Journal of Memory and Language 149 (2026) 104751

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Manipulating language models' training data to study syntactic constraint learning: The case of English passivization☆

Cara Su-Yi Leong [a],*, Tal Linzen [a,b]

[a] Department of Linguistics, New York University, 10 Washington Place, New York, 10003, NY, USA
[b] Center for Data Science, New York University, 60 5th Avenue, New York, 10012, NY, USA

ARTICLE INFO

Dataset link: https://github.com/craaaa/exceptions

Keywords:
Learnability
Language models
Passivization

ABSTRACT

Grammatical rules in natural languages are often characterized by exceptions. How do language learners learn these exceptions to otherwise general patterns? Here, we study this question through the case study of English passivization. While passivization is in general quite productive, there are cases where it cannot apply (cf. the following sentence is ungrammatical: **One hour was lasted by the meeting*). Using neural network language models as theories of language acquisition, we explore the sources of indirect evidence that a learner can leverage to learn whether a verb can be passivized. We first characterize English speakers' judgments of exceptions to the passive, and confirm that speakers find some verbs more passivizable than others. We then show that a neural network language model's verb passivizability judgments are largely similar to those displayed by humans, suggesting that evidence for these exceptions is available in the linguistic input. Finally, we test two hypotheses as to the source of evidence that language models use to learn these restrictions: frequency (entrenchment) and semantics (affectedness). We do so by training models on versions of the corpus that have had sentences of the types implicated by each hypothesis removed, altered, or introduced. We find support for both hypotheses: entrenchment and affectedness make independent contributions to a verb's passivizability. From a methodological point of view, this study highlights the utility of altering a language model's training data for answering questions where complete control over a learner's input is vital.

Introduction

Grammatical generalizations in natural languages are often characterized by systematic exceptions: classes of cases where the generalizations do not apply. What sources of evidence do learns use to acquire these exceptions? In the current study, we address this question through a computational study of neural network language models' learning of the constraints on English passivization. Passivization is in general quite productive in English: Speakers freely use most transitive verbs in both active and passive voice, and English-speaking children who learn novel transitive verbs in the active voice can use those verbs in the passive voice (Brooks & Tomasello, 1999; Pinker et al., 1987). But for a small set of verbs this generalization does not hold; the verb *last*, for example, is acceptable in the active but not in the passive:

(1) a. The meeting lasted one hour.
b. * One hour was lasted by the meeting.

Why do speakers judge (1b) as unacceptable? One tempting explanation may be that they do so because they simply have never heard a sentence such as (1b), where *last* is passivized. But that explanation is most likely insufficient. Consider the verb *defenestrate*:

(2) a. The writer defenestrated the editor.
b. The editor was defenestrated by the writer.

Because *defenestrate* is a rare lemma, and passives are rare in everyday speech (on average, only one out of ten utterances uses the passive voice; Roland et al. 2007), the odds that a particular speaker has heard a sentence like (2b), where *defenestrate* is passivized, are very low, quite possibly as low as for 1b. Yet English speakers judge (2b), but not 1b, as acceptable. How do learners of English consistently arrive at a grammar under which *last* cannot be passivized but *defenestrate* can? This learnability challenge—separating forms that do not occur because they are unacceptable from forms that are not observed simply

☆ This article is part of a Special issue entitled: 'Language Models & Psycholinguistics' published in Journal of Memory and Language.
* Corresponding author.
E-mail addresses: caraleong@nyu.edu (C.S.-Y. Leong), linzen@nyu.edu (T. Linzen).

https://doi.org/10.1016/j.jml.2026.104751
Received 20 February 2026; Accepted 26 February 2026

due to chance—is sometimes referred to in linguistics as "Baker's Paradox" (Baker, 1979).

Could innate constraints account for exceptions to passivization? The answer to this question is unlikely to be entirely positive: because exceptions to passivization vary across languages, they need to be learned. For example, stative verbs like *cost* and *have* cannot be passivized in English, but they can in Kinyarwanda (Keenan & Dryer, 2007):

(3) *A new car is had by John.

(4)

Ibifuungo	*bibiri*	*bi-fit-w-e*	*n-îshaâti*
buttons	two	they-have-PASS-ASP	by-shirt

'Two buttons are had by the shirt.'

(Keenan & Dryer, 2007, 332)

While there are certain regularities across languages in terms of the set of verbs that can and cannot passivize (Ambridge et al., 2023), the particular verbs that are subject to restrictions on passivization are language-specific. Since this information is likely not explicitly taught to children by caregivers, it must be acquired by learners of the language through exposure to *indirect evidence*.

In this paper, we explore two hypotheses concerning the kinds of indirect evidence for these exceptions that learners might rely on. According to the *entrenchment hypothesis* (Braine & Brooks, 1995; Goldberg, 2006; Theakston, 2004), learners use the statistical distribution of verbs in different constructions to determine where a verb can appear and infer where it cannot occur. Under this hypothesis, learners who never encounter a particular verb in the passive but see it consistently in other contexts will conclude that the verb cannot appear in the passive.

A second potential source of indirect evidence for passive exceptions is the *lexical semantics* of the verb (Ambridge et al., 2016; Pinker, 1989). Under this hypothesis, a verb is passivizable when it denotes an action whose theme participant is *affected* (Pinker, 1989), that is, the theme undergoes a change in state, location, or existence caused by the action's agent participant (Beavers, 2011). Verbs inconsistent with these semantics are unacceptable in the passive; 1b **One hour was lasted by the meeting*, for example, is ungrammatical because *one hour* is not affected by *the meeting* lasting an hour.

Both the affectedness of a verb and its degree of entrenchment in the active, measured by how frequently the verb occurs in the active compared to the passive, *correlate* with English speakers' judgments of its passivizability (Ambridge et al., 2016). Yet it is difficult to study if these factors *causally* affect the learning of restrictions on passivization. Since verbs with low-affectedness semantics are used in the passive only very infrequently, affectedness and entrenchment are highly correlated, and it is hard to know whether speakers make passivizability judgments using one of these factors to the exclusion of the other, or possibly using an entirely different source of evidence. In a counterfactual world, we could experimentally disentangle these factors by manipulating the language that a human language learner is exposed to, for example by artificially increasing the frequency of passive forms of low-affectedness verbs. This is, of course, impossible; instead, we investigate the relationship between properties of the input and the outcome of learning using neural networks as models of language acquisition (Baroni, 2022; Warstadt & Bowman, 2022). We discuss these computational models and their role in cognitive modeling in the following section.

Testing learning hypotheses by manipulating language models' training corpora

Neural network language models are systems that learn probability distributions over sequences of words based on a text corpus (for a review, see Linzen and Baroni 2021). While they are not trained or designed to provide acceptability judgments, such judgments can be derived from them through *targeted syntactic evaluation* (Lau et al., 2017; Linzen et al., 2016; Marvin & Linzen, 2018; Warstadt et al., 2020): given a minimal pair of sentences such as 1, one of which is grammatical and one is not, we use the language model to assign a probability score to each sentence. If the model assigns a higher probability to the grammatical sentence 1a than the ungrammatical 1b, we conclude that it shows sensitivity to the underlying syntactic differences between the two sentences. Targeted syntactic evaluation has shown that neural network language models are sensitive to a variety of syntactic and semantic constraints, including subject–verb agreement, constraints on negative polarity items, and island constraints on filler-gap dependencies (Linzen & Baroni, 2021).

By using language models as theories of acquisition, we can address the limitation that acquisition studies with humans are by necessity observational only, as these models allowing for complete control of the input provided to the learner: we can train multiple models on corpora that differ in controlled and targeted ways, and compare how learners with the same initial state and learning objective but different input diverge in their behavior at the end of learning (Jumelet et al., 2021; Misra & Mahowald, 2024; Wei et al., 2021).

Here, we apply this method to study the emergence of exceptions to English passivization. We train neural network language models on approximately the same amount of linguistic data that humans are exposed to, and use these models to answer two questions: first, is this amount of linguistic input sufficient for the models to learn to make judgments that are similar to human judgments? Second, what kinds of information are available in the linguistic input for the learner to come to make those judgments?

In Experiment 1, we answer the first question in the affirmative. We show that neural network language models' verb passivizability judgments are highly correlated with those of English speakers (r = 0.9), a significantly higher correlation that we observed for simpler frequency-based models. Such behavior suggests that language models can use evidence from the linguistic input to learn exceptions.

To answer the second question, we generate counterfactual training corpora that manipulate different sources of evidence for a verb's passivizability. In Experiments 2A and 2B, we withhold the evidence considered to be critical under the affectedness hypothesis and entrenchment hypothesis. We then train language models on both types of modified corpora, and compare the acceptability judgments provided by these to those of models trained on the original corpus. Finally, in Experiment 3, we introduce a novel verb into the corpus and manipulate how often and in what semantic environments it occurs. This makes it possible to measure the interaction between the affectedness and entrenchment hypotheses in a controlled way.

To foreshadow our results, we do not find clear support for one hypothesis to the exclusion of the other. Rather, entrenchment and affectedness each contribute to a verb's passivizability, and we do not find evidence for an interaction between them. We also find that neither of the hypotheses alone can account for the full difference between passivizable and unpassivizable verbs, which suggests that the input contains sources of evidence for a verb's passivizability other than its frequency and affectedness. More broadly, this empirical study illustrates a method by which researchers can examine how controlled changes to the learner's input affect the outcome of learning.

Overview of experiments

In Experiment 1, we compare human passivizability judgments with those of a model we train. This experiment has two parts. In Experiment 1A, we collect acceptability judgments from English speakers on a set of active and passive sentences containing verbs that are reported in the literature as unacceptable in the passive (Bach, 1980; Levin, 1993; Postal, 2004; Zwicky, 1987). In Experiment 1B, we compare our previously-collected human judgments against to judgments derived from a neural network language model which we train on 100 million words of English.

In Experiment 2, we evaluate the causal factors that our model used to learn to approximate human judgments. In Experiment 2A, we evaluate the entrenchment hypothesis by testing if models' acceptability judgments for passive sentences containing a highly passivizable verb change if the models are trained on a corpus in which that verb occurs much less frequently in the passive than in the original corpus. In Experiment 2B, we evaluate the affectedness hypothesis, which predicts a link between the affectedness of the theme argument and the verb's passivizability. We do so by training models on a corpus where an unpassivizable verb co-occurs with arguments that are associated with a canonically passivizable verb. This alters the proportion of the verb's arguments in the corpus that are affected, signaling to the learner that it is higher in affectedness.

Finally, in Experiment 3, we test if a verb's frequency of occurrence and its semantic context interact in determining its passivizability. To do so, we introduce a novel verb which only occurs in active-voice sentences, and vary the semantic contexts and frequency (and therefore active-to-passive ratio) with which the novel verb occurs.

Materials

English passives are subject to some restrictions that are uncontroversial and do not appear challenging to learn; in particular, the entire class of intransitive verbs cannot occur in the passive (Comrie et al., 1977). In this work, we focus on more subtle restrictions on the passivization of verbs that appear to be transitive, in that they are typically followed by a noun phrase; the evidence for the exceptionality of these verbs is less clear. To compare human and language model judgments for such verbs, we created a dataset of 140 pairs of active and passive sentences—280 sentences in total—using 28 different verbs, 10 of which were ordinary passivizable verbs (*control verbs*), and 18 verbs characterized as unpassivizable in the linguistics literature (*critical verbs*; Bach 1980, Levin 1993, Postal 2004, Zwicky 1987). Each sentence pair consisted of an active transitive sentence (e.g. *A boy dropped the cup*) and a corresponding passive sentence, where the passive is introduced by a form of the verb *be* (e.g. *The cup was dropped by a boy*). Passive sentences always included an explicit *by*-phrase that matched the subject of the active sentence (i.e. *The cup was dropped by a boy*, but not *The cup was dropped*).

The critical verbs belonged to five verb classes. The meanings of the verbs in each class were sufficiently similar that they could all be substituted into the same frame. We constructed five frames for each class; for examples of the frames, see Table 1, and for the full list of critical and control sentences see Appendix A.1. The verbs in each verb class were:

- Advantage verbs: *benefit*, *help*, *profit*, *strengthen*
- Price verbs: *cost*, *earn*, *fetch*
- Ooze verbs: *discharge*, *emanate*, *emit*, *radiate*
- Duration verbs: *last*, *require*, *take*
- Estimation verbs: *approximate*, *match*, *mirror*, *resemble*

Some of the test verbs have multiple senses, of which only one is exceptional. Our sentences used the sense of the verb reported as unpassivizable in the literature. For *take*, for instance, sentences did not use the sense in (5a), only the one illustrated in (5b):

(5) a. The photo was <u>taken</u> by the boy.
b. *Two days was <u>taken</u> by the meeting.

All verbs in the class were inserted into each sentence frame, resulting in 90 total test sentence pairs: 20 pairs each from the advantage, ooze and estimation classes, which have four test verbs each, and 15 pairs from the price and duration classes, which have three test verbs each. Example (6) demonstrates a sentence pair generated from the sentence frame in Table 1 using the verb *benefit*:

(6) a. The gift <u>benefited</u> my organization.
b. My organization was <u>benefited</u> by the gift.

In addition to the five critical verb classes, which contained verbs expected to be unacceptable in the passive, we created stimuli for two control verb classes which we expected to be acceptable in both the active and the passive voice:

- Agent-patient: *hit*, *push*, *wash*, *drop*, *carry*
- Experiencer-theme: *see*, *hear*, *know*, *like*, *remember*

Given the diverse semantics of the verbs in these groups, we used unique sentence frames for each verb. This resulted in 50 control test sentence pairs, and a total of 140 sentence pairs across both critical and control verbs.

The experiment also included filler sentences with similar lengths to the critical items (for a list, see Appendix A.2). Since the passives of control sentences were expected to be acceptable, we included a larger number of ungrammatical than grammatical fillers (52 vs. 26) such that the experiment as a whole contained the same number of grammatical and ungrammatical sentences.

Experiment 1A: English speakers find some verbs more passivizable than others

We first conducted a human acceptability judgment study. This study had two goals: first, to verify the linguists' judgments reported in the syntax literature; and second, to measure any gradient differences in the degree to which different verb classes and individual verbs can be passivized, providing a fine-grained benchmark against which language model judgments can be compared.[1]

Procedure

We collected acceptability judgments from English speakers for the active and passive sentences described in the Materials section. Each participant rated either the active or the passive version of any given sentence pair. Specifically, we divided the 140 sentence pairs into two groups of 70 sentence pairs (i.e. 140 sentences) such that each group contained either two or three sentence frames per verb. We then split each group into two sets of 70 sentences such that the active and passive versions of each item were in different sets.

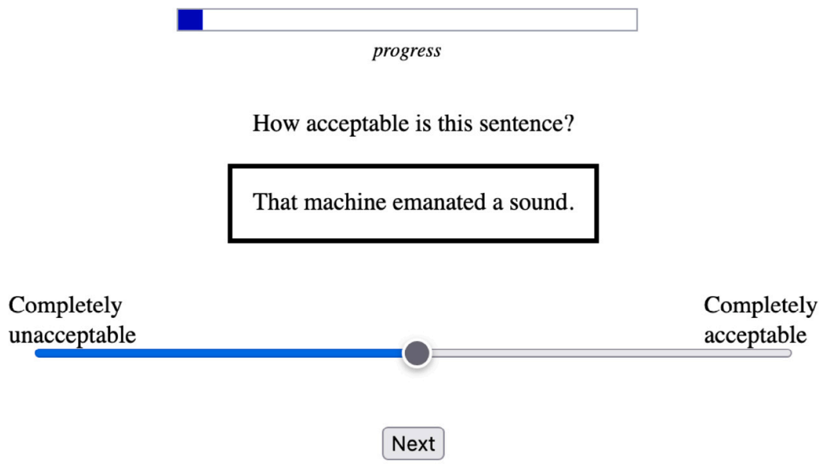
We further counterbalanced the presentation order by creating four ordered lists for each group, as follows. We organized each group into two lists such that an item that appeared in the first half of one list appeared in the second half of the other list. We pseudorandomized the order of items within those lists to avoid more than two consecutive active sentences, more than two consecutive passive sentences, and more than two consecutive sentences from the same verb class. Each experimental sentence (critical or control) was followed by at least one filler sentence. We then created reversed versions of these two lists, resulting in four sentence lists per group (eight sentence lists in total).

Participants were instructed to rate the acceptability of each sentence based on their "gut reaction", and were told that there were no right or wrong answers (for the full instructions, see Appendix B). They rated sentences by moving a slider from "completely unacceptable" to "completely acceptable"; the location of the slider corresponded to an integer score between 0 and 100. This score was not made visible to participants. Participants could not rate a sentence with a score of 50 (the initial location of the slides): they had to move the slider at least slightly to the right or left on each trial such that the score was either lower or higher than 50. Before the experiment began, participants were familiarized with the experimental setup by rating two practice

[1] This work was originally reported in Leong and Linzen (2023).

Table 1
Example sentence frames. Each verb in the verb class was substituted into frames specific to the class (the table shows one of the five frames used for each class).

Verb class	Active sentence frame	Passive sentence frame
Advantage	The gift <u>0.2in</u> my organization.	My organization was <u>0.2in</u> by the gift.
Price	Your book <u>0.2in</u> thirty dollars.	Thirty dollars was <u>0.2in</u> by your book.
Ooze	My machine <u>0.2in</u> a sound.	A sound was <u>0.2in</u> by my machine.
Duration	Her speech <u>0.2in</u> seventeen minutes	Seventeen minutes was <u>0.2in</u> by her speech.
Estimation	Your friend <u>0.2in</u> my brother.	My sketch was <u>0.2in</u> by your friend.



How acceptable is this sentence?

That machine emanated a sound.

Fig. 1. An example of a trial in the human acceptability judgment experiment.

sentences, one of which was expected to be fully acceptable (*The mirrors reflected light*) and one expected to be unacceptable (*The teacher was spoke*). In one of the practice trials, participants were told that many people find the sentence acceptable, and if they agree they should move the slider to the right edge of the scale. The other practice trial was similar except the sentence was unacceptable. Fig. 1 shows an example of the interface that participants used to rate sentences.

Participants

We used the Prolific crowdsourcing platform to recruit 84 participants whose IP addresses were located in the US and who self-reported as native English speakers. Each participant rated 140 sentences (70 test sentences and 70 filler sentences) and was paid US$3.50. The experiment took an average of 12 min to complete.

Results

Participants were excluded from analysis if they rated more than 15 filler sentences differently than expected, either by giving ungrammatical sentences scores above 50 or giving grammatical sentences scores below 50. This resulted in the exclusion of 24 participants. In the analysis that follows, each sentence was rated by at least 13 participants.

