Language Models Learn Constructional Semantics, *Not To Mention* Syntax: Investigating LM Understanding of PAIRED-FOCUS Constructions

Wesley Scivetti[1] **Ethan Wilcox**[1] **Nathan Schneider**[1]
Kanishka Misra[2] **Leonie Weissweiler**[3]
[1]Georgetown University[1], [2]The University of Texas at Austin
[3]Leipzig University & ScaDS.AI Dresden/Leipzig
wss37@georgetown.edu

Abstract

Grasping the semantics of rare *constructions* (form–meaning pairings) has been shown to be a challenging problem that has currently only been solved by the largest LLMs. It remains an open question if open-source models have robust constructional understanding, and if so, what learning dynamics underlie the acquisition of this knowledge. Focusing on a set of rare PAIRED-FOCUS constructions in English (e.g. "let alone", "much less"), we construct a novel dataset to test their meanings using both scalar adjectival semantics and general world knowledge. Testing a wide range of models differing in parameter count, architecture, and pretraining dataset size, we find that several modestly sized models are sensitive to both the forms and the meanings of PAIRED-FOCUS constructions, though models trained on human-scale data fail at all meaning evaluations. Turning to training dynamics for a set of open-checkpoint models, we find that PAIRED-FOCUS understanding emerges later in training than PAIRED-FOCUS syntactic knowledge, and that learning of PAIRED-FOCUS semantics is correlated with gains in some domains of world knowledge. Overall, our empirical results support the conclusion that modestly sized open-source models can grasp the rare PAIRED-FOCUS constructions, and demonstrate a connection between knowledge of PAIRED-FOCUS constructions and other meaning domains.[1]

1 Introduction

Language models (LMs) have great potential as tools for studying language. The success of a domain-general statistical learner on many linguistic tasks has prompted some researchers to argue that LMs should be more central in evaluating and developing linguistic theory (Warstadt and Bowman, 2022; Futrell and Mahowald, 2025). In this work, we focus on constructionist approaches, or Construction Grammar (Goldberg, 1995, 2006; Croft, 2001, *inter alia*), a family of linguistic theories which posit that language is primarily structured as form-function mappings of gradient complexity. Construction Grammar accounts for not only the most integral parts of linguistic structure (e.g. verb argument structures; Goldberg, 1995), but also rare phenomena traditionally relegated to the "periphery", arguing that both can be accounted for with similar representations (Fillmore et al., 1988). Construction Grammar's ability to describe and account for rare linguistic phenomena makes it a valuable framework for studying language models, as mastery of rare and complex constructions is a crucial part of humans' linguistic knowledge, and may be particularly challenging for language models due to scarcity in their input.

At the same time, the general success of neural language models, which do not explicitly distinguish between syntax and semantic information, has led some researchers to argue that Construction Grammar is more aligned with LM processing of language relative to other frameworks (Weissweiler et al., 2023; Goldberg, 2024; Piantadosi, 2024). In order to fully evaluate these claims, it is worth investigating the extent to which LMs can serve as "model learners" (Warstadt and Bowman, 2022) of constructionist theories of language.

Because Construction Grammar posits that form and meaning are not separable parts of linguistic knowledge, a constructionist "model system" should learn and understand constructions with respect to both their forms and functions. However, up to this point, there are somewhat mixed results regarding LMs' syntactic and semantic knowledge of rare constructions. There is ample evidence that even relatively small language models learn formal properties of rare constructions, including ARTICLE-ADJECTIVE-NUMERAL-NOUN (Misra and Mahowald, 2024), LET-ALONE (Rozner et al., 2025b; Scivetti et al.,

[1]https://github.com/WesScivetti/Meaning_Alone

Cxn	Olmo3	Pythia	COCA	Prop.	Cxn Freq. COCA
LET-ALONE	4261	2874	8631	1.00	8311–8631
MUCH-LESS	8671	5521	13184	0.44	4570–7089
NOT-TO-MENTION	3862	2183	8803	0.38	2561–4207
NEVER-MIND	807	796	2974	0.31	677–1209

Table 1: **PAIRED-FOCUS construction frequencies in LM pretraining data and COCA.** String frequencies are reported for Olmo3 pretraining data, Pythia pretraining data (The Pile) and COCA. All counts are normalized per billion tokens. Prop. refers to the proportion of strings in COCA that were true instances of the construction out of 100 sampled instances. Cxn Freq. refers to the approximate frequency of the construction based on 95% confidence intervals of the sampled proportion in COCA. LM pretraining counts computed using infini-gram (Liu et al., 2024).

2025a), and the COMPARATIVE-CORRELATIVE (Weissweiler et al., 2022). However, in many cases, these smaller models seem unable to grasp the semantics of these rare constructions (Weissweiler et al., 2022; Scivetti et al., 2025a). On the other hand, extremely large LLMs have been shown to have relatively nuanced semantic understanding of a range of constructions (Mortensen et al., 2024; Scivetti et al., 2025b). Since most LLM evaluation has been done in the prompting setting on closed-source LLMs, it remains unclear whether semantic understanding of constructions can be observed in the raw probabilities of (smaller) language models, particularly for exceedingly rare constructions.

Additionally, it is not clear that all current evaluation datasets are well suited for testing constructional semantic understanding. Past work relies on unnatural sounding test items (Weissweiler et al., 2022; Scivetti et al., 2025a) or more complex metalinguistic tasks which may limit LM performance (Bonial and Tayyar Madabushi, 2024). Additionally, particularly for rare constructions, it is likely that constraints on their behavior are related to more general phenomena in language (Potts, 2024; Misra and Mahowald, 2024), and it is reasonable to expect that the semantics of constructions are shaped and constrained by their interactions with other domains of semantics.

In this work, we address the above gaps with a new dataset for testing a rare[2] family of four constructions: LET-ALONE, MUCH-LESS, NOT-TO-MENTION, and NEVER-MIND, collectively known as PAIRED-FOCUS constructions because they conjoin two phrases that are both in focus:

(1) He doesn't like shrimp, let alone squid. (Fillmore et al., 1988)

We evaluate PAIRED-FOCUS syntax and semantics on a much wider range of models than has been tested in past work on constructional semantics. We ground the creation of our dataset in other factors in language production which are likely to inform learning of the constructions' semantics: scalar adjectives and general world knowledge.

This allows us to ask and answer detailed research questions about the acquisition of PAIRED-FOCUS semantics.

1) How do training data, parameter count, and pretraining objective impact knowledge of PAIRED-FOCUS meaning? We show that a range of medium-sized open-source models show sensitivity to the construction-level semantics in their raw probability distributions. However, we do not observe any above-chance knowledge of PAIRED-FOCUS semantics for models trained on human-scale data (BabyLMs, Warstadt et al., 2023). **2) When are PAIRED-FOCUS form and meaning learned throughout pretraining?** We show that PAIRED-FOCUS form is consistently acquired prior to PAIRED-FOCUS semantics. Furthermore, we show that inferences early in learning are often influenced by typicality (world knowledge) rather than constructional meaning, particularly for weaker models. We find that the different PAIRED-FOCUS constructions we test are highly correlated with one another in terms of both syntactic and semantic performance. **3) How does PAIRED-FOCUS meaning correlate with performance on other linguistic benchmarks?** We examine the learning dynamics of both the form and meaning of the PAIRED-FOCUS constructions, in relation to performance on a range of existing linguistic benchmarks. We find that performance on our syntactic evaluations of the constructions plateaus early in pretraining (mirroring learning curves for the BLiMP (Warstadt et al., 2020) grammatical benchmark), while functional knowledge is acquired much later and more closely mirrors the learning trajectory of the world-knowledge-based

[2]For approximate corpus frequencies see Table 1.

EWoK (Ivanova et al., 2025) benchmark.

Overall, our results show that PAIRED-FOCUS constructions can be learned by relatively small models (under 400M parameters). They also shed light on the interrelatedness of PAIRED-FOCUS constructions with scalar semantics, and with domains of world knowledge in which such scalar semantics are relevant. Furthermore, the failure of all "human-scale" models on our semantic evaluation points calls into question the ability of human-scale pure text LMs to serve as model systems of linguistic theories that posit joint acquisition of form and meaning.

