Unpacking *Let Alone*: Human-Scale Models Generalize to a Rare Construction in Form but not Meaning

Wesley Scivetti Tatsuya Aoyama Ethan Wilcox Nathan Schneider
Georgetown University
{wss37, ta571, ethan.wilcox, nathan.schneider}@georgetown.edu

Abstract

Humans have a remarkable ability to acquire and understand grammatical phenomena that are seen rarely, if ever, during childhood. Recent evidence suggests that language models with human-scale pretraining data may possess a similar ability by generalizing from frequent to rare constructions. However, it remains an open question how widespread this generalization ability is, and to what extent this knowledge extends to *meanings* of rare constructions, as opposed to just their *forms*. We fill this gap by testing human-scale transformer language models on their knowledge of both the form and meaning of the (rare and quirky) English LET-ALONE construction. To evaluate our LMs we construct a bespoke synthetic benchmark that targets syntactic and semantic properties of the construction. We find that human-scale LMs are sensitive to form, even when related constructions are filtered from the dataset. However, human-scale LMs do not make correct generalizations about LET-ALONE's meaning. These results point to an asymmetry in the current architectures' sample efficiency between language form and meaning, something which is not present in human language learners.[1]

1 Introduction

The ability of neural network–based language models (LMs) to learn human language has profound implications for our theories of learning and cognitive science of language (Warstadt and Bowman, 2022; Wilcox et al., 2024). Of particular interest is whether LMs trained on human-scale data can learn nuanced humanlike generalizations about linguistic form and meaning (Wilcox et al., 2025). Recent studies have found that models have remarkable success at both, but that human-scale models appear to make better generalizations about linguistic form (Warstadt et al., 2023). This is particularly true when it comes to rare constructions, where models have been shown to learn formal constraints much more robustly than constraints about the constructions' meanings (Weissweiler et al., 2022).

These results pose a potential problem for construction-based theories of grammar. Construction grammar is a family of theories, which propose that linguistic knowledge is stored in templatic packets (constructions), and that a constructions' form and meaning are learned simultaneously (Goldberg, 2006). These theories predict that if constructionist learning is happening in LMs, form and meaning should be learned simultaneously. However, previous studies that test form and meaning of specific constructions in a controlled way (e.g., Weissweiler et al., 2022) do so only in large-scale LMs, which limits their cognitive interpretation (Wilcox et al., 2025). Other studies, which assess human-scale LMs, investigate formal and semantic competence on different phenomena (Warstadt et al., 2023).

In this work, therefore, we test for a form–meaning learning asymmetry on human-scale LMs. We focus on the LET-ALONE construction (§2), a rare construction of English that is subject to a nuanced but well-studied set of syntactic, semantic and pragmatic constraints (Fillmore et al., 1988). We train human-scale models from scratch and assess them using a bespoke behavioral test suite that probes various facets of LET-ALONE, including conjunction, negation, and scalar properties (§3). Our experiments show strong evidence for a form–meaning asymmetry: Human-scale models learn LET-ALONE's formal constraints almost perfectly (§4), but fail to learn any semantic constraints (§5). Moreover, we find that our models still learn LET-ALONE's formal constraints *even when examples of* LET-ALONE *and related constructions are filtered from pretraining data* (§6). These results underscore the claim that indirect evidence can be crucial for learning constraints on

[1] Code and data: https://github.com/WesScivetti/BabyAlone

rare constructions at human-scale (Misra and Mahowald, 2024). However, we find that removing instances of individual *let* and *alone* tokens from pretraining altogether destroys sensitivity to most formal constraints. An emerging theme in the literature on LM learning of constructions (§7), the performance disparity between form and meaning calls into question how much human-scale models can learn about the meaning of rare constructions.

2 Background: *Let Alone*

In this work, we focus our attention on the LET-ALONE construction. This construction was analyzed in detail by Fillmore et al. (1988). The LET-ALONE construction joins two constituents, which can be of various types, as shown in 1–3:

(1) Max won't eat shrimp, let alone squid. [NP CONJUNCTION]

(2) I barely got up in time to cook lunch, let alone cook breakfast. [VP CONJUNCTION]

(3) They couldn't make John eat the shrimp, let alone Lucille the squid. [ELIDED VP CONJUNCTION]

However, unlike the prototypical conjoiner, *and*, LET-ALONE cannot join two full, unelided independent clauses, as in (4).

(4) *I couldn't afford the red sunglasses let alone I couldn't afford the black sunglasses.

Additionally, LET-ALONE resists movement and fronting in situations which generally allow it. For example, LET-ALONE can't be inserted into cleft sentences, like in (5).

(5) *It is the red sunglasses let alone the black sunglasses that I couldn't afford.

LET-ALONE is also considered a negative polarity item (NPI), and generally ungrammatical when not under the scope of negation, as in (6).[2]

(6) *I could lift the orange crate let alone the green crate.

Regarding the semantics of LET-ALONE, Fillmore et al. (1988) argue that it is best understood as one member of a family of paired focus constructions, including "never mind", "much less", and "not to mention". These constructions have two phrases that are simultaneously placed in focus and are semantically connected via a comparison. Specifically, LET-ALONE implies that the two phrases in focus are in a scalar relationship. The two phrases are thus interpreted as being "two points on a scale" (Fillmore et al., 1988), with the second phrase being higher than the first phrase on whatever scale is evoked. In this way, LET-ALONE has some semantic shared properties with more general comparative constructions, which also place two entities on either explicit or implied scales.

Importantly for our studies, LET-ALONE is exceedingly rare. In our pretraining corpus of ≈100 million words, LET-ALONE occurs fewer than 400 times. To our knowledge, LET-ALONE is the most infrequent construction to be examined in a study such as this at human-scale.

In the experiments that follow, we test both syntactic and semantic properties of LET-ALONE. While descriptions of our test items are given in the respective experimental sections, at a high level, we test four properties: For formal properties, we test NPI licensor sensitivity, clefting, and restrictions on conjunction. For semantic properties, we test scalar sensitivity, specifically, whether models prefer contexts whose properties match the scales of LET-ALONE.

