Disentangling Statistical Preemption from Entrenchment in Language Models' Avoidance of Overgeneralization

Anonymous ACL submission

Abstract

How do learners avoid overgeneralizations such as *Tom laughed me* without explicit negative evidence? Constructionists have posited two proposals that describe indirect negative evidence against overgeneralizations: preemption (which privileges exposure to near-synonymous construction—e.g., *she made him laugh*) vs. entrenchment (all exposures to a verb's grammatical usages, including cases like *He laughed*). We disentangle these hypotheses by running controlled rearing experiments on LMs trained on child-caregiver conversations, where we systematically remove preemptive vs. non-preemptive evidence. We find that while LMs avoid overgeneralizations, they do not show preemption at a verb-specific level, instead showing weak but non-zero evidence of abstract preemption. Combined with results from analyzing the LMs' training dynamics, we find that LMs treat competing structures as indirect positive—as opposed to negative—evidence in the verb-specific condition. Insofar as preemption is the more plausible route to avoiding overgeneralizations in humans, our results point the need for there to be sensitivities to indirect negative evidence in neural network learners, and suggest new human experiments to test abstract preemption.

1 Introduction

One of the hallmarks of human cognition is our ability to generalize productively. For instance, on hearing (1a), competent speakers of English use their experience of intransitive and transitive verbs (*break, roll*) to effortlessly produce (1b). But their generalization of this pattern is also constrained—e.g., while we are perfectly fine with (2a), we do not produce (2b).

(1) a. The water is boiling.
b. Dad boiled some water.

(2) a. The baby laughed.
b. * Mom laughed the baby.

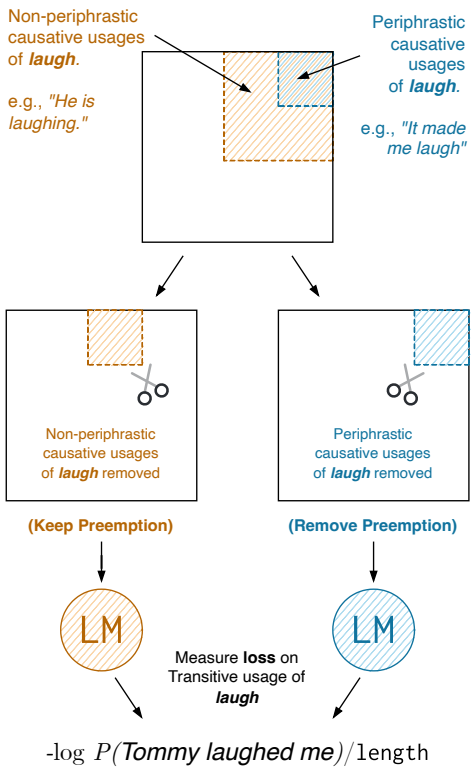

Figure 1: Overview of our experimental paradigm. For a given verb (*laugh*), we train LMs on two versions of our corpus, created by removing: 1) the preemptive evidence for that verb; and 2) same amount of non-preemptive evidence. We then compare the behavior of these LMs to disentangle preemption from entrenchment in explaining that verb's overgeneralization.

Children often tend to *overgeneralize* the above pattern by indeed producing sentences like (2b) (Bowerman, 1988; Braine and Brooks, 1995; Akhtar and Tomasello, 1997; Brooks and Tomasello, 1999, i.a.), but eventually *retreat* and constrain their generalizations. They do so *without* much explicit negative evidence (Baker, 1979)—caregivers rarely ever correct children's grammatical errors systematically. How, then, are such constraints and restrictions learned?

In the realm of constructionist and usage-based

approaches to language acquisition (Goldberg, 1995; Tomasello, 2003; Goldberg, 2005; Rowland, 2013), there are two closely related proposals that attempt to explain this avoidance of overgeneralizations: *preemption* and *entrenchment*. Preemption suggests that given two near-synonymous constructions A and B, exposure to a verb used in A in contexts where the verb could have been used in either construction serves as evidence *against* the verb's usage in B (Goldberg, 1995; Brooks and Zizak, 2002; Boyd and Goldberg, 2011). In the case of *laugh*, preemption predicts that the transitive **Mom laughed the baby* is *preempted* by the periphrastic causative usage *Mom made the baby laugh*, because in situations where *X* did something that caused *Y* to *laugh*, the utterance *X made Y laugh* was used instead of *X laughed Y*. Entrenchment, on the other hand, does not privilege near-synonymous constructions or specific contexts (Theakston, 2004; Stefanowitsch, 2008; Ambridge et al., 2008). Here, the usage of the verb in *any* attested construction is evidence against the unattested usage—entrenchment not only considers periphrastic causative usages of *laugh* in blocking its transitive usage, but also other constructions like the intransitive (as in (2a)).

There are more similarities between these proposals than there are differences: 1) both explanations rely on *indirect* negative evidence, since they posit *other* constructions as evidence against overgeneralization; 2) both are statistical in nature, since they rely on the frequency of exposure to the indirect evidence—more frequent this exposure is, stronger is the blocking effect; and most importantly, 3) entrenchment is a strict superset of preemption, since the evidence that falls under the purview of preemption (near-synonymous construction) is always part of the evidence considered by entrenchment. The latter point explains why attempts to disentangle the two proposals in the laboratory (Brooks and Zizak, 2002; Ambridge et al., 2015, 2018; Bidgood et al., 2021) often suffer from multi-collinearity issues in both variables. In these studies, participants' acceptability judgments on unconventional forms are predicted using the frequencies of the preemptive usages of a verb (preemption) as well as that of the total frequency of the verb (entrenchment) sourced from openly available corpora, rather than the learners' own experience (which is unobtainable for humans). Therefore, apart from multicollinearity (due to the overlap in preemptive and entrenching evidence), these experiments and analysis are also unable to shed light on the causal status of these proposals.

To adjudicate between these two proposals, we conduct "controlled rearing" (Frank, 2023; Misra and Mahowald, 2024) experiments on Language Models (LMs), where we systematically manipulate the learning experience of LMs to disentangle between preemption and entrenchment. As a case study, we apply our methods to explaining transitive overgeneralizations of (usually) intransitive verbs such as *laugh, smile, cry*, etc., as this is among the quintessential examples of entrenchment and preemption analyses over the past 30 years (Akhtar and Tomasello, 1997; Ambridge et al., 2015; Bidgood et al., 2021). While analyses on LMs cannot *directly* be taken to have implications for humans (cf. Warstadt and Bowman, 2022; Portelance and Jasbi, 2024; Misra and Kim, 2026; Futrell and Mahowald, 2025), such experiments allow us to conclude about the viability of the two proposals and how (and whether) they might arise from the general purpose statistical learning of the kind employed by LMs. Controlled rearing allows us to do this by considering counterfactual learning scenarios—e.g., "what if there was no preemptive evidence in the learner's training data?" Evidence of a privileged role of preemptive evidence would suggest that LMs acquire restraints on their generalization behavior from fine-grained sensitivities to meaning, since preemption privileges contexts where either construction could have been used to communicate the same intent or semantics. More broadly, this allows us to understand if similar constraints arise within LMs trained on cognitively plausible amounts of language-only exposure, or if explicit evidence of semantics is needed (Abend et al., 2017; Yedetore and Kim, 2024).

