraging representations that capture contextual compatibility. Decoder-only models (henceforth decoders), such as GPT-style models (Brown et al., 2020; OpenAI et al., 2024; Grattafiori et al., 2024), are trained autoregressively to predict the next token given preceding context, directly modeling the probability of word sequences. Recent work, using controlled comparisons between paired encoder and decoder models trained on matched data and training recipes, has shown that differences in training objectives alone lead to systematic differences in downstream behavior, such that encoders are better at classification and retrieval, whereas decoders are better at generation. (Weller et al., 2026).

This paper asks whether the difference in the training objective also shapes the linguistic bias of LLMs differently. We focus on two tasks that require understanding of verb semantics, namely semantic role assignment and structural choice between actives and passives. Previous work that evaluated LLM performance on these two tasks mostly focused on BERT-variants (Petty et al., 2022; Papadimitriou et al., 2022; Ettinger, 2020; Wilson et al., 2023). Among those that included decoder-only models (Lee et al., 2024; Kauf et al., 2023), systematic comparison with controlled training data, training recipe, and number of parameters has not yet been established. It thus remains less clear if the reported sensitivity of LLMs in these works is associated with encoders, especially given the evidence that encoders and decoders encode semantic representations differently (Ethayarajh, 2019). Leveraging the paired encoder–decoder suite developed by Weller et al. (2026), this paper extends this line of work by understanding the source of LLMs' biases, especially the role of training objectives, in these tasks.

Semantic roles refer to the roles of participants that involved in the event, e.g., who deliberately initiates and experiences the event (proto-agent) vs. who undergoes the event (proto-patient) (Dowty, 1991). When encoding an event from the conceptual level into a linguistic message, verb-specific constraints determine how semantic roles map onto grammatical functions. For instance, an action verb *chase* always places the agent-like role (the chaser) to the subject position (e.g., The dog chased John), while an emotion verb *amaze* always places the more agent-like experiencer to the object position (e.g., The bird amazed John). Thus, the correct mapping between semantic role and grammatical

	Animate noun first	Inanimate noun first
Active	1. The activist protested/#angered the conflict.	2. The conflict #protested/angered the activist.
Passive	3. The activist was #protested/angered by the conflict.	4. The conflict was protested/#angered by the activist.

Table 1: Example frames constructed with the noun pair (*the activist - the conflict*), with the action verb (*protested*) or the emotion verb (*angered*). Frames that are unacceptale by human speakers are labelled with #.

function requires knowledge of verb semantics.

Verb semantics also plays a crucial role in determining structural choice. Psycholinguistic research has found that human speakers tend to put the agent-like role to the subject position, and the patient-like role to the object position (Ferreira, 1994; Do and Kaiser, 2022). Consequently, with emotion verbs that place the experiencer in object position, speakers are more likely to use passives to promote the experiencer to subject position (e.g., *John was amazed by the bird*) than with action verbs that inherently align agents with subjects.

There are three major findings: First, LLMs exhibit a consistent animacy-first bias regardless of the training objective. Second, when animacy cues are removed, encoders generalize better than decoders, and larger models outperform smaller ones. This suggests that training objective still matter for acquiring linguistic bias. Third, encoders show a substantially strong preference for passives compared to decoders, despite the infrequency of passives in human speakers. We suggest this is because the training objective of encoders does not penalize the infrequency of passives: as long as it is compatible with the context, passives are acceptable.

2 Experiments

2.1 Models

We evaluated a broad set of encoder-only and decoder-only language models to compare how architectural differences relate to the linguistic biases under investigation. Encoder models include RoBERTa-BASE/LARGE (Liu et al., 2019) and MODERNBERT-BASE/LARGE (Warner et al., 2024), covering both widely used BERT-style encoders and more recent encoder architectures. Decoder models include GPT2-SMALL/XL (Radford et al., 2019), OLMO-2-7B-BASE (Team OLMo et al., 2025), and OLMO-3-7B-BASE (Team Olmo et al., 2026), spanning earlier autoregressive models as well as recent open-weight decoder-only LLMs. In addition, to enable a controlled head-to-head comparison between architectures, we included the full Ettin series of paired encoders and decoders, ranging from 17M to 1B parameters (Weller et al., 2026). This model suite allows us to examine linguistic biases across architecture type, model family, scale, and training setup, while distinguishing broad encoder-decoder differences from patterns that may be specific to individual model families.