2 Background

2.1 Paired-Focus Constructions

In this work, we focus on four related PAIRED-FOCUS constructions: LET-ALONE, MUCH-LESS, NEVER-MIND, and NOT-TO-MENTION. We focus on these constructions because their syntactic and semantic properties are well established in Construction Grammar theory (Fillmore et al., 1988), and have also been studied in past computational work (e.g. Bonial and Tayyar Madabushi, 2024; Rozner et al., 2025a,b; Scivetti et al., 2025a), though past work on investigating language model understanding of PAIRED-FOCUS semantics has yielded mostly negative results.

Fillmore et al. (1988) provide a comprehensive treatment of the LET-ALONE construction, though much of their analysis can be extended to the other PAIRED-FOCUS constructions in this work. Syntactically, these PAIRED-FOCUS constructions function somewhat similarly to coordinating conjunctions (Examples 1 and 2), but resist various types of syntactic movement (Example 3), and generally behave like negative polarity items (Example 4).

(2) He doesn't like shrimp, or squid.

(3) ??It's shrimp, not to mention squid, that he doesn't like.

(4) ??I like shrimp, much less squid.

Regarding PAIRED-FOCUS semantics, the constructions indicate a relationship between two conjoined and focused elements which are being compared. The comparison between the focused elements evokes a scalar relationship with the two elements representing "two points on a scale" (Fillmore et al., 1988). Generally, the second focused element has a higher value on the scale than the first focused element. Overall, the semantics of PAIRED-FOCUS constructions are related to scalar semantics more generally (in the invocation of a scale) and to world knowledge, which informs what scalar properties are natural for the focused items in the construction.

2.2 Related Work

2.2.1 Linguistic Capabilities of LMs

Broadly, this work follows in a long line of research seeking to investigate LM knowledge of linguistic phenomena by examining LM probabilities for grammatical and ungrammatical sequences (Hu et al., 2020, *inter alia*). Among the most relevant of these works are those which seek to design datasets to assess broad domains of linguistic abilities, whether that be syntactic, conceptual, or world knowledge. We leverage several of these datasets as comparison points for learning dynamics of PAIRED-FOCUS constructions. Furthermore, investigations of scalar semantics of adjectives (Garí Soler and Apidianaki, 2020; Schuster et al., 2020; Lin et al., 2024; Nizamani et al., 2024) and adverbs (Lorge and Pierrehumbert, 2023) are of particular relevance to this present work, though most past work has focused on scalar implicature and relative intensity of adjectives, which is orthogonal to the present study. Contributing to this line of work are several datasets containing scalar adjectives of varying intensities (de Melo and Bansal, 2013; Wilkinson and Tim, 2016; Cocos et al., 2018). We rely on the dataset of Wilkinson and Tim (2016) for the scales and adjectives in our dataset.

2.2.2 LM Understanding of Rare Constructions

Past work on evaluating LM understanding of constructions has primarily focused on evaluating various constructions that have been outlined in theoretical linguistics, such as argument structure constructions (Li et al., 2022; Veenboer and Bloem, 2023; Sung and Kyle, 2024), the COMPARATIVE-CORRELATIVE (Weissweiler et al., 2022), CAUSAL-EXCESS (Zhou et al., 2024), NPN (Scivetti and Schneider, 2025), PIPP (Potts, 2024), AANN (Mahowald, 2023; Chronis et al., 2023; Misra and Mahowald, 2024), and LET-ALONE (Scivetti et al., 2025a). Other work opts for a more general investigation of a range of constructions (Tayyar Madabushi et al., 2020; Rozner et al., 2025b; Scivetti et al., 2025b), or discovers constructions using unsupervised methods (Dunn, 2017;

Idx	Name/Description	Feat.	Sent.1	Feat.	Sent. 2: ___ *than lifting a huge one.*	Pairwise Feat.
6	Cxn Entails Plaus.	+Cxn	I couldn't lift a tiny rock, let alone a huge one.	+Plaus.	Lifting a tiny rock is **easier**	Entailment
7	Cxn Contradicts Implaus.	+Cxn	I couldn't lift a tiny rock, let alone a huge one.	−Plaus.	Lifting a tiny rock is **harder**	Contradiction
8	Neutral Plaus. Control	−Cxn	I couldn't lift a tiny rock, or a huge one.	+Plaus.	Lifting a tiny rock is **easier**	Neutral
9	Neutral Implaus. Control	−Cxn	I couldn't lift a tiny rock, or a huge one.	−Plaus.	Lifting a tiny rock is **harder**	Neutral
10	Cxn Contradicts Plaus.	+Cxn	I couldn't lift a huge rock, let alone a tiny one.	+Plaus.	Lifting a tiny rock is **easier**	Contradiction
11	Cxn Entails Implaus.	+Cxn	I couldn't lift a huge rock, let alone a tiny one.	−Plaus.	Lifting a tiny rock is **harder**	Entailment

Table 2: PAIRED-FOCUS semantics dataset overview. Sent. 1 serves as the context and Sent. 2 as the target.

Beuls and Van Eecke, 2024; Verheyen et al., 2025). While most work on constructional understanding in LMs has focused on English, there is growing work on evaluation of constructions in languages beyond English (Tseng et al., 2022; Bunzeck et al., 2025; Huang et al., 2025; Yang, 2025).

PAIRED-FOCUS constructions have been addressed in several past works. Rozner et al. (2025a) and Rozner et al. (2025b) use Global Affinity measures to show that both RoBERTa and a range of BabyLMs have knowledge of collocations for the PAIRED-FOCUS constructions LET-ALONE and MUCH-LESS. Our results robustly replicate theirs. Bonial and Tayyar Madabushi (2024) and Scivetti et al. (2025b) include LET-ALONE as one of several constructions and find that large, closed-source LLMs can perform a metalinguistic grouping task and natural language inference successfully on examples of these constructions. Our work diverges from these approaches in that we test smaller models using direct, probability-based evaluations.

Our work is most similar in approach to Scivetti et al. (2025a), who design probability-based measures for syntactic and semantic properties of LET-ALONE. As discussed previously, we argue that their dataset for semantics relies on arbitrary contrasts which do not clearly implicate scalar semantics. Furthermore, they only test a single model architecture (OPT-125M) and data scenario (100M tokens of BabyLM), and only address LET-ALONE and not the highly related MUCH-LESS, NOT-TO-MENTION, and NEVER-MIND.

3 Paired-Focus Dataset

We construct a novel dataset to test the semantics of the PAIRED-FOCUS constructions. Past work (Scivetti et al., 2025a) used a fully arbitrary dataset, where there was no obvious scalar relationship between the focused elements (outside of what may be communicated by the construction itself):

(5) I couldn't lift the orange crate, let alone the green crate. (Scivetti et al., 2025a)

Failure on fully arbitrary evaluations could be due to several factors, apart from the models truly not understanding the construction. It is also possible that they may be unable to map the arbitrary relationship between the two focused elements onto any scale, which would also lead to a failure to understand the construction overall. If this is the case, we would expect the models to fail to comprehend arbitrary examples (like Example 5) but succeed at understanding examples with a coherent scale. Furthermore, it is possible that the models have only limited knowledge of the construction, where the constructions can only be interpreted in contexts where the scalar relationship between the focused elements is already obvious due to world knowledge. We would then expect the models to succeed in cases where there is a clear scalar relationship that makes sense in the context of world knowledge but fail in cases where there is a clear scalar relationship which conflicts with world knowledge.

To distinguish between these possibilities, we construct a new dataset that contains PAIRED-FOCUS examples where the focused elements have clear scalar relationships between them. We use adjectives and scales from Wilkinson and Tim (2016) as a starting point for the focused elements that will be compared within the construction. In all examples, we pair adjectives from their dataset with a set of common nouns that are natural sounding with the chosen scalar adjectives.[3] In total, we generate 198 starting templates across 4 unique scales from Wilkinson and Tim (2016), and then fill the resulting templates with a set of appropriate adjectives and nouns, resulting in a dataset of 3.5k example sentence pairs per construction. Example sentences for all scales are shown in Table 4 in Appendix A.

[3] PAIRED-FOCUS constructions can focus other syntactic phrases besides noun phrases. To control for the confound of how syntactic structure could influence understanding, we only place noun phrases in the focused slot of the construction.

