Beyond Memorization: Assessing Semantic Generalization in Large Language Models Using Phrasal Constructions

Wesley Scivetti[1]*, Melissa Torgbi[2]*, Austin Blodgett[3], Mollie Shichman[4], Taylor Pellegrin[3], Claire Bonial[3], Harish Tayyar Madabushi[2]

[1]Georgetown University, [2]University of Bath,
[3]DEVCOM U.S. Army Research Laboratory, [4]University of Maryland, College Park
wss37@georgetown.edu, mat66@bath.ac.uk

Abstract

The web-scale of pretraining data has created an important evaluation challenge: to disentangle linguistic competence on cases well-represented in pretraining data from generalization to out-of-domain language, specifically the dynamic, real-world instances less common in pretraining data. To this end, we construct a diagnostic evaluation to systematically assess natural language *understanding* in LLMs by leveraging Construction Grammar (CxG). CxG provides a psycholinguistically grounded framework for testing generalization, as it explicitly links syntactic forms to abstract, non-lexical meanings. Our novel inference evaluation dataset consists of English phrasal constructions, for which speakers are known to be able to abstract over commonplace instantiations in order to understand and produce creative instantiations. Our evaluation dataset uses CxG to evaluate two central questions: first, if models can 'understand' the semantics of sentences for instances that are likely to appear in pretraining data less often, but are intuitive and easy for people to understand. Second, if LLMs can deploy the appropriate constructional semantics given constructions that are syntactically identical but with divergent meanings. Our results demonstrate that state-of-the-art models, including GPT-o1, exhibit a performance drop of over 40% on our second task, revealing a failure to generalize over syntactically identical forms to arrive at distinct constructional meanings in the way humans do. We make our novel dataset and associated experimental data, including prompts and model responses, publicly available.[1]

1 Introduction

Understanding the extent to which Large Language Models (LLMs) generalize from relatively frequent phenomena well-represented in pretraining data to creative, novel usages of language has important implications for LLM development. Identifying the precise nature and limits of LLM generalization can inform decisions about architectures and training regimes (Li et al., 2023; Zhang et al., 2023). This becomes especially relevant as models move toward 'reasoning'-based systems and the inevitable widespread deployment of AI agents (DeepSeek-AI et al., 2025). Identifying failure patterns will enable targeted improvements at different stages of development. However, testing LLMs' ability to generalize is particularly challenging because they are trained on vast web-scale data (Lu et al., 2024). Even if pretraining datasets were fully accessible, ensuring that a test example is truly independent would remain nontrivial, as a model may not have encountered that instance but could have been exposed to related cases that provide indirect information (Tayyar Madabushi et al., 2025).

Model	Constructional Semantics	Construction Distinction
GPT-4o	0.88	0.58
GPT-o1	0.90	0.46
Llama 3 70B	0.74	0.52
Human	0.90	0.83

Table 1: We demonstrate a drop in performance, even in the latest models, as we move from evaluating functional understanding of constructional semantics to understanding syntactically identical but semantically distinct constructions. We report accuracy on NLI tasks leveraging distinct constructional premises.

Therefore, this work introduces a novel evaluation dataset grounded in the theory of Construction Grammar (CxG) (Goldberg, 1995; Croft, 2001) (see Appendix A for an overview of CxG). Specifically, we focus on phrasal constructions (Cxns) because of the body of psycholinguistic research demonstrating that speakers are able to abstract over syntactic slots of these Cxns in order to inter-

*Equal Contribution

[1] https://github.com/melissatorgbi/beyond-memorization

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
December 20-24, 2025 ©2025 Association for Computational Linguistics