2.2 Materials

We constructed 88 different frames with the animate-inanimate noun pairs based on Ferreira (1994)'s experimental materials. Each frame comes with four versions, varying in noun order and sentence voice (active vs. passive). Each version is paired with both an action and an emotion verb, yielding 88 $\times$ 8 = 704 sentences. Table 1 illustrates all eight versions of sentences. This set of frames with animate-inanimate nouns were used for both the semantic role assignment task and the structural choice task. We further constructed 48 frames with two animate nouns (e.g., Ray stunned Max). These animate-only frames were used to test the possibility that the structural choice bias is due to animacy. Semantic role assignemnt was not evaluated because when both nouns are animate, there is no clear bias to assign the semantic role (both *Ray stunned Max* and *Max stunned Ray* are acceptable). All of the animate-only frames come with the eight versions listed in Table 1. The full set of frames can be found in Appendix E.

2.3 Evaluation

We evaluated the performance of encoder-only models by obtaining the masked probability of the verb token of each version of the animate-inanimate frame listed in Table 1, and the performance of decoder-only models by comparing the sentence probability of these frames.

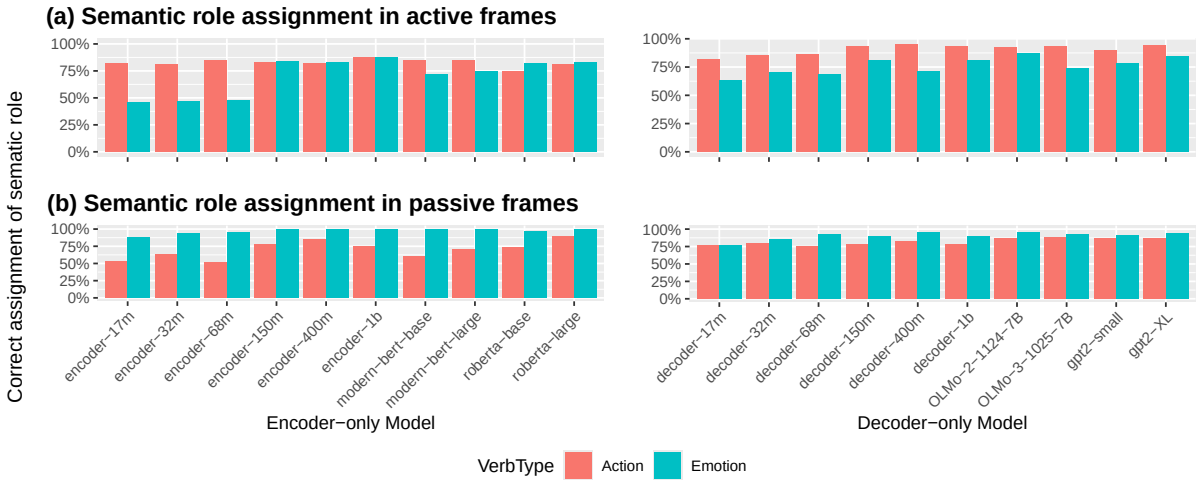

Figure 1: Correctness of semantic role assignment by verb type (action vs. emotion), model type (encoder vs. decoder) and frame type (a: active frames vs. b: passive frames). In active frames, the correctness of semantic role assignment is higher with action verbs thanwith emotion verbs. In passive frames, the correctness is higher with emotion verbs than with action verbs.

For semantic role assignment, the masked probability of the verb or the sentence probability should be higher whenever the animate entity is placed in line with the verb-specific constraints than when it violates the verb-specific constraints. That is, the masked probability of the verb token or the sentence probability should be higher when the animate noun is the subject than when it is the object for action verbs (i.e., 1 > 2; 4 > 3 in Table 1), and when it is the object than when it is the subject for emotion verbs (i.e., 2 > 1; 3 > 4 in Table 1). For structural choice, we used passive bias as the evaluation metric, defined as a higher masked probability of the verb or sentence probability of the correct passive frame than in the correct active frame (i.e., 4 > 1 for action verbs; 3 > 2 for emotion verbs). Recall that human speakers produce passives more frequently with emotion verbs than with action verbs (Ferreira, 1994). Thus, if semantic roles are correctly assigned, passive bias should be stronger for emotion verbs than for action verbs.

3 Results and Discussion

Figure 1 plot the correctness of semantic role assignment by model and verb type, in active and passive frames respectively. Figure 2 plot the passive bias of each model, i.e., the percentage of stimuli that passive frames are more probable than active frames, respectively with animate-inanimate and animate-only noun pairs.