3.1 Evaluation of Dataset

To test model sensitivity to PAIRED-FOCUS semantics, we measure the effect that the constructions have on model output probabilities on a follow-up sentence that would be entailed by the PAIRED-FOCUS construction. Because PAIRED-FOCUS constructions imply that the second focused element is a higher value on the scale than the first focused element (Fillmore et al., 1988), we include follow-up sentences which either entail or contradict this scalar relationship. Models are expected to prefer the follow-up which reinforces the semantics of the construction to one which contradicts the construction (compare 6 and 7 in Table 2).

Directly comparing the model probabilities of follow up-sentences is confounded by the logical plausibility of the follow-ups independent of the construction. We expect that, irrespective of the presence of a PAIRED-FOCUS construction, a good model will assign a higher probability to a follow-up that aligns with world knowledge, all other things being equal. To control for this fact, we compare example pairs with a PAIRED-FOCUS construction to examples *without* a construction with scalar semantics, specifically the simple conjunction "or" (see 8 and 9, Table 2). Intuitively, if the model is sensitive to PAIRED-FOCUS semantics, we expect a model to assign higher probability to a follow-up that aligns with the PAIRED-FOCUS construction, beyond what is assigned to the follow-up due to world knowledge.

Concretely, given a base sentence $S_{+\mathbf{Cxn}}$ and a follow-up sentence $T_{\pm\mathbf{Plausible}}$, we calculate the difference ΔP as follows. Here, $\imath_\theta = -\log p_\theta$ denotes the surprisal value derived from a language model parameterized by θ.

$$\Delta P(+\mathbf{Cxn}) = \imath_\theta(T_{-\mathbf{Plaus.}} \mid S_{+\mathbf{Cxn}}) - \imath_\theta(T_{+\mathbf{Plaus.}} \mid S_{+\mathbf{Cxn}}) \quad (1)$$

$\Delta P(-\mathbf{Cxn})$ for the control condition with "or" is computed similarly. Finally, given a dataset with k example pairs, we then derive an accuracy score by comparing the PAIRED-FOCUS condition and the control ("or") condition:

$$\mathbf{Acc_{PF}} = \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}[\Delta P(+\mathbf{Cxn}) > \Delta P(-\mathbf{Cxn})] \quad (2)$$

There are other confounding factors (beyond the PAIRED-FOCUS construction and world knowledge) that could influence LM probability associated with the follow-up sentences. Since the two follow-ups are minimal pairs (with only one word different between them) there is a potential confound of lexical bias if the correct sentence has a more frequent/probable word irrespective of context (e.g. since "easier" is more frequent than "harder" overall). Another potential confound is if there is an ordering effect of the two focused elements, where a follow-up sentence is preferred or dispreferred for presenting focused elements in the same order as the PAIRED-FOCUS sentence. We control for both of these confounds by balancing both the plausible and implausible follow-ups in our dataset by ordering and by the lexical items that appear in the entailed follow-up.

3.2 Syntactic Evaluation

Beyond evaluating PAIRED-FOCUS semantics, we also wish to test if models have knowledge of the *forms* of PAIRED-FOCUS constructions. We create a syntactic evaluation suite which adapts the syntactic tests from Scivetti et al. (2025a) to our dataset. Specifically, we select three syntactic manipulations from Scivetti et al. (2025a) which are ungrammatical for PAIRED-FOCUS constructions but grammatical for simple conjunctions (see Table 3). Similarly to our PAIRED-FOCUS semantic tests, we evaluate the ΔP values between grammatical and ungrammatical sentences compared to simple conjunctions. For more details on the syntactic evaluation suite, as well as evaluations of Global Affinity (Rozner et al., 2025a), see Appendix C.

4 Experiment 1: Impact of Model Size and Training Data

4.1 Models and Evaluation

We test a total of 36 models, varying in parameter count, pretraining data size, pretraining objective (masked vs. autoregressive),[4] and model family. See Table 9 in Appendix E for model details. In addition to computing accuracy, we run a linear mixed effects model analysis to test the impact of model size, pretraining data, and architecture across all models. The dependent variable is PAIRED-FOCUS accuracy (see Equation 2). We include pretraining data (log # of tokens), parameters (log # of parameters), and architecture (causal vs. MLM) as fixed effects, and a random intercept for model name.

[4]For MLM models, we evaluate the probability at the target region (e.g. "easier" vs. "harder"), whereas for the decoder models, we evaluate the probability of the entire follow-up sentence conditioned on the first sentence. See Appendix B for a replication with MLM scoring using Psuedo-log-likelihood.

Manipulation	Feat.	Manipulated Sentence	Base
Clause Conjunction	+Cxn	*I couldn't lift a tiny rock, let alone I couldn't lift a huge one.	I couldn't lift a tiny rock, let alone a huge one.
Clause Conjunction	−Cxn	I couldn't lift a tiny rock, and I couldn't lift a huge one.	I couldn't lift a tiny rock, and a huge one.
NPI	+Cxn	*I could lift a tiny rock, let alone a huge one.	I couldn't lift a tiny rock, let alone a huge one.
NPI	−Cxn	I could lift a tiny rock, and a huge one.	I couldn't lift a tiny rock, and a huge one.
Pseudocleft	+Cxn	*A tiny rock, let alone a huge one, I couldn't lift.	I couldn't lift a tiny rock, let alone a huge one.
Pseudocleft	−Cxn	A tiny rock, and a huge one, I couldn't lift.	I couldn't lift a tiny rock, and a huge one.

Table 3: PAIRED-FOCUS syntactic tests. Tests adapted from evaluation paradigm of Scivetti et al. (2025a). Accuracy is measured by comparing ΔP between **+Cxn** and **−Cxn** manipulated sentences versus Base sentences.

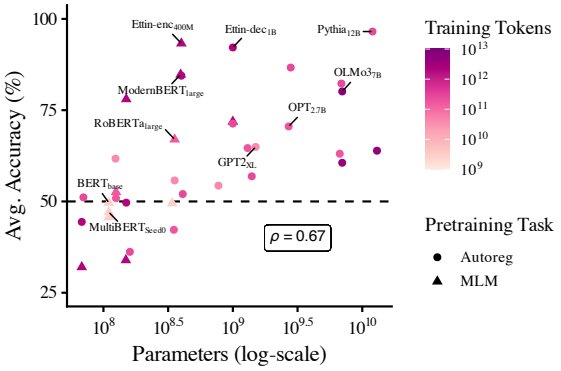


Figure 1: PAIRED-FOCUS Semantic Results by Model. Accuracy is averaged across our four PAIRED-FOCUS constructions: LET-ALONE, MUCH-LESS, NOT-TO-MENTION, and NEVER-MIND. We observe a high rank-order correlation between model parameters and average accuracy (Spearman's $\rho = 0.67$).

4.2 Results

Figure 1 visualizes the PAIRED-FOCUS accuracy[5] as a function of model size and training data amount. Below a parameter threshold of roughly 400M, model performance is generally at or below chance regardless of the amount of training data (with the exception of Ettin-encoder-150M). Beyond this parameter threshold, there is a generally positive relationship between training data and accuracy as well as parameter count and accuracy. However, there is substantial variation between individual models, regardless of parameter count, training data, and training objective.

For our linear modeling analysis, we find that only parameter count has a significant main effect ($\beta = 6.055$, $p = .011$). As a follow-up analysis, we additionally fit a series of linear mixed effects models assessing each predictor independently. For single predictor models (again with a random intercept for model name), we find that both parameter count ($\beta = 6.607$, $p = 003$) and pretraining data ($\beta = 2.651$, $p = .034$) are significant, though the effect of parameter count remains larger. While we note that MLM and autoregressive scoring is not directly comparable, we find no significant effect of pretraining objective.

Regarding our PAIRED-FOCUS form evaluations, we find that even very small models have strong performance on our syntactic evaluation suite, and assign high Global Affinity scores to fixed slots in the construction (see Tables 7 and 8 in Appendix C), echoing past results (Rozner et al., 2025a,b; Scivetti et al., 2025a). The dissociation between tasks for small models underscores the need to evaluate both form and semantics when assessing LMs' overall knowledge of rare constructions.