	Exp. 1 Dataset: CxNLI	**Exp. 2 Dataset: CxNLI-Distinction**
CxNLI Premise	I brushed my hair smooth.	A famous emperor buried scholars alive.
Entrenched Variant	I made my hair smooth (by brushing).	*NA: no entrenched variants for Exp2 Cxns*
Cxn Form	NP1 V NP2 ADJ	NP1 V NP2 ADJ
Cxn Meaning	NP1's action of V causes NP2 to become ADJ	NP2 is ADJ during NP1's action of V
CxNLI Hypothesis	My brushing caused my hair to be smooth.	Burying the scholars caused them to be alive.
CxNLI Relation	Entailment	Contradiction

Table 2: Example items illustrating experiments. Exp. 1 uses Natural Language Inference (NLI) to test abstraction of the semantics of frequent, entrenched Cxns (usually realized with the verb of the Entrenched Variant) when realized as a creative, infrequent instantiation (CxNLI Premise). Exp. 2 uses NLI to test if models can distinguish and apply the appropriate, distinct semantic interpretation of Cxns that are syntactically identical to the entrenched Cxns of Exp. 1 (with grammatical phrase types of Noun Phrase (NP), Verb (V), a second NP, and an Adjective (ADJ)).

pret and produce creative and novel Cxn instantiations (Johnson and Goldberg, 2013; Tomasello, 2003). Speakers recognize the syntactic structures of a familiar Cxn in order to interpret the meaning, despite the fact that the speaker may have never encountered that set of lexical items within the Cxn. The Exp. 1 column of Table 2 provides an example premise where our human annotations show that people can easily recognize the (RESULTATIVE) constructional semantics despite the fact that the verb ("brush") is relatively atypical in this Cxn, which is realized with a relatively limited set of verbs (frequently "make"). Additionally, speakers can balance knowledge of lexical semantics against the constructional semantics in order to distinguish Cxns that are syntactically identical but have different meanings. The Exp. 2 column of Table 2 provides an example premise involving a (DEPICTIVE) Cxn that is syntactically identical to the RESULTATIVE, but has a divergent meaning, as evidenced by the fact that a templatically similar hypothesis holds the opposite relation.

In addition to providing experimentally validated explanations of human language acquisition and use, CxG is uniquely suited to evaluating whether LLMs primarily derive meaning from the composition of lexical meanings (as would be the view of Generative Grammar (Chomsky, 2014a)), or if the recognition of particular syntactic structures can cue constructional meaning (this would be evidenced by strong performance on Exp. 1), or finally, if LLMs can balance both lexical meaning and constructional meaning together to recognize constructional meaning while distinguishing between syntactically identical constructions based on lexical semantics (this would be evidenced by strong performance on Exp. 2).

We evaluate whether or not models can generalize knowledge of highly frequent Cxns of English to creative instantiations of those Cxns with lexical items unlikely to have been encountered within that structure in pretraining data (Exp. 1), and we evaluate model ability to recognize when the lexical items are so distinct that this cues a different Cxn with a different meaning (Exp. 2).

To investigate LLM generalization, we conduct two experiments (described in §3) leveraging the Natural Language Inference (NLI) task. We show a summary of results in Table 1: all models lag behind human performance in multiple experimental settings. Exp. 1 (§4) shows us that some aspects of constructional semantics are ascribed adequately for success on an NLI task; however, Exp. 2 (§5) shows that even state-of-the-art reasoning models like GPT-4o and GPT-o1 show a significant performance drop when the models are asked to ascribe distinct constructional semantics to syntactically identical Cxns. In combination, our results and error analysis (§6) highlight key differences between human and model linguistic capabilities (§7 and §8). This study differs from previous research in significant ways highlighted by the following contributions:

1. We create a manually-validated, diagnostic evaluation dataset of 534 NLI triples testing if LLMs ascribe the appropriate semantics to phrasal Cxns.

2. Rather than focusing on the metalinguistic task of identifying Cxns, as most prior works have done, we use the well-established NLI task to evaluate LLM 'understanding' of the underlying meaning communicated by a Cxn.