Finding 1: Animacy-first bias across architectures. LLMs consistently place animate entities in subject position regardless of verb semantics, such that they made more mistakes in semantic role assignment for emotion verbs in active frames and action verbs in passive frames, where the verb only allows the animate experiencer or agent role to be put in the object position. This pattern holds across both encoder and decoder architectures, suggesting it reflects a general property of LLM training rather than a specific architectural feature. For structural choice, when the noun pairs differ in animacy, all models exhibit a stronger passive bias in emotion verbs than in action verbs, in line with human results. Yet, such as bias could still be attributed to the animacy-first heuristics. This is because when the animacy cue is removed, the passive bias only exists in a few models with a much smaller magnitude. This differs from human strategies, in which the passive bias is found independent of the animacy effect (Ferreira, 1994).

Finding 2: Encoders generalize better and scaling matters. When animacy cues are removed for the structural choice task, larger models outperform smaller ones, and encoders are more likely to replicate the stronger passive bias in emotion verb than decoders overall.

Finding 3: Encoders prefer passives. Encoders exhibit a much stronger preference for passive constructions than decoders (>80%), regardless of the model size or the presence of animacy cues.

Why do encoders generalize better? Several factors may contribute to the generalization advantage of encoders. First, the bidirectional attention

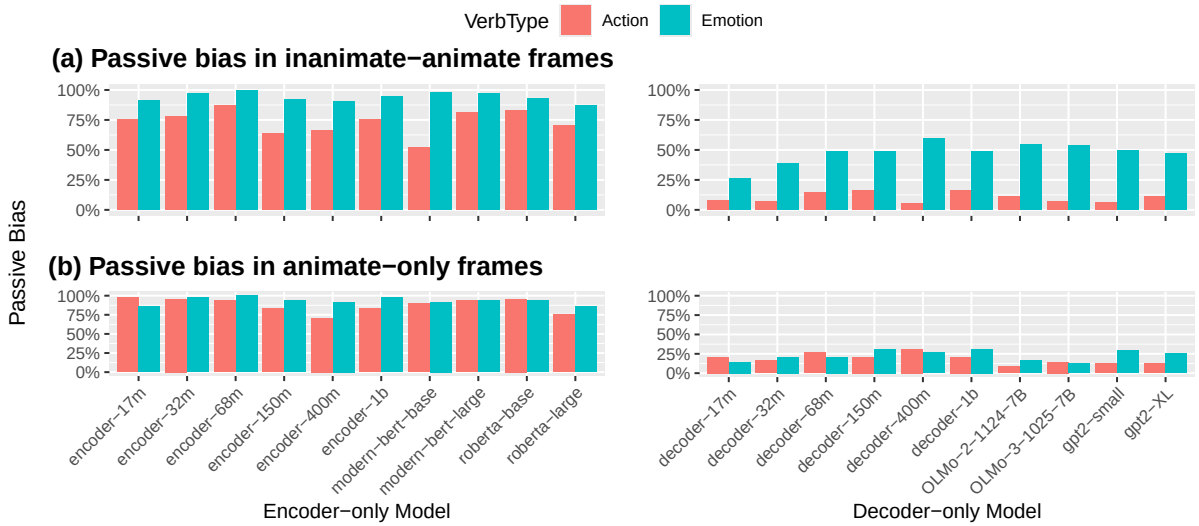

Figure 2: Preference to use passive frames over active frames (passive bias) by verb type (action vs. emotion), model type and the animacy of noun pair (inanimate-animate vs. animate-only). Only frames that models are correct with semantic role assignment are evaluated. In line with human bias, the passive bias is larger for emotion verbs than action verbs, but only when the noun pair differs in animacy.

mechanism in encoders allows verb representations to be simultaneously informed by both the subject and object positions, potentially enabling encoders to develop more robust mappings between verbs and their thematic roles. Decoders, processing tokens left-to-right, may encode verbs in a way that is more strongly conditioned on the preceding subject, making them more susceptible to animacy cues at that position. Second, the masked language modeling objective encourages encoders to resolve semantic ambiguity using all available context, which may incidentally train them to rely more on verb-specific thematic role information rather than general animacy heuristics. Third, it is possible that encoder and decoder models develop systematically different verb embedding spaces, where encoder verb embeddings more clearly distinguish between verbs that strongly prefer animate subjects for semantic reasons and those that do not. This conjecture is partially confirmed by visualizing the embedding space using PCA and t-SNE projections of the largest paired encoders and decoders in the full Ettin series (Weller et al., 2026) in Appendix D.