Across all verb classes, participants gave higher scores on average to active sentences (mean: 88.5 points) than passive sentences (mean: 66.4 points). Participants were also in greater agreement with each other in their judgments for active than passive sentences (Fig. 2, left). The right panel of Fig. 2 illustrates this pattern with one particular pair of sentence frames: *Her speech <u>0.2in</u> seventeen minutes* and *Seventeen minutes was <u>0.2in</u> by her speech*. Active sentences with this frame received ratings that were close to maximal, regardless of the verb, while passive sentences showed more variance: judgments for *lasted* in this frame were unimodal and close to 0, judgments for *took* were spread relatively uniformly across the entire scale, and judgments for verb *required* were bimodal. We leave an investigation of the variability across participants to future work.

We next report differences in judgments between the active and passive sentences. In what follows, we focus on comparisons between the active and passive sentences within a sentence pair; because these sentences contained the same lexical items except for the auxiliary *was/were* and *by*, which are common across all sentences, comparing the sentences within a pair isolates the effect of passivization from any frequency effects that might increase the acceptability of sentences with more common verbs such as *helped* compared to low-frequency ones such as *profited*. We define the **passive drop** of a sentence pair as the difference in mean acceptability rating between its active and passive version. The results are reported in Fig. 3; a steeper downward slope corresponds with a larger average passive drop for sentences with the verb in question.

Although the average passive drop was positive for all verbs, its magnitude differed considerably across verb classes. The duration class showed the largest mean passive drop (61.9 points). The ooze class showed the lowest mean passive drop (8.4 points); in fact, although verbs from this class are considered in the linguistics literature to be unpassivizable, their mean passive drop was comparable to that observed for the agent-patient class of control verbs, which are considered to be passivizable (8.9 points). In summary, this experiment confirmed linguists' judgments for a majority of verbs, but not all of them.

To determine whether the difference in passive drop between verb classes was significant, we fit a linear mixed-effects model to predict sentence score from the following predictors: as fixed effects, sentence type and verb class as well as their interaction; frame and verb as random intercepts; and by-participant random slopes and intercepts for sentence type. We used the canonically passivizable agent-patient verb class as the reference level. We found significant interactions between sentence type and verb class in four cases: estimation verbs, price verbs, duration verbs and experiencer-theme verbs (all $p < 0.001$). This indicates that the passive drop for these verb classes was significantly different from the passive drop for the canonically passivizable agent-patient verbs. On the other hand, there was no significant interaction between sentence type and verb class in the sentence scores obtained from agent-patient verbs and ooze verbs ($p = 0.75$) or advantage verbs ($p = 0.1$).

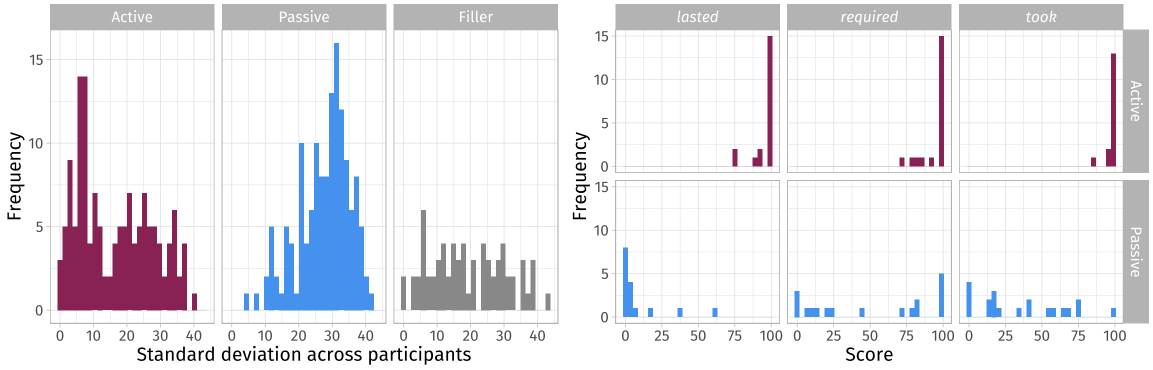


Fig. 2. *Distribution of human acceptability judgments*. Left: Histogram of standard deviations of acceptability ratings across participants. Active sentences received more consistent ratings than passive sentences. Right: Histogram of the ratings of the sentence pairs with the active frame *Her speech 0.2in seventeen minutes* and passive frame *Seventeen minutes was 0.2in by her speech*, for the three verbs *lasted*, *required* and *took*, illustrating differences in the spread of ratings across sentences.

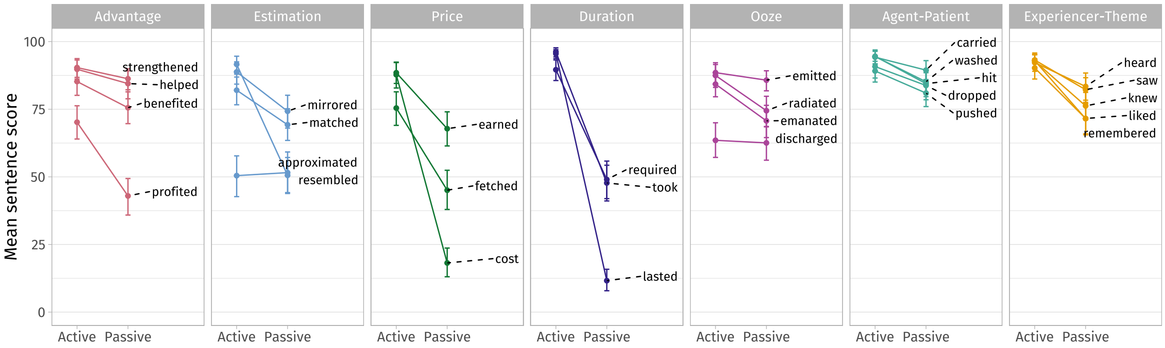


Fig. 3. *Passive drop in human acceptability judgments of active and passive sentences by verb*. The steeper the downward gradient between active and passive conditions, the larger the passive drop. Error bars indicate bootstrapped 95% confidence intervals.

Within the verb classes that were significantly less passivizable than agent-patient verbs, some verbs were more passivizable than others. To examine this variation, we fit separate linear mixed-effects models within each verb class, predicting SENTENCE SCORE from VERB and SENTENCE TYPE as fixed effects, with random intercepts for FRAME and by-participant random slopes for SENTENCE TYPE. We used the verb with the lowest passive drop in the class as the reference level. We found a significant interaction between SENTENCE TYPE and VERB in some but not all cases. Within the duration class, for example, *last* was significantly less passivizable than *required* ($p < 0.001$), but *took* was not ($p = 0.36$). Likewise, within the price class, *cost* was less passivizable than *earned* ($p < 0.001$) but *fetched* was not ($p = 0.26$). These results point to the fact that, even among verbs that can occur in the same frames, some verbs are more passivizable than others. This suggests either that verb-specific learning is necessarily (Zwicky, 1987), or that speakers are sensitive to fine-grained semantic differences across verbs in the same class.

Estimating judgment reliability across participants

How much of the variance in acceptability judgments across verbs can we hope to explain using a computational model, and how much of it is due to inherent variability in our measurements? To answer this question, we conducted a split-half reliability analysis, following Huang et al. (2024); this analysis is predicated on the premise that a computational model cannot be expected to show a higher correlation with the empirical data than two randomly sampled halves of the data are expected to show with each other.

We repeated the following analysis ten times, and averaged the ten sets of results. In each instance of the analysis, participants were randomly split into two groups. We computed the mean acceptability judgment score of each item within each half. We then calculated the Pearson correlation coefficient between the sets of estimates derived from each half of the results. The mean of these ten correlation coefficients was then entered into the Spearman-Brown prophecy formula (Spearman, 1910) to calculate the reliability coefficient. We found that the reliability of the collected acceptability judgment scores was high across all test items (0.93 on average). While reliability was somewhat lower for some verb classes (the lowest was 0.71, for agent-patient; for the full results by verb class, see Table 2), overall we conclude that participants' judgments were largely consistent for most items.

Table 2
Spearman-Brown-corrected split-half reliability for each verb class. Acceptability judgments showed high reliability on all test items as well as within verb classes.

Verb class	Split-half reliability
All items except fillers	0.93
Advantage	0.88
Estimation	0.91
Price	0.95
Duration	0.97
Ooze	0.84
Agent-Patient	0.71
Experiencer-Theme	0.83
Fillers	0.99

Discussion

Overall, we found that for the vast majority of verbs, including canonically passivizable agent-patient ones, participants rated active sentences more highly than passive ones (the only exception was *approximated*). This difference may reflect pragmatic factors: each sentence in the acceptability judgment task was presented to participants without any surrounding context. Because the passive construction is more pragmatically marked than the active, and requires more contextual support (Comrie, 1988), this setting might have caused participants to rate passive sentences as less acceptable than their active counterparts even in the control verb classes.

Notably, although experiencer-theme verbs are often thought of as passivizable, they showed a significant difference in passive drop from the other class of canonically passivizable verbs, the agent-patient verbs. Conversely, despite being reported as unpassivizable in the literature, the advantage and ooze verb classes did not differ in their passive drop from the canonically passivizable agent-patient class. As we showed in the previous section, these patterns are robust across participants. Overall, the pattern of gradient judgments that emerges from this experiment is considerably richer than that captured by the binary judgments from the linguistics articles that identified the verb classes in question, pointing to the value of formal acceptability judgment experiments for complex phenomena (Sprouse & Almeida, 2017).

In summary, Experiment 1A demonstrated that some verbs in the verb classes being tested are degraded in the passive voice, and that the degree of unacceptability was graded across verbs. For a model to adequately approximate the human pattern of behavior, then, it must capture the following patterns:

- **Class-level exceptionality**: Some verbs classes (e.g. duration verbs) exhibit passive drops that are significantly higher than the baseline passive drop expected of the canonically passivizable agent-patient verbs.
- **Verb-level exceptionality**: Some verbs within a verb class (e.g. last and cost) display passive drops that are significantly different from the passive drops of other verbs in the same class.
- **Gradience**: Acceptability is gradient rather than categorical. Verbs' passive drops span a broad range of values and do not obviously cluster into two groups (passivizable and unpassivizable).

Experiment 1B: Comparing language model and human judgments

In the previous section, we established that English speakers judge unpassivizability on a cline, rating some verbs such as *cost* as highly unpassivizable, and other verbs such as *pushed* as highly passivizable. In this section, we use a computational model to test if there is indirect evidence for the unpassivizability of these verbs. We do so by comparing the human judgments to those derived from a neural network language model. To adequately capture the indirect evidence available to humans, we train the language model on a corpus of 100 million words, comparable in size to the linguistic input available to English speakers by adolescence (Linzen, 2020; Warstadt et al., 2023; Wilcox et al., 2025).

What can we expect the results of this experiment to be? Some positive indications that exceptions can be learned from indirect evidence come from Bayesian modeling of the dative alternation (Perfors et al., 2010). Exceptions to the dative alternation in English follow a similar pattern to exceptions to passivization. Most ditransitive verbs can occur in both the double object construction (e.g. *Lucy gave Divya a bag*) and the prepositional dative construction (e.g. *Lucy gave a bag to Divya*), but not all verbs can appear in both constructions: *donate*, for example, can only occur in the prepositional object construction (e.g. *Lucy donated the car to Divya*, but **Lucy donated Divya the car*). Perfors et al. (2010) find that a hierarchical Bayesian model can learn to identify verbs that participate in this alternation after exposure to a subset of the CHILDES child-directed speech corpus (Brown, 1973; MacWhinney, 2000).

These results indicate that exceptions to an otherwise productive constraint can be learned from indirect evidence with a sufficiently powerful learning algorithm. However, it is unclear if such findings would extend to the language models we use in this work, which, like the majority of contemporary language models, are based on the transformer architecture (Vaswani et al., 2017): these models may not have strong enough inductive biases to implement the explicit inference procedure implemented by a Bayesian model, and as such might not be sufficiently sensitive to indirect evidence. Indeed, there is evidence that transformers sometimes *over-generalize*, for instance by translating English idioms like "kick the bucket" compositionally instead of treating such multi-word expressions as exceptions that should resist the compositional interpretation rule (Dankers et al., 2022). Even if neural network language models are sensitive to indirect evidence, their weaker biases may lead them to require much more data than humans to learn effectively; quite generally, neural network models are less data-efficient learners than humans (Warstadt & Bowman, 2022), and in practice are usually trained using large corpora that are vastly larger than the input available to humans (Frank, 2023; Linzen, 2020). It is thus unclear how similar to humans a neural network will be if it is trained on a corpus that approximates the amount of access to the passive that a human might have.

Model architecture

The models we trained were based on the transformer architecture as implemented in GPT-2 (Radford et al., 2019), specifically GPT-2 small, which has 117M parameters. We used the Adam optimizer (Kingma & Ba, 2015) with an initial learning rate of 0.0006. The maximum input length was 512 tokens and the batch size was 16. Models were trained for up to 50 epochs, with early stopping if validation loss did not decrease for three consecutive evaluation steps. The initial weights of the neural network and other aspects of training vary based on random seed, which could lead to different judgments on our test sentences. Such variability is particularly common for tests of linguistic generalization (e.g. McCoy et al. 2020). We thus trained five different models with a different random seed each.

We adopted the transformer architecture not only because it is the dominant language modeling architecture at the time of writing (e.g., Achiam et al. 2023, Touvron et al. 2023), but also because there is reason to believe, based on prior work, that transformers may be able to learn exceptions to passivization: the original GPT-2 models, albeit trained on a much larger corpus than our models, produced judgments that correlated well with human acceptability judgments of passive exceptions (Leong & Linzen, 2023). GPT-2 is also sensitive to more general restrictions on English passives; Warstadt et al. (2020) showed that GPT-2 assigns lower scores to passive sentences containing intransitive verbs (e.g. *Jeffrey's sons are smiled by Tina's supervisor*) than sentences containing transitive verbs (e.g. *Jeffrey's sons are insulted by Tina's supervisor*), demonstrating that the model is sensitive to the fact that intransitive verbs cannot be passivized in English. Finally, GPT-2 demonstrates sensitivity to exceptions to verb argument structure rules; for example, it can differentiate between verbs which do and do not participate in the dative alternation (Hawkins et al., 2020). That being said, since we train our models on substantially fewer words than were used to train OpenAI's GPT-2, it is an empirical question whether our models will behave similarly.

Training corpus

GPT-2 was trained on OpenAI's proprietary WebText corpus, which contains 40 GB of data from the web—approximately 8B words, assuming an encoding such as UTF-8 with one byte per English character and an average of 5 character per word. By contrast, English-speaking children are exposed to 2–7M words per year (Gilkerson et al., 2017), or 26M–91M words by the age of 13. As our goal is to determine what can be learned from the data available to humans, we trained our models using a significantly smaller training corpus than Radford et al. (2019). Rounding to the nearest order of magnitude, we trained our models on a corpus of 100M words. This corpus was a subset of OpenWebText (Gokaslan & Cohen, 2019), an open-source reproduction of the OpenAI Web Text corpus, which contains websites linked from Reddit with at least three upvotes. This selection method aims to choose a wide range of web text curated by humans. We recognize, of course, that this corpus is not entirely plausible as a representation of the input available to an English learner; future experiments using our methodology could use more cognitively plausible corpora as they become available (Wilcox et al., 2025).

Trigram model

To what extent can human passivizability judgments be reduced to a consequence of the frequency of particular sequences of words (for example, “was lasted by” is likely to be a very low-frequency sequence)? To address this question, we also trained a trigram word-level language model on the same 100M word corpus we used for our transformers. A trigram model predicts the upcoming word from the most recent two words, based on simple computations related to the frequency of sequences of three or fewer words in the corpus; unlike transformers, it does not construct a semantic representation of words. We used Kneser-Ney smoothing as implemented in KenLM (Heafield, 2011).

Procedure

We used a modified version of the targeted syntactic evaluation paradigm (Lau et al., 2017; Linzen et al., 2016; Marvin & Linzen, 2018) to derive acceptability judgments from our language models. Many studies in this paradigm derive binary judgments by comparing the probabilities assigned by the model to a minimal pair of sentences, and taking the sentence with the higher probability to be more acceptable. Here, to capture gradient judgments, we obtained the log-probability score for each sentence, defined as the sum of the log-transformed probabilities of all of the tokens in the sentence, and computed the passive drop of each sentence pair by subtracting the score of the active sentence from the score of the passive one. As in the human case, this procedure controls for the lexical effects of the open-class words in each sentence pair.

In addition to the models’ gradient passivization judgments, we also characterize the models’ broader syntactic competence by deriving binary acceptability judgments for a range of grammatical phenomena included in the Benchmark of Linguistic Minimal Pairs for English (BLiMP; Warstadt et al. 2020); these include, among others, subject–verb agreement, restrictions on the distribution of quantifiers like *at least*, and, most pertinently, argument structure restrictions such as the unpassivizability of intransitive verbs.

Materials

We compare our models’ passivizability judgments to human judgments for two sets of experimental materials. The first set of materials is the one we used in Experiment 1A, our human experiment above. The second is drawn from Ambridge et al. (2016). Ambridge and colleagues collected judgments from 20 human participants for active and passive sentences containing 475 verbs. Their study differed from ours in a number of ways. First, Ambridge and colleagues covered a wide range of verbs with a smaller number of participants per item. They also included a large number of verbs that they expected to be passivizable, as well as some phrasal verbs (e.g. *succumb to, look like*). This dataset thus provides a broader but potentially noisier sense of our models’ passivizability judgments.

Table 3
Spearman-Brown-corrected split-half reliability for each verb class in the Ambridge et al. (2016) human experiment. Human judgments in their experiment showed moderate reliability on all test items and moderate reliability within verb classes.

Verb class	Split-half reliability
All items	0.73
Agent-Patient	0.61
Theme-Experiencer	0.41
Experiencer-Theme	0.70
Other Passivizable	0.69
Non-Passivizable	0.92

To quantify the variability in the Ambridge et al. dataset that we can hope to account for with a computational model, we conducted a split-half reliability analysis using the same method we used for Experiment 1A. The results are reported in Table 3.

The reliability of human judgments in Ambridge et al. is considerably lower than in our Experiment 1A (see Table 2); consequently, there is a lower ceiling for the amount of variance that a computational model can account for.

Results

Overall syntactic competence

We first assess our models’ broad syntactic competence. Fig. 4 shows our five models’ performance on BLiMP, compared with OpenAI’s GPT-2 trained on the 40 GB WebText corpus, and the LSTM model trained by Gulordava et al. (2018) on 83 million words from the English Wikipedia (GPT-2 and LSTM results were obtained from Warstadt et al. 2020). Across a wide variety of syntactic and semantic phenomena, our models preferred grammatical sentences to ungrammatical ones an average of 76.7% of the time, suggesting that the sentence scores produced by our models are sensitive to syntactic and semantic constraints. The variability across the five models is fairly limited. Our models performed better than the Gulordava et al. (2018) LSTM model, although both types of model were trained on approximately the same amount of data. Finally, our models were only marginally worse than OpenAI’s GPT-2 despite being trained on 80 times less data.