3. We test on a variety of common English phrasal Cxns instantiated by relatively unexpected words, thereby testing the ability of

models to perform understanding tasks without the aid of memorization and pattern matching from large pretraining datasets.

4. We test some of the largest models currently available, including GPT-o1 and Llama 3 70B.

2 Related Work

The evaluation of LLMs with datasets that consider pretraining data has largely focused on identifying and mitigating test set data leaks (Balloccu et al., 2024; Sainz et al., 2023a). Examples of this include work by Golchin and Surdeanu (2024) and Sainz et al. (2023b) who identify data contamination using prompting. Furthermore, considerable effort has been directed toward designing datasets that minimize such leakage (Zhou et al., 2025). While this is an important concern, preventing direct data leakage does not eliminate the possibility that models can infer answers using related information from pretraining data (Tayyar Madabushi et al., 2025). Another relevant line of research involves counterfactual reasoning, where standard rules of the world are slightly altered. By design, counterfactuals provide an effective way to test LLMs in scenarios where pretraining data offers little advantage, and it has been shown that the performance of LLMs does in fact drop in these cases (Wu et al., 2024; Lewis and Mitchell, 2024).

Starting with CxGBert (Tayyar Madabushi et al., 2020), there has been substantial past work on probing language models' understanding of Cxns. These works have typically focused on either a single Cxn or a handful of Cxns, like AANN (Mahowald, 2023; Chronis et al., 2023; Misra and Mahowald, 2024), COMPARATIVE-CORRELATIVE (Weissweiler et al., 2022), LET-ALONE (Scivetti et al., 2025), NPN (Scivetti and Schneider, 2025) and more schematic phrasal Cxns (Li et al., 2022; Veenboer and Bloem, 2023). Tseng et al. (2022) focus on Cxns in Taiwanese Mandarin, a notable exception to the works above which focus on English (see also Weissweiler et al. 2024; Bunzeck et al. 2025). Zhou et al. (2024) introduce NLI as a proxy task for understanding Cxns, though their results are limited to the CAUSAL-EXCESS and related Cxns. We expand on the use of NLI to a broad set of new Cxns, and create NLI examples which utilize both entrenched (Exp. 1) and syntactically-identical (Exp. 2) Cxns.

Additionally, recent studies show that small LMs (e.g., BabyLMs, Warstadt et al. 2023) can learn the forms of rare constructions (Misra and Mahowald, 2024; Rozner et al., 2025b; see Rozner et al. (2025a) for strong formal results using RoBERTa (Liu et al., 2019)). In contrast, we focus on generalization of constructional *understanding*, which has been shown to be difficult even for LLMs in few-shot settings (Bonial and Tayyar Madabushi, 2024b; Zhou et al., 2024; see Mackintosh et al. 2025 regarding impact of fine-tuning). **To our knowledge, no prior work has explored the use of cognitive linguistic principles to generate human-generalizable datasets for assessing the generalization capabilities of LLMs.**

3 Experimental Design

We conduct experiments exploring two questions: **Research Question (RQ) 1: To what extent can models generalize constructional semantics to relatively infrequent instantiations of common Cxns?** Exp. 1 uses NLI to evaluate understanding of constructional semantics in cases where high-frequency constructional templates are instantiated by words not commonly found within that Cxn. We select 8 Cxns that are roughly balanced across two types: *argument structure Cxns* with no fixed words but clear syntactic slots (e.g., CAUSED-MOTION in Table 3) and phrasal Cxns with two or more fixed words that clearly identify the Cxn (e.g., LET-ALONE in Table 3). To evaluate if the appropriate constructional semantics are associated with these Cxns, we create a novel NLI dataset, where premises are derived from corpus instances of Cxns and an understanding of constructional semantics is required to determine entailment.

RQ 2: To what extent can models distinguish the semantics of Cxns that are syntactically identical, but have different meanings? Exp. 2 uses NLI to evaluate abstraction of distinct constructional semantics given identical syntactic phrasings. We select five Cxns that are syntactically identical to the five argument structure Cxns of Exp. 1. We create a second set of NLI instances again using corpus instances of the five semantically distinct, but syntactically identical Cxns.