Why do encoders prefer passives more? We attribute this to differences in training objectives: decoders disfavor low-frequency constructions like passives, whereas encoders are more sensitive to contextual compatibility, making them more accepting of passives in appropriate contexts. This account is supported by the representational analyses in Appendix D: encoders encode animacy/order and verb type as continuous, globally structured semantic dimensions, whereas decoders encode these distinctions in a more discrete, locally clustered manner. Although neither architecture has a dimension for voice, this suggests that passive preference is not directly encoded but instead emerges from how semantic factors are structured: encoders' sensitivity to contextual compatibility is reflected in an integrated representational space that supports flexible mapping from semantics to structure.

4 Conclusion

This paper examined how encoder-only and decoder-only models differ in linguistic biases in tasks involving semantic role assignment and structural choice. Across tasks, LLMs exhibit a robust animacy-first bias, diverging from human preferences, which further take the alignment between grammatical position and semantic role into account. However, the two architectures differ beyond animacy: when animacy cues are absent, encoders generalize better but show a much stronger passive bias, suggesting a greater influence of verb semantics on structural choice. Decoders, by contrast, rely more on surface frequency and show a weaker sensitivity to verb semantics. These findings indicate that architectural differences shape not only performance but also the linguistic bias acquired by models.

Limitations

First, the present tasks are primarily rooted in semantic understanding. Both semantic role assignment and structural choice require models to use verb semantics, animacy, and thematic-role information to evaluate sentence-frame compatibility. Our conclusions should therefore be interpreted as applying to semantic-role- and structure-related biases, rather than to linguistic bias in general.

Second, the model comparison is limited by the models included and by how behavior is operationalized. Although the Ettin suite provides controlled encoder-decoder pairs, other models differ in training data, tokenization, scale, and model family. In addition, encoder and decoder probabilities are computed with architecture-specific procedures, so absolute effect sizes should be compared with caution.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Monica L Do and Elsi Kaiser. 2022. Sentence formulation is easier when thematic and syntactic prominence align: evidence from psych verbs. *Language, Cognition and Neuroscience*, 37(5):648–670.

David Dowty. 1991. Thematic proto-roles and argument selection. *language*, 67(3):547–619.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Allyson Ettinger. 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.

Fernanda Ferreira. 1994. Choice of passive voice is affected by verb type and animacy. *Journal of Memory and Language*, 33(6):715–736.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Carina Kauf, Anna A Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan Selena She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2023. Event knowledge in large language models: the gap between the impossible and the unlikely. *Cognitive Science*, 47(11):e13386.

Eun-Kyoung Rosa Lee, Sathvik Nair, and Naomi Feldman. 2024. A psycholinguistic evaluation of language models' sensitivity to argument roles. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3262–3274.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Isabel Papadimitriou, Richard Futrell, and Kyle Mahowald. 2022. When classifying grammatical role, BERT doesn't care about word order... except when it matters. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 636–643, Dublin, Ireland. Association for Computational Linguistics.

Jackson Petty, Michael Wilson, and Robert Frank. 2022. Do language models learn position-role mappings? *CoRR*, abs/2202.03611.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, and 50 others. 2026. Olmo 3. *Preprint*, arXiv:2512.13961.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 24 others. 2025. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, and Benjamin Van Durme. 2026. Seq vs seq: An open suite of paired encoders and decoders. *Preprint*, arXiv:2507.11412.

Michael Wilson, Jackson Petty, and Robert Frank. 2023. How abstract is linguistic generalization in large language models? experiments with argument structure. *Transactions of the Association for Computational Linguistics*, 11:1377–1395.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Computational Experiment Details

All model evaluations were conducted with deterministic scoring rather than free-form generation, implemented with the Hugging Face Transformers library (Wolf et al., 2020). For encoder-only masked-language models, we used `AutoModelForMaskedLM` to compute the probability of the target verb in the masked verb position. When a target verb was split into multiple subword tokens, we inserted the corresponding number of mask tokens and computed the joint probability of the verb span by sequentially unmasking the subword tokens from left to right. For decoder-only models, we used `AutoModelForCausalLM` to compute full-sequence log probabilities for the active and passive sentence frames. Because this procedure evaluates the likelihood of fixed strings, no decoding hyperparameters such as temperature, top-p, or maximum generation length were used.