Focusing on the subsets of BLiMP that test for models’ sensitivity to restrictions on passivization, we find that all five models were able to determine, with an accuracy substantially higher than chance, that intransitive verbs are less acceptable than transitive verbs in the passive voice (see the PASSIVE 1 and PASSIVE 2 rows in Fig. 4). They also showed a preference for animate over inanimate subjects in passive sentences (ANIMATE SUBJECT PASSIVE test in Fig. 4). Overall, we conclude that our models gleaned from their training data considerable information about English grammar broadly and, more specifically, about the broad generalizations about the environments in which the passive construction is acceptable.

We do not report tests of the trigram model’s syntactic competence: Many of the phenomena in the BLiMP dataset span more than three words and as such this model, which only has access to a window of three words, has no hope of capturing them; in fact, Warstadt et al. (2020) show that even a 5-gram model is unable to capture most of the phenomena in BLiMP.

Comparison to experiment 1A

We next compare the models’ judgments for our test sentences to those of humans. Fig. 5 graphs the models’ average passive drop for each verb against the passive drop observed in our human experiment. We found a high Pearson’s correlation coefficient of $r = 0.91$ between the human and model passive drop at the verb level. At the individual

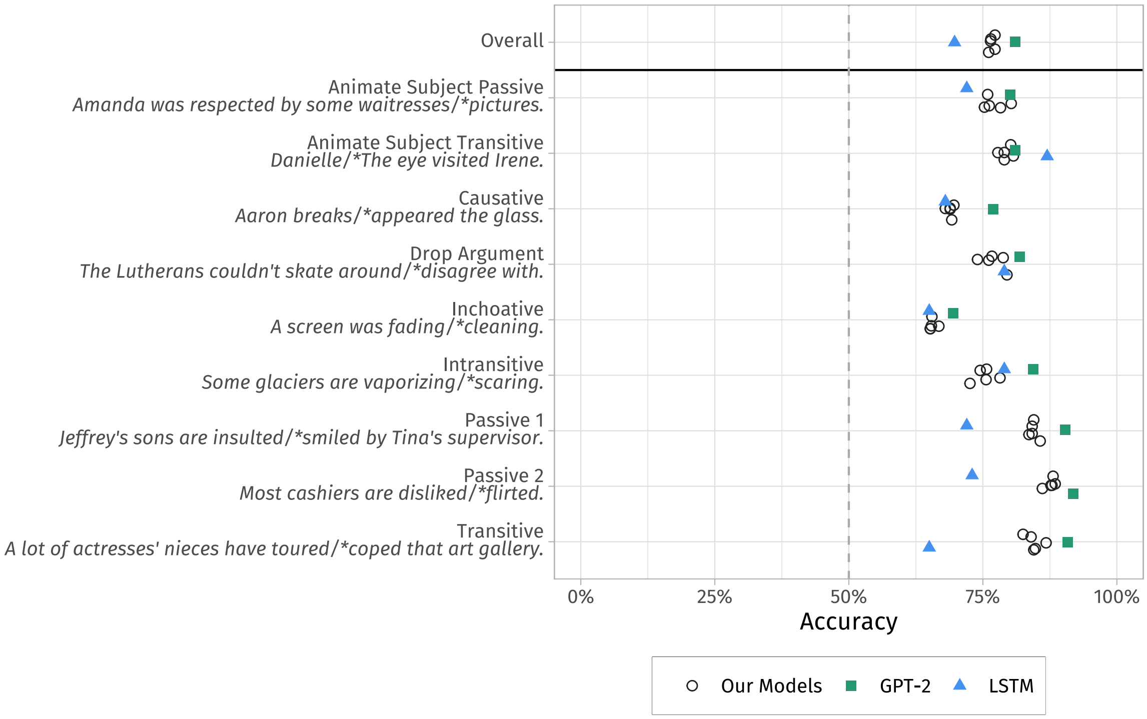


Fig. 4. *Accuracy of our models' acceptability judgments on the Benchmark of Linguistic Minimal Pairs (BLiMP, Warstadt et al. 2020)*. We report overall accuracy over the entire benchmark, as well as detailed accuracy on the subset of constructions in BLiMP that are related to argument structure. Our models perform better than the Gulordava et al. (2018) LSTM model and marginally worse than OpenAI's GPT-2, which was trained on considerably more data. The dashed line indicates chance-level accuracy.

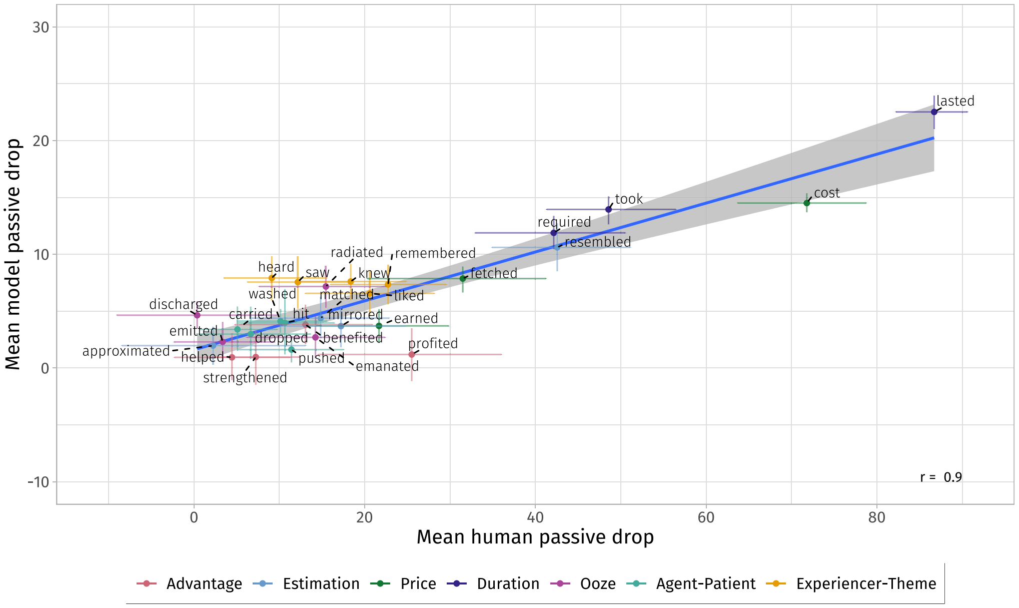


Fig. 5. *Passive drop in humans (Experiment 1A) vs. neural network language models*. At the verb level, the correlation between human and model judgments is $r = 0.9$. Each point represents the average passive drop of a verb in five sentence frames scored by five models. Horizontal error bars indicate bootstrapped 95% confidence intervals over human judgments for sentence frames; vertical error bars indicate bootstrapped 95% confidence intervals over model judgments for sentence frames. All error bars are corrected for between-participant and between-model variance (Bakeman & Mcarthur, 1996).

item level, the correlation was $r = 0.65$; while this correlation is still fairly high, it remains below the upper bound of 0.93 derived from the split-half reliability analysis (Table 2). The models matched humans' judgments of **exceptionality** within verb classes: among verbs with similar meanings, the same verbs displayed high passive drops for across humans and models. For instance, for both humans and models, the passive drops of *earned* and *discharged* were low compared to other verbs in their respective classes. Likewise, like humans, our models predicted high passive drops for *lasted*, *resembled* and *cost*, compared to other verbs in the respective classes. Finally, our models also largely matched human judgments of **gradience**: the sentence scores obtained from our models predict not only low and high passive drops, but also intermediate levels of passive drop in verbs such as *took* and *required*.

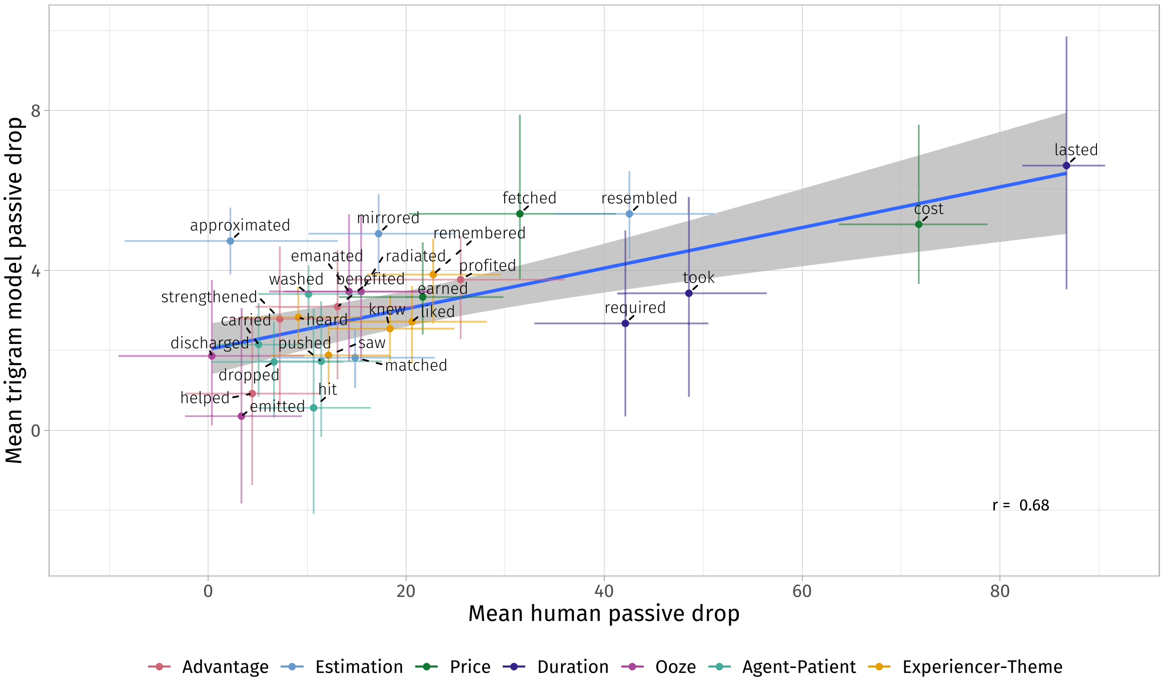


Fig. 6. Passive drop in humans (Experiment 1A) vs. passive drop predicted by the trigram model estimated from the training corpus.

Moving to the trigram model, we found a moderate-to-strong linear correlation with human passive drop at the verb level ($r = 0.68$; Fig. 6) and a moderate correlation at the item level ($r = 0.34$). These results suggest that although some information about the passivizability of the verbs in our materials is recoverable from the frequency of short sequences of words, transformer models are able to capture more information about passivizability than can be attributed to trigram frequency alone. We attribute the moderately high performance of the trigram model to the relatively simple structure of the materials from the human experiments, which included low-probability trigrams such as *was lasted by*; if the materials were modified to introduce additional words around the verb, for example *was previously lasted in total by*, judgments derived from the trigram model would likely no longer be able to capture the unpassivizability of the verb.

A data contamination analysis found that one of our active test sentences, *The journey lasted three days*, occurred verbatim in the training dataset, which would have inflated the probability assigned to the sequence and thereby the gap between the active and passive versions of this particular sentence. Because no other items were present in the training data, the effect of data contamination on our results is minimal.

Exceptions to passivization: comparison to Ambridge et al. (2016)

Fig. 7 shows the transformers' average passive drop for each verb against the passive drop calculated based on human scores reported by Ambridge et al. (2016). While there was a positive linear correlation between human and model passive drop ($r = 0.42$ at the verb level, $r = 0.19$ at the item level), this correlation is substantially weaker than the correlation we observed for our stimuli from Experiment 1A. Qualitatively, the models predicted poorly the extent to which different verbs resist passivization: they overpredicted the unpassivizability of some verbs, such as *serve* and *fit*, and underpredicted the unpassivizability of others (e.g. *last*, *lack*).

We should emphasize that the correlation coefficients cannot be directly compared across Experiment 1A and Ambridge et al. (2016), because, as we described above, the human judgments from Ambridge et al. (2016) are much more variable, which leads to a lower upper bound on any model's potential correlations (this variability is reflected in the width of the horizontal error bars in Fig. 7). That being said, the gap between the empirical correlation at the item level and the ceiling reflected by the split-half reliability is larger for the Ambridge et al. (2016) materials (0.20 vs. 0.75) than for Experiment 1A (0.64 vs. 0.93).

The trigram model performed much worse than the transformers on the data provided by Ambridge et al. (2016), with a *negative* linear correlation between human and model passive drop at the verb level ($r = -0.27$). We hypothesize that this behavior arose from the use of proper nouns (e.g. *Marge*, *Wendy*) in the test sentences: approximately 32% of the tokens in this dataset were out of vocabulary for the trigram model. Such data sparsity-related issues, in combination with the short length of the test sentences (e.g. *Bob eluded Wendy*), likely affected the trigram model's accuracy.

Discussion

Broadly, transformer language models captured human patterns of verb passivizability well when evaluated on the test items from Experiment 1A. Qualitatively, our models showed gradient judgments that varied across individual verbs in each verb class as well as across verb classes. Quantitatively, the transformers' judgments were strongly correlated with human judgments ($r = 0.91$). We note, however, that the highly passivizable verbs in our test set were all relatively frequent, and likely occurred in the passive a substantial number of times in the training corpus. As such, none of them posed the problem we referred to in the introduction as Baker's paradox, which would be posed by an infrequent transitive verb such as *defenestrate*.

In comparison with the transformers, judgments based on a simpler frequency-based trigram language were only moderately correlated with human judgments. In other words, the trigram model was unable to predict the human results of Experiment 1A as effectively as the transformers, even though our test sentences were structurally simple, and did not have adverbs that intervened between the auxiliary and the verb, such that 'was/were [VERBed] by' was a contiguous sequence.

Evaluating our models against the materials of Ambridge et al. (2016), in contrast, showed a discrepancy between our language models' and humans' passivizability judgments that was not present in our data. Although positive, the correlation between model and human judgments of passivizability was weak, and our models predicted the

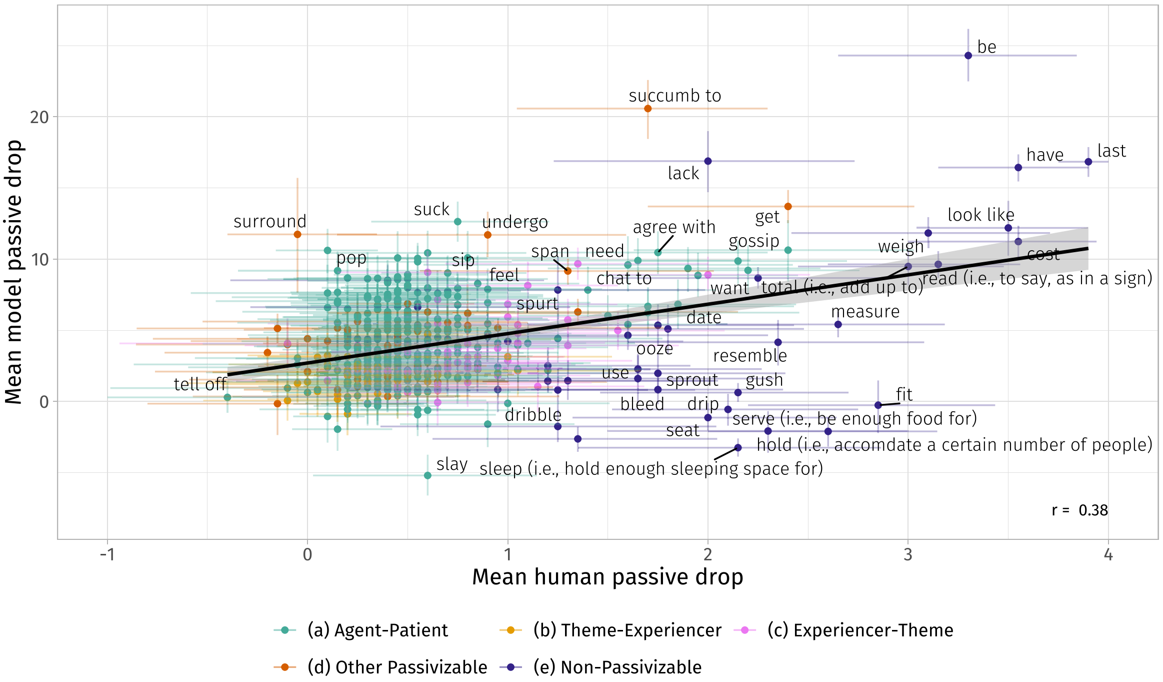


Fig. 7. *Passive drop in humans from Ambridge et al. (2016) vs. neural network language models.*

passive drop of non-passivizable verbs more poorly than verbs in other classes. Some of this difference in performance can be attributed to the large variance in human judgments. However, a further study with higher power comparing human and model judgments on more verbs might point to divergences between model and human judgments.

Experiment 2: What indirect evidence do models use to learn restrictions on passivization?

In the previous sections, we showed that 100 million words of English text provide sufficient evidence for a transformer language model to produce judgments of passive exceptions that align to a significant extent, though not fully, with those of humans. In the following sections, we turn to our second research question: which aspects of the linguistic input serve as evidence for models to learn these patterns? To answer this question, we take inspiration from work that has used controlled interventions on a model's training corpus to draw causal links between aspects of the training data and the model's behavior (Jumelet et al., 2021; Misra & Mahowald, 2024; Patil et al., 2024; Wei et al., 2021). This approach manipulates particular sources of evidence in the training corpus and compares models trained on the original dataset with models trained on the modified dataset; for example, to assess the causal effect of verb frequency on a language model's subject–verb agreement prediction accuracy, Wei et al. (2021) removed an increasing number of occurrences of the verb from the corpus, retrained the model, and measured the resulting change in its accuracy on this task.

We apply this method to test two hypotheses proposed in the literature as to the evidence that supports humans' acquisition of passive exceptions: the *entrenchment hypothesis* (Braine & Brooks, 1995; Demuth, 2011; Theakston, 2004) and the *affectedness hypothesis* (Ambridge et al., 2016; Darmasetiyawan et al., 2022; Messenger et al., 2012; Pinker, 1989). We tested each hypothesis by first altering or removing elements of the corpus that are crucial for learning under each hypothesis and retraining models on these modified corpora. If models trained on a modified corpus consistently differ from models trained on the original corpus in their acceptability judgments on passive exceptions, then we can attribute the change in behavior to the particular intervention that we made on the training data.