3 Methods

3.1 Evaluation Dataset

We experiment on a templatically constructed dataset of LET-ALONE minimal pair instances. This approach follows other minimal pair datasets which are used to test LM processing of grammatical phenomena (e.g. SyntaxGym, Gauthier et al., 2020; BLiMP, Warstadt et al., 2020; COMPS, Misra et al., 2023). For each of our experiments, we create a test suite of k templatically generated items. Each item comes in four different conditions, which cross a *grammatical manipulation* (±MANIP) with the *presence or absence* of LET-ALONE (±LTALN). (7) serves as an example:

(7) a. *Max **could** lift the red box **let alone** the blue box. [+MANIP, +LTALN]
b. Max **couldn't** lift the red box **let alone** the blue box. [−MANIP, +LTALN]
c. Max **could** lift the red box **and** the blue box. [+MANIP, −LTALN]

[2] Though see Fillmore et al. (1988) for some examples where NPI licensors are not needed, such as "You have enough material for a semester, let alone a week."

Test Type	Property	N	Manipulation	Example
Formal	Conjunction (Clause)	5217	*conjoin* independent clauses	*I couldn't lift the blue crate let alone I couldn't lift the red crate.
Formal	Conjunction (VP)	5217	*conjoin* VPs	I couldn't lift the blue crate let alone lift the red crate.
Formal	Conjunction (Gap)	5217	*conjoin* elided VP clauses	I couldn't lift the blue crate let alone you the red crate.
Formal	Clefting	5217	*cleft* S	*It is the blue crate let alone the red crate that I couldn't lift.
Formal	NPI	5217	*remove* "not"	*I could lift the blue crate let alone the red crate.
Meaning	Scalar Semantics	16887	*contradictory follow-up*	#I couldn't lift the blue crate let alone the red crate. The blue crate is heavier than the red crate.

Table 1: Number of examples (N) for each test in our test set for LET-ALONE. These tests are inspired by the properties of LET-ALONE as described in Fillmore et al. (1988). An example for each manipulation is shown relative to the base sentence *I couldn't lift the blue crate let alone the red crate.* Some of the manipulations induce an ungrammatical LET-ALONE sentence where *and* would be acceptable, while others are equally acceptable in the base and manipulated versions.

d. Max **couldn't** lift the red box **and** the blue box. [−Manip, −LtAln]

Note that +Manip is a grammatical manipulation that makes LET-ALONE sentences (+LtAln) *ungrammatical* (for some manipulations), but does not affect the grammatically of non-LET-ALONE sentences (−LtAln); hence, only the [+Manip, +LtAln] configuration is ever ungrammatical. The specific manipulations we test involve conjoining independent clauses, clefting the sentence, and removing the negative licensor ("not"), respectively, as shown in Table 1. We also experiment with two additional conjunction experiments for which the manipulations of LET-ALONE are grammatical: conjoining verb phrases (VPs) and conjoining elided VPs. From the sentences in these five conditions, we calculate an *accuracy score*, which is described in the next section in detail. We opt for template-based examples, as opposed to natural corpus data, in order to control for several factors, including the frequency of lexical items, the scalar semantics invoked by LET-ALONE, and the syntactic context in which the LET-ALONE sentence occurs.

There is evidence that human-scale models struggle with world knowledge (Ivanova et al., 2024; Hu et al., 2024), and thus their performance on interpreting the scalar semantics of LET-ALONE in examples like (1–3) may be bottlenecked by their lack of reasoning over the properties which are being compared on the scale (e.g., the unusualness of eating shrimp versus squid). Because of this, we design templates around scalar properties from domains which involve quantitative scales, such as height, weight, distance, and price. For consistency, all of our LET-ALONE focus elements are direct objects. To control for the possibility of lexical biases inherent to the objects, our focus elements are always the same lexical noun (e.g. "box" in the example above), which is modified with different neutral adjectives, such as color terms. Table 1 reports example counts for the dataset.

The meaning of LET-ALONE implies that the two phrases in the construction have some shared scalar property, and the second one has a higher value than the first. This allows us to probe the semantics of the construction by designing minimal pairs in which there is a follow-up sentence which either directly follows from the scalar semantics of the construction, or contradicts it, while still maintaining our overall 2x2 manipulation:

(8) I couldn't afford the red sunglasses **let alone** the black sunglasses.

a. #The **red** sunglasses are more expensive than the **black** sunglasses. [+Manip, +LtAln]

b. The **black** sunglasses are more expensive than the **red** sunglasses. [−Manip, +LtAln]

(9) I couldn't afford the red sunglasses **and** the black sunglasses.

a. The **black** sunglasses are more expensive than the **red** sunglasses. [+Manip, −LtAln]

b. The **red** sunglasses are more expensive than the **black** sunglasses. [−Manip, −LtAln]

The only difference in the manipulated examples is that the color items have been swapped. This shouldn't impact models' predictions when the context sentence contains *and*. However, the LET-ALONE sentence makes clear which object is higher on the expense scale, and so only one follow-up sentence is valid, while the other is infelicitous.

