CxMP: A Linguistic Minimal-Pair Benchmark for Evaluating Constructional Understanding in Language Models

Miyu Oba[1] **Saku Sugawara**[2,3]
[1]Nara Institute of Science and Technology
[2]National Institute of Informatics [3]The University of Tokyo
oba.miyu.ol2@is.naist.jp saku@nii.ac.jp

Abstract

Understanding language acquisition in language models remains an open question, yet many benchmarks focus on grammatical acceptability, with far less attention to interpreting meanings conveyed by grammatical forms. We introduce the Linguistic Minimal-Pair Benchmark for Evaluating Constructional Understanding in Language Models (CxMP), grounded in Construction Grammar, which treats form–meaning pairings (constructions) as fundamental linguistic units. It evaluates whether models interpret the semantic information implied by constructions, using a controlled minimal-pairs across nine types. Our results show that constructional understanding develops more gradually and remains limited for some constructions even in large language models (LLMs), whereas performance on grammatical acceptability emerges earlier, with shallow heuristics in CxMP exhibiting a U-shaped pattern. These findings highlight the need to broaden existing linguistic evaluations to capture meanings encoded in linguistic form. [1]

1 Introduction

Recent work increasingly analyzes language models from a linguistic perspective to better understand the sources of their success and their potential for language acquisition (Kallini et al., 2024; Huebner et al., 2021; Constantinescu et al., 2025). Such analyses involve evaluations of the linguistic abilities of these models. These evaluations primarily focus on grammatical acceptability (Huebner et al., 2021; Warstadt et al., 2019, 2020), assessing whether they distinguish acceptable from unacceptable sentences based on grammatical well-formedness, often emphasizing rule-oriented, less context-dependent cues (Weissweiler et al., 2025). Such methods reflect a linguistic tradition that often treats syntax as largely independent from meaning,

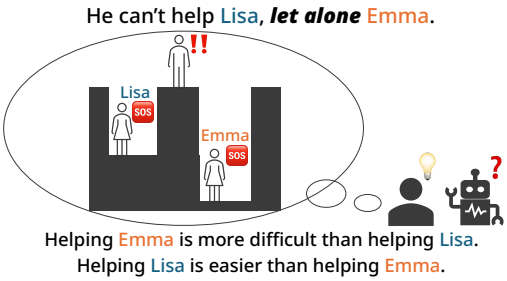


Figure 1: Overview of the let-alone construction. Humans capture the relation between the two entities from sentences containing this construction.

thereby focusing on the grammaticality of linguistic expressions. However, language ability also requires interpreting the meanings implied by form. For instance, *Mary ran into John* not only satisfies grammatical constraints but also conveys that Mary moved toward John, who serves as the endpoint of that motion. To bridge this gap, we adopt the framework of Construction Grammar (Goldberg, 1995, 2006), which treats form–meaning pairings (constructions) as the basic units of linguistic analysis. In this view, syntax and meaning are inherently interconnected, enabling analyses that focus on meanings conveyed by linguistic form.

Previous studies have evaluated language models from a constructional perspective (Weissweiler et al., 2022; Misra and Mahowald, 2024; Potts, 2024; Scivetti et al., 2025a,b). However, existing approaches remain limited for fine-grained linguistic analysis of constructions, as they offer limited support for comprehensive and systematically controlled comparisons across models and constructions. Most focus on a single construction (Weissweiler et al., 2022; Misra and Mahowald, 2024; Potts, 2024; Scivetti et al., 2025a), making it difficult to analyze relations among constructions or connect constructional understanding to other linguistic abilities. In addition, evaluation formats such as natural language inference (NLI) (Mackin-

[1]CxMP is available at https://github.com/nii-cl/cxmp.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 45949–45963

tosh et al., 2025; Scivetti et al., 2025b) present a pair of sentences with a label and require a direct response. Such settings can make the linguistic cues underlying model judgments difficult to interpret and may limit consistent comparisons across models that span from developmentally plausible models to modern large language models (LLMs). Smaller models, in particular, often struggle with prompt interpretation or label classification.

We introduce the Linguistic Minimal-Pair Benchmark for Evaluating Constructional Understanding in Language Models (CxMP), which evaluates whether language models interpret the semantic relationships and roles implied by constructions. For example, as illustrated in Figure 1, in the let-alone construction (*He can't help Lisa, let alone Emma*), humans infer that helping Emma is more difficult than helping Lisa. Going beyond evaluations focused on the well-formedness of sentences, we examine whether language models capture the meanings encoded in linguistic form. CxMP covers nine constructions, uses a controlled minimal-pair design to isolate linguistic cues, and supports evaluation across models ranging from developmentally plausible small models to large-scale LLMs.

We evaluate a range of language models that differ in model size, data scale, and training objectives. We find that developmentally plausible small models make few correct judgments, and even recent 70B-scale open models still struggle with some constructions, indicating substantial room for improvement in construction-level understanding. Performance varies across construction types: formally complex constructions are more affected by model size, whereas argument-structure constructions are more sensitive to instruction tuning.

To demonstrate the usefulness of our dataset for comparison with human language acquisition, we analyze learning trajectories, spurious heuristics, and pseudoword generalization. BLiMP (Warstadt et al., 2020), a benchmark of sentence acceptability, reaches high accuracy early in training, whereas CxMP improves more gradually in interpreting syntactically encoded meaning. CxMP reveals stage-like learning transitions from near-random responses to spurious heuristic use and eventual stabilization, along with a U-shaped pattern in bias magnitude (low–high–low). Under controlled conditions where lexical items are replaced with pseudowords, models show a substantial decline in constructional understanding, unlike humans.

These findings highlight the need for evaluation beyond grammatical acceptability to capture the semantic contributions of constructions. They also establish the proposed dataset as a foundation for future research on constructional grammar.

2 Background and Motivations

2.1 Construction Grammar

Construction Grammar is a family of approaches within cognitive linguistics that, despite differences in detail, take constructions, that is, pairings of form with meaning or function, to be basic units of grammar (Croft, 2001; Fillmore et al., 1988; Goldberg, 1995, 2006; Kay and Fillmore, 1999). Constructions form a continuum from idiomatic expressions to highly abstract schemas, capturing a wide range of linguistic phenomena.

This broad scope has motivated recent attempts in computational linguistics to incorporate constructional notions into model-based analyses of language (Misra and Mahowald, 2024; Potts, 2024; Scivetti et al., 2025a; Weissweiler et al., 2022). In turn, this has given rise to benchmark-oriented resources, from Bonial and Tayyar Madabushi (2024)'s corpus of constructions varying in schematicity to targeted diagnostic benchmarks for constructional understanding (Mackintosh et al., 2025; Scivetti et al., 2025b). However, current resources are fragmented and offer only limited support for systematic comparison across constructions and models.