For representational analyses, we extracted contextual verb embeddings using `AutoModel`. The full sentence was provided to the model without masking, and the hidden states corresponding to the verb span were mean-pooled across all subword tokens. Unless otherwise specified, embeddings were extracted from the final hidden layer. For encoder models, the resulting verb representation is conditioned on both left and right context; for decoder models, the verb-position representation is conditioned only on the verb and its left context due to causal attention. All experiments were run on a single NVIDIA H100 GPU and required approximately one GPU hour in total.

B License Statement

Materials. Our experimental stimuli are based on the psycholinguistic materials of Ferreira (1994), which have been widely used to study thematic role assignment and active/passive structural choice. We adapt these materials for controlled model evaluation by constructing active and passive sentence frames with animate-inanimate and animate-only noun pairs. The resulting stimuli and evaluation scripts will be released for non-commercial academic research with attribution to the original source.

Models. We evaluate publicly released pretrained language models under their respective model licenses and terms of use. The encoder models include RoBERTa-base/large (Liu et al., 2019), ModernBERT-base/large (Warner et al., 2024), and the encoder members of the Ettin suite (Weller et al., 2026). The decoder models include GPT2-small/XL (Radford et al., 2019), OLMo-2-7B-Base (Team OLMo et al., 2025), OLMo-3-7B-Base (Team Olmo et al., 2026), and the decoder members of the Ettin suite (Weller et al., 2026). All models are used only for non-commercial academic research.

C GenAI Statement

We disclose the use of generative AI tools in this work in accordance with the ACL Policy on the

Use of AI Writing Assistance.

AI use in research artifacts. Generative AI tools were not used to create the experimental stimuli, compute model scores, extract embeddings, generate statistical results, or produce the figures. The stimuli were manually constructed from prior psycholinguistic materials, and all model evaluations and visualizations were conducted using the scripts described in this paper.

AI use in experiments. The language models evaluated in this work are themselves the subject of study. Their use is fully described in Section 3 and Appendix A. They were not used as auxiliary annotators or data-generation tools.

AI use in writing. We used AI assistants, including ChatGPT, for surface-level writing support, such as grammar correction, sentence-level rephrasing, LaTeX formatting suggestions, and code organization. All scientific claims, experimental design choices, analyses, interpretations, and conclusions were authored and verified by the human authors. AI assistants were not used to generate citations, statistical results, or factual claims reported in this paper.

Responsibility. The authors take full responsibility for the content of this paper, including any text or code that may have been drafted or edited with AI assistance.

D Plots of verb embeddings

Figure 3 shows a PCA projection of the representations for the stimuli in the largest paired encoders and decoders, which have 1B parameters. Figure 4 shows a t-SNE visualization of the same representations. In both active and passive conditions, representations are primarily separated along PC1, with decoders forming a compact cluster on the left and encoders occupying a broader region on the right. Within the encoder space, points form a diagonal gradient running from the upper-left (lower PC1, higher PC2) to the lower-right (higher PC1, lower PC2). Along this axis, conditions vary systematically: configurations in which the animate entity appears earlier tend to lie toward one end of the gradient, while those in which the inanimate entity appears earlier are shifted toward the opposite end. This reflects a continuous displacement of points along both PC1 and PC2, rather than separation into discrete clusters. Verb type (AP = action vs. TE = emotion) induces smaller, approximately orthogonal or local shifts within this space, producing partial offsets among points with the same animacy/order configuration but different verbs. In contrast, decoder representations remain tightly clustered with minimal displacement along either axis. The overall geometry is highly similar across voice conditions. This pattern indicates that encoders encode animacy/order as a continuous, low-dimensional semantic direction in representation space, with verb type modulating this structure in a secondary way, consistent with their stronger generalization and more systematic structural biases, whereas decoders do not exhibit comparable linear organization of these distinctions.

In both plots, encoders and decoders are separated along the horizontal axis, with encoders on the right and decoders on the left. Within the decoder space, points form two vertically separated clusters, and conditions with different animacy/order configurations tend to group within the same cluster, indicating that animacy/order is reflected as local grouping structure rather than a continuous gradient. In contrast, encoder points are more diffusely distributed, and differences in animacy/order appear as gradual shifts in local neighborhoods rather than forming discrete clusters, with nearby points sharing similar configurations. The spatial organization remains largely unchanged between active and passive conditions. This suggests that animacy/order is encoded in a more clustered, locally separable way in decoders, but as a continuous, globally structured dimension in encoders.