3.2 Evaluation Metric

Following Misra and Mahowald (2024), instead of comparing raw probabilities between conditions, we use the Syntactic Log Odds Ratio (SLOR; Pauls and Klein, 2012; Lau et al., 2017). We calculate SLOR over sentences, $\mathbf{w} = [w_1 \dots w_N]$, where w is drawn from a vocabulary of words Σ. A sentence $\mathbf{w}$ is potentially conditioned on a context $\mathbf{c}$. In our form experiments, $\mathbf{c}$ is empty, and $\mathbf{w}$ is the LET-ALONE sentence. In our meaning experiment, the context $\mathbf{c}$ is our LET-ALONE sentence, and $\mathbf{w}$ is a following sentence that is either felicitous or non-felicitous given the scalar properties of the LET-ALONE context. We assume an LM with parameters θ that can produce probability $p_\theta(\cdot)$, and a model $p_U(\cdot)$ of the unigram distribution over $w \in \Sigma$. SLOR can then be defined as:

$$\mathcal{S}(\mathbf{w}) = \frac{1}{N} \log \frac{p_\theta(\mathbf{w} \mid \mathbf{c})}{\prod_{w \in \mathbf{w}} p_U(w)} \tag{1}$$

SLOR is designed to control for the fact that more frequent words are inherently less surprising for an LM. To turn by-sentence SLOR scores into accuracy scores, we compare conditions in our test suites. Every item in our test suites come in four conditions, corresponding to the examples in (7a)–(7d). For a given item, i, we notate its conditions with sub and superscripts: $i^{+\text{LTALN}}_{+\text{MANIP}}$ refers to i's +MANIP, +LTALN condition. First, we define what we refer to as the delta SLOR, or $\Delta\mathcal{S}$, which is simply the difference in SLOR due to our grammatical manipulation:

$$\Delta\mathcal{S}_i(\ell) = \mathcal{S}(i^{\ell}_{-\text{MANIP}}) - \mathcal{S}(i^{\ell}_{+\text{MANIP}}) \tag{2}$$

where ℓ can either be +LTALN or −LTALN. To calculate accuracy, we inspect the *differences* in $\Delta\mathcal{S}$ scores. Namely, we predict that the effect of grammatical manipulation should be greater (reflecting lower grammaticality) when LET-ALONE is used, compared to when a vanilla conjunction is used. With this prediction, our accuracy scores for a test suite of k items can be defined as:

$$\frac{1}{k} \sum_{i=1}^{k} \mathbb{1}[\Delta\mathcal{S}_i(+\text{LTALN}) \geq \Delta\mathcal{S}_i(-\text{LTALN})] \tag{3}$$

This corresponds to the effect of the manipulation in the LET-ALONE case (between (7a) vs. (7b)) above and beyond that observed with the non-LET-ALONE control ((7c) vs. (7d)).

Intuitively, a model achieving high accuracy on this task has correctly understood that these manipulations are more ungrammatical for LET-ALONE than they are for *and*, and has thus learned a core part of the idiosyncratic nature of the construction. We also control for the possible bias related to ordering of colors by swapping the orders of the colors for each example, and an example is only considered correct if *both* orders are correct. Since this involves two pairwise comparisons (both orderings), chance performance on our tasks is 25%.

3.3 Model Architecture and Pretraining

We use the training split of the BabyLM-strict 100M dataset (Warstadt et al., 2023) for pretraining. For all experiments, we follow Misra and Mahowald (2024) in utilizing the OPT architecture (Zhang et al., 2022). We utilize identical hyperparameters to those reported in Misra and Mahowald (2024) where possible.[3] For all model settings, we pretrain two models with identical hyperparameters and different random seed, and report the average results.

In all experiments, we use the minicons library (Misra, 2022) for calculating sequence-level surprisals. Following Misra and Mahowald (2024), we calculate SLOR using the surprisal values from minicons combined with the unigram frequencies from the BabyLM training set.[4]

4 Experiment 1: Formal Constraints

We first test constraints on the formal characteristics of LET-ALONE. We focus on restrictions of LET-ALONE which differ from those for simple conjunctions, namely conjunction of clauses, clefting, and NPI licensor sensitivity. Additionally, we add two conjunction conditions (conjunction of VPs and conjunction of elided VP clauses), which are grammatical for both LET-ALONE and simple conjunctions. These two additional conditions serve as a control to see if the models not only recognize constraints on LET-ALONE, but also recognize valid syntactic variations.

[3] Full hyperparameters are reported in Table 5 in Appendix A.

[4] In filtered pretraining experiments, we calculate unigram frequencies on the filtered training dataset.

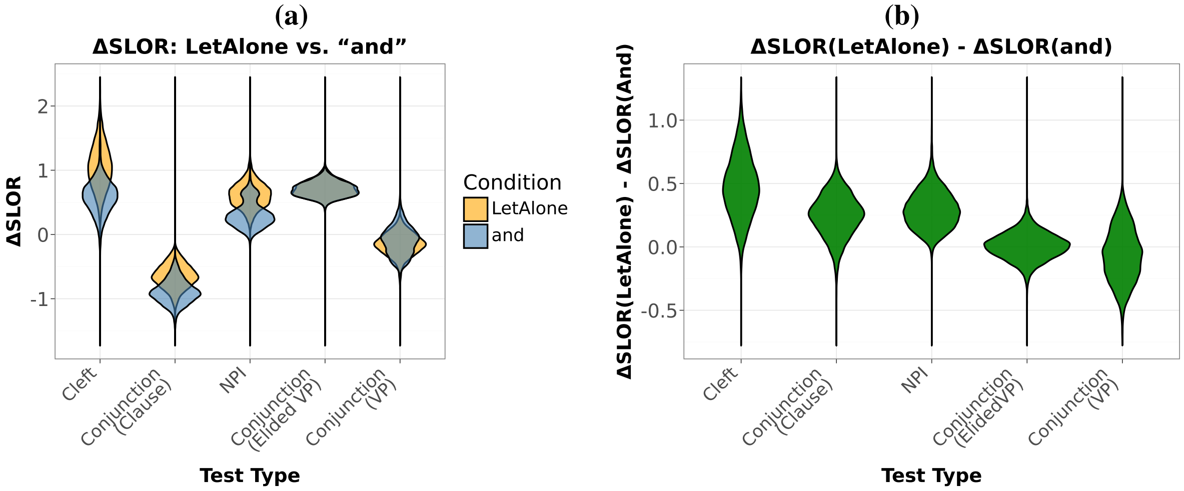

Figure 1: Results on Syntactic Tests: (a) shows $\Delta\mathcal{S}$ where higher delta values indicate a greater effect of the constraint. Patterns are consistent with the grammaticality of the syntactic manipulation. **(b)** shows $\Delta\mathcal{S}$ (LetAlone)-$\Delta\mathcal{S}$ (And).