Our methods and research questions contribute to the burgeoning program of "controlled rearing" (Frank, 2023; Misra and Mahowald, 2024) or "filtered corpus training" (Patil et al., 2024), which has established itself as an important paradigm in understanding the role of input in shaping a statistical learner's generalizations. Most applications of this method have primarily demonstrated how LMs generalize to "novel" usages in the absence of direct or impoverished evidence (Jumelet et al., 2021; Misra and Mahowald, 2024; Patil et al., 2024; Yao et al., 2025; Yang et al., 2026). That is, they only focus on indirect *positive* evidence, since they investigate what facilitates generalization, and not what *blocks* it. An exception to this is work by

Leong and Linzen (2026), who focus on explicitly understanding signals that *block* models' usage of a structure, but focus on a different construction (active/passive), and do not consider the difference between preempting and non-preempting evidence (i.e., they consider entrenchment as a whole). This positions our work uniquely within this literature, since we aim to causally disentangle preemption and entrenchment, and explicitly measure their viability as sources of indirect *negative* evidence.

1.1 Summary of Findings

Before testing what explains models' avoidance of overgeneralization, we first measure if they learn the appropriate generalizations in the first place—i.e., finding transitive usages of intransitive verbs less probable than their periphrastic causative counterparts. Finding that they do (§3) for six different verbs, we then run our controlled rearing studies to disentangle preemption from entrenchment (§4). Here, we test if preemptive usages of a given verb (i.e. their periphrastic causative usages like in *Tom made me laugh*) have a privileged role over non-preemptive ones (e.g., *I laughed.*), by training LMs on corpora where each of them are removed in equal amounts separately. Using LMs' negative log-likelihood loss (NLL) on the unconventional, unseen transitive usages of these verbs as our measure of overgeneralization, we considered two different manipulation types: 1) **verb-specific**, where for a verb only its own usages were manipulated; and 2) **abstract**, where usages of all verbs were manipulated, to quantify if the evidence of preemption comes from higher order sources.

We do not find clear evidence for preemption in the verb-specific condition—LMs' NLL of the unconventional transitive usage was greater when preemptive exposures were removed than when they were kept in. By contrast, preemption would predict that in the absence of preemptive exposures, models should find transitive usages *more* likely. The abstract condition, however, did show relatively larger effects of preemption, though in absolute terms these were generally weak. This finding is consistent with the hypothesis that models treat the periphrastic causative construction as indirect positive evidence, common in other controlled rearing studies. That is, the models might have inferred the relation between transitive and periphrastic causatives via other verbs for which both are acceptable (e.g., *She boiled the water* vs. *She got the water to boil*), which explains why this is not observed when all periphrastic causatives are removed (in the abstract condition).

To further explore this finding of positive—as opposed to negative—evidence, we investigated the change in the model's behavior on unseen transitive and periphrastic causative sentences before and after a periphrastic causative usage was encountered by the model during training. We find the model's generalization behavior on both constructions to be highly correlated, further supporting the indirect positive evidence hypothesis—the presence of periphrastic causatives is tightly associated with the likelihood of transitive usages.

Insofar as preemption occurs and drive retreat from overgeneralization in humans (Samara et al., 2025), our verb-specific results indicate the need for LM learners to model fine-grained indirect negative evidence. We speculate that this could come from more explicit semantic signals about the context in which either construction is used, or from morphological constraints, where the concept of preemption originates from (Kiparsky, 1973; Aronoff, 1976). At the same time, our finding of abstract preemption motivates new experiments that test this finding in humans via an artificial language learning design, since preemption for argument structure alternations is often treated as a verb-specific effect (Goldberg, 2005; Ambridge et al., 2015, 2018; Bidgood et al., 2021). Our code and data are available at (url).

2 Data, Model, and Methods

2.1 Training Data

To remain as close as possible to the input and constraints available to children, we use the English subset of the Child Language Data Exchange System (CHILDES) dataset (MacWhinney, 2000) as the training dataset for our models. CHILDES consists of human-transcribed conversations between children and caregivers. The corpus consists of approximately 5M utterances, amounting to a total of 26M space-delimited words. Importantly, CHILDES also comes with morphological and syntactic annotations.[1] This allows us to precisely detect phenomena we are interested in, avoiding any potential errors from existing automatic tools (cf. Yang et al., 2025; Padovani et al., 2026).

As a preprocessing step, we lemmatize the corpus and strip away morphological inflections. This allows models to represent tense-inflected versions

[1] https://talkbank.org/0info/manuals/CHAT.pdf

of verbs using the same token (*laughed* → *laugh*, *laughing* → *laugh*), thereby giving them the maximum chance—if any—to use inflected forms of the same verb as collective evidence.[2] All subsequent experimental data, including evaluation, use the lemmatized sentences.

2.2 Model

We use a word-level tokenizer for the lemmatized training data, with a vocabulary size of 37,482. Our learners are autoregressive LMs with the GPT-2 architecture (Radford et al., 2019). Each LM has 12 layers and 12 attention heads per layer, with a hidden size of 768, resulting in a total of 115M parameters. Each model is trained for 3 epochs, and training is repeated with 5 random seeds to marginalize nondeterminism, including parameter initialization and training data shuffling.

2.3 Disentangling Preemption from Entrenchment

Because preemption is necessarily a subset of entrenchment (see §1), disentangling the two boils down to testing whether the hypothesized preemptive evidence has a *privileged* role compared to those not considered as part of preemption. This effectively partitions the set of usages of a particular verb to preemptive and non-preemptive. In the context of transitive overgeneralizations (*she laughed me*), the preemptive evidence is the periphrastic causative construction (*she made me laugh*), and the non-preemptive ones are all other types of usages (*he is laughing, she laughed at me,* etc.). This give us two kinds of manipulations for our controlled rearing experiments: 1) **REMOVE-P**, where we remove *all* occurrences of the periphrastic causative construction usages for a given verb (or a set of verbs); and 2) **KEEP-P** (or REMOVE-NON-P), where we remove the same number of occurrences of a given verb (or set of verbs) as in REMOVE-P, that *do not* use it in the periphrastic causative construction.

The raw frequency of a verb of interest in both cases is reduced by the same amount, but the structures being manipulated are different. Insofar as models assign a special role to preemptive evidence, we should expect them to retreat further from overgeneralizations in KEEP-P, which retains the preemptive evidence, than in REMOVE-P. In other words, the model should find transitive overgeneralizations more likely under REMOVE-P than under KEEP-P, since the hypothesized critical evidence against transitive usages is absent in REMOVE-P.

[2]We show results with no lemmatization in Appendix A.

Verb	PC Usage	Verb Usage
go	1295	316515
laugh	99	1781
cry	49	6166
sleep	33	8614
sneeze	31	461
smile	15	724

Table 1: Usage statistics of the verbs of interest in the training set, including the number of times each word appears in a Periphrastic Causative (PC) construction, and the total number of times the word is used as a verb.