Experiment 2A: Testing the entrenchment hypothesis

The first hypothesis we tested is the *entrenchment hypothesis* (Ambridge et al., 2015; Regier & Gahl, 2004; Theakston, 2004). According to this hypothesis, learners track the frequency of verbs in particular constructions (Gordon & Chafetz, 1990), and if that verb never appears in a particular context but does appear with substantial frequency in other contexts, they conclude the verb is ungrammatical in that context. In the passivization case, the relevant factor is the relative frequency of the active compared to the passive: as learners are exposed to more and more occurrences of a verb in the active but not in the passive, they gradually conclude that the passive form is not just rare but ungrammatical. While the classic entrenchment hypothesis assumes the unacceptable form has to be completely absent from the corpus, here we test a softer, gradient version of the hypothesis, where different degrees of frequency asymmetries between the active and the passive forms of a verb result in gradient levels of acceptability for the passive form.

To test this hypothesis, we identified verbs which appeared relatively frequently in the active and highly infrequently in the passive. We chose the verbs *last*, *cost* and *resemble* as our TARGET VERBS, that is, the verbs whose relative frequencies we would try to emulate. All three verbs were highly unpassivizable, and had large *active–passive ratios* (henceforth A/P ratio); they occurred much more frequently in the active compared to the passive. We then chose MUTATING VERBS, whose frequency we would modify in the training corpus. These verbs were drawn from the highly passivizable agent-patient class. We computed the A/P ratio of each target verb, then matched the mutating verb's A/P ratio to the target verb's A/P ratio by removing as many passive sentences with that verb as necessary (and in some cases removing a small number of active sentences to match the ratio more closely).

The main motivation for this design is as follows: if A/P ratio serves as a cue for passivizability, we expect this intervention to cause the originally passivizable MUTATING VERB to decrease in passivizability. In fact, if the *only* cue to a verb's passivizability is its relative frequency of occurrence in the active voice compared to the passive, then we expect the MUTATING VERB would become just as unpassivizable as the TARGET VERB.

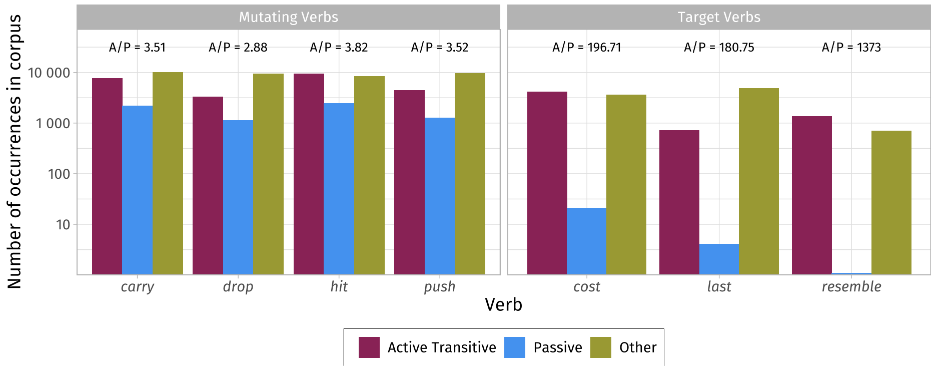


Fig. 8. *Frequency of occurrence of mutating and target verbs in the original corpus.* We include sentences that were parsed as transitive and active in our count of active sentences, and sentences parsed as passive in our count of passive sentences. All other sentences were labeled other.

Estimating corpus frequencies

We used spaCy's `en_core_web_trf` pipeline (Honnibal et al., 2020) to obtain dependency parses for each sentence in the training corpus. We then counted the number of times our mutating and target verbs were used in transitive active sentences and passive sentences. Specifically, sentences where the verb had a dependency edge to a passive auxiliary (`auxpass`), a passive nominal subject (`nsubjpass`) or a passive clausal subject (`csubjpass`) were classified as passive sentences, while sentences with a direct object (`dobj`) or a clausal complement (`ccomp`) dependency edge from the verb were classified as active sentences. All other sentences were classified as other. The results of this analysis are shown in Fig. 8.

The active category only included sentences where the verb appeared in an unambiguously transitive frame, since such sentences expressed events that could also be expressed in the passive voice. Sentences with other syntactic structures, or that the filter failed to identify as active transitive, such as those in (7) for *drop*, were classified as other and excluded from analysis:

(7) a. Realtors believe home resales, which dropped in September, peaked in July and August.

b. I apologize for that pun, but its definitely not worse than the ones Arnold Schwarzenegger drops.

This filtering mechanism prioritized precision over recall; that is, we prioritized the accuracy of the classification of a sentence as active or passive, potentially at the expense of classifying a larger number sentences as other. To assess the success of this strategy for the passive class, we used a regular expression to extract from the corpus 500 strings with auxiliaries such as *had been* and *was* followed by a past participle within the same paragraph. We found that only 24 (or 4.9%) of these strings were not classified as passives by our filter (binomial 95% confidence interval: 3.1–6.8%), indicating that recall was only minimally affected. We additionally verified the precision of the filtering mechanism by hand-checking 500 sentences that were labeled as passive sentences. We found that only 10 out of the 500 sentences (2.0%) were misidentified as passive (binomial 95% confidence interval: 0.9–3.3%), indicating high precision. Examples of misparsed passive sentences for the verb *last* are given in Appendix C.

Does relative corpus frequency predict human judgments?

Before we discuss the corpus intervention experiment, we test if a verb's A/P ratio on its own can predict human passivizability judgments, as predicted by the entrenchment hypothesis. We find only partial support for this hypothesis (Fig. 8). Consistent with the hypothesis, all four mutating verbs, which had relatively low A/P ratios (2.88–3.82), were judged as highly passivizable, and all three target verbs, which had high A/P ratios, also displayed high passive drops in the human experiments. At the same time, the target verb with the highest A/P ratio, *resemble*—which we expect to have the largest passive drop if corpus frequency alone drives passivizability—was instead judged by both our models and human participants as more passivizable than the other two target verbs with lower A/P ratios.

Moving to the full set of verbs, we observe a moderate positive correlation ($r = 0.62$) between log-transformed A/P ratio and human passive drop (Fig. 9). This correlation is similar to the one we observed between the human judgments and those derived from the trigram model, which similarly reflects count-based information. The substantial gap between this correlation coefficient and the one obtained by the transformers ($r = 0.90$) suggests that frequency asymmetries alone do not fully account for transformers' ability to predict human judgments; in particular, frequency does not explain patterns at the verb class level—it underpredicts the passive drop associated with all three duration verbs while overestimating the passive drop associated with advantage verbs.

Corpus intervention procedure

We created modified corpora for all combinations of the four mutating and three target verbs (a total of 12 verb pairs), as follows. In each corpus, we used the A/P ratio of the target verb A/P_{target} to obtain the corresponding number of occurrences of the mutating verb in the active and passive that should be kept in the modified corpus so that, keeping as many active sentences as possible, we had $A/P_{mutating} \approx A/P_{target}$. For example, in the original corpus the mutating verb *drop* occurred 3279 times in transitive active sentences and 1146 times in the passive ($A/P_{drop} = 2.88$), and the target verb *last* occurred 723 times in transitive active sentences and four times in the passive ($A/P_{last} = 180.75$). To match the A/P ratios of the two verbs, we randomly chose 3253 active occurrences and 18 passive occurrences of *drop* in the training corpus and discarded all other active and passive occurrences of *drop*, such that $A/P_{drop} \approx 180.75$ in the modified corpus. Fig. 10 illustrates the distribution of verbs in the training corpora before and after we performed the intervention for this particular verb pair (*last* and *drop*). Raw counts for each of the 12 verb pairs are available in Appendix D.

We trained five models, each with a different random seed, on each of the 12 modified corpora, using the same training procedure outlined in Experiment 1B. We then obtained acceptability judgments from these models and compared them to the judgments of the models trained on the original corpus in Experiment 1B.

If a verb's A/P ratio significantly affects its passivizability, then we expect the mutating verb to be judged as less passivizable (i.e. be given a higher passive drop) by models trained on the modified corpora compared to models trained on the original corpus. We expect the passive drops of all verbs other than the mutating verb to remain the same.

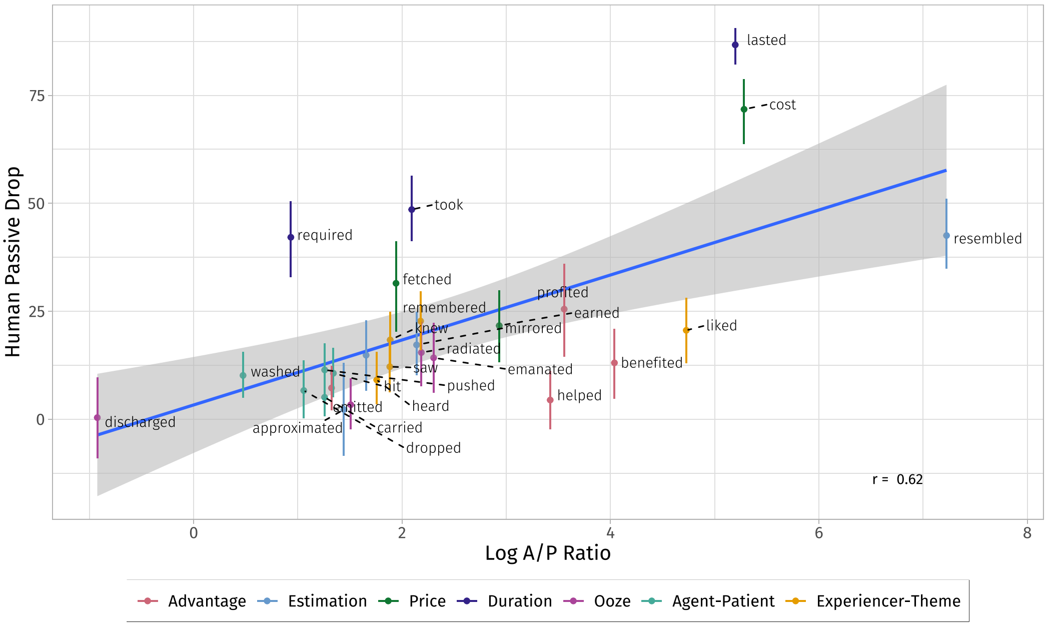


Fig. 9. Correlation between a verb's relative corpus frequency in the active vs. passive and human passive drop.

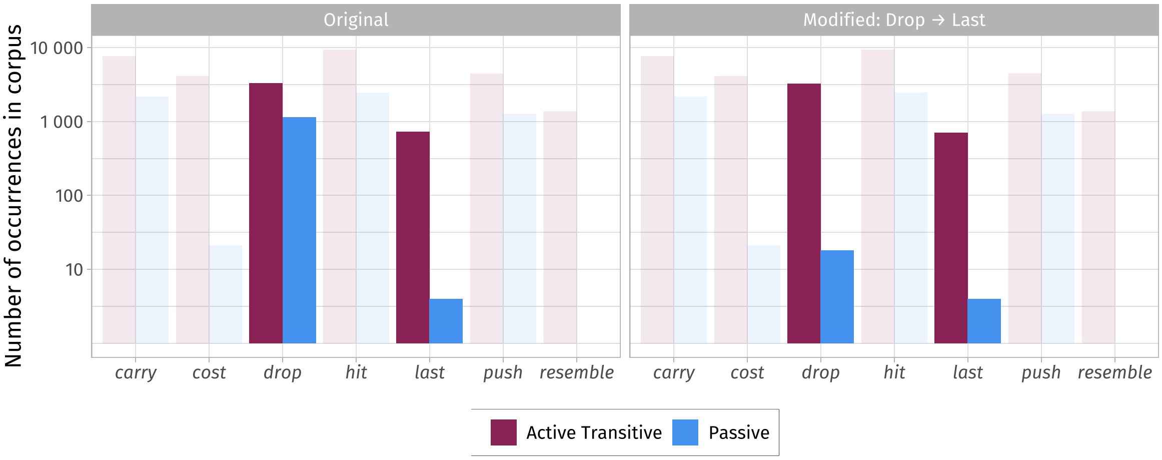


Fig. 10. *Corpus statistics before and after the intervention used for Experiment 2A. In this example, the mutating verb is drop and the target verb is last*. The frequency of the mutating verb *drop* is decreased in both the active and passive to match the relative frequency of the target verb *last*. All other verbs do not undergo any change.

Results

The results of Experiment 2A are shown in Fig. 11. Across all target and mutating verb pairs, the mutating verb showed an increase in passive drop as a result of the intervention. To test whether the intervention resulted in a larger increase in the passive drop of the mutating verb compared to any overall difference in passive drop across all verbs, we fit two linear mixed-effects models: one predicting PASSIVE DROP as a function of TRAINING CORPUS, with by-MODEL, by-VERB, by-VERB CLASS and by-FRAME random intercepts; and another model that included all of these predictors as well as an additional fixed effect indicating whether the verb is MUTATING in the corpus. A likelihood-ratio test indicated that mutating verbs showed increases in passive drop not shown by other verbs ($\chi^2(1) = 141.2$,p< 0.001).

The intervention did not significantly affect the passive drop for verbs that did not undergo mutation (black and purple lines in Fig. 11). To determine that, we fit two linear mixed-effects models for the non-mutating verbs: one predicting PASSIVE DROP as a function of VERB and VERB CLASS, with random intercepts for FRAME and SEED; and another model that further included a fixed effect indicating whether the corpus was modified. A likelihood-ratio test indicated that there was no significant effect of modifying the corpus ($\chi^2(1) = 0.52, p = 0.478$).

Although each mutating verb's passive drop increased when it was mutated, in the majority of cases the intervention did not cause the mutating verb to become as unpassivizable as the target verb. Across all verb pairs, the mutating verbs in all twelve conditions showed an increase in passive drop of 2 to 4 points after intervention, regardless of the passive drop of the target verb. In the one case where the two verbs' passive drops converged after the intervention (for the mutating verb *hit* and the target verb *resemble*), this was because the target verb's passive drop unexpectedly decreased. Overall, while the intervention reliably increased the mutating verb's passive drop, in general it did not fully close the gap between the mutating verb and the target verb.

Discussion

In this experiment, we found that increasing the A/P ratio of a verb consistently led language models to judge the verb as less passivizable.

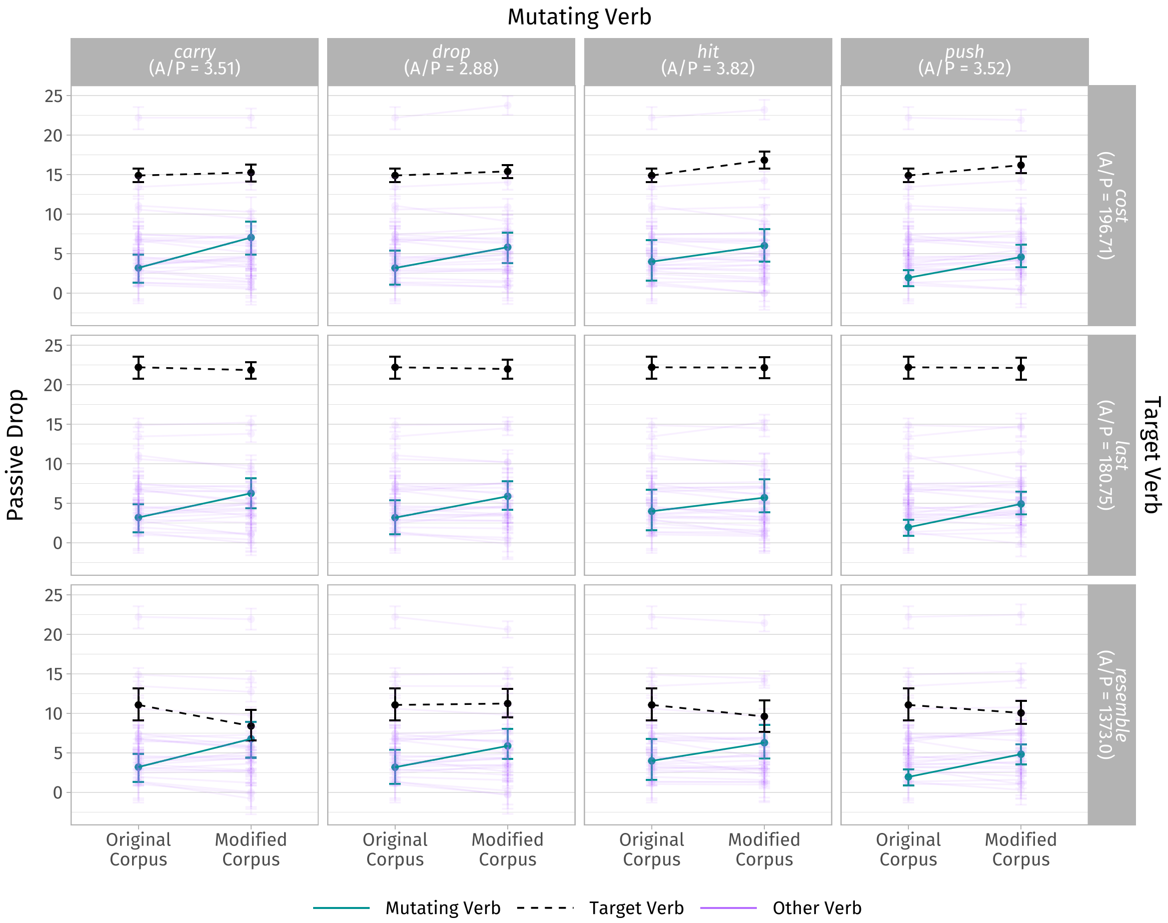


Fig. 11. *Change in passive drop as a result of training on a modified corpus with a higher active-to-passive frequency ratio for the mutating verb.* Passive drop of mutating verbs (in green) increases when their distribution is modified, but only reaches the same level as the target verb when that verb is *resemble*. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

This suggests that the A/P ratio is one potential source of evidence by which models learn whether a verb is or is not passivizable. This is most likely not models' *only* source of evidence for passivizability, however: if it were, we would expect each mutating verb to behave exactly like the target verb after our intervention. That was not the case: although the mutating verbs occurred in the same A/P ratio as the target verbs, the passive drop of these mutating verbs did not increase to the level of unpassivizability of their respective target verbs. This source of evidence is also most likely not the only one that humans use to make passivizability judgments: if it were, we would expect corpus frequencies to be just as predictive of human passive drop as were the transformers' predictions, which was not the case (Section “Does relative corpus frequency predict human judgments”).

We note that although we reduced the number of passive sentences that our mutating verbs appeared in, we did not eliminate all instances of the passive—each verb appeared at least once in the passive in each training corpus (see Appendix D for details). This meant that we tested a weaker version of the entrenchment hypothesis: while the model was exposed to a smaller number of passives of the mutating verb than in the original corpus, this number was not zero. The reason we matched the A/P ratio of an existing verb was to test the hypothesis that doing so would lead to convergence between the passive drop of the mutating verb and the target verb. We also reasoned that this weaker version of the entrenchment hypothesis mirrors natural language acquisition more faithfully: speech errors and misparses may well lead a learner to conclude that a particular verb was observed in the passive even when the sentence was not intended by the speaker to be a passive one (indeed, some of the sentences that our parser classified as passives may have been misparses). In the novel verb experiment below (Experiment 3) we test a stronger version of the entrenchment hypothesis, in that the modified corpora contain no passive occurrences at all.