2.2 Benchmark Requirements

To enable such comparisons, we identify four key properties that a benchmark should satisfy to assess constructional understanding in language models.

Model Agnostic Recent work has examined developmentally plausible small models to investigate language acquisition under limited data conditions (Charpentier et al., 2025). Evidence from LLMs, however, shows that generalization can arise even without strong constraints, driven by inherent simplicity biases that offer insights for linguistic theory (Futrell and Mahowald, 2025). Existing benchmarks for constructional understanding rely on NLI (Mackintosh et al., 2025; Scivetti et al., 2025b), which requires understanding task-specific notions such as entailment and contradiction, making zero-shot evaluation difficult for smaller models that struggle with prompt interpretation and label classification. Fine-tuning further obscures the boundary between pretrained and newly acquired

knowledge. Hence, benchmarks are needed for consistent evaluation across language models of varying sizes.

Comprehensive across Constructions Most prior studies focus on individual constructions such as the comparative correlative (Weissweiler et al., 2022), Article+Adjective+Numeral+Noun (Misra and Mahowald, 2024), let-alone (Scivetti et al., 2025a), or preposing in PP (Potts, 2024), examining how language models recognize, reason about, and generalize over these specific patterns. However, Construction Grammar covers a wide range of constructions, from lexically grounded substantive ones to highly abstract schematic ones. A benchmark that covers diverse constructions supports more comprehensive analyses of the relationships among constructions and their connections to other aspects of linguistic knowledge.

Controlled Design Datasets such as SNLI (Bowman et al., 2015) and CoLA (Warstadt et al., 2019), which consist of sentences paired with labels, evaluate models in diverse ways but can induce shallow heuristics or surface-level biases in model judgments. For example, sentences with high vocabulary overlap are more likely to be classified as entailment (McCoy et al., 2019), and those containing negation tend to receive negative labels (Gururangan et al., 2018). The rationale for labeling is often implicit, obscuring which linguistic cues the model relies on. To mitigate such issues, a minimal-pair design that isolates controlled differences is desirable, with precise evaluation of the linguistic factors to which models are sensitive (Warstadt et al., 2020). Furthermore, common nouns, while natural and frequent, often exhibit strong lexical associations; for instance, teacher co-occurs with teach far more often than with learn (Sakaguchi et al., 2020). To control for such biases, introducing multiple noun variants and alternating participant roles (e.g., swapping subject and object) supports a more fine-grained analysis of underlying model biases.

Multiple Sentences Single-sentence minimal pairs are effective for evaluating the formal naturalness of sentences but insufficient for assessing the semantic contributions of constructions, motivating a multi-sentence design. Among existing datasets adopting such designs, EWoK (Ivanova et al., 2025) is a representative example: it evaluates world knowledge using pairs consisting of a context describing a situation and a target proposition whose plausibility is judged relative to that context. However, benchmarks applying this framework to language acquisition in language models from a linguistic perspective remain limited. As sentence meaning may involve inferences from its surrounding context during training, a multi-sentence minimal-pair design enables investigation of how well language models capture semantic relations between sentences.

3 CxMP

We construct a dataset that integrates these motivations into a unified framework. We assume that a model demonstrates constructional understanding when it correctly identifies the meaning, role, or relation of entities within a construction, reflecting successful form–meaning mapping. We design CxMP to assess this.

3.1 Task Design

Examples of CxMP for each construction are shown in Table 1. Each instance consists of a sentence containing a target construction (constructional example) and a sentence probing its associated meaning (diagnostic sentence). The diagnostic sentence is realized as a minimal contrast between plausible and implausible interpretations; together with the constructional example, it defines an evaluation pair. Each construction has two variants (A and B) that probe complementary perspectives.

3.2 Constructions

Following Bonial and Tayyar Madabushi (2024), we adopt nine types of constructions (Table 1), ranging from partially fixed to fully schematic.

LET-ALONE The two elements linked by let alone represent different points on the same scalar dimension, with the second expressing a stronger degree, typically a stronger form of negation, than the first. Previous linguistic studies, including Fillmore et al. (1988), analyze this construction in detail, as it exhibits unique properties that combine a negative polarity environment with a comparative relation. We assume that actions higher on the negative scale are more difficult, and design sentences to probe which action is harder. While a wide range of syntactic elements (e.g., nouns, verbs, clauses) can appear around let alone, we focus on the simplest and most controllable case: nominal comparison. Because the construction requires a negative polarity licenser, we include negation in the stimuli

Construction	Constructional Example	Diagnostic Sentence
LET-ALONE	He can't help N1, let alone N2.	Helping N2/*N1 is more difficult than helping N1/*N2. Helping N1/*N2 is easier than helping N2/*N1.
CAUSATIVE-WITH	N1 covered N2 with a coat.	The one who provided the coat is N1/*N2. The one who received the coat is N2/*N1.
WAY-MANNER	N1 pushed his way into the room.	He did so while moving/*staying. He did so with/*without effort.
COMPARATIVE-CORRELATIVE (CC)	The harder you work, the stronger you become. N1 works harder than N2.	The one who becomes stronger is N1/*N2.
	The stronger you become, the harder you work. N1 becomes stronger than N2.	The one who works harder is N1/*N2.
CONATIVE	N1 kicked at the ball.	Whether she succeeded in kicking the ball is unclear/*clear. Whether the ball was affected is unclear/*clear.
DITRANSITIVE	N1 bought N2 a gift.	The one who gave the gift is N1/*N2. The one who received the gift is N2/*N1.
CAUSED-MOTION	N1 pulled N2 out of the water.	The one who moved her is N1/*N2. The one who got moved is N2/*N1.
RESULTATIVE	N1 pushed N2 awake.	The one who caused her to become awake is N1/*N2. The one who became awake is N2/*N1.
INTRANSITIVE-MOTION	N1 ran into N2.	The one who moved is N1/*N2. The one who he moved toward is N2/*N1.

Table 1: Constructions that we focus on. * indicates implausible versions. N1 and N2 denote the entities involved.

using negative auxiliaries such as *cannot* and *don't*. Although Bonial and Tayyar Madabushi (2024) include MUCH-LESS, it differs from let alone only in surface form and is therefore omitted.

CAUSATIVE-WITH This construction involves a *with*-phrase following the direct object. The direct object denotes the affected entity, and the *with*-phrase indicates what is applied or provided. We restrict both the subject and the object to human nouns to test whether models correctly interpret the supply–recipient relation between entities.

WAY-MANNER This expresses movement achieved by creating one's own path, often despite external difficulty or through indirect means (Goldberg, 1995). Formally, a fixed phrase one's way appears in the direct-object position, followed by a prepositional phrase indicating the path. We design sentences that probe whether the event involves motion or rest and whether it is carried out with strong intentionality.