PCA and t-SNE jointly indicate that, while both architectures exhibit an animacy-first bias, encoders encode animacy/order (with verb type as a secondary factor) as continuous, globally structured semantic dimensions, whereas decoders encode these distinctions in a discrete, locally clustered manner. The absence of a voice-related dimension suggests that passive preference is not explicitly represented, but instead emerges from how semantic information is structured. In this respect, the more integrated representational geometry of encoders supports flexible mapping from semantics to syntactic choice, leading to stronger passive bias and improved generalization in the absence of animacy cues, while decoders' less structured representations favor reliance on surface frequencies.

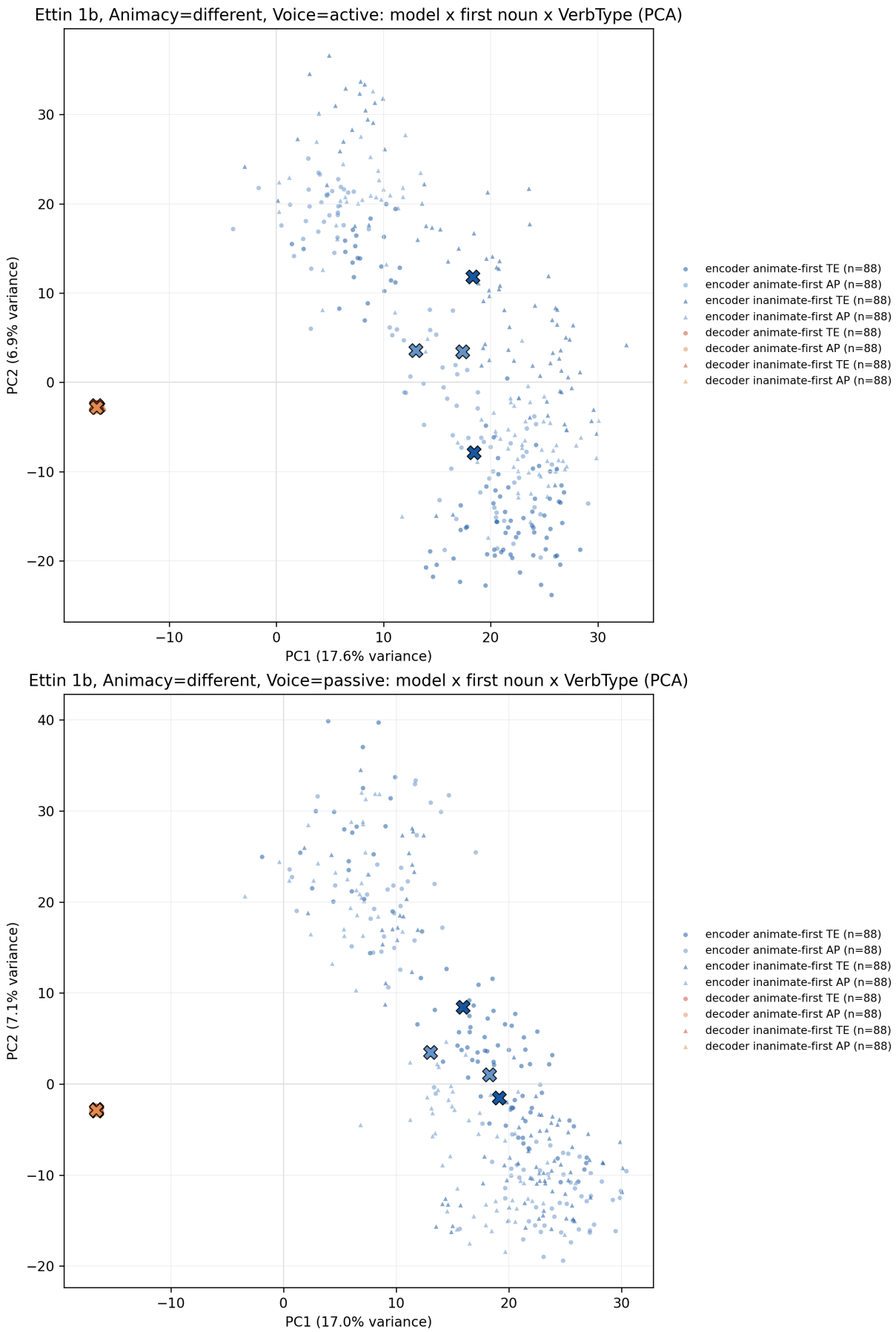

Figure 3: PCA of sentence representations for encoder and decoder (1B) models under active and passive conditions. Representations are primarily separated by architecture along PC1, with decoders forming a compact cluster and encoders occupying a broader, structured space. Within the encoder space, conditions are organized along a diagonal gradient reflecting animacy/order, with verb type introducing smaller local shifts. The overall geometry is highly similar across voice conditions.