By removing many or most of the occurrences of the mutating verb in the passive forms, we naturally also reduced the total number of passives in the corpus, as well as the total number of words in the corpus. While in principle either of these factors could confound our results, we suspect that the effect is negligible: we eliminated a few thousand sentences out of many millions of sentences in the original corpus, of which hundreds of thousands were passive sentences. Indeed, empirically we do not see changes in the passive drops of verbs other than the mutating verb, suggesting that the manipulation did not materially affect the models' ability to learn the passive. That being said, we note this as a potential limitation of the filtering methodology we use; in cases where filtering is more aggressive, it may be appropriate to replace the sentences that were filtered with new sentences from the corpus that are not relevant to the hypothesis, so that the total number of sentences is matched between the original and modified corpus (Misra & Mahowald, 2024).

Experiment 2B: Testing the affectedness hypothesis

We next consider the hypothesis, proposed by Pinker (1989, 1987), that the passivizability of a verb can be predicted from its lexical semantics, specifically the *affectedness* (or lack thereof) of the theme argument of the verb. In the sentence *The apple was eaten by the boy*, for example, the by-object *the boy* can be said to *affect* the subject of the passive (e.g. *the apple*), in that it causes a change of state, location, or existence to the subject. Support for this hypothesis comes, for instance, from Ambridge et al. (2016), where participants rated verbs on a series of proxies for affectedness in sentences where the arguments were not explicitly specified; for the sentence *A likes B*, for example, participants might be asked to rate the statements *A is responsible*, or *A is doing something to B*. By aggregating these proxies, Ambridge et al. computed a measure of the verb's prototypical affectedness (that is, its affectedness with the prototypical arguments inferred by participants when arguments are not explicitly mentioned). They found a positive correlation between a verb's prototypical affectedness and participants' acceptability judgments for sentences where it was used in the passive. Similar effects of affectedness on passivizability have been documented across languages (Ambridge et al., 2023; Aryawibawa & Ambridge, 2018; Bidgood et al., 2020; Darmasetiyawan et al., 2022; Liu & Ambridge, 2021).

Experiment 2B tests the hypothesis that affectedness causally explains language models' passivizability judgments. We modify the training corpus by intervening on the same pairs of verbs as in Experiment 2A, but in contrast with Experiment 2A, here we use the highly-unpassivizable verb as the MUTATING VERB, and the highly-passivizable verb as the TARGET VERB.

In transformer language models, as in most neural networks, lexical semantics is encoded by word vector representations (embeddings) that reflect the contexts in which the word appeared in the training corpus. Because the individual dimensions of this vector are not interpretable, it is not straightforward to modulate the degree of affectedness of a verb by intervening on its embedding. Instead, we aim to shift the semantic representation of the mutating verb in a more or less affected direction by changing the contexts in which this verb appeared. We do so by placing the mutating verb in active sentences that originally contained the target verb; this allows the mutating verb, which is in general highly unpassivizable, to co-occur in the active with the agent-like subjects and patient-like objects that are normally associated with the target verb (see examples in the next section). Crucially, we only intervened on active sentences: we did not manipulate the passive sentences containing the mutating verb, or add new passive sentences to the corpus.

Procedure

For each pair of mutating and target verbs, we randomly selected a portion of active transitive sentences (identified using the procedure outlined in Section "Estimating corpus frequencies") in the training corpus containing the target verb and replaced the target verb with the mutating verb (e.g. *dropping* → *lasting*; *dropped* → *lasted*), thus making the mutating verb appear in contexts that previously contained the target verb while making minimal changes to the syntactic environments in which the verb occurs. Examples of sentences that underwent intervention for the experiment where the mutating verb was *last* and the target verb was *drop* are given in (8):

(8) a. You know, people are always ~~dropping~~ <u>lasting</u> off samples of gluten-free products at our office.

b. The pilot, worried the bomb might break loose from the damaged plane, ~~dropped~~ <u>lasted</u> it into the water outside of Savannah, Ga. near Wassaw Sound.

Placing the mutating verb in environments that previously contained the target verb caused the mutating verb to co-occur with some of the subjects and objects that previously co-occurred with the target verb, which in turn caused the distribution of its arguments to more closely resemble that of the target verb. Since we only accounted for syntax, phrasal verbs and idiomatic expressions were also affected by this process, as illustrated in (8a).

Note that unlike in Experiment 2A, in which our intervention changed the distribution of the mutating verb but not the target verb in the training corpus, the intervention in Experiment 2B affected the both target and mutating verbs. The mutating verb appeared in the sentences it originally occurred in as well as the newly mutated sentences, while the target verb was seen less frequently in the corpus, since some of its occurrences were mutated. For this reason, the target verb's passive drop after the intervention may not be representative of its passive drop before the intervention.

The entrenchment hypothesis, for which we found support in Experiment 2A, predicts that increasing the A/P ratio of a verb would make the mutating verb less passivizable. To counteract this, we removed the same number of active transitive sentences that originally contained the mutating verb. Thus, in both corpora, the mutating verb appears the same number of times in active transitive sentences, but in different contexts.

We created two such modified corpora for each pair of mutating and target verbs (12 pairs in total), varying the proportion of a verb's occurrences that were altered. In one corpus, we replaced a moderate amount (30%) of the active transitive sentences that originally contained the mutating verb. In the other, we replaced a large proportion (70%) of the active transitive sentences containing the mutating verb. Our particular choice of how many sentences to replace (either 30% or 70%) was arbitrary, and future work could consider modulating or interpolating this proportion, and thereby modulating the similarity between the distribution of the mutating verb and target verb. We trained five models on each modified corpus, and obtained acceptability judgments from those models following the procedure outlined in Experiment 1B.

Results

The results of Experiment 2B are shown in Fig. 12. When 30% of the mutating verb was replaced, the passive drop of verbs with no change across training corpora ("other" verbs, purple lines in Fig. 12) decreased by a mean of 0.19 points, while the passive drop of mutating verbs decreased by a mean of 1.25 points and the passive drop of target verbs increased by 0.38 points. When 70% of the mutating verb was replaced, the passive drop of 'other' verbs decreased by a mean of 0.16 points, while the passive drop of mutating verbs decreased by a mean of 2.53 points, and the passive drop of target verbs decreased by a mean of 0.06 points.

Non-mutating verbs did not show a significant change in passive drop across training corpora. To verify this, we compared two linear mixed-effects models predicting the passive drop of non-mutating verbs. The full model used CORPUS TYPE, VERB, and VERB CLASS as main effects and random intercepts for FRAME and MODEL SEED, while the reduced model did not include CORPUS TYPE as a fixed effect. A likelihood ratio test showed that including corpus type did not significantly improve the fit, $\chi^2(1) = 2.50, p = 0.11$ in the 30% replacement scenario, and $\chi^2(1) = 2.15, p = 0.14$ in the 70% replacement scenario.

We next turn to the significance of our intervention on the passive drop of mutating verbs. We used as a reduced model a linear-mixed effects model with CORPUS TYPE as a fixed effect. We compared the reduced model to a full model that also used whether the verb was a MUTATING verb as a fixed effect. Both models included VERB, VERB CLASS, FRAME, and MODEL SEED as random effects. In the 30% replacement condition, a likelihood ratio test showed that the full model was a significantly better fit, $\chi^2(1) = 7.14, p = 0.008$, suggesting that argument affectedness does affect model judgments of passivizability. We found similar results in the 70% condition, $\chi^2(1) = 70.0, p < 0.001$.

Unlike in Experiment 2A, where a proportionate change in the frequency of occurrence affected every mutating verb's passive drop similarly, in Experiment 2B we found that altering a verb's distribution did not always result in a decrease in passive drop: intervening on the verbs *resemble* and *cost* consistently reduced passive drop, while intervening on *last* did not. To test whether verb identity significantly moderated the effect of our intervention, we compared two mixed-effects models. Both models included main effects of CORPUS TYPE, VERB,

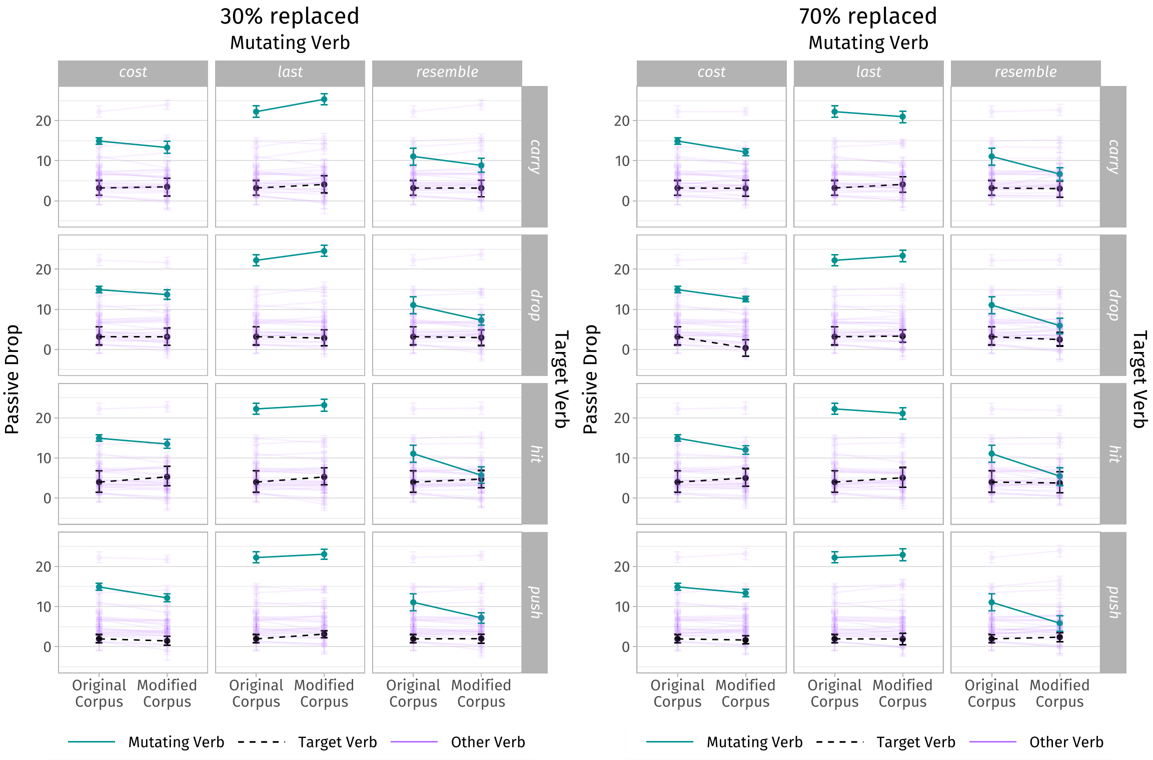


Fig. 12. *Change in passive drop as a result of training on data with target verb-like arguments*. Mutating verbs which were altered to appear with the arguments of target verbs in the training data did not show a consistent decrease in passive drop.

whether the verb was MUTATED, and manipulation PERCENTAGE, as well as random intercepts for each FRAME and MODEL SEED. The full model added an interaction term between mutation status and verb identity (MUTATING × VERB), allowing different verbs to show different responses when their distributions were altered. A likelihood ratio test revealed that the full model provided a significantly better fit to the data than the reduced model, $\chi^2(2) = 47.00, p < 0.001$. This indicates that verb identity significantly moderated the effect of the manipulation.

Was there a difference between intervening on a moderate or large proportion of a verb's occurrences? We investigated this question by comparing the passive drop assigned to verbs at 30% and 70% replacement rates. We compared a reduced linear mixed-effects model with whether the verb was MUTATING as a fixed effect against a full model that additionally included the PERCENTAGE of occurrences altered as a discrete variable. Both models included VERB, VERB CLASS, FRAME, and MODEL SEED as random effects. A likelihood ratio test showed that the full model did not provide a significantly better fit to the data, $\chi^2(1) = 0.52, p = 0.47$, suggesting that passive drop was not significantly affected by increasing how extensive our intervention was.

Discussion

The goal of Experiment 2B was to test if the semantics of a verb has a causal effect on language models' passivizability judgments for the verb. We manipulated distributional cues to the semantics of the verb: we placed the mutating verb in sentences with arguments that occur with agent-patient verbs, and are consequently more likely to be affected. We found a significant main effect of our intervention on our mutating verbs' passivizability, suggesting that affectedness impacts our model's judgments of verb passivizability. We additionally found that changing how many of a verb's active transitive occurrences we intervened on did not significantly affect the verb's change in passivizability.

While we found a significant main effect of affectedness, we found that the magnitude of the effect interacted with the mutating verb's identity, with *cost* and *resemble* demonstrating consistent decreases in passive drop, while *last* did not. One possible account for why *last* behaved differently from *cost* and *resemble* arises from the original distributional statistics of the verbs. Whereas the majority of the total verbal occurrences of *cost* (53.3%) and *resemble* (66.3%) were classified as active transitive in the training corpus, the same is not true for *last*; only 13.0% of the sentences that *last* appeared in were classified as active transitive, while 86.9% of the occurrences of *last* were classified as OTHER. Some instances of *last* classified as OTHER are given in (9):

(9) a. Don't put something in just one form and expect it to last.
b. There's a very large body of research that says that more generous benefits and benefits that last longer ...

Thus, although we manipulated an equal proportion of the active transitive sentences that each verb occurred in, our intervention changed 16.0% of the *total* occurrences of *cost* and 19.9% of the total occurrences of *resemble*, but only 3.9% of the total occurrences of *last*. As the neural networks we trained used a single vector embedding to represent all uses of a word, as is standard, this finding could indicate that in neural networks passivizability can "spill over" from uses of a verb in contexts not immediately relevant to passivization such as intransitive sentences. This raises the question of whether similar spillover effects might occur in the other direction: would changing the affectedness of intransitive uses of *last* affect the passivizability of transitive uses of the verb?

Our choice to intervene only on active transitive sentences may also explain why we found no significant difference between a moderate (30%) and a more extensive (70%) manipulation of the verb's active transitive occurrences. Although the extensive manipulation intervened on 70% of the active transitive uses of the mutating verbs, occurrences of the mutating verbs in other contexts were not replaced, resulting in more than half of the total occurrences of each mutating verb being left in their original contexts. Specifically, in the condition labeled as 70%, we intervened on 37.3% (vs. 16.0% in the ‘moderate’ manipulation) of the total occurrences of *cost*, 46.4% (vs. 19.9%) of the total occurrences of *resemble*, and just 9.1% (vs. 3.9%) of the total occurrences of *last*. While we more than doubled the number of sentences we intervened on, then, it is possible that the more extensive manipulation did not shift the overall distribution of each verb’s contexts enough to produce a detectable change in affectedness. Future work can consider manipulations or interpolations over *all* occurrences of a verb without limiting the syntactic context of intervention. Doing so would allow for more extensive changes to the distributions of verbs, although this procedure would require strict control over which target verb sentences are used to replace mutating verb sentences, in order to avoid changing the syntactic distribution of the verb (e.g. by replacing a transitive occurrence of a verb with an intransitive occurrence).

Overall, results from this experiment suggest that changing the distribution of a verb’s arguments affects its passivizability, although the extent to which this occurs is clouded by limitations in the filtering process. This finding is consistent with the body of experimental and computational work showing that the affectedness of a verb is a significant predictor of human acceptability ratings across languages (e.g. Ambridge et al. 2023, 2016, Liu and Ambridge 2021), and suggests that model judgments of acceptability may be driven by similar factors as human judgments.

Experiment 3: Testing the interaction between entrenchment and affectedness

Experiments 2A and 2B were designed to test, by manipulating the distribution of the active and passive forms of verbs in the training corpus, how passivizability judgments are affected by the frequency of active and passive forms of the verb and by the lexical semantics of the verb. Experiment 2A found evidence that a verb’s active-to-passive frequency ratio, its A/P ratio, was a significant predictor of passivizability. Experiment 2B found that the affectedness of a verb’s semantics affects its passivizability, although this effect may have been appeared verb-dependent. These studies tested each hypothesis independently, and were operationalized by filtering an existing corpus.

Experiment 3 presents a complementary method to test both the entrenchment and the affectedness hypothesis that further allows us to explore the relationship between the entrenchment and affectedness hypotheses. In this experiment, we ask whether one factor is more primary than the other in determining a verb’s passivizability, and whether entrenchment and affectedness interact.

Instead of ablating or altering existing parts of our training corpus, we introduced a **novel** verb into the corpus that *only* occurred in active sentences. We then tested if passivizability judgments from models trained on this corpus were consistent with the predictions of the entrenchment and affectedness hypotheses. We repeated this experiment a number of times, manipulating two factors: first, whether the verb occurred in sentences that had high or low affectedness; and second, the number of times the verb appeared in the training corpus (again, only in the active voice). This design allowed us to test whether the lexical semantics of the novel verb’s context changed its passivizability in the absence of confounds caused by verbs’ naturally differing syntactic distributions as well as limitations in our filtering procedure. It also allowed us to test for an interaction between the entrenchment and affectedness hypotheses.

The two hypotheses make the following predictions for this design. The entrenchment hypothesis, which attributes unpassivizability to a high relative frequency of the active compared to the passive, predicts that, since the novel verb never appears in the passive voice, increasing the number of times it appears in the active voice should cause its passive drop to increase, regardless of the semantic contexts in which the verb appears. On the other hand, the affectedness hypothesis predicts that a novel verb will be more passivizable if it occurs only in high-affectedness contexts than if it occurs only in low-affectedness contexts.

What interaction patterns can we expect between entrenchment and affectedness? One possible outcome could be entrenchment effects that arise only when the verb occurs in high-affectedness contexts. If a verb occurs in high-affectedness contexts, learners may expect the novel verb to be potentially passivizable, and will then rely on the active-to-passive ratio to determine if that is the case. By contrast, if a verb occurs in low-affectedness contexts, learners may have weaker expectations about its likelihood of appearing in the passive, and as such may be less sensitive to the relative frequency of the active compared to the passive.

Procedure

We generated sets of sentences with high and low affectedness using GPT-4o (OpenAI et al., 2024). We first prompted the model to generate carrier sentences based on a set of seed verbs used by Ambridge et al. (2016). We next prompted GPT-4o to rate the affectedness of each carrier sentence based on questions adapted from Ambridge et al. (2016) and Reisinger et al. (2015), who used the questions to collect sentence affectedness ratings from human raters on a scale of one to five. We only kept low-affectedness sentences, defined as sentences whose average affectedness rating was below two, and high-affectedness sentences, whose average affectedness rating was above four. Table 4 provides examples of high-affectedness and low-affectedness carrier sentences, and Appendix E.1 gives a list of the verbs used in the generated sentences. The prompts we used are given in Appendix E.2. All stimuli were generated in October 2024.