In our experiments, we primarily focus on six verbs that have been described to be 'intransitive-only' by Levin (1993): *go, laugh, cry, sleep, sneeze,* and *smile*. Using them transitively almost always constitutes an overgeneralization.[3] Table 1 shows the usage statistics of these verbs in CHILDES, both as a verb in general, and also more specifically in the periphrastic causative construction. The periphrastic usages of these verbs are detected using the gold-dependency parses provided in the CHILDES dataset. A periphrastic causative construction is identified if the matrix verb (e.g., *make*, *let*, etc.) has a complement child (COMP) and a patient in the dependency parse. Technical details about the identification of periphrastic causative constructions can be found in Appendix B.

2.4 Measuring Overgeneralization Effects

Evaluation Data Our evaluation data constitutes a pair of transitive and periphrastic causative sentences containing the abovementioned verbs and a set of nouns that occur in the CHILDES dataset. We construct lemmatized versions of these sentences by using the template "S V O" (e.g., *Tommy laugh Mary*) for the transitive structure, and "S make O V" (e.g., *Tommy make Mary laugh*) for the periphrastic causative structure. Here, S, O, and V are the subject, object, and the verb lemma, respectively, such that none of the sentences occur in CHILDES. In total, we generate an average of 70 pairs for each verb. We denote transitive sentences as s^-, and periphrastic causative ones as s^+.

[3]There can be creative, usages of these verbs that appear transitive—e.g., in the CAUSED-MOTION construction: *She sneezed the foam off the cappuccino* (Goldberg, 2005), but these do not have the same meaning as the causative context—the agent (she) is not making the object (foam) sneeze.

Measure We measure the overgeneralization effect for a given verb as the average negative log-likelihood loss (NLL) of the transitive sentences for that verb in our evaluation data. Specifically, given an LM $P^{\mathcal{D}}$, trained on a corpus $\mathcal{D}$, and a transitive sentence, $s^- = x_1^- \dots x_{|s^-|}^-$, we measure:

$$\ell(s^-; P^{\mathcal{D}}) = -\frac{1}{|s^-|} \sum_{i=1}^{|s^-|} \log P^{\mathcal{D}} \left(x_i^- \mid x_{<i}^- \right)$$

This allows us to compare models trained on corpora with different amounts of indirect negative evidence in terms of how (un)likely they find transitive usages of a verb. If the evidence contained in a corpus $\mathcal{D}_1$ better suppresses the overgeneralization of a verb than does the evidence in $\mathcal{D}_2$, then we expect to observe a higher NLL on $P^{\mathcal{D}_1}$ than on $P^{\mathcal{D}_2}$, i.e., $\ell(s^-; P^{\mathcal{D}_1}) > \ell(s^-; P^{\mathcal{D}_2})$.

It might be tempting to treat overgeneralization effect as the *preference* of the periphrastic causative evaluation sentence s^+ over its overgeneralized transitive form s^-, by measuring the difference between their NLLs. Afterall, these two constructions are hypothesized to compete (Goldberg, 2005). However, we argue that this metric does not allow us to explain overgeneralization effects, which are explicitly concerned with what allows the learner to conclude that the unattested form (here, transitive) is unacceptable. This is because our experiments (as we will show in §4.1) directly manipulate the presence of the periphrastic causative exposures available to the model. Therefore, changes to this alternate measure could be completely driven by the likelihood of the conventional, periphrastic causative sentence (whose exposure has been manipulated) rather than the unlikelihood of the transitive counterpart. Indeed, such a measure has been used in previous work (Guo et al., 2026), but due to the above reason, it can paint an illusory picture of preemption (see more in Appendix E and Appendix F where we compare our results to the one had we used this alternate measure).

3 Preconditions to Not Overgeneralizing

Before analyzing how indirect negative evidence might facilitate our LMs to avoid overgeneralization, we first establish if the model trained on the full CHILDES corpus learns the appropriate generalization behavior—that the transitive usages of our target verbs is indeed unacceptable, relative to their periphrastic causative usages. This is the minimal precondition the model must satisfy before we can use it to conclude about the factors that affect its generalization, following Misra and Kim (2026).

To test this criterion, we compare the NLL of the transitive evaluation sentences (i.e., the unconventional forms) to their periphrastic causative counterparts. Insofar as the model has learned the right generalizations, we expect this measure to be a positive value, since the loss on the transitive should be greater than that on the periphrastic causative, since the latter is more acceptable. We measure this across training steps (every 50 batches), and show results per verb in Figure 2. For all verbs, we observe positive values in the majority of the cases throughout the course of training. This makes our model well-positioned to answer questions about the factors that lower or increase its (un)likelihood on the unconventional form, since the model does, in fact, treat it as relatively unconventional.

4 Verb-specific and Abstract Preemption

Having shown that the models indeed learn the appropriate generalization, we now turn to disentangling preemption from entrenchment. We compare the role of preemptive and non-preemptive evidence in the model's behavior on the unconventional, overgeneralized, transitive usages of our target verbs. That is, we compare models trained on manipulations specified by KEEP-P vs. REMOVE-P.

4.1 Manipulations

We consider different levels at which preemption (or non-preemption) might affect model behavior:

Verb-specific This condition specifies the canonical form of preemption (Goldberg, 2005; Ambridge et al., 2015, 2018; Bidgood et al., 2021). Here, preemption for a given verb is only measured with respect to the exposure of the learner to the preemptive evidence for that particular verb, regardless of the learner's experience with usages of other verbs. Therefore, KEEP-P and REMOVE-P in this condition are only concerned with the periphrastic causative usages of a given verb at a time. This results in a total of 60 manipulations, resulting in 5 random seeds, 2 manipulation types, and 6 verbs.

Abstract The abstract condition considers the effect of other verbs' preemptive usages in the learner's experience. Here, we measure the effect of the whole category that is denoted by the structure specified as the preemptive evidence. That is,

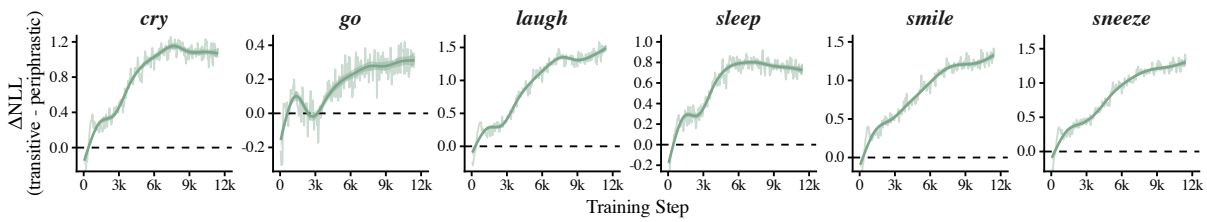

Figure 2: Difference in NLL of transitive and periphrastic causative sentences from our evaluation dataset at every 50 steps during training. Smoothed curves are estimated using a generalized additive model (Hastie and Tibshirani, 1986). Across all verbs, models show greater loss on transitive sentences than on their periphrastic causative counterparts, suggesting that they have acquired the right preference for these verbs.

for REMOVE-P, we remove periphrastic causative usages of *all* verbs (N=7,452), and for KEEP-P, we remove equivalent amounts of (randomly sampled) non-periphrastic causative usages of all verbs. This results in a total of 10 manipulations, resulting from 5 seeds and 2 conditions.