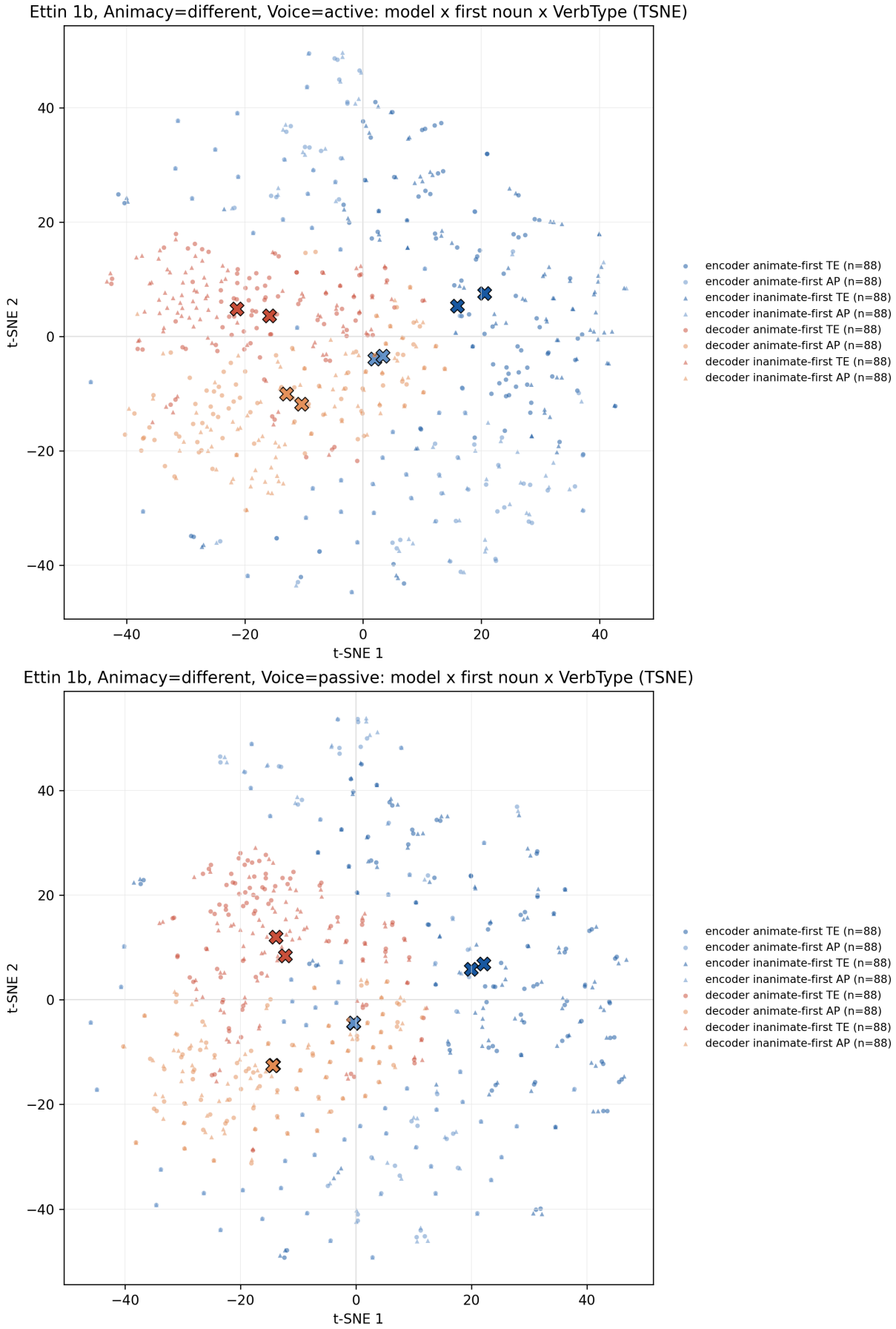

Figure 4: t-SNE visualization of sentence representations for encoder and decoder models (1B) under active and passive conditions. Encoders and decoders occupy distinct regions, with decoders forming two discrete clusters that group conditions by animacy/order and verb type, while encoder representations are more diffusely distributed with gradual local variation. The spatial organization remains largely consistent across voice conditions.