Using these carrier sentences, we formed a training set containing 2000 high-affectedness carrier sentences and 2000 low-affectedness carrier sentences. Each carrier sentence was surrounded by two sentences of GPT-4o generated context before and after the carrier sentence, forming a five sentence-long paragraph. Next, we substituted a novel verb into some of the carrier sentences. We varied whether the novel verb occurred in high-affectedness or low-affectedness contexts, as well as the number of sentences n the novel verb occurred in, $n \in \{0, 1, 2, 4, 8, 10, 16, 32, 64, 100, 128, 256, 500, 512, 1000, 1024, 2000\}$. Carrier sentences where the verb was not replaced with the novel verb retained their original verb, such that the total size of the corpus was matched across conditions (all of the conditions other than $n = 2000$ included such sentences). We then shuffled this training set into the same 100 million word training corpus we have used in all of our experiments so far, and trained five randomly initialized models with different seeds on each combined corpus using the same setup as in Experiments 2A and 2B.

We used a new set of carrier sentences, distinct from those in the training set, to form test sets of 100 sentences from which to obtain passivizability judgments. In order to isolate the changing effect of the novel verb’s passivizability on passive drop from variability in passive drop attributable to other parts of the carrier sentences, we created test sets that substituted the *original* verb or the *novel* verb into the carrier sentence. Low-affectedness test sets contained 100 carrier sentences where the verb was seen in low-affectedness contexts, while high-affectedness test sets that contained 100 carrier sentences with high affectedness. We tested the novel verb’s passivizability only on the test sets that corresponded to the training condition: models trained on a dataset where the novel verb was seen in high-affectedness contexts during training were tested on the 100 sentences where the novel verb was also seen in high-affectedness contexts, and likewise for low-affectedness contexts. We made this choice to avoid an adversarial testing setting where the semantic material surrounding the verb sometimes differed sharply between the training and test sets.

Table 4
Examples of high and low affectedness carrier sentences generated by GPT-4o.

High affectedness	The mob boss **murdered** the rival gang leader.
	The performer **frightened** the audience with a sudden scream. The thief **robbed** the jewelry store.
Low affectedness	The property **bordered** the national park. The dog **feared** the sound of thunder. The musician **heard** the applause from the audience.

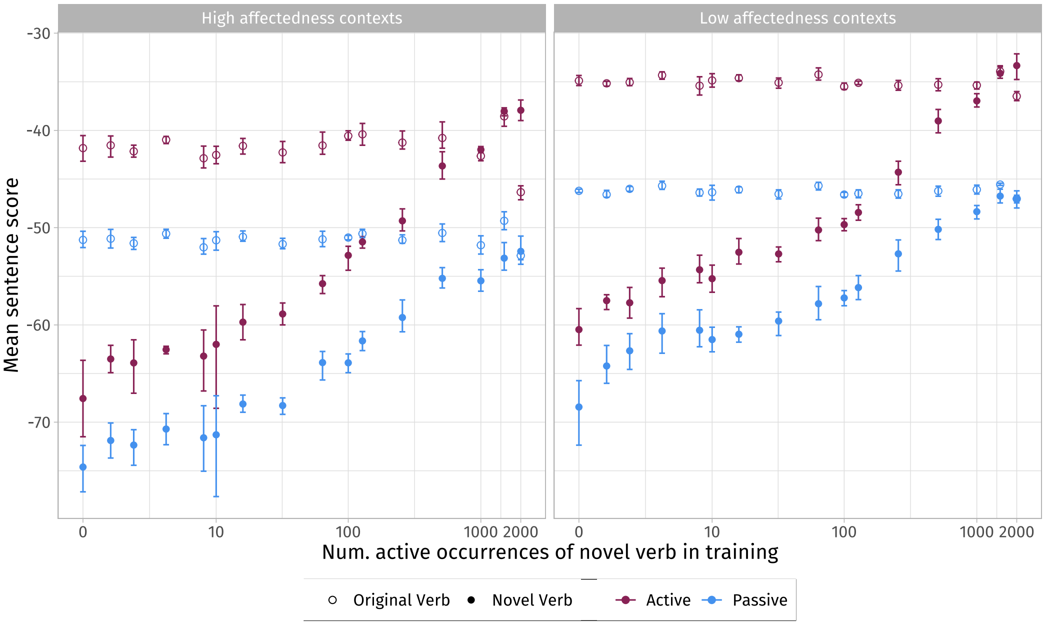


Fig. 13. *Mean sentence scores assigned to test sentences by models with differing levels of exposure to a novel verb in the active (the novel verb never occurred in the passive)*. Sentence scores assigned to sentences including the novel verb increased as the novel verb was used more frequently in the corpus. Scores assigned to original versions of the same test sentences (hollow dots) are shown as a baseline.

Results

We first examined the mean sentence scores assigned to active and passive test sentences containing the novel verb (filled points in Fig. 13) compared to the original verb (hollow points). Across all conditions, active sentences were given higher sentence scores than passive sentences. When the novel verb was not seen in the training data at all—i.e. none of the verbs in the corpus generated by GPT-4o were replaced with the novel verb—test sentences containing the novel verb were rated worse than the original versions of the test sentences (hollow points). In both high and low-affectedness conditions, the sentence scores assigned to sentences containing the novel verb increased as the novel verb was seen more frequently in the training corpus and, at around 1000 occurrences converged with the scores assigned to the original test sentences.

Sentences with high-affectedness contexts had lower sentence scores than sentences with low-affectedness contexts, regardless of whether they included the original verb or the novel verb. We suspect this is due to the fact that test sentences in the high affectedness condition were on average two words longer than test sentences in the low affectedness condition. Language model acceptability judgments are known to be sensitive to sentence length, though the relationship may be complex (Tjuatja et al., 2025). That being said, because passive drop is computed within-item, comparing the active and passive versions from the same minimal pair, any difference in length across sentence pairs should cancel out.

Next, we consider how the passive drop associated with the novel verb varied as models were exposed to more occurrences of the novel verb in the active voice, and no occurrences of this verb in the passive (Fig. 14). Regardless of affectedness, passive drop increased as the number of occurrences increased. To evaluate the statistical significance of this pattern, we fit a linear mixed-effects model predicting passive drop from the AFFECTEDNESS of the context as a fixed effect and random intercepts for MODEL SEED and TEST ITEM. We compared this model to a model that additionally included NUMBER OF OCCURRENCES of the novel verb as a fixed effect. A likelihood ratio test showed that number of occurrences was a highly significant predictor of passive drop ($\chi^2(1) = 74.25, p < 0.001$).

The novel verb's passivizability was also impacted by affectedness: for a given number of occurrences of the novel verb in the training corpus, sentences containing the verb were assigned a higher passive drop in high-affectedness contexts than in low-affectedness ones. To verify the significance of this effect, we fit a linear mixed-effects model predicting passive drop from NUMBER OF OCCURRENCES of the novel verb as a fixed effect and random intercepts for MODEL SEED and TEST ITEM, and compared this model to a full model that additionally included AFFECTEDNESS as a fixed effect. A likelihood ratio test showed that high affectedness significantly increased passive drop, though this effect was not as dramatic as that of frequency ($\chi^2(1) = 3.88, p = 0.049$).

Although there were significant main effects of both affectedness and frequency on passive drop, we did not find a significant interaction between the two. We compared a linear mixed-effects model predicting passive drop using AFFECTEDNESS and FREQUENCY as fixed effects and MODEL SEED and TEST ITEM as random effects against a full model that additionally included the interaction between affectedness and frequency as a fixed effect. A likelihood-ratio test showed that there was no

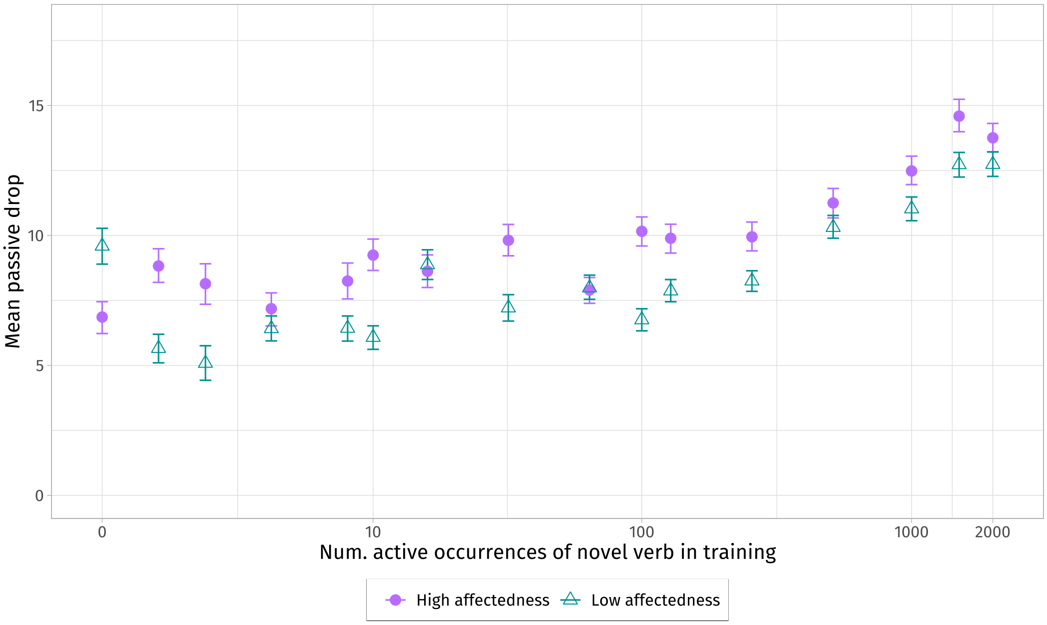


Fig. 14. *Change in passive drop as a verb is seen more often in the active and in more semantically affected contexts.* Passive drop increased as models were exposed more frequently to the novel verb in active sentences while not seeing the verb in passive sentences. The novel verb also had a higher passive drop when trained in high-affectedness contexts.

significant difference between the full model and the reduced model ($\chi^2(1) = 0.22, p = 0.645$).

Discussion

Experiment 3 found support for both the entrenchment and affectedness hypothesis. As we entrenched a novel verb in the active voice, by increasing the number of active occurrences of the verb without including any passive ones in the corpus, the novel verb became increasingly unpassivizable. At the same time, the novel verb was also assigned higher passive drop if it was seen in high-affectedness contexts than if it was seen in low-affectedness contexts. The affectedness effect, while consistent, was smaller than the effect of frequency.

These results are consistent with our findings in Experiment 2A, There, we showed that entrenchment effects significantly affect passive drop—the higher the active-to-passive ratio, the less passivizable a verb was. The current experiment, where there were no passive occurrences of the verb at all, found a similar relationship between entrenchment and passive drop.

The results of this experiment are also consistent with the results of Experiment 2B, which found a main effect of affectedness on passivizability. Experiment 3, where each verb *only* occurred in active transitive sentences, found that passive drop was consistently higher when the novel verb was placed in high-affectedness contexts. Experiment 3 further found that frequency and affectedness independently affect a verb's passivizability, without a clear interaction between them.

An extension of the method we introduce here could explore other facets of input context. For example, by selectively introducing a small number of passive sentences, one could test how entrenchment effects diminish as there is increased competition between the two structures. Another potential extension would study graded affectedness effects: for Experiment 3 we selected semantic contexts at the two extremes of the affectedness spectrum, to maximize the likelihood of detecting an effect. But the novel verb's lexical semantics could also be varied in a more fine-grained way, for instance by allowing the novel verb to appear in both high- and low-affectedness contexts to differing proportions, or by generating sentences with gradient affectedness.

General discussion

Even highly productive grammatical rules often have environments in which they cannot apply. Since learners typically do not receive direct evidence that a particular sentence is *not* grammatical, they must infer such exceptions to generalizations from indirect evidence. In this study, we explored the sources of indirect evidence that a learner might use to acquire the restrictions on English passivization. We focused on two hypotheses: the entrenchment hypothesis, which attribute the unpassivizability of particular verbs to frequency asymmetries between the active and passive forms of those verbs; and the affectedness hypothesis, which attribute unpassivizability to an incompatibility between the semantics of the verb and the passive construction, a construction where the subject of the sentence is expected to be affected by the event described in the sentence.

To test these hypotheses, we used transformer language models trained on a corpus that approximates the amount of exposure to language that human learners receive. This aspect of our design is critical as mainstream language models trained on much larger amounts of data have access to considerably more indirect evidence than humans (Wilcox et al., 2025). We found that language models' judgments of verb passivizability correlate highly with human judgments ($r = 0.9$). Then, by performing targeted interventions on the model's training data, we showed that the relative frequency with which a verb appears in the active and passive constructions in the training data provided indirect evidence for the model to learn exceptionality, though this factor did not fully explain the magnitude of the difference between highly passivizable verbs and highly unpassivizable ones. We also found a smaller effect of affectedness, where verbs observed during training in high-affectedness contexts were rated as more passivizable.

Human judgments of exceptions to passivization

We used the English passive as a case study of the acquisition of exceptions to syntactic generalizations. As much of the existing literature on these exceptions relies on linguists' acceptability judgments, in Experiment 1A we sought to first verify these judgments with naive participants. A further goal of the experiment was to identify any gradience in these judgments that could serve as a benchmark for quantitative computational modeling. In general, this experiment reproduced the linguists' judgments for a number of individual verbs and classes of verbs, but there were discrepancies in some cases—specifically, verbs in the *ooze* and *advantage* classes did not significantly differ in their passivizability from canonically passivizable verbs. We further showed that English speakers' judgments of passive exceptions are more nuanced than binary acceptability judgments might suggest. Not all verbs reported to be unpassivizable were equally unacceptable: for instance, although *last* and *resemble* are both reported as unacceptable in the literature, we found a much larger average passive drop for *last* than *resemble*.

Using neural networks as models of human learners

What benefits do neural network language models bring to the study of human language acquisition? Existing studies of children's acquisition of the restrictions on passivization (e.g. Fox and Grodzinsky 1998, Gordon and Chafetz 1990, Maratsos 1985) are limited by researchers' inability to exert full control over a child's linguistic input: It would be difficult, for instance, to ensure that a child never hears a specific verb in the passive voice, as we did in Experiment 3. These issues are addressed, to a limited extent, by human artificial language learning experiments, which target specific hypotheses through controlled experimentation on constructed languages. But the artificial languages used in these experiments are by necessity much simpler than natural languages—often with vocabularies of fewer than 50 items—and it is unclear if they engage the same cognitive mechanisms as first language acquisition.

Using neural networks as model learners addresses these existing methodological lacunae in acquisition studies. Neural networks, unlike human learners, are trained on data over which we have full control. Unlike the symbolic models sometimes used in language acquisition research, which are often simplified proof-of-concept systems, neural networks are broad-coverage models that can be trained on a corpus that is as close to the input to human children as possible (Vong et al., 2024; Warstadt et al., 2023). The greater degree of researcher control afforded by computational experiments is not limited to intervention experiments: while we have not explored this direction in the present work, the neural network paradigm makes it possible to probe a model's internal processes to understand which mechanisms are vital to the model's learning process and form hypotheses about how humans may learn (Baroni, 2022; Lakretz et al., 2021).

Neural network models can also be used in future work to shed light on cross-lingual typological patterns of exceptions. Models can be trained on mixes of corpora in different languages which can be strictly controlled to explore the relationship between the languages (e.g. Constantinescu et al. 2024). Mechanistic analysis of these multilingual models may reflect shared representations and mechanisms, which would support the argument that passive constructions across different languages reflect a common semantic universal (Ambridge et al., 2023; Papadimitriou et al., 2021). Finally, withholding instances of the passive in one language and not another may allow researchers to explore whether models extend the passive construction across languages (Papadimitriou et al., 2023).

While neural network language models have substantial potential as model learners for grammatical phenomena, there are a number of methodological limitations to this modeling approach. First, the value of modeling is limited by the interpretability and cognitive plausibility of the models we use (Baroni, 2022). Without a clear understanding of the inductive biases of the particular neural network chosen for comparison, we cannot make a fair comparison between these models and our theories of human cognition. Although we highlighted some similarities between the GPT-2 architecture and human language learning and processing, this architecture is clearly not a perfect model for human language learning (for example, transformers' working memory constraints are fundamentally different from those of humans; Armeni et al. 2022, Timkey and Linzen 2023), and care should be taken to make fair comparisons between the two.

Secondly, the methodology of intervening on a corpus by removing particular classes of examples, or swapping some words out for others in particular syntactic contexts, relies heavily on the accuracy of the linguistic analysis tools we have at our disposal. Without the ability to precisely and accurately parse and alter a corpus containing heterogeneous data, it is difficult to ensure the feasibility and reliability of any intervention. For instance, the filters we used to make our intervention in Experiment 2A may have introduced new confounds in the training data by removing clear examples of passive sentences while failing to identify and filter out more complex examples. While our goal in Experiment 2B was to transplant verbs into agent-patient environments, our intervention was limited to replacing active transitive sentences with other active transitive sentences, and so could not account for verb-level differences in syntactic distribution. Refining the filtering process would allow us to test hypotheses that are more narrowly defined, and thus to make conclusions at a more granular level.

The paradigm we used in Experiment 3, where we construct new sentences with a novel verb and insert them into the corpus, may serve as a cleaner and easier to control alternative to the syntactic filtering paradigm. However, this method is contingent on using another language model to create human-like language input. While generative models can create much larger-scale datasets than can be hand-crafted, the distribution of the data formed by these models may not align completely with that of human-created text. Any analyses based on these synthetic datasets may thus reflect artifacts of the language model's generative procedure, rather than features of human language use.

Finally, the computational cost of training models on each modified corpus can be high. We trained a total of 125 models for Experiment 2, each requiring two days to train and using approximately 3e15 floating point operations (FLOPs). These training regimes are highly computationally expensive and their potential environmental impacts should be considered. These limitations notwithstanding, we hope that the use of targeted interventions on naturalistic data can be used to compare the plausibility of hypotheses in situations where such interventions are not possible in human research.

Implications for human learners

What are the consequences of our results to the study of human learners? We have shown in causal experiments using sentence deletion (Experiment 2A) and novel verbs (Experiment 3) that models can leverage the relative frequency of the active and passive constructions in training data to learn exceptions. This finding is consistent with usage-based approaches to human language acquisition (Goldberg, 2006; Tomasello, 2000), and can be taken as an existence proof demonstrating that a learner could display a human-like pattern as a consequence of tracking the statistics of verb-construction co-occurrences.