COMPARATIVE-CORRELATIVE This construction has been widely studied as a challenge for generative approaches and a prominent construction in English (Fillmore, 1986; Hoffmann, 2017). Recent studies have examined how language models understand it (Weissweiler et al., 2022). It consists of two clauses, both beginning with *the* followed by a comparative adjective or adverb (Goldberg, 2003). The second clause functions as a dependent variable determined by the first, often implying causal or temporal relations. Following Weissweiler et al. (2022), our dataset uses constructional examples in which the first clause establishes a relational condition between entities, and the diagnostic sentence asks which entity exhibits the resulting property.

The following five constructions are identified as representative argument-structure constructions by Goldberg (1995). Following Goldberg (1995), we use Subj (subject), V (verb), Obj (object), Obl (oblique), and Xcomp (open clausal complement).

CONATIVE The construction features a direct object preceded by *at*, indicating that the agent makes an effort to act upon the target without implying success. Half of our constructional examples are conative, and the other half are standard transitive. Diagnostic sentences test whether models can distinguish that the result of the action is uncertain in conative cases, whereas the action is successful and the object is affected in transitive cases.

DITRANSITIVE This construction has the schema Subj V Obj1 Obj2, expressing that the subject transfers or gives the theme (Obj2) to the recipient (Obj1). We evaluate whether models capture this transfer-of-possession meaning by

asking who gave and received the item.

CAUSED-MOTION This construction expresses that the causer causes the theme to move along a path indicated by an oblique phrase. Formally, it follows the schema Subj V Obj Obl, where the oblique phrase always denotes direction or path. We evaluate whether models correctly assign causal semantic roles using diagnostic sentences that ask who caused the motion and who was moved.

RESULTATIVE This construction follows the schema Subj V Obj Xcomp, indicating that the agent's (Subj's) action causes the patient (Obj) to undergo a change of state (Xcomp). The Xcomp is typically an adjective phrase or a prepositional phrase denoting the resulting state. We evaluate whether models can identify which participant underwent the change and which one caused it, testing their understanding of the causal and resultative semantics inherent to the construction.

INTRANSITIVE-MOTION This construction follows the schema Subj V Obl and denotes that the agent (Subj) moves autonomously. In our dataset, the oblique phrase also denotes an animate entity, and diagnostic sentences ask which participant moved and which one was the goal of the movement, assessing whether the model captures directionality and goal relations.

Entities and Generation Procedure We use four types of names that represent entities: *female name*, *male name*, *name+alphabet* (e.g., *Name B*), and *common noun*. Common nouns yield the most natural expressions but can introduce strong lexical biases. Name+alphabet provides the most controlled condition, although it is artificial. Human names mitigate lexical biases and artificiality but may still exhibit gender bias. For constructions involving two entities, we also create swapped versions by reversing their positions to test whether models rely on shallow heuristics. These correspond to sentences with N1 and N2 swapped in Table 1. All four entity types are used in equal proportion, and swapped and non-swapped versions are equally balanced. Details of name selection and switching procedures are provided in Appendix A.

3.3 Data Generation with LLMs

We generate constructional examples for each construction using GPT-5 (Singh et al., 2025) based on predefined templates designed to elicit corresponding diagnostic sentences. For each construction, we prepare templates for constructional examples and plausible diagnostic sentences. Constructional examples consist of slots (e.g., {Subj; N1; male name} {V} ...), and plausible diagnostic sentences are predefined. We provide these templates to the model and prompt it to fill the slots with contextually appropriate words. We then filter the generated sentences using heuristic criteria and LLM-based assessments of semantic coherence, and manually inspect a subset of the dataset to validate the coherence of the evaluation pairs. We create implausible diagnostic sentences by converting specific words into implausible alternatives. In total, we create 43k evaluation pairs. Details of the generation process are provided in Appendix A.

Human Data Validation We conduct a human validation with native English speakers using Prolific (https://prolific.com). Participants read a preceding sentence and select the more plausible continuation. We collect three independent annotations per item. Across 896 items, the majority-vote accuracy reaches 96.65%, and full agreement among annotators reaches 83.59%, suggesting that the automatically generated items are generally reliable. The number of items varies by construction (typically 112 and 56 for constructions without switched variants). Detailed per-construction results and annotation statistics are provided in Appendix C.

4 Experiment

We evaluate how well language models perform on the CxMP using models under various settings.

4.1 Settings

Models We use models motivated by developmental considerations, including the BabyLM Challenge (10M and 100M words) (Charpentier et al., 2025), 124M-parameter models trained on child-directed corpora such as CHILDES, and three BabyBERTa models (5M params) trained with different random seeds (Huebner et al., 2021). To examine data size effects, we use Open-sci-ref-0.01 (Nezhurina et al., 2025), which provides 1.3B and 1.7B models trained on C4 (50B and 300B tokens) and Nemotron (300B and 1T tokens), respectively. For model size effects, we use the Pythia suite (Biderman et al., 2023), which includes ten models from 14M to 12B parameters. To compare training objectives, we contrast causal and masked language models (CLM and MLM) across GPT-2-medium, RoBERTa-base, RoBERTa-large, and

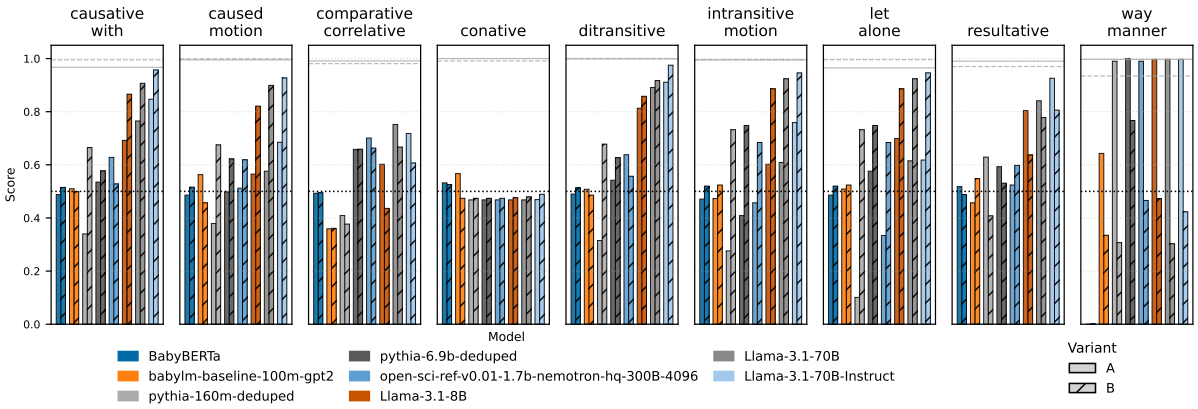


Figure 2: Scores of each language model across constructions and variants. Solid and dashed lines at the top of the figure represent the scores of GPT-5 for variants A and B, respectively.