Formal Property	Prediction	Accuracy
Conjunction (Clause)	near 100%	88.1 ± .8%
Clefting	near 100%	96.5 ± .5%
NPI	near 100%	98.6 ± .3%
Conjunction (VP)	near 25%	31.1 ± 1.3%
Conjunction (Gap)	near 25%	37.5 ± 1.3%

Table 2: Results for Syntactic Tests: *Prediction* column indicates expected accuracies if the model had made the humanlike generalization. Model accuracies are means over two random seeds. We report 95% confidence intervals over the means of the two runs.

4.1 Results

Results for these tests are reported in Table 2 and visualized in Figure 1. Across all 3 formal constraint tests involving ungrammatical manipulations, accuracy is very high over two randomly seeded pretraining runs. For control conditions (VP conjunction and elided VP clause conjunction), accuracy is near chance, as expected. Looking at SLOR differences at the example level in Figure 1, we see strong separation in the $\Delta\mathcal{S}$ scores for LET-ALONE versus *and*. We can see in Figure 1 that the few negative $\Delta\mathcal{S}$ difference values for non-control conditions tend to be clustered very close to 0. For the control conditions, $\Delta\mathcal{S}$ differences are generally evenly distributed above and below 0. These results provide evidence that human-scale models are able to capture the formal properties of LET-ALONE well. This strong performance is despite the fact that the LET-ALONE construction only occurs roughly 300 times in the BabyLM training corpus. With so few training examples, it seems likely that for LET-ALONE, indirect evidence is far more important than direct evidence for the recognition of these formal constraints. See §6 for more discussion of filtered pretraining of LET-ALONE and related constructions.

5 Experiment 2: Semantic Constraints

Having shown that human-scale models have sensitivity to a range of formal constraints on the LET-ALONE construction, we evaluate whether BabyLM scale language models are sensitive to the scalar semantics of LET-ALONE. As stated in §2, we test the scalar semantics by supplying additional follow-up sentences, which are either felicitous or non-felicitous with the scale set up by the LET-ALONE construction. We then compare the SLOR values of the two target sentences, conditioned on the LET-ALONE prefixes.

5.1 Results

Table 3 reports results on the semantic tests. We find no evidence that BabyLMs are sensitive to the semantics of LET-ALONE, as performance on this minimal pair task is below chance for both random seeds. In contrast, we include as a skyline comparison point GPT-4.1 (OpenAI, 2024), which achieves extremely high accuracy (94%) on a prompting version of our task on the same dataset.[5] In §5.1, we visualize the $\Delta\mathcal{S}$ values between the correct and incorrect followups, showing that they cluster very close to zero. This indicates that the BabyLM

[5] See Appendix B for details on the GPT prompting experiment.

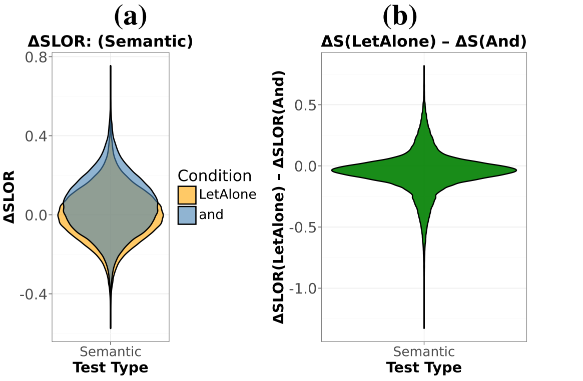

Figure 2: Results on Semantic Tests: (a) shows $\Delta\mathcal{S}$ where higher delta values indicate a greater effect of the constraint. Patterns are consistent with the grammaticality of the syntactic manipulation. **(b)** shows $\Delta\mathcal{S}$ (LetAlone)- $\Delta\mathcal{S}$ (And).

Model	Property	Prediction	Accuracy
BabyLM	Scalar Semantics	near 100%	4.9 ± 0.32%
GPT-4.1	Scalar Semantics	near 100%	94.0 ± 0.361%

Table 3: Accuracies for the Semantic Tests. BabyLM models demonstrate no sensitivity to the scalar semantics of LET-ALONE. In a metalinguistic prompting formulation of our task, GPT-4.1 achieves strong performance, indicating the dataset is solvable with sufficient input.

model has very little preference between the two alternatives, pointing to a general lack of sensitivity to the scalar properties of the construction.

5.2 Analysis

We perform an analysis to see what contributes to the poor semantic performance for BabyLM models. We hypothesize that the specific template we use may impact our results. In Figure 3, we graph the top 10 highest performing templates, represented by the predicate, noun, and comparative in the sentence. We find that several templates do exhibit above chance performance, though accuracy is still generally low and confidence intervals are quite large (each template has dozens of examples). This finding indicates the models may have some limited semantic knowledge of LET-ALONE, but do not encode an abstract LET-ALONE construction. Instead, the construction is context dependent, and the intended meaning is only accessed alongside some specific lexical items and semantic frames.

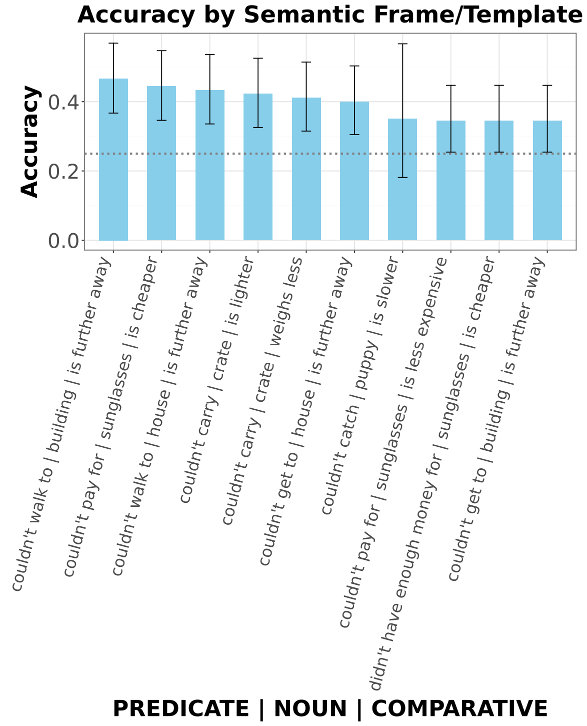

Figure 3: Top 10 Accuracies on the Semantic Tests when separated by predicate, noun, and comparative that fill the template. Error bars indicate 95% confidence intervals over the results of two random seeds. Above-chance accuracies indicate that the model has some nontrivial semantic performance on that template.