At the same time, humans and language models clearly differ in their learning mechanisms, goals, and resources. It is possible, then, that while our models' acceptability judgments are largely similar to those of humans, they achieve such behavior via a very different developmental pathway. Indeed, since our neural networks learn the task of next word prediction by tracking word co-occurrences across a corpus, the fact that they are highly sensitive to frequency statistics is unsurprising. Although human learners can also track statistical information in their linguistic input (Saffran et al., 1996; Thompson & Newport, 2007), they do not rely solely on distributional statistics: language acquisition is also shaped by interactive and communicative social pressures. When a child produces an utterance that an adult finds unacceptable, the adult may repeat that utterance but produce a change in the erroneous portion which the child may then take up in their next utterance if the correction matches their intended meaning (Chouinard & Clark, 2003). Children can learn from these reformulations in order to further their communicative goals—goals which our models lack.

Humans can learn the concept of affectedness—a key factor in Experiment 2B and Experiment 3—from their direct experiences with the world. By contrast, our models were trained without access to sensorimotor input, and thus may lack these conceptual primitives (see, e.g., Lake et al. 2017); instead, they may have learned concepts that approximate, but are not identical to, human concepts of affectedness. Consequently, while in Experiment 2B and Experiment 3 our models were sensitive to the semantic context in which a verb occurred, we cannot conclude that people will be sensitive to this factor in the same way or to the same extent. For instance, while frequency played a larger causal role than affectedness in increasing unpassivizability in our models, semantic cues could be more predictive of passivizability for people, for whom the meaning of a sentence is key. The relative influence of affectedness and frequency on passivizability in humans thus remains an open question. That being said, we take the behavior exhibited by our models in Experiment 3 to constitute a hypothesis for how humans might behave in a learning task, which should be tested in future work.

In sum, while we have illustrated a potential pathway by which a neural network learner might conclude that a verb is unpassivizable, the same learning mechanisms might not be at play in human learners. Repeating these experiments using models that differ in architecture, and in particular models that have access to interaction and causal primitives, may help to disentangle which capacities are required to learn about passivizability.

Other sources of indirect evidence

While we have found support for the role of both entrenchment and affectedness in language models' judgments of passivizability, neither factor alone *fully* accounted for the models' judgments: our interventions in both Experiments 2A and 2B did not consistently make a mutating verb as passivizable as its corresponding target verb. We showed in Experiment 3 that the effects of entrenchment and affectedness are additive, but did not compare these effect sizes to the effect sizes we found in actual unpassivizable verbs. Future work could use the generative method we propose in Experiment 3 to reconstruct the passivizability of a verb from these two sources to explore whether entrenchment and affectedness together can explain the full degree of a verb's passivizability.

If the effects of entrenchment and affectedness alone are insufficient to arrive at the full amount of unpassivizability we found in Experiment 1B, what other sources of evidence could language models—and, potentially, humans—rely on? One potential such source of evidence is the existence of similar constructions to the passive with a *by*-phrase. These include (10b), which differs from the passive (10a) by just one word, or the active construction (10c), which is used in functionally similar contexts:

(10) a. Two hours were required by the meeting.
b. Two hours were required for the meeting.
c. The meeting required two hours.

The existence of alternations like those illustrated in (10) could affect the acquisition of passive exceptions in two ways. First, if learners often hear (10b) or (10c) in contexts where they might otherwise expect to hear (10a), they may conclude that (10a) is unacceptable (Ambridge et al., 2015; Boyd & Goldberg, 2011; Clark, 1987; Goldberg, 1995), as we hypothesized was the case for the frequency asymmetry between actives and passives in Experiment 2A. Second, the existence of alternatives like (10b) for some but not all verbs may also help to explain the gradience in acceptability that we see across verbs within a verb class in Experiment 1A, if viewed through the lens of the noisy channel theory (Gibson et al., 2013; Levy, 2008). If English speakers find (10b) to be unacceptable but have access to an alternative like (10a) that is acceptable, they may process (10b) as a corruption of the acceptable (10a), and thus judge it as more acceptable than a similar passive sentence for which there is no corresponding alternative. These hypotheses and their interactions can be explored through similar training data interventions to the ones we implemented in this paper.

Conclusion

How is knowledge of grammar related to the learner's linguistic input? In this paper, we studied how exceptions to passivization in English, which must be learned through indirect evidence, can be learned by a neural network language model, a broad-coverage model that can learn from amounts of data comparable to those that humans learn from. We manipulated the training corpus of the neural network language models to test the causal links between input and the models' behavior. We first showed that passivizability judgments extracted from a language model match human acceptability judgments to a substantial extent (Experiment 1B). We then made targeted changes to the models' training data to measure the effects of entrenchment and affectedness, two factors that have been argued to be implicated in the learning of passives in humans, on the models' learning of these patterns. Through our interventions, we found that changing the verb's relative frequency of occurrence in the active and passive voice affected the models' judgments of its passivizability, as predicted by the entrenchment hypothesis (Experiment 2A). We also found that manipulating a verb's semantics by changing the arguments it appears with in active transitive sentences significantly affected the passivizability of the verb, although the magnitude of this effect interacted with verb identity (Experiment 2B). In Experiment 3, a more tightly-controlled novel verb experiment, we found further evidence that frequency and affectedness both significant affect passivizability judgments, and additionally found that these two factors do not interact. These findings illustrate a method for testing hypotheses about how large-scale linguistic input affects learning, and raise new questions that can be tested with human participants.

CRediT authorship contribution statement

Cara Su-Yi Leong: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Tal Linzen:** Formal analysis, Funding acquisition, Methodology, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Tal Linzen reports financial support was provided by National Science Foundation. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. BCS-2114505. We thank the audience of Society for Computation in Linguistics 2023 and members of the NYU Computation and Psycholinguistics Lab for comments. This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

Appendix A. Stimuli

This section lists the materials for acceptability judgments.

A.1. Test sentences

Verb class	Sentence frame
Advantage	My donation0.2inmany communities.
	Your actions0.2inyour son.
	Our friendship0.2inour relationship.
	The gift0.2inmy organization.
	The treaty0.2inboth countries.
Price	Your dish0.2inninety dollars.
	The painting0.2in2000 dollars.
	My initiative0.2insome money.
	Your book0.2inthirty dollars.
	His actions0.2inthe medal.
Ooze	my friend0.2inconfidence.
	The lightbulb0.2insome light.
	My machine0.2ina sound.
	The teacher0.2inwisdom.
	The trash0.2inan odor.
Estimation	The caricature0.2inan actor.
	Your friend0.2inmy brother.
	The sketch0.2inmy design.
	Her son0.2inher father.
	The copy0.2inthe original.
Duration	The journey0.2inthree days.
	My meeting0.2intwo hours.
	The surgery0.2insome time.
	Her speech0.2inseventeen minutes.
	His recovery0.2ina month.
hit	My brother hit your friend.
	A boy hit my bag.
	Your dog hit the toy.
	The child hit a monkey.
	The arrow hit the target.
kicked	My brother kicked your friend.
	A boy kicked my bag.
	Your dog kicked the toy.
	The child kicked a monkey.
	My friend kicked the wall.
carried	A boy carried my bag.
	My brother carried your friend.
	The dog carried the toy.
	Your mother carried the child.
	The donkey carried the load.
pushed	A boy pushed the cup.
	My brother pushed a child.
	A child pushed the bag.
	The mother pushed my toy.
	Your sister pushed your friend.
washed	A boy washed the cup.
	My brother washed my plate.
	A child washed the bag.
	The mother washed my toy.
	Your sister washed a towel.
dropped	A boy dropped the cup.
	My brother dropped my plate.
	A child dropped the bag.
	The mother dropped my toy.
	Your sister dropped a book.

A.2. Filler sentences

Type	Sentence
Acceptable	She was worried about the problem.
	Your knife needs to be sharpened.
	Her sister failed her test.
	The bank is located across the road.
	His mother thought that your friendship was strong.
	Attention check: select ’Completely acceptable’.
	The dog bit its owner.
	The girl was unexcited about the trip.
	Her father said that your recovery was quick.
	The meeting ended quickly.
	Your sister claimed that the machine broke.
	A woman sang beautifully.
	The ship was sunk by the enemy.
	The opportunity presented itself.
	His sister slept at my house.
	Your child played the game.
	My brother sold your friend a plate.
	It rained yesterday at noon.
	Attention check: select ’Completely acceptable’.
	Your job requires concentration.
	My mother read the child a book.
	The goldfish died alone.
	The monkey wanted to eat a banana.
	Glass bottles are very fragile.
Unacceptable	A bottle breaking last night.
	The company lent the employee.
	A cat met either mouse.
	My sister said a word all night.
	Your friend is walks home.
	On a book the floor sat.
	The driver handed the keys.
	An infant asleep.
	My friend liked your car at all.
	A doctor was give the dog a toy.
	A ball hit with great force.
	Puppy my bit hand your.
	The teacher bought for the students.
	Candlesticks a picnic.
	A student playing piano well.
	My bottle holds.
	Sat on the floor your sister.
	Her daughter will watches a movie.
	Attention check: select ’Completely unacceptable’.
	My key a cabinet.
	The boy saw anyone.
	The chicken killed.
	Snack this delicious taste.
	That wall are green.
	The car the light.
	Your friend lifted a finger to help.
	His friend is painted his grandmother a portrait.
	The child brought to school.
	The boy looked the picture.
	The cow are grazing in the field.
	The classroom silent.
	Any girls passed the test.
	My feelings were hurting by my brother.
	The class went to on Tuesday.
	The singer are practicing a song.
	The opportunity some wallpaper.
	There is every fly in my soup.

	Attention check: select ’Completely unacceptable’.
	The car driven.
	The doctor disliked last week.
	Box a opened the boy.
	This plates has been chipped.
	The bank will lend me.
	Your backpack heavy.
	Your mother bought any cups.
	The essay was wrote by a genius.

Appendix B. Human acceptability judgment task instructions

In this experiment, you will rate English sentences based on how acceptable they sound to you. Try to answer based on your gut reaction, without analyzing the sentences. There are no ‘right’ or ‘wrong’ answers. The first two questions will be practice questions to familiarize you with the task.

< Participant clicks *Next* button >

S1: How acceptable is this sentence? The mirrors reflected light.

Hint: For many people, this sentence is completely acceptable. Move the slider to the right corner of the scale to rate the sentence if you agree. Then, click the Next button or press the spacebar to continue.

< Participant rates S1 and clicks *Next* button to continue >

S2: How acceptable is this sentence? The teacher was spoke.

Hint*:* For many people, this sentence is completely unacceptable. Move the slider to the left corner of the scale to rate this sentence if you agree. Then click the Next button or press the spacebar to continue

< Participant rates S2 and clicks *Next* button >

Appendix C. Sample parsing errors

These four sentences were parsed by the spaCy model `en_core_webtrf` as passive uses of the verb *last*:

1. It’s lasted for 52 years so far, whether on television or in spin-off media, and that’s in no small part because of the original idea to recast the title character in 1966, thus creating the concept of regeneration.
2. But everyone saying this should have to add “however, he’s certainly lasted much longer than we originally predicted”.
3. It’s lasted 75 years.
4. Fans of the living dead have one man to thank for the birth of the modern cinematic zombie genre: George A. Romero, the filmmaker who made Night of the Living Dead on the cheap in 1968 and kicked off a zombie obsession that’s lasted for decades.

Appendix D. Raw verb counts in original and frequency-altered corpora

Table 5
Verbs’ frequency of occurrence in original corpus and after interventions to match the relative active–passive ratio of the target verb.

	Original		Carry		Drop		Hit		Push	
	Active	Passive	Active	Passive	Active	Passive	Active	Passive	Active	Passive
Original			7660	2181	3288	1143	9372	2452	4461	1269
last	723	4	7591	42	3253	18	9218	51	4338	24
cost	4131	21	7475	38	3147	16	9245	47	4327	22
resemble	1373	1	6865	5	2746	2	8238	6	4119	3

Appendix E. Experiment 3 stimuli details

E.1. Verbs used in experiment 3 stimuli

Verbs used in high-affectedness training sentences. executed (129), drowned (128), crushed (126), stabbed (118), burned (113), murdered (105), suffocated (77), terrorized (76), broke (69), strangled (69), exterminated (64), frightened (61), knifed (59), demolished (55), robbed (44), pushed (42), shot (36), hit (28), wrecked (28), killed (26), boiled (24), kicked (24), smashed (20), destroyed (19), shattered (16), bit (15), hammered (14), punched (14), shook (13), crafted (10), designed (10), annihilated (9), assassinated (9), eradicated (9), painted (9), constructed (8), built (7), discovered (7), invented (7), obliterated (7), programmed (7), thumped (7), composed (6), devastated (6), developed (6), sculpted (6), baked (5), dismantled (5), engineered (5), exploded (5), investigated (5), knocked (5), repaired (5), rescued (5), saved (5), carved (4), cooked (4), extinguished (4), solved (4), wrote (4), collapsed (3), completed (3), decimated (3), eliminated (3), filmed (3), planted (3), stole (3), terrified (3), vaporized (3), arrested (2), brewed (2), burnt (2), captured (2), caught (2), collected (2), created (2), devoured (2), directed (2), engulfed (2), fixed (2), formulated (2), hacked (2), harvested (2), ignited (2), incinerated (2), launched (2), massacred (2), plowed (2), pruned (2), published (2), shelved (2), struck (2), taught (2), toppled (2), tried (2), trimmed (2), uncovered (2), abandoned (1), abducted (1), altered (1), ambushed (1), analyzed (1), apprehended (1), aspired (1), assailed (1), attacked (1), attempted (1), authorized (1), axed (1), bashed (1), battered (1), beat (1), blew (1), bombed (1), breathed (1), bulldozed (1), capsized (1), chiseled (1), choked (1), chopped (1), coded (1), comforted (1), conducted (1), cracked (1), crashed (1), crossed (1), cultivated (1), curated (1), cut (1), defeated (1), delivered (1), detonated (1), discarded (1), displaced (1), drew (1), drove (1), edited (1), emptied (1), enacted (1), erased (1), evaporated (1), explained (1), fired (1), flattened (1), forged (1), found (1), grilled (1), hung (1), imagined (1), influenced (1), injured (1), inked (1), innovated (1), inspected (1), kneaded (1), lit (1), looted (1), manufactured (1), mapped (1), mugged (1), neutralized (1), ordered (1), organized (1), photographed (1), picked (1), pinned (1), plucked (1), poisoned (1), polished (1), popped (1), prepared (1), pulverized (1), pummeled (1), purchased (1), questioned (1), recorded (1), removed (1), replaced (1), ruptured (1), sank (1), scared (1), sentenced (1), set (1), sewed (1), shaped (1), shocked (1), shoved (1), slaughtered (1), slayed (1), sliced (1), squashed (1), steered (1), studied (1), suspected (1), swallowed (1), synthesized (1), tailored (1), thwarted (1), torched (1), tore (1), trampled (1), translated (1), unearthed (1), unraveled (1), violated (1), won (1), wracked (1)

Verbs used in high-affectedness test sentences. executed (11), exterminated (7), crushed (6), burned (5), drowned (5), knifed (5), terrorized (5), broke (4), pushed (4), stabbed (4), demolished (3), frightened (3), suffocated (3), assassinated (2), built (2), murdered (2), wrecked (2), believed (1), bit (1), conducted (1), crashed (1), crippled (1), destroyed (1), developed (1), discovered (1), fried (1), hammered (1), hit (1), incinerated (1), kicked (1), obliterated (1), ordered (1), photographed (1), punched (1), robbed (1), saved (1), shot (1), slaughtered (1), smothered (1), stole (1), strangled (1), transformed (1), weakened (1), wrote (1)

Verbs used in low-affectedness training sentences. heard (113), believed (108), resembled (108), feared (103), noticed (101), recognized (95), lacked (94), trusted (89), missed (85), cost (82), overheard (72), respected (68), saw (67), needed (66), bordered (65), liked (65), dreaded (64), remembered (63), had (57), understood (56), knew (52), abutted (44), fit (41), received (39), spotted (28), totaled (28), lasted (23), looked (19), sensed (19), underwent (13), slept (12), admired (3), appeared (3), contained (3), exceeded (2), included (2), overlooked (2), possessed (2), provided (2), represented (2), accommodated (1), appreciated (1), approached (1), arrived (1), belonged (1), carried (1), cautioned (1), conveyed (1), created (1), delivered (1), described (1), disliked (1), displayed (1), enclosed (1), ended (1), envied (1), evoked (1), felt (1), fitted (1), hated (1), held (1), led (1), matched (1), misunderstood (1), observed (1), owned (1), packed (1), produced (1), published (1), ran (1), recalled (1), reflected (1), regretted (1), seated (1), seemed (1), sounded (1), spanned (1), was (1), weighed (1), wore (1)

Verbs used in low-affectedness test sentences. liked (7), missed (6), recognized (6), trusted (6), abutted (5), believed (5), feared (5), heard (5), knew (5), resembled (5), cost (4), fit (4), had (4), remembered (4), dreaded (3), lacked (3), noticed (3), overheard (3), respected (3), understood (3), needed (2), spotted (2), contained (1), lasted (1), received (1), saw (1), sensed (1), slept (1), underwent (1)