In summary, these results indicate that models far smaller than frontier LLMs can grasp PAIRED-FOCUS constructions, as we find that a few models as small as 400M parameters succeed at learning (with 90%+ accuracy) both the form and semantics of the constructions, and most of the larger models are above chance. There seems to be a broadly positive relationship between parameter count and semantic accuracy, while the effect of training data independent of parameter count is not well supported by our results. Knowledge of the form of PAIRED-FOCUS constructions is present even in the smallest models we test.

5 Experiment 2: Learning Trajectory Experiments

In this experiment, we investigate when PAIRED-FOCUS constructions are learned in training in relation to other linguistic knowledge. We hypothesize that prior to model knowledge of PAIRED-FOCUS semantics, we will observe knowledge of constructional PAIRED-FOCUS forms. We further hypothesize that PAIRED-FOCUS examples which align with appropriate scalar semantics (given general world knowledge) will be acquired first, especially for models that are less proficient at understanding

[5] We average accuracy scores across each of our four PAIRED-FOCUS constructions.

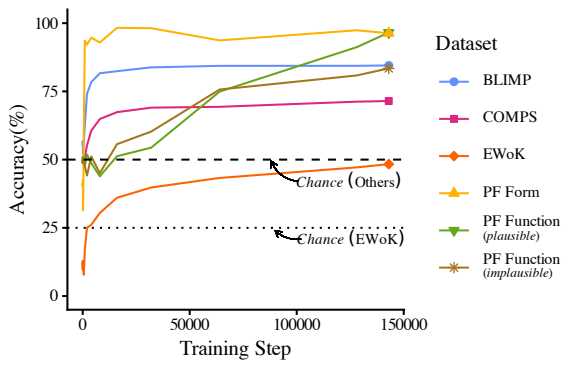


Figure 2: Training dynamics of Pythia-12b on PAIRED-FOCUS evaluations as well as other linguistic benchmarks. Chance performance on EWoK is 25%, while chance performance on all other evaluations is 50%.

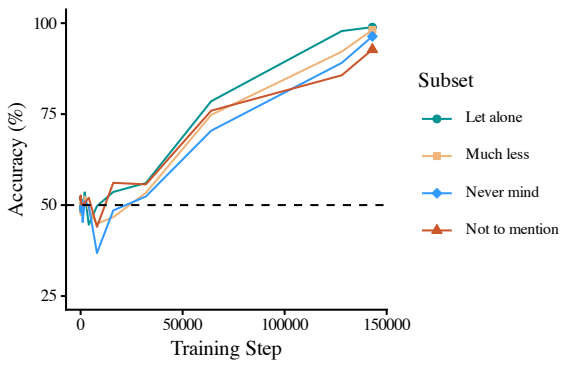


Figure 3: Training dynamics of Pythia-12b for each of the four individual constructions.

the constructions.

To operationalize semantic and formal knowledge of PAIRED-FOCUS constructions, we use PAIRED-FOCUS accuracy for semantics (Equation 2) and our syntactic test suite identical to Experiment 1. In order to test how alignment of scalar semantics with world knowledge interacts with learning of PAIRED-FOCUS constructions, we further test models on sentences where constructional examples entail implausible follow-up sentences and contradict plausible sentences (see 10 and 11 in Table 2). Like in Experiment 1, we consider an example correct if the presence of the construction shifts increases the probability of the follow-up entailed by the construction, relative to a baseline conjunction "or". Intuitively, if a model has a fully abstract understanding of PAIRED-FOCUS semantics, we expect the presence of a PAIRED-FOCUS construction to increase the probability of an otherwise implausible statement that is consistent with the scalar relationship implied by the construction.

5.1 Comparison with Linguistic Benchmarks

We further evaluate the learning dynamics for three other linguistic benchmarks as comparison points throughout training. We evaluate BLiMP (Warstadt et al., 2020), COMPS (Misra et al., 2023), and EWoK (Ivanova et al., 2025). BLiMP tests model knowledge of a range of syntactic constructions. We expect this dataset to be learned relatively early. COMPS tests knowledge of conceptual properties of noun classes. Knowledge of some properties tested in COMPS may be relevant to linking the noun phrases in our dataset with appropriate scalar semantics, and thus we expect high accuracy on COMPS to precede PAIRED-FOCUS semantic accuracy. EWoK tests a range of domains of world knowledge, including physical and material properties. Understanding of these properties is crucial to interpreting the scales in our dataset, and thus we expect that high performance on EWoK will correlate with PAIRED-FOCUS semantic accuracy.

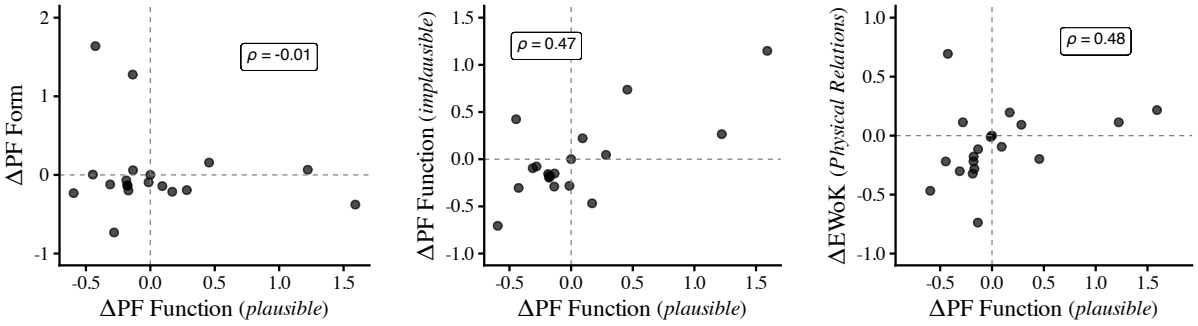
5.2 Models

We focus on learning dynamics of three models in Experiment 1: Pythia-12b (Biderman et al., 2023), Ettin-encoder-400m, and Ettin-decoder-1b (Weller et al., 2025). We select these models because they are the top three performing models at their final checkpoints in Experiment 1. For Pythia-12b, we sample 19 logarithmically spaced checkpoints from start to finish in training. For Ettin models, where logarithmically spaced early checkpoints are not available, we evaluate the first 50 checkpoints[6] for each model. For each of these model checkpoints, we evaluate PAIRED-FOCUS semantic accuracy (for both plausible and world-knowledge implausible examples), syntactic accuracy, and accuracy on each of the comparisons.

5.3 Results

For the purpose of clarity, we primarily discuss and visualize results for Pythia-12b in the main body of the paper. Results on Ettin models (for details see Appendix D) are broadly qualitatively similar, but display noisier and more inconsistent learning trajectories. We also find that the Ettin models display a more substantial difference in constructional effect on plausible compared to implausible follow-ups, especially early in training. In contrast to the Ettin models, which seem more reliant on

[6] Each checkpoint is equivalent to roughly 8-9 billion tokens of pretraining data, and 50 checkpoints is thus equivalent to approx. 400 billion pretraining tokens.



(a) Form vs. plausible accuracy differences.

(b) Implausible vs. plausible accuracy differences.

(c) EWoK physical relations vs. plausible semantic accuracy.

Figure 4: **Learning trajectory correlation scatterplots for Pythia-12b.** Each point represents, for a given checkpoint, how much the model improved over the previous checkpoint with respect to a pair of criteria. EWoK physical relations and PAIRED-FOCUS semantic accuracy show moderate correlation.

world knowledge cues for their interpretations of PAIRED-FOCUS constructions, Pythia-12b displays a sensitivity to the construction which is robust to implausible contexts.

Results for Pythia-12b are visualized in Figure 2. We find strong evidence that PAIRED-FOCUS form is learned prior to PAIRED-FOCUS semantics. Syntactic accuracy reaches a peak much earlier in training than PAIRED-FOCUS accuracy. We find performance on PAIRED-FOCUS semantic evaluations follows similar trajectories regardless of the plausibility of the follow-up sentence, though plausibility does lead to higher absolute performance at the end of training.

For both BLiMP and COMPS, accuracy plateaus relatively early in training, substantially before the PAIRED-FOCUS accuracy rises above chance. For EWoK, performance gains are more gradual, similar to gradual gains on PAIRED-FOCUS semantics.[7]

We report results for each PAIRED-FOCUS construction individually and graph results for Pythia-12b in Figure 3. We observe that performance across the constructions is generally strongly correlated, with all constructions following broadly similar trajectories. Of the four, LET-ALONE is learned more linearly across training, while performance curves for the other constructions are more logarithmic and peak later in training. While we cannot establish a causal relationship, we note that LET-ALONE is the most frequent and least ambiguous of the four (see Table 1) and thus it is perhaps unsurprising that it is learned earlier in training.