5.3 Discussion

These results, combined with those outlined in §4.1, point to a strong divide between formal and functional competence regarding the LET-ALONE construction for our human-scale models. This finding aligns well with past work on constructions (Weissweiler et al., 2022), which has shown that syntactic competence often far outpaces semantic competence in controlled environments for a given construction.

Most constructionist accounts of language contend that constructions are stored as form/meaning pairings in the human mind, and posit that form and meaning are learned in conjunction by humans (Goldberg, 2006). The apparent lack of functional knowledge of LET-ALONE that we observe is compatible with a formal vs. functional distinction of language competence in language models (Mahowald et al., 2024).

6 Experiment 3: Filtered Pretraining

Having shown that language models have at least some sensitivity to the formal constraints, we now test how this capability is impacted by filtering the

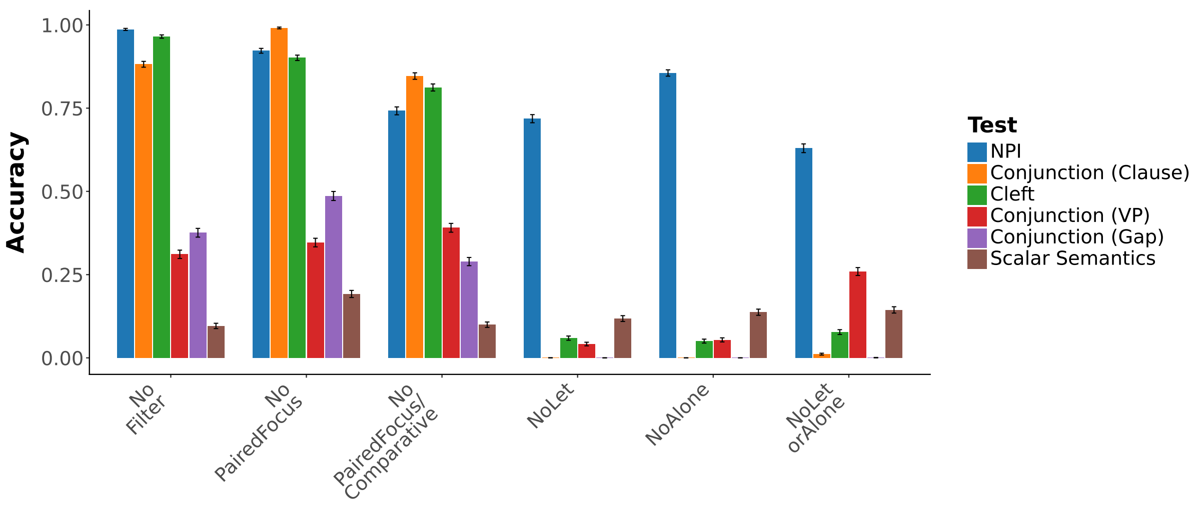

Figure 4: Filtered Pretraining Results. Accuracies are calculated according to Equation 3. Error bars are 95% confidence intervals over the mean accuracies across two randomly seeded runs. Table 4 presents the same data.

pretraining dataset to exclude LET-ALONE and related constructions. We experiment with 5 "filtered pretraining" (Patil et al., 2024) scenarios:

Excluding all paired-focus constructions ("let alone", "much less", "never mind", "not to mention"). These paired-focus constructions generally follow many of the same syntactic constraints as LET-ALONE. These constructions combine to account for 2312 total sentences in the BabyLM train set.

Excluding paired focus AND comparatives. Since semantically, LET-ALONE is thought to be related to comparative constructions, simple comparatives "more than", "less than" are removed as well. This accounts for ≈70k sentences, or roughly .5% of the entire BabyLM 100M train set.

Excluding all instances of *let*. This means *let* is not seen as a token during pretraining. This accounts for ≈165k sentences in BabyLM train.[6]

Excluding all instances of *alone*. This excludes ≈16,000 sentences in BabyLM train.

Excluding all instances of *let* AND of *alone*. This means that neither token is seen (in any context) during pretraining (≈180k sentences).

For all filtering, we use case-insensitive regular expression query matching over the BabyLM training corpus to remove examples. If a target construction is found, then the entire sentence containing that construction is removed from the training corpus. We use SLOR as the evaluation metric as in previous experiments, and calculate unigram probabilities independently for each filtered pretraining set. We test all filtered models on the minimal pair datasets from experiments in §4 and §5.

6.1 Results

We visualize the accuracies for each filtering scenario in Figure 4. We additionally visualize the changes in ΔSLOR due to filtering in Figure 5. Filtering out paired-focus constructions, including LET-ALONE, seems to have little impact on the performance on formal tests, and we observe similarly high performance when additionally filtering out simple comparatives. However, when filtering out *let* or *alone*, performance drops substantially.

Interestingly, even when filtering out both *let* and *alone*, performance on the NPI constraint remains nontrivial. This likely points to some other heuristic that allows models to solve the task without any knowledge of the construction. We hypothesize that because negating a sentence with *and* sometimes results in a sentence with *or*, *and* may be mildly biased against negative contexts.

6.2 Discussion

This work has shown that BabyLM scale models are sensitive to several formal constraints on LET-ALONE. This sensitivity remains even when all instances of LET-ALONE, related paired focus constructions, and even simple comparatives are removed from training, meaning the models are learning from indirect evidence of some kind beyond LET-ALONE or seemingly related constructions. We hypothesize that the syntactic patterns tested here, while seemingly idiosyncratic to LET-ALONE and similar constructions, are likely related

[6] Inflectional variants like "letting" are not removed.