Pythia-410m. Their sizes are comparable, while data sizes increase in the order GPT-2 (40G) < RoBERTa (160G) < Pythia (825G). For larger open LLMs, we use Llama-3.1 (Meta, 2024) (8B and 70B, and base and instruction-tuned) to assess the effect of instruction tuning. Finally, we include GPT-5, a closed-source commercial model, as an approximate upper bound for open models.

Evaluation To evaluate whether language models grasp the construction and its meaning, we compute the probabilities of the plausible and implausible evaluation pairs for each set. We use length-normalized log probability for CLM and pseudo log-likelihood (Salazar et al., 2020) for MLM. GPT-5 is not directly accessible to probabilities; we estimate its likelihoods through prompting. Details of the evaluation procedure are in Appendix D.

4.2 Results

Overall We evaluate eight models of varying sizes from the series used in this work, as shown in Figure 2. Small-scale models trained on developmentally appropriate corpora perform at near-chance levels. While previous work has shown that even small models can perform reasonably well on formal grammatical benchmarks such as BLiMP (Huebner et al., 2021), our results indicate that understanding constructional meaning is considerably more challenging. In contrast, the large closed-source GPT-5 accurately interprets most constructions. Even open models with 70B parameters struggle with certain constructions, some of which remain entirely unsolved. Mid-sized models, in particular, show a mix of very high and very low accuracies across constructions, suggesting reliance on biased heuristics. A detailed analysis of such biases is provided in Section 5.2.

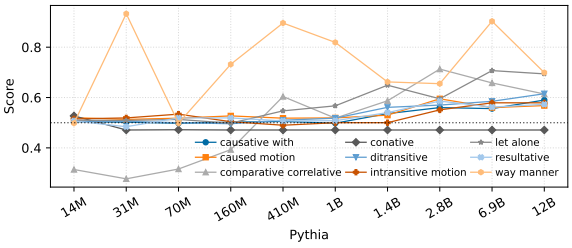


Figure 3: Scores of Pythia models by model size.

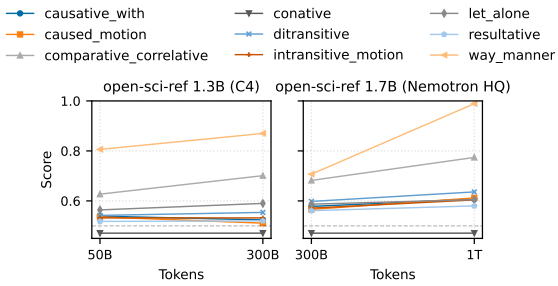


Figure 4: Scores of open-sci-ref models by data size.

Model Size Figure 3 shows the accuracy for each construction across ten Pythia model sizes. Most constructions, except for CONATIVE, exhibit a clear upward trend as model size increases. Less canonical constructions such as LET-ALONE and CC show a strong dependence on model capacity, with substantial improvements observed only in larger models. For the remaining constructions, scores begin to rise around 1.4B parameters, while smaller models remain near chance. WAY-MANNER shows no consistent trend across model sizes.

Data Size Figure 4 presents results from Open-sci-ref-0.01 trained on different data scales. Over-

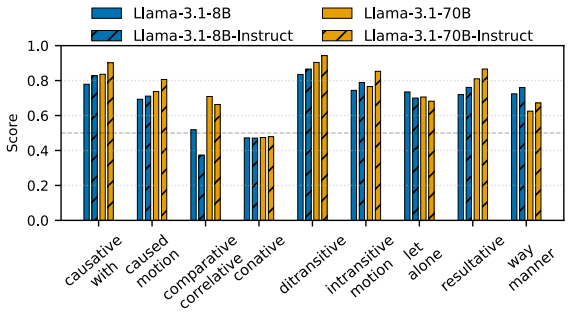


Figure 5: Scores of Llama3.1 base/instruction models.

all, performance improves with larger training datasets, though gains from 50B to 300B tokens remain limited. At the level of individual constructions, most show minimal change or slight decreases between 50B and 300B. However, all constructions improve from 300B to 1T, suggesting that larger-scale training data substantially improve constructional understanding under this setup.

Instruction Tuning Figure 5 shows scores for base and instruction-tuned variants of Llama3.1-8B and -70B. Overall, the instruction-tuned models tend to achieve higher scores than their base counterparts. However, this effect varies across constructions: while consistent improvements are observed for argument-structure constructions, non-standard constructions such as LET-ALONE and CC show decreased performance. These results suggest that instruction tuning does not uniformly enhance constructional meaning understanding.

Objectives Table 2 compares MLM and CLM of comparable size. Though RoBERTa-large is slightly larger, it outperforms Pythia despite being trained on over five times more data. Similarly, RoBERTa-base achieves higher scores than Pythia, even when Pythia has advantages in model and data scale. Overall, within the examined range, MLMs tend to outperform CLMs. Recent theoretical work (Zhang et al., 2024) suggests that MLMs induce richer co-occurrence patterns and stronger semantic associations, enhancing classification performance, whereas their fixed masking rate may limit generalization in generative tasks with variable-length inputs. Although our task differs from probability-based classification, the results indicate that MLMs may exploit bidirectional context to better integrate form and meaning. Consistent with this trend, several BabyLM Challenge systems (Hu et al., 2024) have adopted hybrid MLM–CLM objectives (Charpentier and Samuel, 2024; Yu et al., 2024). Exploring which linguistic abilities benefit from the MLM objective remains a promising direction for future work.

Model	Model size	Data size	Score
RoBERTa-base (MLM)	125M	160G	0.573
RoBERTa-large (MLM)	355M	160G	**0.674**
GPT-2-medium (CLM)	345M	40G	0.537
Pythia-410m (CLM)	410M	825G	0.544

Table 2: Scores of MLM and CLM models.

5 Model Analysis for Linguistic Insights

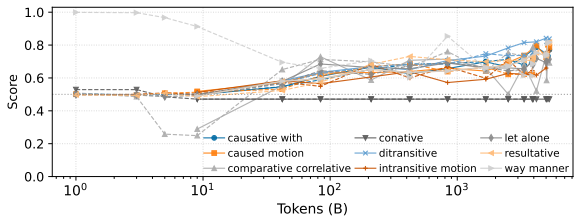
We conduct three analyses to examine the linguistic abilities of models using our dataset.