4.2 Results and Analysis

We compute the NLLs (as described in §2.4) for the transitive-sentence subset of our evaluation dataset across all our manipulations. We then analyze our results using a linear mixed-effects regression (Bates et al., 2015), using the NLL as our dependent variable, the target verb (`verb`) and the training set variant ($\mathcal{D}$, KEEP-P vs. REMOVE-P) and their interaction as fixed effects, and the random seed (`seed`) and evaluation sentence (s_i^-) as random effects. This is specified by the following equation:

$$\ell(s_i^-) \sim \texttt{verb} \times \mathcal{D} + (1 \mid s_i^- : \texttt{seed})$$

We compute this separately for the verb-specific and abstract conditions, and analyze the contrast between KEEP-P and REMOVE-P across verbs. Insofar as the model shows an effect of preemption, we should expect this contrast to be positive—i.e., the average NLL for REMOVE-P should be lower than that of KEEP-P. This is because the preemptive evidence is not present in REMOVE-P, so the unconventional form should no longer be blocked by the model and therefore should be likely.

Figure 3 shows our results across verbs and manipulation conditions. For the verb-specific condition, instead of seeing a privileged (positive) effect of preemption, we instead see largely negative effects across verbs. That is, the models' NLL on the unconventional, transitive usages for our target verbs is generally *lower* when the hypothesized preemptive evidence is removed, than when it is kept. This is in direct opposition to what is specified by the preemption proposal (Goldberg, 1995; Brooks and Zizak, 2002; Boyd and Goldberg, 2011; Ambridge et al., 2015).

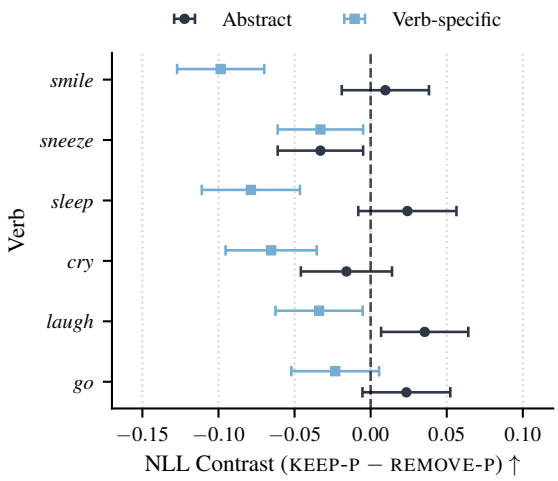

Figure 3: Contrast of estimated marginal means between KEEP-P and REMOVE-P on the linear mixed-effects model, with 95% confidence intervals, across verb-specific (blue, square) and abstract (black, circle) conditions. Positive values indicate preemption.

For the abstract condition, we see relatively stronger effects of preemption than in the verb-specific case, though in absolute terms these are generally weak, with the exception of *laugh* (where the contrast is greater than 0).[4] That is, our LMs are more affected by the *accumulated* indirect negative evidence from periphrastic causative usages of *all* verbs than by the evidence only for a single verb at a time. This suggests that preemption effects (if any) are more likely to be an abstract effect rather than item-specific, at least from the statistics contained in CHILDES.

[4] In our results on models trained on data *without* lemmatization, we find even stronger effects of abstract preemption across 5/6 verbs (though not for verb-specific conditions).

4.3 Interim Discussion

Our verb-specific results from the previous experiment suggest that CHILDES-trained LMs find transitive usages of (generally) intransitive verbs (like *cry* and *smile*) *less*—as opposed to *more*—likely when their periphrastic causative usages are removed. That is, the models do not treat the periphrastic causative as an indirect negative evidence.

A plausible explanation of why this might be the case comes from earlier controlled rearing results, where models were shown to be sensitive to indirect *positive* evidence. For instance, Misra and Mahowald (2024) show that knowledge of constructions such as *a beautiful five days* can come from other cases where a plural measure noun phrase (*6 months*) is construed as a singular unit (e.g., *6 months* ***is*** *all I need...*) even in the absence of direct evidence (see also Jumelet et al., 2021; Patil et al., 2024; Yao et al., 2025). Therefore, instead of encoding the *competition* between periphrastic causative and transitive constructions, the models might only be encoding their similarity.[5] This can arise from verbs such as *tickle, boil, melt, move, grow*, etc. that *can* be used in both constructions. This is perhaps why we see a much reduced effect of non-preemption (i.e., relatively more preemption) in the abstract condition, where REMOVE-P effectively removes evidence that the transitive and periphrastic causative can be related.

We will explore this hypothesis of indirect *positive*—and not *negative*—evidence, by turning to a training dynamics analysis in the next section.

5 Prospective Moments of Indirect Negative (or Positive?) Evidence

In this section, we explore model generalization behavior at an extremely fine-grained level, by testing it at specific time-steps during pretraining. Specifically, we identify 'Prospective Moments of Indirect Negative Evidence', which we define as training batches that contain constructions that may serve as the hypothesized negative evidence against overgeneralization. On identifying such cases, we then measure the model's behavior before and after encounter with a batch of interest, to quantify its effect (given encounters with previous batches) on the model's overgeneralization behavior.

[5] It is important to note that encoding similarity is a precursor to encoding competition—the learner must recognise what structures are similar before concluding that they also compete (Suttle and Goldberg, 2011; Goldberg, 2016).

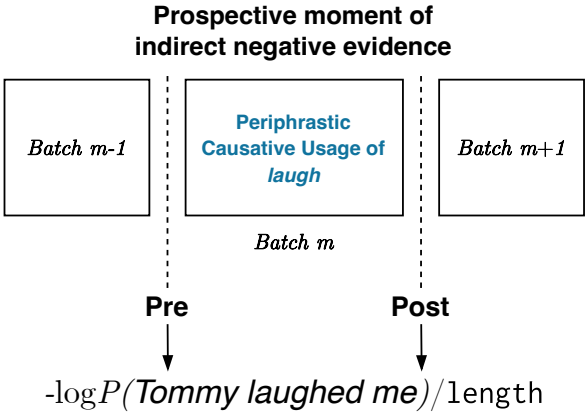

Figure 4: Depiction of a *Prospective Moment of Indirect Negative Evidence*. We identify batches that contain the periphrastic causative usage of a verb (here, *laugh*) and then measure the NLL of an evaluation sentence before (pre) and after (post) encountering the batch.

In the context of the previous experiment's findings, we specifically narrow in on batches containing the periphrastic causative usages of verbs. This is because, according to preemption, these were meant to be sources of indirect negative evidence, but our results suggest that the opposite is true. That is, models trained on corpora where this evidence was removed also found transitives to be *less* likely than when this evidence was not removed. This suggests that these moments might in fact be supplying models with indirect *positive* evidence that the two structures are similar.