[7]Chance accuracy for EWoK is 25%.

5.4 Correlation Analysis of Learning Trajectories

We run a first-difference correlation between PAIRED-FOCUS semantic performance and performance on other benchmarks. We find negligible correlation between PAIRED-FOCUS form and semantics, and between PAIRED-FOCUS semantics and BLiMP. We find a moderate correlation between PAIRED-FOCUS semantic performance on plausible and implausible follow-ups, further showing Pythia-12b has substantial knowledge of scales implied by PAIRED-FOCUS constructions beyond the scalar relationships evident through world-knowledge alone. When subdividing EWoK into its different world knowledge domains, we find a moderate correlation between PAIRED-FOCUS semantics and the physical relations domain (ρ = .48). The first-order correlations between PAIRED-FOCUS formal accuracy, semantic accuracy, and EWoK physical relations accuracy are shown in Figure 4.

5.5 Discussion

Overall, these results provide strong evidence that PAIRED-FOCUS form and meaning are learned with vastly different amounts of training input. We find that performance across individual PAIRED-FOCUS constructions is strongly correlated, and that overall PAIRED-FOCUS performance is moderately correlated with performance on relevant world knowledge domains in EWoK. We find no evidence that PAIRED-FOCUS form and meaning acquisition are correlated in any way, nor is PAIRED-FOCUS semantics significantly correlated with the more syntactic BLiMP benchmark. Regarding

model comparison, we find that the learning trajectories of Pythia-12b are much more stable than the trajectories of the smaller Ettin models, which are more susceptible to spikes and valleys in performance across training, and are less proficient at PAIRED-FOCUS semantics when follow-up sentences are implausible.

6 Discussion

In this work, we have shown that medium-sized models (≈400M parameters) can acquire knowledge of PAIRED-FOCUS semantics. However, small models, and models trained on human-scale data, show a clear gap in performance on PAIRED-FOCUS form and meaning at their final checkpoints. Furthermore, we have shown that even models which eventually learn PAIRED-FOCUS meaning only do so much later in training relative to PAIRED-FOCUS form. The performance gap we observe for PAIRED-FOCUS constructions echoes past results on COMPARATIVE-CORRELATIVE (Weissweiler et al., 2022) and CAUSAL-EXCESS (Zhou et al., 2024) constructions, which similarly found that LMs fail at semantic evaluations which target those constructions. However, we note that our syntactic tests are inherently different from our semantic tests, and interact with different parts of grammar (e.g., other constructions like conjunction and pseudoclefting for syntactic tests vs. scalar semantics for semantic tests). While our results do not provide definitive evidence that form and meaning are necessarily learned separately by LMs, they underscore that modeling joint learning of constructional form and meaning (as would be ideally possible for a "model system" of constructional acquisition) has not been clearly demonstrated by LMs thus far.

While we do not claim that are results prove that LMs *could never* jointly acquire form and meaning via text-only language modeling, there are obvious reasons why LMs would be unlikely to model a constructionist account of language development. Humans learn language in social settings in order to perform communicative goals. We are also exposed to rich non-linguistic input in the form of embodied experience. Pure text language models do not have access to these rich sources of input, which are likely particularly relevant for extracting semantic and pragmatic features of language. We look towards future work on multimodal human-scale models (Hu et al., 2024; Wang et al., 2025) and models that incorporate human feedback (Ziegler et al., 2020) or learning in situated contexts (Beuls and Van Eecke, 2024; Botoko Ekila et al., 2025) as potential avenues for future work.

Finally, we highlight the importance of evaluation design. By improving on past evaluations, we find evidence of PAIRED-FOCUS semantic learning in models that are much smaller than previously reported. Further innovation in evaluation methods for constructional meaning may yet reveal that functional abilities emerge earlier in training, and at smaller parameter scales, than found here.

7 Conclusion

In this work, we have investigated if LMs can learn nuanced semantic interpretations for a rare family of PAIRED-FOCUS constructions. We find that models larger than ≈400M parameters—both encoders and decoders—are broadly successful at the task, with larger models generally performing better. We find that smaller models completely fail at our novel semantic benchmark despite robust knowledge of the constructions' forms. This seeming divergence underscores a broader lack of evidence that text-only LMs are jointly modeling form and meaning of constructions. Turning to an analysis of learning dynamics, we show that learning of PAIRED-FOCUS semantics occurs after learning formal knowledge of the construction, and is correlated with learning of relevant world knowledge. Our largest high performing model is robust to changes in plausibility when interpreting PAIRED-FOCUS examples, while smaller models are more sensitive to the plausiblity of the construction. Overall, this work shows that relatively modest-sized models can acquire nontrivial semantic knowledge of rare constructions, and finds correlational evidence relating learning of PAIRED-FOCUS semantics to learning trajectories of other realms of linguistic knowledge.

Limitations

This paper is limited in terms of its coverage of constructions. While we evaluate a range of PAIRED-FOCUS constructions in English, it is not clear if our concrete findings would generalize to other constructions which may be less closely related than the set we test here. In addition, while we hope that our work serves as a test case of linking the semantics of a set of rare constructions to other semantic properties of language, our evi-

dence in this paper is not causal: while performance on PAIRED-FOCUS semantics is correlated with certain realms of world knowledge, it is not clear how learning would be impacted if such knowledge were perturbed or removed altogether from training. Furthermore, while models displayed some robustness to implausible follow-up sentences, it is not clear how plausibility is causally linked to learning of PAIRED-FOCUS constructions. The number of scales that we used to create our dataset was somewhat small; increasing the number of scales used (and using less common scales) could impact the results here, especially for smaller models, which display some borderline knowledge of PAIRED-FOCUS semantics. Finally, PAIRED-FOCUS constructions are not unique to English, but our evaluations were limited to English constructions and primarily monolingual models.

Acknowledgments

Leonie Weissweiler was supported by a postdoctoral fellowship from the German Research Foundation (DFG, WE 7627/1-1). This research was supported in part by NSF award IIS-2144881. We thank CoNLL reviewers and members of the NERT and PiCOL labs for their insightful comments which improved the final version of this paper.

References

Katrien Beuls and Paul Van Eecke. 2024. Humans learn language from situated communicative interactions. what about machines? *Computational Linguistics*, 50(4):1277–1311.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.

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.

Jérôme Botoko Ekila, Lara Verheyen, Katrien Beuls, and Paul Van Eecke. 2025. Constructions all the way up: From sensory experiences to construction grammars. In *Proceedings of the Second International Workshop on Construction Grammars and NLP*, pages 84–95, Düsseldorf, Germany. Association for Computational Linguistics.

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.

Gabriella Chronis, Kyle Mahowald, and Katrin Erk. 2023. A method for studying semantic construal in grammatical constructions with interpretable contextual embedding spaces. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 242–261, Toronto, Canada. Association for Computational Linguistics.

Anne Cocos, Skyler Wharton, Ellie Pavlick, Marianna Apidianaki, and Chris Callison-Burch. 2018. Learning scalar adjective intensity from paraphrases. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1752–1762, Brussels, Belgium. Association for Computational Linguistics.

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

Gerard de Melo and Mohit Bansal. 2013. Good, great, excellent: Global inference of semantic intensities. *Transactions of the Association for Computational Linguistics*, 1:279–290.

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

Jonathan Dunn. 2017. Computational learning of construction grammars. *Language and Cognition*, 9(2):254–292.

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.

Richard Futrell and Kyle Mahowald. 2025. How linguistics learned to stop worrying and love the language models. *Behavioral and Brain Sciences*, page 1–98.

Aina Garí Soler and Marianna Apidianaki. 2020. BERT knows Punta Cana is not just beautiful, it's gorgeous:

Ranking scalar adjectives with contextualised representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7371–7385, Online. Association for Computational Linguistics.

Adele E. Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press.

Adele E. Goldberg. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press.

Adele E. Goldberg. 2024. Usage-based constructionist approaches and large language models. *Constructions and Frames*, 16(2):220–254.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

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. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21, Miami, FL, USA. Association for Computational Linguistics.