E.2. GPT-4o prompt for creation of affected sentences

The prompt used to generate high-affectedness sentences is given below. The scores reported in the example below are the mean scores given to the sentence by humans in the Semantic Proto-Roles dataset (Reisinger et al., 2015). The low-affectedness prompt uses the same prompt format with different target values (`1s` instead of `5s`) and a different example adapted from the dataset that received low affectedness scores.

```
  You will be given a number. Generate that number of examples by following these steps.
1. Find a verb to use. Try to pick a verb that will score all 5s on the tasks below.
2. Use the verb in a transitive sentence in the past tense. This means that the sentence
contains a subject and an
object, for instance 'After going to the supermarket, the man ate two bagels after dinner'.
Let's call this the target
sentence. Here, the subject is 'the man who loved watching television', and the object is 'two bagels'.
An example of a sentence that
is not transitive is 'The worry ate at the man' or 'The man ate'. Try to make a sentence that will
score all 5s on the tasks below.
3. Find the subject and object of the target sentence.
4. Create 2 sentences of context before the target sentence and 2 sentences after the target
sentence.
5. Rate the sentence on a scale of 1-5 for how likely it is that the subject of the target sentence
caused the action
to happen.
6. Rate the sentence on a scale of 1-5 for how likely it is that the subject of the target sentence
chose to be involved
in the action.
7. Rate the sentence on a scale of 1-5 for how likely it is that the subject of the target sentence
was aware of the action.
8. Rate the sentence on a scale of 1-5 for how likely it is that the subject of the target sentence
was sentient.
9. Rate the sentence on a scale of 1-5 for how likely it is that the object of the target sentence
was altered or somehow
changed during or by the end of the action in the target sentence.
10. Rate the sentence on a scale of 1-5 for how likely it is that the action in the target sentence
caused a change in the
object of the target sentence.
11. Rate the sentence on a scale of 1-5 for how likely it is that the object of the target sentence
changed possession
during the action in the target sentence.
12. Rate the sentence on a scale of 1-5 for how likely it is that the object of the target sentence
changed location during the action in the target sentence.
13. Repeat steps 1-13 using verbs and sentences that will score all 5s on the ratings. Do this until
there are as many
```

```
examples as was requested by the user.
14. Report the scores for all of the sentences in json format. In the json, include the following
fields: ’subject’,
’object’, ’verb’, ’sentence’, ’paragraph’, and ’scores’, as follows:
[{’subject’: ’...’, ...},
...,
{’subject’: ’...’, ...},
{’subject’: ’...’, ...}
]

------
EXAMPLE
Let’s think step by step. First, we find a verb that we want to use.
Verb: killed.
Next, we make a sentence using that verb.
Sentence: Saddam killed every month more people than all those who died from suicide murders since the Coalition
occupation of Iraq.
Now we know the subject and object of the sentence: the subject is ’Saddam’, the object is ’more
people than
all those who died from suicide murders since the Coalition occupation of Iraq’, and the action
is ’killed’. Then, we create two sentences of context before and after the sentence.
Sentence in context:
Sept. 11 was quantitatively much less lethal than many earthquakes. More people die from AIDS in
one day in
Africa than all the Russians who died at the hands of Chechnya-based Moslem suicide murderers
since that
conflict started. Saddam killed every month more people than all those who died from suicide
murders since the Coalition
occupation of Iraq. So what is all the fuss about suicide killings? It creates headlines.
Finally, we rate the sentence on a scale of 1-5 for each of the questions.
First, rating for how likely it is that the subject of the target sentence, ’Saddam’, caused
the action,
’killing’, to happen.
Rating: 5
Rating for how likely it is that the subject of the sentence,’Saddam’, chose to be involved
in the action ’killing’.
Rating: 5
Rating for how likely it is that the subject of the target sentence, ’Saddam’, was aware of
the action ’killing’.
Rating: 5
Rating for how likely it is that the subject of the target sentence, ’the criminal’, was sentient.
Rating: 5
Rating for how likely it is that the object of the target sentence, ’more people... than since
the conflict started’,
was altered or somehow changed during or by the end of the action ’kidnapping’.
Rating: 4.5
Rating for how likely it is that the action in the target sentence caused a change in the
object of the target sentence.
Rating: 4.5
Rating for how likely it is that the object of the target sentence changed possession during
the action
in
the target sentence.
Rating: 2.5
Rating for how likely it is that the object of the target
sentence changed location during the action in the
target sentence.
Rating: 4

Now we report our answer as a json.
‘‘‘
[
{’subject’: ’Saddam’,
’object’: ’more people than all those who died from suicide murders since the Coalition occupation of Iraq’,
’verb’: ’killed’,
’sentence’: ’Saddam killed every month more people than all those who died from suicide murders since the
Coalition occupation of Iraq.',
'paragraph': 'Sept. 11 was quantitatively much less lethal than many earthquakes. More people die from AIDS
in one day in Africa than all the Russians who died at the hands of Chechnya-based Moslem suicide murderers since
that conflict started. Saddam killed every month more people than all those who died from suicide murders
since the Coalition occupation of Iraq. So what is all the fuss about suicide killings? It creates headlines.',
'ratings': [5,5,5,5,4.5,4.5,2.5,4]
}
...
]
‘‘‘
```

Data availability

Data and code available at https://github.com/craaaa/exceptions.

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 abs/2303.08774, URL: https://arxiv.org/abs/2303.08774.

Ambridge, B., Arnon, I., & Bekman, D. (2023). He was run-over by a bus: Passive – but not pseudo-passive – sentences are rated as more acceptable when the subject is highly affected. New data from Hebrew, and a meta-analytic synthesis across english, Balinese, Hebrew, Indonesian and Mandarin. *Glossa Psycholinguistics*, *2*(1), http://dx.doi.org/10.5070/G6011177.

Ambridge, B., Bidgood, A., Pine, J. M., Rowland, C. F., & Freudenthal, D. (2016). Is passive syntax semantically constrained? Evidence from adult grammaticality judgment and comprehension studies. *Cognitive Science*, *40*(6), 1435–1459. http://dx.doi.org/10.1111/cogs.12277.

Ambridge, B., Bidgood, A., Twomey, K. E., Pine, J. M., Rowland, C. F., & Freudenthal, D. (2015). Preemption versus entrenchment: Towards a construction-general solution to the problem of the retreat from verb argument structure overgeneralization. *PLoS One*, *10*(4), Article e0123723. http://dx.doi.org/10.1371/journal.pone.0123723.

Armeni, K., Honey, C., & Linzen, T. (2022). Characterizing verbatim short-term memory in neural language models. In A. Fokkens, & V. Srikumar (Eds.), *Proceedings of the 26th conference on computational natural language learning* (pp. 405–424). Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.conll-1.28.

Aryawibawa, I. N., & Ambridge, B. (2018). Is syntax semantically constrained? Evidence from a grammaticality judgment study of Indonesian. *Cognitive Science*, *42*(8), 3135–3148. http://dx.doi.org/10.1111/cogs.12697.

Bach, E. W. (1980). In defense of passive. *Linguistics and Philosophy*, *3*(3), 297–341, arXiv:25001027.

Bakeman, R., & Mcarthur, D. (1996). Picturing repeated measures: Comments on loftus, morrison, and others. *Behavior Research Methods, Instruments & Computers*, *28*(4), 584–589. http://dx.doi.org/10.3758/BF03200546.

Baker, C. L. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry*, *10*(4), 533–581, arXiv:4178133.

Baroni, M. (2022). On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. arXiv:2106.08694.

Beavers, J. (2011). On affectedness. *Natural Language & Linguistic Theory*, *29*(2), 335–370, arXiv:41475291.

Bidgood, A., Pine, J. M., Rowland, C. F., & Ambridge, B. (2020). Syntactic representations are both abstract and semantically constrained: Evidence from children's and adults' comprehension and production/priming of the english passive. *Cognitive Science*, *44*(9), Article e12892. http://dx.doi.org/10.1111/cogs.12892.

Boyd, J. K., & Goldberg, A. E. (2011). Learning what **NOT** to say: The role of statistical preemption and categorization in A-adjective production. *Language*, *87*(1), 55–83. http://dx.doi.org/10.1353/lan.2011.0012.

Braine, M. D. S., & Brooks, P. J. (1995). Verb argument structure and the problem of avoiding an overgeneral grammar. In M. Tomasello, & W. E. Merriman (Eds.), *Beyond names for things: Young children's acquisition of verbs* (pp. 353–376). Hillsdale, N.J: L. Erlbaum.

Brooks, P. J., & Tomasello, M. (1999). Young children learn to produce passives with nonce verbs. *Developmental Psychology*, *35*(1), 29. http://dx.doi.org/10.1037/0012-1649.35.1.29.

Brown, R. (1973). *A first language: the early stages*. Cambridge, MA: Harvard University Press.

Chouinard, M. M., & Clark, E. V. (2003). Adult reformulations of child errors as negative evidence. *Journal of Child Language*, *30*(3), 637–669. http://dx.doi.org/10.1017/S0305000903005701.

Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. In *Mechanisms of language aquisition* (pp. 1–33). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Comrie, B. (1988). Passive and voice. In M. Shibatani (Ed.), *Typological studies in language*, *Passive and voice* (pp. 9–24). John Benjamins Publishing Company, http://dx.doi.org/10.1075/tsl.16.04com.

Comrie, B., Cole, P., & Sadock, J. M. (1977). In defense of spontaneous demotion: The impersonal passive. In *Grammatical relations* (pp. 47–58). Brill, http://dx.doi.org/10.1163/9789004368866_004.

Constantinescu, I., Pimentel, T., Cotterell, R., & Warstadt, A. (2024). Investigating critical period effects in language acquisition through neural language models. http://dx.doi.org/10.48550/arXiv.2407.19325, arXiv:2407.19325.

Dankers, V., Lucas, C., & Titov, I. (2022). Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 3608–3626). Dublin, Ireland: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.acl-long.252.

Darmasetiyawan, I. M. S., Messenger, K., & Ambridge, B. (2022). Is passive priming really impervious to verb semantics? a high-powered replication of Messenger Et al. (2012). *Collabra: Psychology*, *8*(1), 31055. http://dx.doi.org/10.1525/collabra.31055.

Demuth, K. (2011). The role of frequency in language acquisition. In *The role of frequency in language acquisition* (pp. 383–388). De Gruyter Mouton, http://dx.doi.org/10.1515/9783110977905.383.

Fox, D., & Grodzinsky, Y. (1998). Children's passive: A view from the by-phrase. *Linguistic Inquiry*, *29*(2), 311–332, arXiv:4179020.

Frank, M. C. (2023). Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, *27*(11), 990–992.

Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., & Saxe, R. (2013). A noisy-channel account of crosslinguistic word-order variation. *Psychological Science*, *24*(7), 1079–1088. http://dx.doi.org/10.1177/0956797612463705.

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough, O. D., Hansen, J. H. L., & Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, *26*(2), 248–265. http://dx.doi.org/10.1044/2016_AJSLP-15-0169.

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

Goldberg, A. E. (1995). Constructions: A construction grammar approach to argument structure. In *Cognitive theory of language and culture*, Chicago: University of Chicago Press.

Goldberg, A. E. (2006). Constructions at work: The nature of generalization in language, In *Oxford linguistics*, (1st publ ed.). Oxford: Oxford University Press.

Gordon, P., & Chafetz, J. (1990). Verb-based versus class-based accounts of actionality effects in children's comprehension of passives. *Cognition*, *36*(3), 227–254. http://dx.doi.org/10.1016/0010-0277(90)90058-R.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers)* (pp. 1195–1205). New Orleans, Louisiana: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N18-1108.

Hawkins, R., Yamakoshi, T., Griffiths, T., & Goldberg, A. (2020). Investigating representations of verb bias in neural language models. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 4653–4663). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-main.376.

Heafield, K. (2011). KenLM: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation* (pp. 187–197). Edinburgh, Scotland: Association for Computational Linguistics.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength natural language processing in Python.

Huang, K.-J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2024). Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, *137*, Article 104510.

Jumelet, J., Denić, M., Szymanik, J., Hupkes, D., & Steinert-Threlkeld, S. (2021). Language models use monotonicity to assess NPI licensing. http://dx.doi.org/10.48550/arXiv.2105.13818, arXiv:2105.13818.

Keenan, E. L., & Dryer, M. S. (2007). Passive in the world's languages. In T. Shopen (Ed.), *Language typology and syntactic description* (2nd ed.). (pp. 325–361). Cambridge University Press, http://dx.doi.org/10.1017/CBO9780511619427.006.

Kingma, D. P., & Ba, J. (2015). Adam: a method for stochastic optimization. arXiv:1412.6980 [cs].

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, Article e253. http://dx.doi.org/10.1017/S0140525X16001837.

Lakretz, Y., Hupkes, D., Vergallito, A., Marelli, M., Baroni, M., & Dehaene, S. (2021). Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, *213*, Article 104699. http://dx.doi.org/10.1016/j.cognition.2021.104699.

Lau, J. H., Clark, A., & Lappin, S. (2017). Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, *41*(5), 1202–1241. http://dx.doi.org/10.1111/cogs.12414.

Leong, C. S.-Y., & Linzen, T. (2023). Language models can learn exceptions to syntactic rules. In *Proceedings of the society for computation in linguistics*: *Vol. 6*, Amherst: University of Massachusetts Amherst, http://dx.doi.org/10.7275/H25Z-0Y75.

Levin, B. (1993). *English verb classes and alternations: a preliminary investigation*. Chicago: University of Chicago Press.

Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the conference on empirical methods in natural language processing* (p. 234). Honolulu, Hawaii: Association for Computational Linguistics, http://dx.doi.org/10.3115/1613715.1613749.

Linzen, T. (2020). How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5210–5217). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.465.

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*, 195–212.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535. http://dx.doi.org/10.1162/tacl_a_00115.

Liu, L., & Ambridge, B. (2021). Balancing information-structure and semantic constraints on construction choice: Building a computational model of passive and passive-like constructions in Mandarin Chinese. *Cognitive Linguistics*, *32*(3), 349–388. http://dx.doi.org/10.1515/cog-2019-0100.

MacWhinney, B. (2000). *The CHILDES project: tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Maratsos, M. (1985). Semantic restrictions on children's passives. *Cognition*, *19*(2), 167–191. http://dx.doi.org/10.1016/0010-0277(85)90017-4.

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 1192–1202). Brussels, Belgium: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D18-1151.

McCoy, R. T., Min, J., & Linzen, T. (2020). BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In A. Alishahi, Y. Belinkov, G. Chrupała, D. Hupkes, Y. Pinter, & H. Sajjad (Eds.), *Proceedings of the third blackboxNLP workshop on analyzing and interpreting neural networks for NLP* (pp. 217–227). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.blackboxnlp-1.21, URL: https://aclanthology.org/2020.blackboxnlp-1.21.

Messenger, K., Branigan, H. P., McLean, J. F., & Sorace, A. (2012). Is Young children's passive syntax semantically constrained? Evidence from syntactic priming. *Journal of Memory and Language*, *66*(4), 568–587. http://dx.doi.org/10.1016/j.jml.2012.03.008.

Misra, K., & Mahowald, K. (2024). Language models learn rare phenomena from less rare phenomena: the case of the missing AANNs. arXiv:2403.19827.

OpenAI, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A. J., Welihinda, A., Hayes, A., Radford, A., Mądry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., ... Malkov, Y. (2024). GPT-4o system card. http://dx.doi.org/10.48550/arXiv.2410.21276, arXiv:2410.21276.

Papadimitriou, I., Chi, E. A., Futrell, R., & Mahowald, K. (2021). Deep subjecthood: Higher-order grammatical features in multilingual BERT. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume* (pp. 2522–2532). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.eacl-main.215.

Papadimitriou, I., Lopez, K., & Jurafsky, D. (2023). Multilingual BERT has an accent: evaluating english influences on fluency in multilingual models. In A. Vlachos, & I. Augenstein (Eds.), *Findings of the association for computational linguistics: EACL 2023* (pp. 1194–1200). Dubrovnik, Croatia: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.findings-eacl.89.

Patil, A., Jumelet, J., Chiu, Y. Y., Lapastora, A., Shen, P., Wang, L., Willrich, C., & Steinert-Threlkeld, S. (2024). Filtered corpus training (FiCT) shows that language models can generalize from indirect evidence. arXiv:2405.15750.

Perfors, A., Tenenbaum, J. B., & Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, *37*(3), 607–642. http://dx.doi.org/10.1017/S0305000910000012.

Pinker, S. (1989). Learnability and cognition: The acquisition of argument structure, In *Learning, development, and conceptual change*, (1st paperback ed., 4th print ed.). Cambridge, Mass: MIT Press.

Pinker, S., Lebeaux, D. S., & Frost, L. A. (1987). Productivity and constraints in the acquisition of the passive. *Cognition*, *26*(3), 195–267. http://dx.doi.org/10.1016/S0010-0277(87)80001-X.

Postal, P. M. (2004). *Skeptical linguistic essays*. Oxford ; New York: Oxford University Press.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*: *Technical report*, OpenAI.

Regier, T., & Gahl, S. (2004). Learning the unlearnable: The role of missing evidence. *Cognition*, *93*(2), 147–155. http://dx.doi.org/10.1016/j.cognition.2003.12.003.

Reisinger, D., Rudinger, R., Ferraro, F., Harman, C., Rawlins, K., & Van Durme, B. (2015). Semantic proto-roles. *Transactions of the Association for Computational Linguistics*, *3*, 475–488. http://dx.doi.org/10.1162/tacl_a_00152.

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. http://dx.doi.org/10.1016/j.jml.2007.03.002.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928. http://dx.doi.org/10.1126/science.274.5294.1926.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 1904-1920*, *3*(3), 271–295. http://dx.doi.org/10.1111/j.2044-8295.1910.tb00206.x.

Sprouse, J., & Almeida, D. (2017). Design sensitivity and statistical power in acceptability judgment experiments. *Glossa: A Journal of General Linguistics*, *2*(1), http://dx.doi.org/10.5334/gjgl.236.

Theakston, A. L. (2004). The role of entrenchment in children's and adults' performance on grammaticality judgment tasks. *Cognitive Development*, *19*(1), 15–34. http://dx.doi.org/10.1016/j.cogdev.2003.08.001.

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, *3*(1), 1–42. http://dx.doi.org/10.1207/s15473341lld0301_1.

Timkey, W., & Linzen, T. (2023). A language model with limited memory capacity captures interference in human sentence processing. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the association for computational linguistics: EMNLP 2023* (pp. 8705–8720). Singapore: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.findings-emnlp.582.

Tjuatja, L., Neubig, G., Linzen, T., & Hao, S. (2025). What goes into a LM acceptability judgment? Rethinking the impact of frequency and length. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Proceedings of the 2025 conference of the nations of the americas chapter of the association for computational linguistics: human language technologies (volume 1: long papers)* (pp. 2173–2186). Albuquerque, New Mexico: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2025.naacl-long.109.

Tomasello, M. (2000). Do Young children have adult syntactic competence? *Cognition*, *74*(3), 209–253. http://dx.doi.org/10.1016/S0010-0277(99)00069-4.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. ArXiv Preprint abs/2307.09288, URL: https://arxiv.org/abs/2307.09288.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. arXiv:1706.03762 [cs].

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*, http://dx.doi.org/10.1126/science.adi1374.

Warstadt, A., & Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language* (pp. 17–60). CRC Press.

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., & Cotterell, R. (2023). Findings of the babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. In A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, & R. Cotterell (Eds.), *Proceedings of the babyLM challenge at the 27th conference on computational natural language learning* (pp. 1–34). Singapore: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.conll-babylm.1.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, *8*, 377–392. http://dx.doi.org/10.1162/tacl_a_00321.

Wei, J., Garrette, D., Linzen, T., & Pavlick, E. (2021). Frequency effects on syntactic rule learning in transformers. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 932–948). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.emnlp-main.72.

Wilcox, E. G., Hu, M. Y., Mueller, A., Warstadt, A., Choshen, L., Zhuang, C., Williams, A., Cotterell, R., & Linzen, T. (2025). Bigger is not always better: The importance of human-scale language modeling for psycholinguistics. *Journal of Memory and Language*, *144*, Article 104650.

Zwicky, A. M. (1987). Slashes in the passive. *Linguistics*, *25*(4), http://dx.doi.org/10.1515/ling.1987.25.4.639.