To shed light on this hypothesis, we measure the correlation between the NLL on our evaluation data's periphrastic causative and transitive usages for each verb, using the models (5 different seeds) trained on the full CHILDES dataset. Specifically, for a batch containing a given verb's usage in the periphrastic-causative construction, we compute the difference between a model's average NLL on an evaluation set before and after encountering that batch. We do this for both periphrastic causative and transitive evaluation sentences. We then measure the Pearson's correlation and the pairwise agreement between the NLL differences for both constructions. Agreement is measured as the percentage of time both values had the same sign. A positive correlation and high agreement would indicate that the losses on both structures change in the same way upon encounters with a periphrastic causative usage during training. For reference, according to preemption, we should instead expect a negative correlation and agreement—the loss on

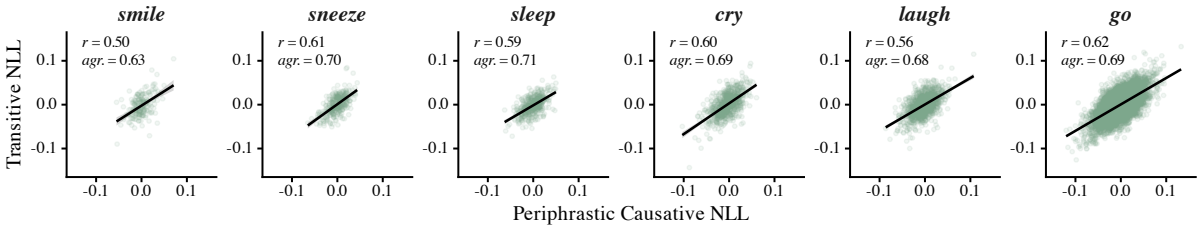

Figure 5: Statistical relationship between transitive and periphrastic causative NLLs, measured at every batch during training that contains a periphrastic causative exposure (indicated by individual points) to the verb of interest.

transitives should increase with exposure to a periphrastic causative, whereas loss on periphrastic causatives should decrease.

Figure 5 shows our results. We see generally high correlation and agreement scores across all verbs, corroborating the indirect positive evidence hypothesis. That is, on encountering a periphrastic causative usage of a verb during training, the change in model's behavior on both periphrastic causatives *and* transitives are positively correlated.

6 Implications and Speculations

We find that at least at the verb-specific level, LMs trained on CHILDES treat the periphrastic causative construction as indirect evidence *for*, and not *against*, transitive usages of intransitive verbs. Below we discuss the broader implications and speculations of these findings:

Semantic cues against overgeneralization Preemption as a mechanism attributes a more special role to the communicative intent of contexts in which the preemptive evidence is used as compared to entrenchment. Models' failure to show effects of preemption in the verb-specific condition shows that the language-only cues in CHILDES are not sufficient to give rise to this sensitivity in models. But this also raises the question about whether providing explicit semantic evidence of events could facilitate preemption in LMs. This could come from environmental annotations—cues to what is going on in the child's environment at the time of the utterance—that are made available in CHILDES (but were not used in our work, though see Wu et al. (2025), who did use them). This could also be induced through more explicit means, like in Yedetore and Kim (2024), who showed evidence of hierarchical generalization in transformers when semantic signals (in the form of semantic parses) were available during training.

Morphological origins of preemption While we investigated preemption for argument structure constructions in this work, the phenomenon traces its origins to morphological 'blocking' (Aronoff, 1976; Kiparsky, 1973; Clark, 1987). For example, *children* preempts *childs* or *went* preempts *goed*. Availability of such a phenomenon during learning could be the key to preemption. Insofar as this is true, a hypothesis for why we might not see strong effects of preemption in our LMs is that tokens like *childs* and *goed* are simply not represented in neither our BPE nor our lemmatized variants, and so the model is never preempted in the morphological sense. This motivates serious consideration of morphologically-informed tokenizers in the development of LMs.

Preemption from abstract cues While we do not observe preemption in our verb-specific experiments, we do see evidence of more abstract signals to preemption via exposures from other verbs, weak though they might be in absolute terms. In the context of human language acquisition, preemption has always been treated as a verb-specific phenomenon (Goldberg, 2005; Ambridge et al., 2015, 2018; Bidgood et al., 2021). Our results suggest the possibility of preemptive signals coming from more accumulated sources of indirect evidence—e.g., transitive usages of *laugh* could come from periphrastic causative usages of *laugh*, combined with those from other verbs (like *giggle*, *smile*, etc.). This motivates *novel* human experiments, perhaps via artificial language learning paradigms (e.g., Samara et al., 2025), to test if this is a viable route for human learners.

Overall, our work joins several others in understanding how the input to a general purpose statistical learner like a transformer LM shapes its generalization, and contributes insight about how subtly different proposals for the indirect sources of generalizations can be disentangled.

Limitations

Single Phenomenon Due to the number of pretraining experiments, we focus on a single phenomenon, although it is among the quintessential examples of this debate which has lasted around 30 years (Akhtar and Tomasello, 1997; Brooks and Tomasello, 1999; Brooks and Zizak, 2002; Theakston, 2004; Ambridge et al., 2008, 2015, 2018; Bidgood et al., 2021). It is possible that models show different generalization behaviors on different constructions. Our methodology, and the methodology of controlled rearing in general can be used to adjudicate between preemption and entrenchment across several different constructions in the future.

Generalization of our conclusion across models Our analyses have only looked at one instance of the entire class of transformer LMs that can be trained on CHILDES, finding lack of evidence for preemption at the level of individual verb exposures. However, this does not necessarily mean that it cannot arise in transformers or neural networks in general. In fact, there is a connectionist model that does show effects of both preemption and entrenchment (Ambridge and Blything, 2016)—however its architecture explicitly bakes in hand-coded semantic features which have to be supplied by the researcher for a given input. That is, this model is at a different level of analysis and theoretical commitment than ours, and cannot learn from naturalistic data. Furthermore, it would also be worthwhile to understand preemption from a theory of computation perspective, in order to provably analyze preemption in classes of transformers (or other types of neural network models). Finally, it is possible for these effects to arise with larger datasets, e.g., the BabyLM challenge (Choshen et al., 2026).

Operationalizing preemptive signals We have assumed that all instances of the competing construction constitute cases where either construction could have been used. But that need not be true. More generally, there is a need for a methodology to detect the possibility of competition vs. non-competition from raw data. Additionally, it might be too narrow of an assumption to treat only near-synonymous constructions as preemptive (Ambridge et al., 2015). For instance, in a scenario where something (a joke) makes someone laugh (John), it might be perfectly ok to simply use the intransitive, *John laughed*. In this sense, the intransitive could have preempted the transitive usage. Therefore, future work could address this by running controlled rearing on specific cases where constructions necessarily compete vs. do not. Finally, what counts as a "semantic context" can itself differ. For instance, in Misra and Kim (2026), the features of the constructional exposure itself ended up showing an effect of preemption. Their case focused on the dative alternation (DO: *John gave me the book* vs. PO: *John gave the book to me*; Levin, 1993) and found that the information structure of the theme (the book) and recipient (me) predicted the degree of preemption. Future work can hopefully consolidate the many different ways preemption can be cued to learners.