5.1 Learning Trajectories

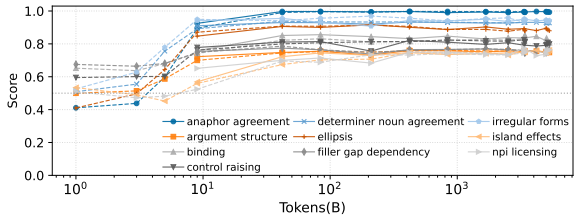
Our work investigates the ability to interpret meanings implied by linguistic forms, which we suggest is not captured by grammatical acceptability alone. This section traces the development of grammatical acceptability and constructional meaning across training and data scale. We hypothesize that grammatical acceptability improves earlier in training, as it often relies on relatively shallow statistical patterns. For instance, in number agreement, a model that learned to associate the number feature of the subject with the verb correctly prefers *The cats annoy Tim* over **The cats annoys Tim*. In contrast, many constructions in CxMP require understanding the meaning encoded in the form, which likely demands more data and emerges later.

Settings We use the base versions of OLMo2-7B and 13B with publicly available checkpoints (OLMo et al., 2024). We select 11 representative checkpoints from Stage 1 of OLMo2-13B and three from Stage 2. For OLMo2-7B, we use checkpoints corresponding to approximately the same cumulative token counts as those of the 13B model. For evaluation datasets, we use BLiMP (Warstadt et al., 2020) in addition to CxMP. BLiMP consists of pairs of (un)grammatical sentences across 12 linguistic phenomena, and we compute scores for each. The scoring method is identical to that used for CxMP.

Results Figure 6 shows BLiMP and CxMP trajectories (log scale) as a function of training token count. For BLiMP, models correctly solve about 80% of phenomena by around 50B tokens, after which performance plateaus with only a one–two point gain. In contrast, CxMP scores improve

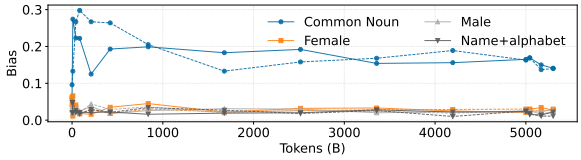


(a) Scores on CxMP

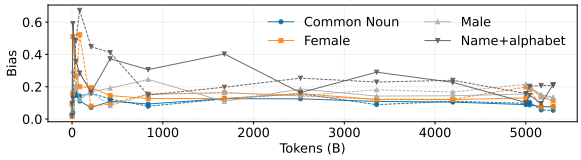


(b) Scores on BLiMP

Figure 6: Scores of OLMo-2 on CxMP and BLiMP across training token counts (log scale). Solid and dashed lines denote the 13B and 7B models.



(a) (i) Bias between original and switched sentences



(b) (ii) Bias across variants

Figure 7: Biases of OLMo-2 on CxMP across training token counts (log scale).

steadily from around 10B tokens through the final training stage. The results suggest that formal linguistic knowledge like grammatical acceptability emerges relatively early, whereas constructional meaning requires longer-term learning. Linguistic performance of language models has mostly been assessed with BLiMP, and reaching a certain score has often been taken as evidence of sufficient linguistic ability. However, results from CxMP show that understanding not only formal grammatical aspects of constructions but also their semantic and functional dimensions requires continued learning.

5.2 Effect of Bias

In Section 4.2, we find that models trained on data sizes and domains intended to approximate developmental conditions remain near chance on many constructions and variants. In contrast, mid-sized models show large score gaps across variants, whereas this tendency is smaller in larger models. These patterns suggest biases driven by shallow heuristics. Building on these observations, this section examines how such biases evolve throughout training, taking a broader view of model behavior.

Definition of Bias We define the following two kinds of bias. **Type (i) Bias between original and switched sentences:** It arises when scores differ between original and switched sentences within the same construction and variant. For example, in the sentences *Lisa left Emma with a receipt. The one who provided the receipt is Lisa/*Emma.*, the model correctly prefers the plausible interpretation. However, in the switched version *Emma left Lisa with a receipt. The one who provided the receipt is Emma/*Lisa.*, it still tends to choose *Lisa*. In such cases, the model may rely on the diagnostic sentence alone, mechanically selecting the subject most strongly associated with the verb. This bias is likely more pronounced with common nouns, as switching reverses typical agent–patient relations (e.g., *The student taught the teacher English*). **Type (ii) Bias across variants:** It occurs when scores differ between variants of the same construction. For instance, while the model prefers the plausible interpretation in *Lisa left Emma with a receipt. The one who received the receipt is Emma/*Lisa.*, it still favors *Emma* in *Lisa left Emma with a receipt. The one who provided the receipt is Lisa/*Emma.*, selecting the implausible interpretation. Such cases suggest that the model may rely on shallow heuristics based on surface cues, such as choosing the nearest or first noun in the sentence.

Quantification of Bias We evaluate each checkpoint of the models used in Section 5.1 (OLMo2-7B and 13B) using the following metrics. Larger values would indicate stronger bias in the model. **(i) Bias between original and switched sentences:** For each noun type, we compute the absolute difference between the mean scores of the original and switched variants. **(ii) Bias across variants:** For each noun type, we calculate the absolute difference between the mean scores of the two variants.

Results Figures 7a and 7b show score trajectories across checkpoints (training token count) for each noun type, corresponding to Type (i) and Type (ii), respectively. In Type (i), the trend is most promi-

nent for *Common noun*s, whereas in Type (ii), it appears more broadly, especially for *Name+alphabet*. In both cases, the scores start low in the early training stages, rise thereafter, and gradually decline again toward the end, exhibiting a U-shaped trajectory. Compared with the results in Section 5.1 (Figure 6a), this pattern suggests that the model initially responds at random without bias, then develops shallow heuristics, and finally converges toward more accurate judgments. As expected, in Type (i), this tendency is particularly evident for common nouns. In contrast, in Type (ii), a similar pattern emerges most clearly for name+alphabet. This may indicate that the model becomes confused by low-frequency input formats and overly relies on nearby cues. Overall, these results demonstrate that different kinds of biases can arise depending on the noun type and sentence structure being tested. They also highlight the importance of incorporating diverse question formulations to comprehensively assess the model's linguistic knowledge.

5.3 CxMP Using Pseudoword

In theoretical linguistics, the meanings of constructions are theoretically defined and systematized (Goldberg, 1995). In contrast, Kako (2006) examines whether speakers can directly perceive such meanings from constructions. Kako (2006) uses stimulus sentences such as *The rom gorped the blickit to the dax*, in which verbs and nouns are replaced with meaningless pseudowords, thereby controlling for lexical semantics and evaluating how humans interpret meaning from constructional form alone. The results reveal significant correspondences between constructions and the semantic attributes assumed in prior work. This suggests that speakers can recover constructional meaning even in the absence of lexical information. In this section, adopting a similar perspective, we examine how well language models capture the correspondence between constructions and their meanings when lexical semantics is controlled. We evaluate whether models can reconstruct meaning based on constructional form, using CxMP.