Xinyao Huang, Yue Pan, Stefan Hartmann, and Yang Yanning. 2025. Assessing minimal pairs of Chinese verb-resultative complement constructions: Insights from language models. In *Proceedings of the Second International Workshop on Construction Grammars and NLP*, pages 144–150, Düsseldorf, Germany. Association for Computational Linguistics.

Anna A. Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi U. Kumar, Setayesh Radkani, Thomas H. Clark, Carina Kauf, Jennifer Hu, R. T. Pramod, Gabriel Grand, Vivian C. Paulun, Maria Ryskina, Ekin Akyürek, Ethan G. Wilcox, Nafisa Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Joshua Tenenbaum, and Jacob Andreas. 2025. Elements of world knowledge (EWoK): A cognition-inspired framework for evaluating basic world knowledge in language models. *Transactions of the Association for Computational Linguistics*, 13:1245–1270.

Carina Kauf and Anna A. Ivanova. 2023. A better way to do masked language model scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935, Toronto, Canada. Association for Computational Linguistics.

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.

Fangru Lin, Daniel Altshuler, and Janet B. Pierrehumbert. 2024. Probing large language models for scalar adjective lexical semantics and scalar diversity pragmatics. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13033–13049, Torino, Italia. ELRA and ICCL.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. In *Proceedings of the First Conference on Language Modeling*.

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

Isabelle Lorge and Janet B. Pierrehumbert. 2023. Not wacky vs. definitely wacky: A study of scalar adverbs in pretrained language models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 296–316, Singapore. 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*, pages 265–273, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

David R. Mortensen, Valentina Izrailevitch, Yunze Xiao, Hinrich Schütze, and Leonie Weissweiler. 2024. Verbing weirds language (models): Evaluation of English zero-derivation in five LLMs. In *Proceedings of*

the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 17359–17364, Torino, Italia. ELRA and ICCL.

Rashid Nizamani, Sebastian Schuster, and Vera Demberg. 2024. SIGA: A naturalistic NLI dataset of English scalar implicatures with gradable adjectives. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14784–14795, Torino, Italia. ELRA and ICCL.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, and 49 others. 2026. Olmo 3. *arXiv preprint*. ArXiv:2512.13961 [cs].

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

Steven Piantadosi. 2024. *Modern language models refute Chomsky's approach to language*, page 353–414. Language Sciences Press, Berlin, Germany.

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

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

Joshua Rozner, Leonie Weissweiler, Kyle Mahowald, and Cory Shain. 2025a. Constructions are revealed in word distributions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2138, Suzhou, China. Association for Computational Linguistics.

Joshua Rozner, Leonie Weissweiler, and Cory Shain. 2025b. BabyLM's first constructions: Causal interventions provide a signal of learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2237–2249, Suzhou, China. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Sebastian Schuster, Yuxing Chen, and Judith Degen. 2020. Harnessing the linguistic signal to predict scalar inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5387–5403, Online. Association for Computational Linguistics.

Wesley Scivetti, Tatsuya Aoyama, Ethan Wilcox, and Nathan Schneider. 2025a. Unpacking *Let Alone*: Human-scale models generalize to a rare construction in form but not meaning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27503–27514, Suzhou, China. 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, Mollie Shichman, Taylor Pellegrin, Austin Blodgett, Claire Bonial, and Harish Tayyar Madabushi. 2025b. Beyond memorization: Assessing semantic generalization in large language models using phrasal constructions. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1184–1201, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Raluca Turc, Jacob Eisenstein, Dipanjan Das, and Ellie Pavlick. 2022. The multiBERTs: BERT reproductions for robustness analysis. In *International Conference on Learning Representations*.

Hakyung Sung and Kristopher Kyle. 2024. Leveraging pre-trained language models for linguistic analysis: A case of argument structure constructions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7302–7314, Miami, Florida, USA. Association for Computational Linguistics.

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.

Yu-Hsiang Tseng, Cing-Fang Shih, Pin-Er Chen, Hsin-Yu Chou, Mao-Chang Ku, and Shu-Kai Hsieh. 2022. CxLM: A construction and context-aware language model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, page 6361–6369, Marseille, France. European Language Resources Association.

Tim Veenboer and Jelke Bloem. 2023. Using collostructional analysis to evaluate BERT's representation of linguistic constructions. In *Findings of the Association for Computational Linguistics: ACL 2023*, page 12937–12951, Toronto, Canada. Association for Computational Linguistics.

Lara Verheyen, Jonas Doumen, Paul Van Eecke, and Katrien Beuls. 2025. You shall know a construction by the company it keeps: Computational construction grammar with embeddings. In *Proceedings of the Second International Workshop on Construction Grammars and NLP*, pages 75–83, Düsseldorf, Germany. Association for Computational Linguistics.

Shengao Wang, Arjun Chandra, Aoming Liu, Venkatesh Saligrama, and Boqing Gong. 2025. Babyvlm: Data-efficient pretraining of vlms inspired by infant learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, page 1380–1390, Honolulu, Hawaii.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. 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, Taiqi He, Naoki Otani, David R. Mortensen, Lori Levin, and Hinrich Schütze. 2023. Construction grammar provides unique insight into neural language models. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, pages 85–95, Washington, D.C. Association for Computational Linguistics.

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.

Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, and Benjamin Van Durme. 2025. Seq vs seq: An open suite of paired encoders and decoders. *Preprint*, arXiv:2507.11412.

Bryan Wilkinson and Oates Tim. 2016. A gold standard for scalar adjectives. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2669–2675, Portorož, Slovenia. European Language Resources Association (ELRA).

Xiulin Yang. 2025. Language models at the syntax-semantics interface: A case study of the long-distance binding of Chinese reflexive ziji. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3808–3824, Abu Dhabi, UAE. Association for Computational Linguistics.

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 preprint*. 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.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences. *arXiv preprint*. ArXiv:1909.08593 [cs].

A Dataset Examples

This Appendix contains examples for each frame and construction. See Table 4. All examples in our evaluations are presented with all four PAIRED-FOCUS constructions: LET-ALONE, MUCH-LESS, NOT-TO-MENTION, and NEVER-MIND.

B Experiment 1 MLM Replication with Pseudo-Log-Likelihood

Here, we replicate the semantic tests from Experiment 1 for MLMs using Pseudo-log-likelihood

Scale	# Num Adjectives	# Templates	Example
beautiful — ugly	6	24	You couldn't paint an ugly picture, **let alone** a gorgeous one.
bright — dim	2	2	They couldn't see a bright light, **let alone** a dim one.
good — bad	8	112	We couldn't cook a bad meal, **let alone** a great one.
small — large	10	60	You couldn't pick up a tiny rock, **let alone** a huge one.

Table 4: Example PAIRED-FOCUS sentences for each scale. The number of templates refers to the number of unique combinations of verbs, adjectives, and nouns that were appropriate for that scale. The number of adjectives on the scale is based on Wilkinson and Tim (2016).

(PLL, Salazar et al., 2020), using the formulation from Kauf and Ivanova (2023). We use the minicons library for computing all PLL scores. Table 5 reports the accuracy on the PAIRED-FOCUS semantic benchmark as measured by PLL (compare with Table 6, which contains the full results for the semantic evaluations from Experiment 1). We see similar trends in accuracy as we do in our main evaluation for Experiment 1, where we compare probabilities of masked tokens directly. All MLMs with less than 150 million parameters achieve chance or near chance performance in both settings. We find that ettin-400m, ettin-1b, and ModernBERT-large are the top performing models in both settings.