References

Omri Abend, Tom Kwiatkowski, Nathaniel J Smith, Sharon Goldwater, and Mark Steedman. 2017. Bootstrapping language acquisition. *Cognition*, 164:116–143.

Nameera Akhtar and Michael Tomasello. 1997. Young children's productivity with word order and verb morphology. *Developmental Psychology*, 33(6):952.

Ben Ambridge, Libby Barak, Elizabeth Wonnacott, Colin Bannard, and Giovanni Sala. 2018. Effects of Both Preemption and Entrenchment in the Retreat from Verb Overgeneralization Errors: Four Reanalyses, an Extended Replication, and a Meta-Analytic Synthesis. 4(1):23.

Ben Ambridge, Amy Bidgood, Katherine E Twomey, Julian M Pine, Caroline F Rowland, and Daniel Freudenthal. 2015. Preemption versus Entrenchment: Towards a Construction-General Solution to the Problem of the Retreat from Verb Argument Structure Overgeneralization. *PloS One*, 10(4):e0123723.

Ben Ambridge and Ryan P Blything. 2016. A connectionist model of the retreat from verb argument structure overgeneralization. *Journal of Child Language*, 43(6):1245–1276.

Ben Ambridge, Julian M Pine, Caroline F Rowland, and Chris R Young. 2008. The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgements of argument-structure overgeneralization errors. *Cognition*, 106(1):87–129.

Mark Aronoff. 1976. Word formation in generative grammar. *Linguistic Inquiry Monographs Cambridge, Mass*, (1):1–134.

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

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67:1–48.

Amy Bidgood, Julian Pine, Caroline Rowland, Giovanni Sala, Daniel Freudenthal, and Ben Ambridge. 2021. Verb argument structure overgeneralisations for the English intransitive and transitive constructions: Grammaticality judgments and production priming. 13(3):397–437.

Melissa Bowerman. 1988. The 'no negative evidence' problem: How do children avoid constructing an overly general grammar? In *Explaining Language Universals*, pages 73–101. Basil Blackwell.

Jeremy K Boyd and Adele E Goldberg. 2011. Learning What NOT to Say: The Role of Statistical Preemption and Categorization in A-Adjective Production. *Language*, 87(1):55–83.

Martin DS Braine and Patricia J Brooks. 1995. Verb argument structure and the problem of avoiding an overgeneral grammar. In *Beyond names for things: Young children's acquisition of verbs*, pages 353–376. Psychology Press.

Patricia J Brooks and Michael Tomasello. 1999. How children constrain their argument structure constructions. *Language*, pages 720–738.

Patricia J Brooks and Otto Zizak. 2002. Does preemption help children learn verb transitivity? *Journal of Child Language*, 29(4):759–781.

Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Jaap Jumelet, Tal Linzen, Aaron Mueller, Suchir Salhan, Raj Sanjay Shah, Alex Warstadt, and Ethan Gotlieb Wilcox. 2026. Babylm turns 4 and goes multilingual: Call for papers for the 2026 babylm workshop. *arXiv preprint arXiv:2602.20092*.

Eve V Clark. 1987. The principle of contrast: A constraint on language acquisition. In Brian MacWhinney, editor, *Mechanisms of Language Acquisition*. Hillsdale.

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

Richard Futrell and Kyle Mahowald. 2025. How Linguistics Learned to Stop Worrying and Love the Language Models. *Behavioral and Brain Sciences*, pages 1–98.

Adele E Goldberg. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Adele E Goldberg. 2005. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press.

Adele E Goldberg. 2016. Partial productivity of linguistic constructions: Dynamic categorization and statistical preemption. *Language and cognition*, 8(3):369–390.

Dongxin Guo, Jikun Wu, and SM Yiu. 2026. Do Language Models Know What Not to Say? Causal Evidence for Statistical Preemption in LLMs. In *30th Conference on Computational Natural Language Learning*.

Trevor Hastie and Robert Tibshirani. 1986. Generalized Additive Models. *Statistical Science*, 1(3):297–310.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Jaap Jumelet, Milica Denic, Jakub Szymanik, Dieuwke Hupkes, and Shane Steinert-Threlkeld. 2021. Language models use monotonicity to assess NPI licensing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4958–4969, Online. Association for Computational Linguistics.

Paul Kiparsky. 1973. "Elsewhere" in Phonology. In Stephen R. Anderson and Paul Kiparsky, editors, *A Festschrift for Morris Halle*, pages 93–106. Holt, Rinehart, and Winston, New York.

Cara Su-Yi Leong and Tal Linzen. 2026. Manipulating language models' training data to study syntactic constraint learning: The case of English passivization. *Journal of Memory and Language*, 149:104751.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

Ziqiao Ma, Shuyu Wu, Yixuan Wang, Joyce Chai, and Freda Shi. 2025. Trabank: A toolkit for computational developmental studies in language models. https://github.com/Mars-tin/PyChildes.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3rd edition. Lawrence Erlbaum Associates, Mahwah, NJ.

Kanishka Misra and Najoung Kim. 2026. A systematic framework for generating novel experimental hypotheses from language models. *arXiv preprint arXiv:2408.05086*.

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 913–929, Miami, Florida, USA. Association for Computational Linguistics.

Francesca Padovani, Xiulin Yang, Bastian Bunzeck, Jaap Jumelet, Yevgen Matusevych, Nathan Schneider, and Arianna Bisazza. 2026. CAIT: A Syntactic Parsing Toolkit for Child-Adult InTeractions. *arXiv preprint arXiv:2605.19718*.

Abhinav Patil, Jaap Jumelet, Yu Ying Chiu, Andy Lapastora, Peter Shen, Lexie Wang, Clevis Willrich, and Shane Steinert-Threlkeld. 2024. Filtered corpus training (FICT) shows that language models can generalize from indirect evidence. *Transactions of the Association for Computational Linguistics*, 12:1597–1615.

Eva Portelance and Masoud Jasbi. 2024. The roles of neural networks in language acquisition. *Language and Linguistics Compass*, 18(6):e70001.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, page 9.

Caroline Rowland. 2013. *Understanding Child Language Acquisition*. Routledge.

Anna Samara, Elizabeth Wonnacott, Gaurav Saxena, Ramya Maitreyee, Judit Fazekas, and Ben Ambridge. 2025. Learners restrict their linguistic generalizations using preemption but not entrenchment: Evidence from artificial-language-learning studies with adults and children. *Psychological Review*, 132(1):1.

Anatol Stefanowitsch. 2008. Negative entrenchment: A usage-based approach to negative evidence. *Cognitive Linguistics*, 19(3):513.

Laura Suttle and Adele E Goldberg. 2011. The partial productivity of constructions as induction. *Linguistics*, 49(6):1237–1269.

Anna L Theakston. 2004. The role of entrenchment in children's and adults' performance on grammaticality judgment tasks. *Cognitive Development*, 19(1):15–34.