Settings We construct an evaluation dataset by replacing content words in CxMP with pseudowords. Our analysis focuses on four major argument-structure constructions identified by Goldberg (1995): CAUSED-MOTION, INTRANSITIVE-MOTION, DITRANSITIVE, and RESULTATIVE. Among these, DITRANSITIVE and

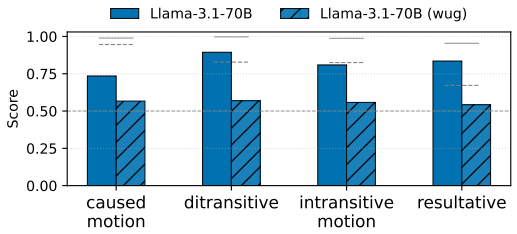


Figure 8: Scores for pseudowords (*wug*) and real words.

INTRANSITIVE-MOTION overlap with the constructions examined by Kako (2006). Pseudowords are generated with Wuggy (Keuleers and Brysbaert, 2010), which produces phonotactically natural and morphologically plausible words (*wug*s). Appendix F shows details. We use Llama3.1-70B as the target model.

Results Figure 8 shows the scores for each construction using either *wug*s or the original words. Across constructions, scores drop markedly when original words are replaced with pseudowords, indicating that the model struggles to infer meaning solely from constructional cues without lexical content. Nevertheless, scores remain slightly above chance, suggesting some sensitivity to constructional information. Kako (2006) reports that the presence or absence of closed-class words had little effect on performance by construction type. Among the constructions tested, CAUSED-MOTION and INTRANSITIVE-MOTION, which include such closed-class words, show smaller pseudoword–original gaps and relatively stable performance. The pseudowords potentially share surface forms with original words and provide weak cues. Following Kako (2006), we also retain determiners and inflectional suffixes such as -ed and -ing, which may provide part-of-speech cues that facilitate construction identification. Future work should further examine which linguistic elements support construction identification under varying levels of lexical control.

6 Conclusion

We present CxMP, a Construction Grammar-based controlled benchmark for evaluating how language models grasp meanings encoded by constructions. Results indicate that there remains considerable room for improvement in constructional meaning understanding, not only in developmentally plausible small models but also in modern LLMs.

Limitations

We prioritized model-agnostic evaluation and therefore adopted a unified likelihood-based comparison across all open models. While this design enables controlled comparison under a common criterion, it is not intended to maximize the capabilities of modern large-scale LLMs. Likelihood-based evaluation does not explicitly specify which factors should be prioritized in model judgments, such as real-world plausibility of the diagnostic sentence, coherence with the constructional context, or grammatical well-formedness. As a result, the basis of model preferences may not always be fully transparent. A variety of evaluation methods have been developed specifically to better elicit the strengths of such models (Ide et al., 2025; Hu and Levy, 2023). When the goal is to assess the full potential of contemporary large-scale LLMs, those evaluation settings may be more appropriate.

We consider constructions along a continuum from substantive to schematic and design the benchmark to cover as much of this range as feasible. However, constructions in Construction Grammar extend more broadly, spanning linguistic units from morphemes to discourse-level structures, which are difficult to fully capture under the constraints of our framework. Our benchmark therefore focuses on a subset of this broader space and does not aim to cover it in its entirety. In addition, our analysis is limited to English, and does not account for cross-linguistic variation in constructions. Extending the benchmark to other languages or to more diverse construction families would allow testing whether the observed developmental patterns generalize across typologically different systems.

Our evaluation introduces a binary distinction between plausible and implausible instances. However, the boundary between uses in which constructions extend or modulate the meaning of lexical items beyond their default interpretations, which is an effect widely noted in Construction Grammar (Goldberg, 1995), and those considered semantically infelicitous is often gradient and context-dependent, and may not always be fully captured by our operationalization.

The diagnostic sentences are constructed by the authors with reference to prior literature on construction–meaning correspondences, and may not fully capture the full range of meanings associated with each construction, with some instances potentially deviating from this range.

Acknowledgments

We would like to express our gratitude to the anonymous reviewers who provided many insightful comments that have improved our paper. This work was supported by JSPS KAKENHI Grant Numbers JP25KJ1824 and JP25K21281, JST BOOST Grant Number JPMJBY24D9, and JST FOREST Grant Number JPMJFR232R.

Ethical Considerations

We designed the annotation process to follow ethical guidelines for crowdwork, including providing clear task instructions, ensuring that no personally identifiable information is collected, and offering compensation at or above recommended fair-pay standards for the estimated task duration. Participation was voluntary, and workers were allowed to withdraw at any time. These measures aim to ensure that the data collection process respects annotator privacy and labor conditions.

In addition, the use of human names (e.g., male and female names) as entities may introduce unintended social or gender-related biases. Although we balance different name types and include alternative entity representations (e.g., common nouns and name+alphabet) to reduce such effects, these biases may still influence model behavior and evaluation results.

References

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 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.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y. Hu, Jing Liu, Jaap Jumelet, Tal Linzen, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Gotlieb Wilcox, and Adina Williams. 2025. Findings of the third BabyLM challenge: Accelerating language modeling research with cognitively plausible data. In *Proceedings of the First BabyLM Workshop*, pages 399–420, Suzhou, China. Association for Computational Linguistics.

Lucas Georges Gabriel Charpentier and David Samuel. 2024. GPT or BERT: why not both? In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 262–283, Miami, FL, USA. Association for Computational Linguistics.

Ionut Constantinescu, Tiago Pimentel, Ryan Cotterell, and Alex Warstadt. 2025. Investigating critical period effects in language acquisition through neural language models. *Transactions of the Association for Computational Linguistics*, 13:96–120.

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

Charles J. Fillmore. 1986. Varieties of conditional sentences. In *Proceedings of the Eastern States Conference on Linguistics (ESCOL)*, volume 3, pages 163–182.

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.

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

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

Adele E. Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Cognitive Theory of Language and Culture Series. University of Chicago Press.

Adele E Goldberg. 2003. Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences*, 7(5):219–224.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Hoffmann. 2017. Construction grammar as cognitive structuralism: the interaction of constructional networks and processing in the diachronic evolution of english comparative correlatives. *English Language and Linguistics*, 21(2):349–373.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. 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.

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

Yusuke Ide, Yuto Nishida, Justin Vasselli, Miyu Oba, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2025. How to make the most of LLMs' grammatical knowledge for acceptability judgments. 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 7416–7432, Albuquerque, New Mexico. 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.

Edward Kako. 2006. The semantics of syntactic frames. *Language and Cognitive Processes*, 21(5):562–575.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14691–14714, Bangkok, Thailand. Association for Computational Linguistics.

Paul Kay and Charles J. Fillmore. 1999. Grammatical constructions and linguistic generalizations: The what's X doing Y? construction. *Language*, 75(1):1–33.