Filtering Scenario	NPI	Conjunction (Clause)	Cleft	Conjunction (VP)	Conjunction (Elided VP)	Scalar Semantics
NoFiltering	98.6 ± 0.3%	88.1 ± 0.8%	96.5± 0.5%	31.1 ± 1.3%	37.5 ± 1.3%	4.9 ± 0.3%
NoPairedFocus	91.9 ± 0.7%	97.6 ± 0.4%	93.7 ± 0.7%	39.8 ± 1.3%	58.1 ± 1.3%	9.3 ± 0.4%
NoPairedFoc/Comp	84.9 ± 0.9%	92.1 ± 0.7%	88.9 ± 0.8%	37.0 ± 1.3%	42.4 ± 1.3%	11.2 ± 0.5%
NoLet	71.8 ± 1.2%	0.0 ± 0.0%	5.9 ± 0.6%	4.1 ± 0.5%	0.0 ± 0.0%	1.8 ± 0.2%
NoAlone	85.5 ± 0.9%	0.0 ± 0.0%	5.0 ± 0.5%	5.4 ± 0.6%	0.0 ± 0.0%	3.8 ± 0.3%
NoLetorAlone	62.9 ± 1.3%	0.0 ± 0.0%	7.7 ± 0.7%	25.9 ± 1.2%	0.1 ± 0.0%	6.4 ± 0.4%

Table 4: Filtered Pretraining Results. Figure 4 visualizes the same data.

to more general patterns which are better represented in pretraining data and thus facilitate learning (Potts, 2024). In our case, the indirect evidence that models may be using is that of the manipulations that we test. For example, even without observing LET-ALONE, our BabyLMs have observed cleft constructions, and may have learned that there is a restricted set of phrase types that can be clefted which does not include LET-ALONE. Since all of our syntactic tests rely on combining LET-ALONE with more common syntactic constructions, it is possible that robust representations of these interacting constructions allows for strong model performance on our tests.

We observe that performance degrades sharply when all *let* or all *alone* tokens are removed from pretraining. These results seem to indicate that LMs are using some compositionality between the embeddings of *let* and *alone* to arrive at the meaning of the construction overall. This is somewhat counterintuitive in that Construction Grammar theory would not necessarily predict such a strong link between lexical items and a construction in which they are used far outside of their canonical distributions. However, we also note that SLOR as a metric inherently involves normalizing language model scores by the probability of a string from a unigram language model. Thus, removing *let* and *alone* entirely has a substantial effect on the unigram-based probability, and ultimately may explain the large drop in performance when *let* and *alone* are filtered. We leave open the possibility of replacing SLOR with a more complex function (e.g. MORCELA; Tjuatja et al., 2025) for future work.

7 Related Work

Human-Scale LMs. There is increasing interest in designing smaller scale LMs which could potentially be more informative to human congition (Dupoux, 2018). The BabyLM challenge (Warstadt et al., 2023) was created to address this interest and has resulted in robust human-scale models (Charpentier and Samuel, 2023). Furthermore, growing evaluation of constructional knowledge in BabyLMs has yielded promising results (Misra and Mahowald, 2024; Bunzeck et al., 2025; Rozner et al., 2025b). Beyond BabyLM, researchers have endeavored to create smaller scale LMs using a variety of training corpora, including pretraining on the British National Corpus (Consortium et al., 2007) and achieving respectable performance on a variety of syntactic and understanding benchmarks (Samuel et al., 2023).

Constructions in LMs. This present work follows in a line of research which seeks to test language model understanding of "constructions" as defined by Construction Grammar (Goldberg, 1995; Croft, 2001), of which LET-ALONE is just one construction out of many. Starting with CxG-BERT (Tayyar Madabushi et al., 2020), there have been a flurry of papers showing that LMs learn a variety of constructions, including abstract argument structures (Li et al., 2022). Weissweiler et al. (2022) is of particular relevance to this work, as they perform a paired syntactic/semantic analysis of a rare construction, though not at human-scale. Using probing tasks, they show that BERT-scale LMs recognize the syntax of the COMPARATIVE-CORRELATIVE but not its semantics. Furthermore, there have been several works that have shown LM reasoning and semantic capabilities are limited when confronted with rare constructions (Zhou et al., 2024; Bonial and Tayyar Madabushi, 2024), though results from other studies are more promising (Mahowald, 2023; Potts, 2024; Rozner et al., 2025a; Scivetti and Schneider, 2025).

This present work most directly builds off of Misra and Mahowald (2024), who show that BabyLM scale models are sensitive to the formal properties of the AANN construction. They also are the first to apply the "filtered pretraining"