C Experiment 1: Syntactic Evaluations

We develop a syntactic evaluation suite which is based upon several tests from Scivetti et al. (2025a). The tests focus on three grammatical properties, specifically targeting alternations which are grammatical with simple conjunctions but generally ungrammatical for PAIRED-FOCUS constructions (see Table 3). Similar to our evaluations on PAIRED-FOCUS semantics, we measure ΔP between two conditions, relative to the difference in those conditions when a simple conjunction is present. Unlike the semantic tests, there are no follow-up sentences; rather, we evaluate the differences in probabilities for sentences with PAIRED-FOCUS constructions directly. More specifically, given a base sentence with a PAIRED-FOCUS construction $S_{-\textbf{Manip.},+\textbf{Cxn}}$, we expect that an ungrammatically manipulated sentence $S_{+\textbf{Manip.},\pm\textbf{Cxn}}$ will have a relatively high surprisal value. Thus, we compute for a PAIRED-FOCUS example:

$$\Delta P(+\textbf{Cxn}) = \iota_\theta\left(S_{+\textbf{Manip.},+\textbf{Cxn}}\right) - \iota_\theta\left(S_{-\textbf{Manip.},+\textbf{Cxn}}\right) \quad (3)$$

We then compare to $\Delta P(-\textbf{Cxn})$, which is computed similarly using "or" sentences with manipulations. We calculate accuracy based on ΔP values identically to the PAIRED-FOCUS semantics dataset:

$$\textbf{Acc}_{\textbf{PF}} = \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}\left[\Delta P(+\textbf{Cxn}) > \Delta P(-\textbf{Cxn})\right] \quad (4)$$

Results on our syntactic evaluation suite are reported in 7.

Additionally, to test general familiarity with the wordforms associated with each construction, we use the *Global Affinity* metric presented by Rozner et al. (2025a). Global Affinity is defined as the probability assigned to a target word when it is masked in a string. Given an original string s and an index i, Global Affinity defines $P_{s\setminus\{i\}}$ as the probability distribution at position i. The Global Affinity is then the probability assigned to the correct word w_i when position i is masked:[8]

$$\textbf{GlobalAff}_{s,w_i} = P_{s\setminus\{i\}}(w_i) \quad (5)$$

In our case, we calculate the Global Affinity for each word in our target PAIRED-FOCUS constructions (e.g., "much" and "less" for "much less") and average across all words in the construction. We find that regardless of model size, most models have a very high global affinity for all fixed words in PAIRED-FOCUS constructions (Table 8). This indicates that models are confident that these PAIRED-FOCUS constructions are collocations and the fixed words in the constructions are relatively easy to predict for models of all sizes.

D Experiment 2 Full Results

In this section, we provide extended results on learning dynamics for Ettin-encoder400m and Ettin-decoder1b. Results are shown in Figures 5 and 6 respectively. In general, the performance trajectories for PAIRED-FOCUS semantics are substantially noisier in these models than in Pythia,

[8] Global Affinity is only defined for Masked Language Models (Rozner et al., 2025a). Thus, we only evaluate a subset of our models on this metric.

Architecture	Model	LET-ALONE	MUCH-LESS	NOT-TO-MENTION	NEVER-MIND	Avg
BERT	base-uncased	44.4	50.1	49.7	50.4	48.7
	large-uncased	32.8	45.6	49.2	43.7	42.8
Ettin Encoder	150m	76.9	83.2	68.9	75.5	76.1
	1b	77.0	94.4	55.7	79.3	76.6
	400m	91.4	91.8	94.5	92.2	92.5
	68m	31.4	45.5	27.6	27.0	32.9
ModernBERT	base	36.6	34.3	34.0	41.3	36.5
	large	87.4	93.6	80.9	83.6	86.4
multiBERTs	seed_0	47.5	45.9	42.7	48.4	46.1
	seed_1	44.3	48.6	43.1	42.5	44.6
RoBERTa	base	50.2	46.3	47.2	51.4	48.8
	large	70.6	65.6	58.4	68.6	65.8

Table 5: Semantic Evaluation Scores for MLMs, as scored using Psuedo-log-likelihood (Kauf and Ivanova, 2023).

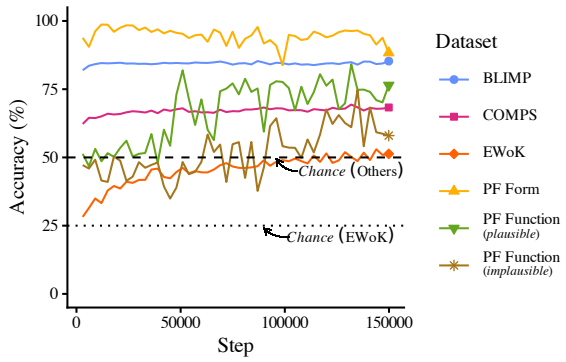


Figure 5: Training dynamics of Ettin-Enc-400M.

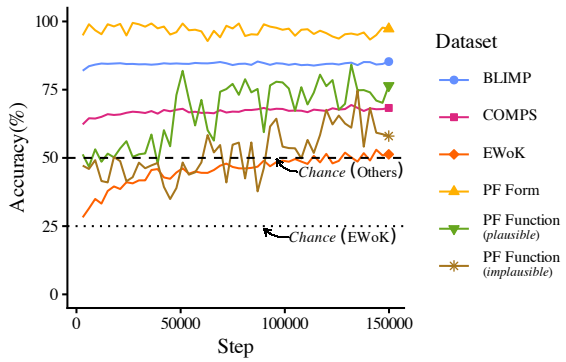


Figure 6: Training dynamics of Ettin-Decoder 1b.

with large peaks and valleys throughout training. For both Ettin models, we observe early spikes in performance on plausible examples that are not accompanied by spikes on implausible examples. Generally, performance on implausible examples does not rise consistently above chance until much later in training. Taken together, these results seem to indicate that the smaller Ettin models do learn nontrivial PAIRED-FOCUS semantics, but may be limited to more natural contexts where the scalar relationship entailed by the construction is further supported by world knowledge. Performance on other benchmarks is generally more stable than on PAIRED-FOCUS semantics, though performance on EWoK is noticeably less stable relative to Pythia.

E Model Details

Table 9 presents details about the models that we test in Experiment 1. In total, we test 36 models in Experiment 1. In Experiment 2, we test 3 models: Pythia-12b, Ettin-encoder400m, and Ettin-decoder1b, which are selected due to their strong performance in Experiment 1 and the availability of their intermediate checkpoints.

Model Type	Family	Model Name	LET-ALONE	MUCH-LESS	NOT-TO-MENTION	NEVER-MIND	Avg.
MLM	BERT (Devlin et al., 2019)	BERT-base-uncased	48.9	49.8	50.1	49.8	49.6
		BERT-large-uncased	41.0	59.8	49.7	47.6	49.5
	ModernBERT (Warner et al., 2025)	ModernBERT-Base	34.1	32.6	31.3	37.7	33.9
		ModernBERT-Large	84.4	92.7	82.3	79.6	84.8
	MultiBERT (Sellam et al., 2022)	MultiBERT seed 0	46.0	49.4	43.4	49.5	47.1
		MultiBERT seed 1	48.2	48.6	44.5	41.5	45.7
	RoBERTa (Liu et al., 2019)	RoBERTa-base	51.5	49.1	56.5	53.3	52.6
		RoBERTa-large	71.7	68.8	56.8	70.7	66.9
	Ettin (Weller et al., 2025)	Ettin-Enc-68M	29.3	45.7	26.5	26.6	32.0
		Ettin-Enc-150M	81.3	84.2	69.3	77.3	78.0
		Ettin-Enc-400M	92.0	91.8	95.9	93.4	93.3
		Ettin-Enc-1B	65.6	94.4	57.4	70.1	71.9
CausalLM	Ettin (Weller et al., 2025)	Ettin-Dec-68M	44.9	41.5	44.4	46.9	44.4
		Ettin-Dec-150M	52.3	50.9	46.	49.5	49.7
		Ettin-Dec-400M	88.4	83.1	84.1	82.2	84.4
		Ettin-Dec-1B	92.3	95.4	87.9	93.1	92.2
	GPT2 (Radford et al., 2019)	GPT2	65.4	75.3	50.1	56.0	61.7
		GPT2-medium	66.3	58.2	48.6	49.9	55.7
		GPT2-large	74.0	54.9	37.9	50.5	54.3
		GPT2-xl	74.1	65.2	57.3	63.1	64.9
	OLMo (OLMo et al., 2025) (Olmo et al., 2026)	Olmo2-7b	64.1	69.0	57.8	51.4	60.6
		Olmo2-13b	73.0	70.8	60.1	51.7	63.9
		Olmo3-7b	79.9	87.4	69.2	84.0	80.1
	OPT (Zhang et al., 2022)	OPT-125M	50.8	56.6	49.0	47.2	50.9
		OPT-350M	50.6	49.7	33.2	35.5	42.2
		OPT-1.3b	69.6	66.7	55.5	66.7	64.6
		OPT-2.7b	72.8	69.5	70.5	69.4	70.5
		OPT-6.7b	65.9	63.9	54.7	67.9	63.1
	Pythia (Biderman et al., 2023)	Pythia-70m	52.7	47.9	45.5	58.5	51.1
		Pythia-160m	36.2	37.2	34.9	36.3	36.2
		Pythia-410m	62.9	65.0	41.4	38.7	52.0
		Pythia-1b	80.6	72.6	68.3	63.6	71.3
		Pythia-1.4b	61.3	63.3	52.5	50.5	56.9
		Pythia-2.8b	94.4	79.1	83.9	89.3	86.7
		Pythia-6.9b	89.0	89.6	67.4	83.2	82.3
		Pythia-12b	98.9	98.1	92.7	96.3	96.5

Table 6: Full Results on our Semantic Test Suite.