Emmanuel Keuleers and Marc Brysbaert. 2010. Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, 42(3):627–633.

Tom Mackintosh, Harish Tayyar Madabushi, and Claire Bonial. 2025. Evaluating CxG generalisation in LLMs via construction-based NLI fine tuning. In *Proceedings of the Second International Workshop on Construction Grammars and NLP*, pages 180–189, Düsseldorf, Germany. Association for Computational Linguistics.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Meta. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Marianna Nezhurina, Jörg Franke, Taishi Nakamura, Timur Carstensen, Niccolò Ajroldi, Ville Komulainen, David Salinas, and Jenia Jitsev. 2025. Open-sci-ref-0.01: open and reproducible reference baselines for language model and dataset comparison. *Preprint*, arXiv:2509.09009.

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 21 others. 2024. 2 OLMo 2 furious. *Preprint*, arXiv:2501.00656.

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

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WinoGrande: An adversarial Winograd Schema Challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8732–8740.

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.

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, 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.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. OpenAI GPT-5 system card. *Preprint*, arXiv:2601.03267.

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.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

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*, pages 10859–10882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Leonie Weissweiler, Kyle Mahowald, and Adele E. Goldberg. 2025. Linguistic generalizations are not rules: Impacts on evaluation of LMs. In *Proceedings of the Second International Workshop on Construction Grammars and NLP*, pages 61–74, Düsseldorf, Germany. Association for Computational Linguistics.

Xinru Yu, Bin Guo, Shiwei Luo, Jie Wang, Tao Ji, and Yuanbin Wu. 2024. AntLM: Bridging causal and masked language models. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 324–331, Miami, FL, USA. Association for Computational Linguistics.

Qi Zhang, Tianqi Du, Haotian Huang, Yifei Wang, and Yisen Wang. 2024. Look ahead or look around? A theoretical comparison between autoregressive and

masked pretraining. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 58819–58839. PMLR.

A Dataset Generation

First, we generate constructional examples for each construction using gpt-5-chat-latest. For each construction, we instruct the model to produce sentences based on predefined templates so that a corresponding diagnostic sentence naturally follows. For example, in the CAUSATIVE-WITH, we use the following template: "{Subj; N1; human common noun (gender-neutral)} {V} {Obj; N2; human common noun (gender-neutral)} {Obl; with-PP}." We then attach reference diagnostic sentences such as *The one who provided {Obl's noun} is N1* and *The one who received {Obl's noun} is N2*, and instruct the LLM to fill each curly brace with contextually appropriate words. For each construction, we generate 50 sentences using at least eight different random seeds and remove duplicates. In constructing the prompts for sentence generation, we paid careful attention to several design considerations. Specifically, we instructed the model to (i) use clear and simple language that is commonly used, (ii) avoid semantically unnatural or implausible sentences, (iii) include verbs whose interpretation depends on the construction rather than only on the verb itself, while ensuring that the reference sentences connect naturally, (iv) avoid reusing words that have already appeared in the output while preserving the intended meaning of the construction, (v) refrain from including adverbs or modifiers not explicitly specified in the template, and (vi) generate only affirmative sentences (except for LET-ALONE). The prompt example is in Figure 9. We then use gpt-5-chat-latest to classify the generated constructional examples into three categories, Prototypical, Extended (Peripheral), and Unacceptable (Anomalous), and exclude those labeled as Unacceptable. The prompt example is in Figure 10.

Next, we generate diagnostic sentences. For each construction and variant, we prepare templates for plausible diagnostic sentences and implausible diagnostic sentences (e.g., *The one who provided {Obl's noun} is {n1}*) and replace the corresponding words in the constructional examples with curly braces.

Finally, we use the model to judge whether each sentence and its diagnostic counterpart are semantically aligned, and we compute the final scores only for pairs confirmed to be semantically consistent. The prompt example is in Figure 11. To reduce the effect of shallow heuristics (McCoy et al., 2019), we design diagnostic sentences so that each of their words occurs an equal number of times across all constructional examples, including those without the corresponding word. We construct approximately 43k evaluation pairs in total (300 pairs $\times$ 9 constructions $\times$ 2 variants $\times$ 4 noun types $\times$ 2 switching types).

```
The sentences below illustrate a {construction} construction in
the order: {pattern}. Fill in the curly braces in the template
with appropriate words, and generate {number} sentences. Make sure
that the following reference sentences can naturally come after
the generated ones.
- Use clear and simple language that is commonly used.
- Avoid semantically unnatural or implausible sentences.
- Include verbs whose meaning comes from the construction, not
only from the verb itself, while ensuring that the reference
sentences connects naturally.
- Avoid reusing words that have already appeared in the output,
while ensuring that the meaning of the construction is preserved.
- Do not include any adverbs or modifiers that are not explicitly
part of the template.
- Generate only affirmative sentences.
- Output only the generated sentences.

Each sentence must exactly follow this format:
{template}
For reference only (do not output):
- {ref1}
- {ref2}
```

Figure 9: Prompt for generating constructional examples from templates. The placeholders {construction}, {pattern}, and {template} specify the construction name, constructional pattern (e.g., Subj V Obj Obj2), and slot structure, respectively. {ref1} and {ref2} provide diagnostic reference sentences.

```
You are a linguist specializing in Construction Grammar. Each of
the sentences below is a candidate example of the {construction}
construction. Classify each sentence into one of three categories:
A. Prototypical: The verb meaning naturally aligns with the
constructional meaning (e.g., ditransitive "X gives Y Z").
Example: "Alex gave Jamie a book."
B. Extended (Peripheral): The verb itself does not encode the
constructional meaning, but the sentence remains interpretable
because the construction supplies that meaning. Example: "Pat
sneezed the foam off the cappuccino."
C. Unacceptable (Anomalous): The sentence is ungrammatical or
unnatural because (i) the verb is incompatible with the
constructional meaning, (ii) the result phrase is redundant with
the verb's lexical meaning, or (iii) the verb–result relation
fails to yield a plausible interpretation. Example: "Pat admired
the book onto the table.", "Pat entertained the customer amused."

Sentences:
{sentences}

Output format:
Sentence number: [A. Prototypical / B. Extended / C. Unacceptable]
Provide the output in exactly this format.
```

Figure 10: Prompt used to filter out unacceptable sentences based on constructional category classification. {sentence} denotes a placeholder replaced with generated sentences.