Michael Tomasello. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Alex Warstadt and Samuel R Bowman. 2022. What Artificial Neural Networks Can Tell Us about Human Language Acquisition. *Algebraic Structures in Natural Language*, page 17.

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

Shuyu Wu, Ziqiao Ma, Xiaoxi Luo, Yidong Huang, Josue Torres-Fonseca, Freda Shi, and Joyce Chai. 2025. The mechanistic emergence of symbol grounding in language models. *arXiv preprint arXiv:2510.13796*.

Xiulin Yang, Arianna Bisazza, Nathan Schneider, and Ethan Gotlieb Wilcox. 2026. A unified assessment of the poverty of the stimulus argument for neural language models. *arXiv preprint arXiv:2602.09992*.

Xiulin Yang, Zhuoxuan Ju, Lanni Bu, Zoey Liu, and Nathan Schneider. 2025. UD-English-CHILDES: A collected resource of gold and silver Universal Dependencies trees for child language interactions. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 52–58, Ljubljana, Slovenia. Association for Computational Linguistics.

Qing Yao, Kanishka Misra, Leonie Weissweiler, and Kyle Mahowald. 2025. Both Direct and Indirect Evidence Contribute to Dative Alternation Preferences in Language Models. In *Second Conference on Language Modeling*.

Aditya Yedetore and Najoung Kim. 2024. Semantic training signals promote hierarchical syntactic generalization in transformers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4059–4073, Miami, Florida, USA. Association for Computational Linguistics.

A Additional Results Without Lemmatization

Due to the nature of our training data and LM size, the commonly used Byte-Pair Encoding (BPE) subword tokenization can introduce extra burden to the training process. For example, the model will have to learn the connection between different word forms of *make* (e.g., *made*, *making*, etc.), which will be tokenized into different subword sequences. Nevertheless, we repeat our main analyses under a BPE subword tokenization setting. We first train a 8,192-token BPE tokenizer on the CHILDES dataset, following Huebner et al. (2021); Misra and Kim (2026). Then in both training and evaluation data pipelines we remove the lemmatization components, and swap the original word-level tokenizer with this BPE tokenizer. We then repeat the verb-specific and abstract condition in §4.1 for 6 verbs and train 70 models (5 seeds $\times$ 6 verbs $\times$ 2 manipulation types for verb-specific, 5 seeds $\times$ 2 manipulation types for abstract), and show results in Figure 6. The results are largely consistent with those using lemmatization, with negative effects on our at the presence of the hypothetical preemptive evidence. Under abstract level manipulation, similar to the phenomenon in the lemmatized setting, the effects also become stronger than the verb-specific case. Interestingly, almost all verbs (except for *sleep*) show non-trivial positive effects.

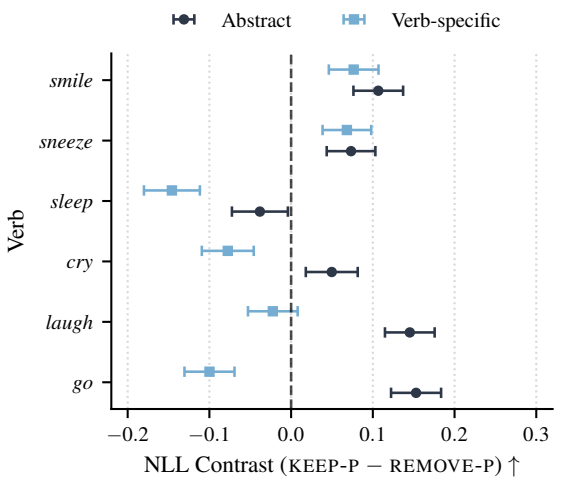

Figure 6: Contrast of estimated marginal means between KEEP-P and REMOVE-P LM variants trained and evaluated on BPE-tokenized versions of datasets.

B Identification of Periphrastic Causative Constructions

We use PyChildes (Ma et al., 2025) to parse the raw CHILDES files of Eng-NA and Eng-UK, and then extract morphological and dependency grammar annotations. These two corpora of CHILDES are open to public access. An example of the dependency parse trees is shown in Figure 7. Then we use the following heuristics to identify periphrastic causative constructions in the CHILDES dataset.

For *let* and *make*, we identify them as matrix verbs in a periphrastic causative construction if they satisfy the following conditions:

- Its complement (COMP) child is a verb not in a to-infinitive (e.g. *make something to eat*).
- Either it has an object (OBJ) child (e.g. *make me laugh*), or its complement verb has a subject (SUBJ) child (e.g. *make him see it*).

Additionally, for *let*, we require it to have a subject (SUBJ) within its predicate to exclude hortatives like *let's go*. This is done by searching through its ancestor chain formed by complement (COMP), coordination (COORD), and conjunction (CONJ) arcs.

For *get*, *cause*, and *force*, we use the same criteria except that the complement verb must be in a to-infinitive (e.g. *get him to sleep*).

Finally, for *have*, we use a similar method as *let* and *make*, and then use an external LLM, Gemini 3 Flash,[6] to filter out false positives. We use the following prompt:

```
In the following sentences, some contain a
↪  causative *have* and some do not.
"May I have more juice please ?" (False)
"I have finished my juice ." (False)
"I had the waiter bring more juice ." (True)
"I had my juice refilled by the waiter ." (True)

Analyze the following sentence and determine
↪  whether it contains a causative *have* or
↪  not. Answer with True if it does and False if
↪  it does not.
"{sentence}"
```

We include the statistics of some more verbs that we identify in their usage in CHILDES as having periphrastic causative constructions in Table 2.

C Details for Evaluation Data Synthesis

We use Gemini 3 Flash to synthesize evaluation sentences. Our system prompt is as follows:

```
You are a child literature author tasked with
↪  creating engaging and easy-to-understand
↪  texts for young children.
```

The main prompt is as follows, where `previous_outputs` contains samples of previously generated sentence for this verb. For `subject` we sample from 273 names and pronouns, and for `object_candidates` we sample from names, pronouns and additionally 499 nouns like *baby*, *mommy*, *toy*. The language model is prompted to choose an object from 25 randomly sampled candidates that fits most naturally with the subject and verb. The chosen object, along with the subject and verb, are then used to form the periphrastic causative sentence in the evaluation set. The ungrammatical transitive counterpart is then generated by mutating the periphrastic causative sentence.

```
### Instructions
1.  Object Selection: Choose an object from the
↪  vocabulary bank that fits naturally with the
↪  subject and verb.
2.  Scene Creation:
    - The scene should be simple and relatable
    ↪  for a young child.
    - In this scene, the subject causes the
    ↪  object to perform the action described by
    ↪  the verb.
3.  Output Generation:
    - First, write a short sentence describing
    ↪  the context of this action. Without it
    ↪  the action should still be
    ↪  understandable. The context sentence
    ↪  should also contain the subject and the
    ↪  object.
```