Inanimate noun	Animate noun	Emotion verb	Action verb
conflict	activist	angered	protested
news	dictator	alarmed	suppressed
insult	fighter	provoked	bellowed
massage	athlete	relaxed	requested
affair	housewife	shamed	shunned
lullaby	baby	soothed	booed
cake	child	tempted	crumbled
evidence	detective	troubled	transformed
lightning	artist	frightened	painted
jokes	audience	entertained	applauded
hairdo	mother	irritated	complimented
thunder	children	terrified	dreaded
theft	judge	disturbed	abhorred
image	painter	haunted	sketched
layoffs	manager	worried	ordered
slayings	citizens	shocked	loathed
gift	mother	pleased	wrapped
sirens	police	alerted	startled
cave	boys	scared	explored
T.V.	student	distracted	ignored
map	explorers	guided	possessed
textbook	student	bored	xeroxed
cheesecake	dieter	enticed	devoured
praise	child	encouraged	awaited
tragedy	family	stunned	mourned
profits	executive	impressed	insured
defeat	troops	demoralized	disregarded
championship	team	excited	secured
crime	community	appalled	detested
fireworks	child	thrilled	feared
delay	customer	annoyed	despised
toy	child	amused	hurled
catcalls	referee	enraged	ignored
puzzle	genius	challenged	unraveled
play	reviewer	captivated	criticized
scandal	politician	disgraced	denounced
phenomenon	scientist	intrigued	analysed
pimple	teenager	embarrassed	camouflaged
belches	socialite	disgusted	disguised
rumors	queen	offended	overlooked

Table 2: Materials with inanimate-animate noun pair

E Materials

Table 2 and 3 list the materials to construct the frames, with inanimate-animate noun pairs and animate-only noun pairs respectively. For inanimate-animate noun pairs, we constructed active frames in the format of [animate noun] + [emotion/action verb] + [inanimate noun] (corresponding to (1) in Table 1) and [animate noun] + [emotion/action verb] + [inanimate noun] (corresponding to (2) in Table 1). Passive frames were constructed in the same format except for the verb form (e.g., protested → was protested by).

For animate-only pair, the same frames were used except that the inanimate noun is replaced by [Noun 2] in Table 3.

Noun 1	Noun 2	Emotion verb	Action verb
constituent	politician	incensed	endorsed
correspondent	ambassador	embarrassed	authorized
pianist	conductor	provoked	denounced
publisher	novelist	scandalized	exploited
sheriff	cowboy	challenged	avoided
sitter	youngster	amused	appreciated
doctor	student	worried	approached
vampire	maiden	frightened	ignored
critic	actor	demoralized	disregarded
clergyman	atheist	offended	mistrusted
lawyer	judge	persuaded	seized
lifeguard	swimmer	shocked	loathed
trainer	boxer	thrilled	timed
supervisor	interpreter	pleased	praised
corporal	guerrilla	alerted	surrendered
administrator	undergraduate	discouraged	interrupted
superintendent	investigator	baffled	inspected
examiner	competitor	encouraged	criticized
grandfather	grandmother	upset	analyzed
hairdresser	socialite	delighted	resented
journalist	dictator	enraged	censored
rescuer	survivor	impressed	evaded
inhabitant	missionary	bored	blessed
performer	spectator	entertained	applauded
instructor	principal	annoyed	despised
teenager	murderer	terrified	resisted
thief	cop	frustrated	escorted
refugee	fisherman	scared	identified
therapist	immigrant	disturbed	heeded
hypnotist	dieter	relaxed	trusted
employer	apprentice	excited	rejected
camper	warden	concerned	halted
Linda	Rachel	distracted	scorned
Helen	Judy	soothed	weighed
Ruth	Ann	angered	disliked
Kathy	Christine	awed	idealized
Lucy	Susan	disappointed	misunderstood
Ray	Max	stunned	concealed
Stephen	Rose	intrigued	embraced
Herman	Michael	appalled	blamed
Phil	Tom	alarmed	excused
Truman	Nelson	perplexed	disobeyed
Emma	Bill	bothered	dreaded
Paula	Chris	amazed	controlled
Mark	Marjorie	tempted	scratched
Roger	Harry	convinced	admired
Dean	Pat	enticed	labeled
Albert	Barbara	astounded	neglected

Table 3: Materials with animate noun pairs