B Entities

We define four types of entities: *common nouns*, *female names*, *male names*, and *name+alphabet*. The *common nouns* type consists of nouns refer-

```
You are a linguist specializing in Construction Grammar.
Each example consists of a target sentence and a follow-up
interpretation statement that attempts to capture the meaning of
the construction.
Classify each pair as either:
A. Semantically appropriate: The interpretation correctly reflects
the meaning of the construction in the target sentence, including
general restatements or logically valid applications to specific
cases, whether the causal link is explicit or implicit.
B. Semantically inappropriate: The interpretation fails to capture
the intended meaning of the construction.

Sentences:
{sentences}

Output format:
Sentence id: [A. Appropriate / B. Inappropriate]
Provide the output in exactly this format.
```

Figure 11: Prompt used for evaluating semantic appropriateness. {sentence} denotes a placeholder replaced with generated sentences.

ring to humans and represents the most natural type. However, certain nouns frequently co-occur with specific semantic roles, making it difficult to determine whether a model's correct answer reflects genuine understanding of the constructional meaning or mere reliance on lexical co-occurrence statistics. For instance, when the verb is *teach* and the entities are *teacher* and *student*, the model may simply infer that the teacher is the agent and the student is the patient, as this pattern is highly frequent in the data. The *female names* and *male names* types maintain naturalness while reducing lexical co-occurrence bias. Nonetheless, name-specific frequency effects may remain; for example, if *Maria* frequently appears as a giver in the training corpus. The names are randomly selected from the top 10 names for each decade in the SSA Baby Names dataset[2] from 1950 to 2020, after removing duplicates. The *name+alphabet* type is the most artificial and controlled type. We randomly sample a letter (e.g., A, B) and insert it into a template such as *Name A*. Because of its low frequency, this type can reduce performance, particularly in smaller models with limited ability to generalize to unseen forms.

C Human Data Validation

We conduct a human validation study to assess the plausibility of automatically generated items. We recruit 168 participants via Prolific and restrict recruitment to native English speakers with an approval rate of at least 97%. Participants read a preceding sentence and select the continuation that is more plausible, and we collect responses using Google Forms. The instruction example for human data validation is in Figure 12. We collect

[2]https://www.ssa.gov/oact/babynames/

```
Your task:
Read the preceding sentence and choose the continuation that
is more plausible as a continuation, not the one that sounds more
natural on its own.

For example:
- Preceding sentence: The child bought a candy for the mother.
- Continuation candidates:
A: The one who gave the child the candy is the mother. → Not
plausible
B: The one who gave the mother the candy is the child. → Plausible
Although A may sound more natural on its own, your task is to
choose the sentence that more plausibly continues the preceding
one, where the child is the agent. Therefore, the correct answer
is B.
```

Figure 12: Instructions for human data validation.

three independent annotations per item, and each participant completes 16 items (approximately 10 minutes). Each construction contains 112 items, except WAY_MANNER and CONATIVE, which contain 56 items because switched variants are not applicable. We include three types of attention checks: (1) a simple subject–verb agreement question, (2) a shuffled sentence detection task, and (3) an instruction-following question matching the example format, and we retain only participants who pass all checks. Participants receive 1.5 GBP (excluding platform fees). We report majority-vote accuracy (the proportion of items where at least two of three annotators select the correct continuation) and full agreement rate (the proportion of items where all three annotators give the same answer).

Results Overall, the human validation results indicate that the automatically generated items are generally plausible. Across 896 items, the majority-vote accuracy reaches 96.65%, and full agreement among annotators reaches 83.59%. Most constructions achieve majority accuracy above 97%. Some constructions (e.g., CONATIVE) show lower agreement, indicating that judgments for these items may be more variable. These statistics provide an additional sanity check for the generated dataset.

D Evaluation Methods

To evaluate whether each language model correctly judge the construction and its associated meaning, we compute the probabilities of the plausible evaluation pair and implausible evaluation pair for each set. To reduce bias due to sentence length (i.e., token count), we use the length-normalized log probability, obtained by dividing the total log probability of the sentence by its number of tokens. A model is considered correct when it assigns a higher probability to the plausible evaluation pair, and the accuracy of these judgments is used as the score. For MLMs, we compute probabilities

Construction	#Items	Maj. (%)	Agr. (%)
LET_ALONE	112	97.32	66.07
CAUSATIVE_WITH	112	97.32	91.96
WAY_MANNER	56	98.21	80.36
CC	112	100.00	91.96
CONATIVE	56	67.86	44.64
DITRANSITIVE	112	100.00	89.29
CAUSED_MOTION	112	97.32	89.29
RESULTATIVE	112	99.11	91.96
INTRANSITIVE_MOTION	112	99.11	85.71
Overall	896	96.65	83.59

Table 3: Human validation results. Majority (Maj.) denotes the proportion of items where at least two of three annotators selected the correct continuation. Agreement (Agr.) denotes the proportion of items where all three annotators gave the same answer.

using the pseudo log-likelihood method (Salazar et al., 2020). In contrast, for chat models such as GPT-5, where probabilities are not directly accessible, we estimate relative likelihoods through prompting. We carefully design the prompts so as not to give the model an advantage over direct probability computation; specifically, the model is instructed to choose the sentence that is higher likelihood. To prevent reliance on local cues rather than holistic sentence comparison, the prompt explicitly requires the model to evaluate the entire pair of sentences. Each evaluation pair is presented individually, with the two options labeled as A and B. To mitigate potential label bias (e.g., a tendency to consistently prefer one label such as A), the assignment of correct and incorrect sentences to the indices is randomized. Note that the scores obtained with gpt-5-chat-latest may vary due to inherent variability in model outputs, and that we use the latest version available as of September 30, 2025.

E OLMo Models

OLMo2-13B is pretrained on 5T tokens and OLMo2-7B on 4T tokens. The training of OLMo2 consists of two pre-training stages. Stage 1 accounts for more than 90% of the total training, while Stage 2 performs additional fine-tuning on high-quality data. Specifically, the 7B model undergoes three runs of 50B tokens each, and the 13B model is further trained on 300B tokens. The final released model is obtained by model souping, and therefore does not correspond to the final checkpoint.

Resource	License
wuggy	MIT license
BLiMP	CC-BY
SSA Baby Names	CC0 / Public Domain
Pythia	Apache-2.0
Llama-3.1	Meta Llama 3.1 Community License
OLMo-2.0	Apache-2.0
BabyLM baselines	Apache-2.0
RoBERTa	MIT license
BabyBERTa	MIT license
GPT2	MIT license

Table 4: Licenses of resources used in this work.

F Wug Generation

We input each original content word to Wuggy (Keuleers and Brysbaert, 2010) to obtain a corresponding pseudoword. Following Kako (2006), determiners and inflectional suffixes (e.g., -ed, -ing) are retained. When Wuggy returns None as its output, we create a pseudoword by randomly generating alphabetic strings with the same number of entities as the original content word.

G Intended-Use and Access-Conditions Compliance

We use all external artifacts (datasets and models) strictly for research benchmarking and analysis, in accordance with their licenses (Table 4). We do not redistribute restricted resources (e.g., model weights), while releasing our dataset (CxMP) for research purposes.