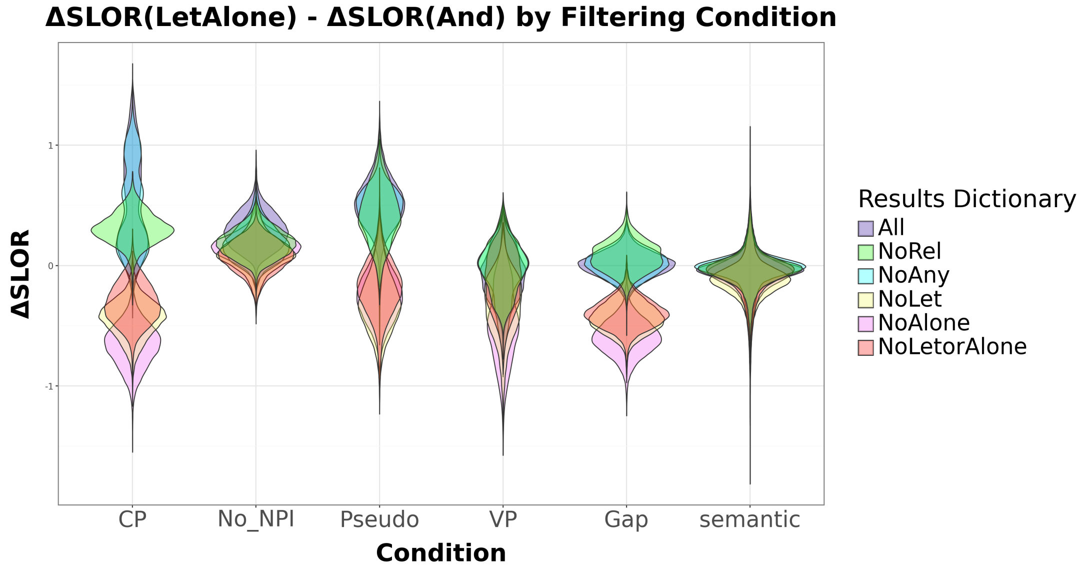

Figure 5: ΔSLOR(Let-Alone) – ΔSLOR(And) for Filtered Pretraining Conditions. Positive Differences indicate sensitivity to the construction. Differences are generally still positive after removing direct evidence from related constructions (NoRel, NoAny) generally, but become negative when all 'let' or 'alone' tokens are removed.

(Patil et al., 2024) technique to construction-based investigations, and show that language models are able to remain sensitive to properties of AANNs even when they don't observe them during training. The findings of this present work on LET-ALONE largely support this claim that properties of constructions can be learned without direct observation in training, which has been shown to be true (to some extent) even for more frequent constructions like the dative alternation (Yao et al., 2025).

Let Alone. Bonial and Tayyar Madabushi (2024), Scivetti et al. (2025), and Rozner et al. (2025a) are the only past works (to our knowledge) which have specifically targeted language model understanding of LET-ALONE in some capacity. Rozner et al. (2025a) introduce *global affinity* and *local affinity* as metrics to quantify the extent to which constructional information is approximated by an LM's output distribution. Using RoBERTa, they show that for corpus instances of LET-ALONE, both *let* and *alone* have high global affinity scores, indicating that the model has at least some distributional knowledge that links the two words as part of a construction. Focusing on natural corpus data from the CoGS dataset, Bonial and Tayyar Madabushi (2024) find that LET-ALONE can be distinguished from distractor constructions at a much higher accuracy by LLMs when compared to fully abstract constructions. Scivetti et al. (2025) extend this work by refashioning the corpus examples into a natural language inference (NLI) dataset, finding that LLMs can perform NLI with very high accuracy for examples which target the semantics of LET-ALONE. Our work diverges from the prior work on LET-ALONE in that we 1) focus on a templatically generated dataset as opposed to corpus data, allowing us to test much finer-grained properties of the construction, and 2) use minimal pair–based evaluations to test both *form* and *meaning* on human-scale models.

8 Conclusion

This work has shown that BabyLM models can learn various formal properties of LET-ALONE. This formal competence is maintained even in the absence of direct evidence of LET-ALONE and related constructions. Such a result points to the crucial nature of indirect evidence for learning a construction as rare as LET-ALONE. On the other hand, BabyLM models seem to have very little grasp of LET-ALONE's meaning. This result underscores the importance of considering both formal and functional competence when assessing LM capabilities regarding a specific linguistic phenomenon, and doing so in a controlled manner. While the formal capabilities of BabyLM models are promising, insofar as we consider meaning to be a central part of language, our results cast doubt on the proposition that robust semantics—including of rare constructions—can be learned from form alone via human-scale pretraining.

Limitations

In this work, we focus on a single construction, LET-ALONE. Future work is needed to determine the extent to which models can learn the form and meaning of various constructions from human-scale pretraining data. Secondly, we only test one type of model architecture. This architecture has been found to be robust at learning from human-scale data in past work, but we cannot be sure that the results would hold for other architectures. Finally, this work only analyzes a single construction in a single language, English, while more work is needed to test a wider variety of rare constructions across languages.

Ethics and Risks

We do not foresee any major ethical concerns or risks associated with this work. All evaluation data is templatically created and thus free of any sensitive or offensive content. The dataset released here is designed for research into the LET-ALONE construction, and the authors do not foresee that it could be deployed for any malicious purpose.

Acknowledgments

We thank the anonymous reviewers, ACs, and SACs for EMNLP and the ARR May Cycle. We also thank members of the NERT and PiCoL labs for their useful comments on early versions of this work. This research was supported in part by NSF award IIS-2144881.

References

Claire Bonial and Harish Tayyar Madabushi. 2024. A Construction Grammar Corpus of Varying Schematicity: A Dataset for the Evaluation of Abstractions in Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 243–255, Torino, Italia. ELRA and ICCL.

Bastian Bunzeck, Daniel Duran, and Sina Zarrieß. 2025. Do Construction Distributions Shape Formal Language Learning In German BabyLMs? In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 169–186, Vienna, Austria. Association for Computational Linguistics.

Lucas Georges Gabriel Charpentier and David Samuel. 2023. Not all layers are equally as important: Every Layer Counts BERT. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 238–252, Singapore. Association for Computational Linguistics.

BNC Consortium et al. 2007. British national corpus. *Oxford Text Archive Core Collection*.

William Croft. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press.

Emmanuel Dupoux. 2018. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59.

Charles J. Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone. *Language*, 64(3):501–538. Publisher: Linguistic Society of America.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An Online Platform for Targeted Evaluation of Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Adele E. Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press. Google-Books-ID: HzmGM0qCKtIC.

Adele E. Goldberg. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press. Google-Books-ID: LHrcqeZmUN4C.

Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. Findings of the Second BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. ArXiv:2412.05149 [cs].

Anna A. Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H. Clark, Carina Kauf, Jennifer Hu, R. T. Pramod, Gabriel Grand, Vivian Paulun, Maria Ryskina, Ekin Akyürek, Ethan Wilcox, Nafisa Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Joshua Tenenbaum, and Jacob Andreas. 2024. Elements of World Knowledge (EWOK): A cognition-inspired framework for evaluating basic world knowledge in language models. ArXiv:2405.09605 [cs].