[6] https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview

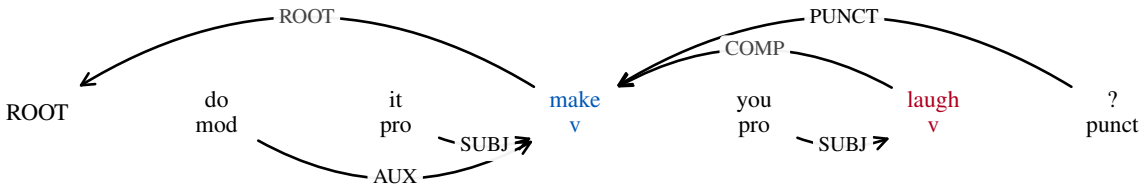

Figure 7: An example of the dependency parse trees on the CHILDES dataset. This is Eng-NA/Bloom/Peter/020303, sentence 1028.

```
    - Then write the main sentence where the
    ↪  subject makes the object perform the
    ↪  action.
4.  Constraints:
    - The main sentence should have the following
    ↪  syntactic structure: "{{subject}} makes
    ↪  {{object}} {{verb}}".
    - Both sentences should be declarative
    ↪  sentences, and quotes are not allowed.

### Output Format
The output should only contain two sentences
↪  separated by a newline character. For example
↪  given subject "I", verb "break", and object
↪  "vase", the output should be:
I knock over the vase.
I make the vase break.

### Previous Outputs
The following outputs have already been generated
↪  for this verb. Prefer to create a different
↪  scene.
{previous_outputs}

### Input
- Subject: {subject}
- Verb: {verb}
- Object Candidates: {object_candidates}
```

Additionally, we lemmatize the evaluation set with Spacy, using model `en_core_web_trf` version 3.8.0.

D Details for Training

Additional details for training hyperparameters are included in Table 3. We pack the consecutive utterances in a single CHILDES file into sequences of 1024 tokens, and the <s> token is used as the beginning of sentence and also the separator between two sentences. We use the GPT-2 implementation provided by Huggingface Transformers (Wolf et al., 2020) and a custom PyTorch Lightning training loop for training our LMs, and a single training run requires around 1.5 hours on 1 NVIDIA L40s GPU.

E A Reassessment of Guo et al. (2026)

Our work is most closely related to that of Guo et al. (2026), who measured preemption and entrenchment effects across three different linguistic phenomena (including transitive overgeneralizations), and across multiple off-the-shelf LMs. They operationalize overgeneralization as the difference in LM surprisal (negative log-probability) of the unconventional (overgeneralized) sentence vs. that of the conventional sentence. In the case of transitive overgeneralizations, their method measures difference in the surprisal of sentences like **Tommy laughed me* and those like *Tommy made me laugh*. Using this measure, they find evidence for preemption over entrenchment in LMs. They do so by first showing high correlation of the difference in surprisals with preemptive usages of verbs, even when controlling for the non-preemptive usages. Then, they fine-tune an off-the-shelf GPT-2 model (Radford et al., 2019) on 5000 instances of conventional vs. unconventional usages, finding overwhelming effects of preemption over non-preemption.

On further scrutiny of Guo et al. (2026)'s methods, we find their results to paint an illusory picture about preemption in LMs. We highlight two concerns. First, they use fine-tuning as their method to claim causal evidence of preemption. However, an LM might have already formed its preferences of avoiding the target overgeneralizations during pre-training (as shown by their results). So, the results from these experiments do not sufficiently shed light on how the dispreference of overgeneralization was learned in the first place—this will require training data manipulations. Second, and more importantly, their method explicitly includes the preemptive construction as part of the overgeneralization measure (e.g., *periphrastic causative* for *transitive*). That is, the conventional form in their measure is exactly the same form that is hypothesized to preempt the unconventional

Verb	PC Usage	Verb Usage
be	272	4150
break	11	10958
change	37	3007
close	18	4822
come	253	88850
cook	15	3538
cough	22	443
cry	49	6166
do	299	153550
drink	19	6209
eat	113	44783
fall	142	15746
feel	240	6237
get	260	178352
go	1295	316515
grow	49	2871
have	439	147278
jump	65	5450
know	106	92370
laugh	99	1781
like	93	54171
look	196	70121
make	163	53323
move	43	7448
open	26	10193
play	198	37279
put	173	99050
roll	16	3555
run	53	9114
see	275	101864
sing	24	7331
sit	150	27679
sleep	33	8614
smile	15	724
sneeze	31	461
stand	132	7124
stay	72	8706
stop	42	5826
swim	12	3901
talk	58	10857
think	94	78822
turn	23	12212

Table 2: Additional verb counts and periphrastic causative usage counts in the CHILDES dataset. This table is not exaustive.

form. Therefore, it is unsurprising that an effect of preemption is observed on fine-tuning LMs on the conventional forms—change in the measure is directly being manipulated in their fine-tuning experiment. So any effect of preemption might just be driven by the increased likelihood of the conventional form rather than the unlikelihood of the unconventional form.

Disentangling preemption from entrenchment will require direct manipulation of the training data, and a more precise measure of overgeneralization that focuses exclusively on the unconventional form. We now describe our methods which embody this conclusion directly.

Hyperparameter	Value
Max Sequence Length	1024
Hidden Dim Size	768
Number of Layers	12
Number of Attention Heads	12
Token per Batch	8192
Optimizer	AdamW
Learning Rate	1e-5
Learning Rate Scheduler	Linear
Warmup Steps	1500
Weight Decay	0
β_1	0.9
β_2	0.95
Gradient Clip Norm	1.0

Table 3: Hyperparameters used for training our LMs.

F Alternate Methods of Measuring Overgeneralization Effects

In this section, we compare our results to those if we had used Guo et al. (2026)'s method of treating the difference in NLL of transitive and periphrastic causative usages as our measure of overgeneralization effects. We replicate the above analysis using this measure (computing NLLs of both subsets of our evaluation data), and show results in Figure 8. We see that on *go*, *laugh*, *sing* there is seemingly positive effect of preemption. But as our previous Figure 3 reveals that the NLL of transitive is actually lower in KEEP-P than in REMOVE-P, we can conclude that the positive effect observed here is driven by the NLL of the periphrastic causative loss being lower in KEEP-P.

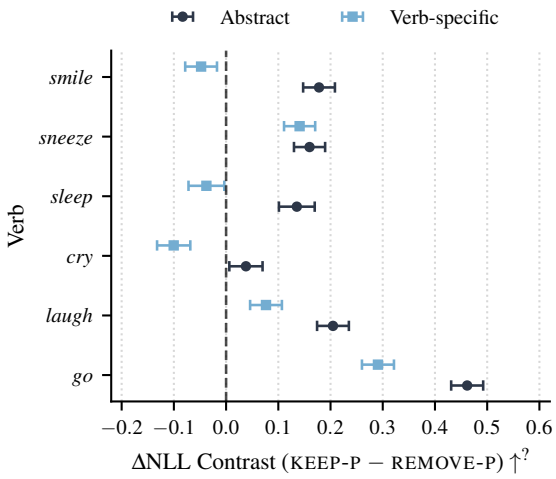

Figure 8: Contrast between KEEP-P and REMOVE-P on a similar linear mixed-effects model as in Figure 3, but with the NLL difference between the transitive and periphrastic causative sentences as the dependent variable. This is the result we would get had we used Guo et al. (2026)'s method.