Model Type	Family	Model Name	LET-ALONE	MUCH-LESS	NOT-TO-MENTION	NEVER-MIND	Avg.
MLM	BERT (Devlin et al., 2019)	BERT-base-uncased	89.8	96.1	93.6	60.2	84.9
		BERT-large-uncased	59.4	56.9	88.2	73.1	69.4
	ModernBERT (Warner et al., 2025)	ModernBERT-Base	97.9	97.8	99.9	99.6	98.8
		ModernBERT-Large	83.0	83.3	94.0	87.1	86.9
	MultiBERT (Sellam et al., 2022)	MultiBERT seed 0	88.0	89.5	84.8	83.9	86.6
		MultiBERT seed 1	89.7	90.8	76.3	67.9	81.1
	RoBERTa (Liu et al., 2019)	RoBERTa-base	93.3	93.6	99.4	93.3	94.9
		RoBERTa-large	82.8	83.8	94.6	89.6	87.7
	Ettin (Weller et al., 2025)	Ettin-Enc-68M	88.3	93.2	93.5	96.7	92.9
		Ettin-Enc-150M	93.2	92.4	98.7	94.9	94.8
		Ettin-Enc-400M	82.5	87.0	99.3	91.4	90.0
		Ettin-Enc-1B	89.2	90.6	98.3	89.4	91.9
CausalLM	Ettin (Weller et al., 2025)	Ettin-Dec-68M	100	100	89.4	95.8	96.3
		Ettin-Dec-150M	99.9	99.9	90.7	97.8	97.1
		Ettin-Dec-400M	96.5	98.9	98.4	99.5	98.4
		Ettin-Dec-1B	86.9	93.5	89.9	94.0	91.1
	GPT2 (Radford et al., 2019)	GPT2	100	100	97.2	100	99.3
		GPT2-medium	98.2	98.6	94.4	98.4	97.4
		GPT2-large	99.8	99.5	97.4	99.7	99.1
		GPT2-xl	99.0	97.7	89.3	98.6	96.2
	OLMo (OLMo et al., 2025) (Olmo et al., 2026)	Olmo2-7b	97.0	96.5	94.9	95.4	96.0
		Olmo3-7b	91.5	99.3	94.0	93.7	94.6
	OPT (Zhang et al., 2022)	OPT-125M	100	100	94.2	99.6	98.4
		OPT-350M	99.9	100	92.7	93.8	96.6
		OPT-1.3b	97.5	98.6	94.7	96.5	96.8
		OPT-2.7b	96.0	98.2	96.1	94.0	96.1
		OPT-6.7b	98.3	98.9	92.8	97.9	97.0
	Pythia (Biderman et al., 2023)	Pythia-68m	100	100	83.3	99.8	95.7
		Pythia-160m	100	100	94.0	98.1	98.0
		Pythia-410m	99.9	100	97.9	99.7	99.4
		Pythia-1b	99.8	99.7	91.4	99.0	97.5
		Pythia-1.4b	99.9	100	87.3	99.8	96.7
		Pythia-2.8b	94.8	98.7	93.2	99.1	96.4
		Pythia-6.9b	99.4	99.2	96.2	99.6	98.6
		Pythia-12b	97.3	97.6	92.3	97.9	96.3

Table 7: Full Results on our Syntactic Test Suite. Results for each construction are averaged across the 3 manipulation types (see Table 3). OLMo2-13b was not run for syntactic tests due to compute constraints.

Architecture	Model	LET-ALONE	MUCH-LESS	NEVER-MIND	NOT-TO-MENTION
BERT	base-uncased	0.999 ± 0.000	0.987 ± 0.000	0.502 ± 0.008	0.933 ± 0.003
	large-uncased	0.999 ± 0.000	0.999 ± 0.000	0.835 ± 0.005	0.999 ± 0.000
Ettin Encoder	150m	0.995 ± 0.000	0.985 ± 0.000	0.958 ± 0.001	0.960 ± 0.001
	1b	0.995 ± 0.000	0.983 ± 0.000	0.924 ± 0.001	0.951 ± 0.001
	400m	0.990 ± 0.000	0.980 ± 0.000	0.932 ± 0.001	0.925 ± 0.001
	68m	0.998 ± 0.000	0.985 ± 0.000	0.779 ± 0.003	0.797 ± 0.005
ModernBERT	base	0.987 ± 0.000	0.992 ± 0.000	0.903 ± 0.002	0.791 ± 0.004
	large	0.992 ± 0.000	0.985 ± 0.000	0.926 ± 0.002	0.923 ± 0.002
multiBERTs	seed_0	0.999 ± 0.000	0.999 ± 0.000	0.418 ± 0.007	0.741 ± 0.006
	seed_1	0.999 ± 0.000	0.995 ± 0.000	0.439 ± 0.007	0.609 ± 0.007
RoBERTa	base	0.998 ± 0.000	0.998 ± 0.000	0.944 ± 0.001	0.998 ± 0.000
	large	0.996 ± 0.000	0.984 ± 0.000	0.963 ± 0.001	0.991 ± 0.000

Table 8: **Global Affinity Scores By PAIRED-FOCUS Construction.** Intervals represent ± 95% confidence intervals. Results are only reported for MLMs as Rozner et al. (2025a) do not provide a definition of Global Affinity for CausalLMs.

Model Type	Family	Model Name	Parameter Count	Pretraining Data (# of Tokens)
MLM	BERT (Devlin et al., 2019)	BERT-base-uncased	110M	3B
		BERT-large-uncased	335M	3B
	ModernBERT (Warner et al., 2025)	ModernBERT-Base	110M	2T
		ModernBERT-Large	375M	2T
	MultiBERT (Sellam et al., 2022)	MultiBERT seed 0	110M	3B
		MultiBERT seed 1	110M	3B
	RoBERTa (Liu et al., 2019)	RoBERTa-base	125M	160B
		RoBERTa-large	355M	160B
	Ettin (Weller et al., 2025)	Ettin-Enc-68M	68M	2T
		Ettin-Enc-150M	150M	2T
		Ettin-Enc-400M	400M	2T
		Ettin-Enc-1B	1B	2T
CausalLM	Ettin (Weller et al., 2025)	Ettin-Dec-68M	68M	2T
		Ettin-Dec-150M	150M	2T
		Ettin-Dec-400M	400M	2T
		Ettin-Dec-1B	1B	2T
	GPT2 (Radford et al., 2019)	GPT2	125M	160B
		GPT2-medium	355M	160B
		GPT2-large	775M	160B
		GPT2-xl	1.5B	160B
	OLMo (OLMo et al., 2025) (Olmo et al., 2026)	Olmo2-7b	7B	4T
		Olmo2-13b	13B	5T
		Olmo3-7b	7B	6T
	OPT (Zhang et al., 2022)	OPT-125M	125M	180B
		OPT-350M	350M	180B
		OPT-1.3b	1.3B	180B
		OPT-2.7b	2.7B	180B
		OPT-6.7b	6.7B	180B
	Pythia (Biderman et al., 2023)	Pythia-70m	70M	260B
		Pythia-160m	160M	260B
		Pythia-410m	410M	260B
		Pythia-1b	1B	260B
		Pythia-1.4b	1.4B	260B
		Pythia-2.8b	2.8B	260B
		Pythia-6.9b	6.9B	260B
		Pythia-12b	12B	260B

Table 9: Information on all models tested in Experiment 1. Parameter Count and # of Pretraining Tokens are approximate measures.