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive science*, 41(5):1202–1241.

Bai Li, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu. 2022. Neural reality of argument structure constructions. In *Proceedings of the 60th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), page 7410–7423, Dublin, Ireland. Association for Computational Linguistics.

Kyle Mahowald. 2023. A Discerning Several Thousand Judgments: GPT-3 Rates the Article + Adjective + Numeral + Noun Construction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, page 265–273, Dubrovnik, Croatia. Association for Computational Linguistics.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

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.

Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. COMPS: Conceptual Minimal Pair Sentences for testing Robust Property Knowledge and its Inheritance in Pre-trained Language Models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, page 2928–2949, Dubrovnik, Croatia. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4.1. `https://openai.com`. Large language model.

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.

Adam Pauls and Dan Klein. 2012. Large-Scale Syntactic Language Modeling with Treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 959–968, Jeju Island, Korea. Association for Computational Linguistics.

Christopher Potts. 2024. Characterizing English Preposing in PP constructions. *Journal of Linguistics*, page 1–39.

Joshua Rozner, Leonie Weissweiler, Kyle Mahowald, and Cory Shain. 2025a. Constructions are Revealed in Word Distributions. ArXiv:2503.06048 [cs].

Joshua Rozner, Leonie Weissweiler, and Cory Shain. 2025b. BabyLM's First Constructions: Causal interventions provide a signal of learning. ArXiv:2506.02147 [cs.CL].

David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. 2023. Trained on 100 million words and still in shape: BERT meets British National Corpus. In *Findings of the Association for Computational Linguistics: EACL 2023*, page 1954–1974, Dubrovnik, Croatia. Association for Computational Linguistics.

Wesley Scivetti and Nathan Schneider. 2025. Construction Identification and Disambiguation Using BERT: A Case Study of NPN. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 365–376, Vienna, Austria. Association for Computational Linguistics.

Wesley Scivetti, Melissa Torgbi, Austin Blodgett, Mollie Shichman, Taylor Hudson, Claire Bonial, and Harish Tayyar Madabushi. 2025. Beyond Memorization: Assessing Semantic Generalization in Large Language Models Using Phrasal Constructions. ArXiv:2501.04661 [cs].

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets Construction Grammar. In *Proceedings of the 28th International Conference on Computational Linguistics*, page 4020–4032, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Lindia Tjuatja, Graham Neubig, Tal Linzen, and Sophie Hao. 2025. What Goes Into a LM Acceptability Judgment? Rethinking the Impact of Frequency and Length. In *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)*, pages 2173–2186, Albuquerque, New Mexico. Association for Computational Linguistics.

Alex Warstadt and Samuel R Bowman. 2022. What Artificial Neural Networks Can Tell Us About Human Language Acquisition. In *Algebraic structures in natural language*, pages 17–60. CRC Press.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The better your Syntax, the better your Semantics? Probing Pretrained Language Models for the English Comparative Correlative. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, page 10859–10882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2024. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, 55(4):805–848.

Ethan Gotlieb Wilcox, Michael Y. Hu, Aaron Mueller, Alex Warstadt, Leshem Choshen, Chengxu Zhuang, Adina Williams, Ryan Cotterell, and Tal Linzen. 2025. Bigger is not always better: The importance of human-scale language modeling for psycholinguistics. *Journal of Memory and Language*, 144:104650.

Qing Yao, Kanishka Misra, Leonie Weissweiler, and Kyle Mahowald. 2025. Both Direct and Indirect Evidence Contribute to Dative Alternation Preferences in Language Models. ArXiv:2503.20850 [cs].

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models. ArXiv:2205.01068 [cs].

Shijia Zhou, Leonie Weissweiler, Taiqi He, Hinrich Schütze, David R. Mortensen, and Lori Levin. 2024. Constructions Are So Difficult That Even Large Language Models Get Them Right for the Wrong Reasons. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3804–3811, Torino, Italia. ELRA and ICCL.

A Pretraining Hyperparameters

Table 5 reports the hyperparameters for BabyLM pretraining, which were kept consistent across all runs. This is the same as those reported in Misra and Mahowald (2024), except that they tune the learning rate between the options of 1e-4, 3e-4, 1e-3, and 3e-3. We opt for 1e-4 after preliminary runs showed that higher learning rates led to development loss bottoming out much before the end of 20 epochs of training. In total, 12 complete pretraining runs were performed, totaling roughly 100 hours runtime on a NVIDIA A100 GPU.

B GPT Prompting Experiment

For the semantic experiment (see §5), we reformulated our task as a prompt-based experiment for input into GPT-4.1. Since this result serves primarily as a comparison point to BabyLM models, we do not perform prompt optimization. Thus, the reported GPT-4.1 accuracy should not be taken as an upper limit on LLM performance on this dataset. GPT-4.1 was accessed through the OpenAI API. The cost of this experiment was roughly $8 USD. Table 6 reports the prompt used for GPT-4.1 experiments.

Model Architecture	OPT (Zhang et al., 2022)
Embed Size	768
FFN Dimension	3,072
Num. Layers	12
Attention Heads	12
Vocab Size	16,384
Max. Seq. Length	256
Batch Size	32
Warmup Steps	32,000
Epochs	20
Learning Rate	1e-4
Total Parameters	97M
Training Size	100M tokens

Table 5: Model hyperparameters for all OPT Models

Prompt template fed to GPT-4.1

```
You are a linguistic annotator who must
understand the ‘let alone‘ construction.  Please
read the following text, and consider its
meaning.  Then choose the sentence that is
most likely to be true, given the text.
Text:  {text}
Question:  Which of the following sentences is
most likely to be true, given the above text?
Choices:  {formatted_choices}
Please respond with only the letter of the
correct choice (e.g. A or B).
```

Table 6: Exact prompt template used in our experiments. Curly-braced items (`{text}`, `{formatted_choices}`) are filled in programmatically with items from our semantic dataset.