Parallel to psycholinguistic research on analogical extension of Cxns (Bybee, 2010), we hypothesize that the frequency and *entrenchment* of the Cxn contribute to model ability to understand the constructional semantics. There is abundant corpus linguistic data from both web and even child language indicating that the 5 argument structure Cxns tested in Exp. 1, CAUSATIVE-WITH, CAUSED-MOTION, CONATIVE, INTRANSITIVE MOTION, RESULTATIVE, are some of the earliest acquired and most frequently used Cxns in the English language (Hoffmann, 2022; Tomasello, 2003; Gries and Stefanowitsch, 2004). Although we lack CxG-annotated resources and access to pretraining data to calculate the precise frequencies of the Cxns tested in Exp. 2 (see Table 5), we can safely assume that these Cxns are less frequent in the language.[2] Most importantly, while the Cxns of Exp. 1 are generally instantiated by a more limited set of verbs, the Cxns of Exp. 2, such as the DEPICTIVE and LOCATIVE, can co-occur felicitously with any verb. As a result, there is no single, *entrenched variant* of the Cxns in Exp. 2 (Gries and Stefanowitsch, 2004). Thus, the Cxns of Exp. 1 have higher type frequency (the frequency of, for example, the RESULTATIVE overall) and have at least one entrenched variant with high token frequency (the frequency of the RESULTATIVE with "make"). Higher entrenchment of one variant means that there is a strong lexico-syntactic signature associated with a specific meaning of a Cxn. Lower entrenchment indicates that there is more variation in the lexico-syntactic features and more variation in how identical lexico-syntactic features are associated with several meanings. **Thus, greater entrenchment may provide critical priors for the model to generalize constructional semantics to novel instantiations, whereas these priors may not be available for the Exp. 2 Cxns, which lack any entrenched, high token-frequency exemplar.**

4 Experiment 1: NLI for Constructional Semantics

4.1 Dataset

In Exp. 1, we leverage the CoGS corpus (Bonial and Tayyar Madabushi, 2024b), which is a collection of about 500 corpus instances, roughly balanced across 10 different Cxns. CoGs consists of carefully curated Cxns chosen for their broad coverage of the basic phrasal Cxns of English (Hoffmann, 2022). The Cxn types collectively represent a significant portion of English usage and provide an effective basis for evaluating LLMs on high-frequency Cxns (such as the CAUSED-MOTION), where instantiated with creative words.

Cxn Name	Example
Causative-With	*Freshly ground coffee beans filled the room with a seductive, earthy aroma.*
Caused-Motion	*But we also exported nickel to the United States.*
Comparative-Correlative	*The more I studied, the less I understood.*
Conative	*Jake sipped at the jug and didn't answer.*
Intransitive Motion	*Armed troops marched to the substations and turned the power back on.*
Let-Alone	*None of these arguments is notably strong, let alone conclusive.*
Resultative	*He hammered the metal flat.*
Way-Manner	*A middle-aged man eased his way into the room.*

Table 3: 8 Cxns tested in Exp. 1, alongside examples.

Broadly, CoGS consists of Cxns of two types: argument structure Cxns, which involve no fixed word forms but have been shown to be the most common Cxns of English as well as other languages (Goldberg, 1995); and phrasal Cxns with multiple fixed words. The latter Cxns are more easily recognizable to LLMs given fixed words cueing that Cxn (Bonial and Tayyar Madabushi, 2024a). We construct our datasets with 8 of the 10 Cxns in CoGS, shown in Table 3.

Overall, our process for creating the constructional NLI dataset can be summarized as:

1. Extract corpus Cxn examples from CoGS.
2. Create general templates for NLI hypotheses for each Cxn type.
3. Generate hypotheses for each example Cxn given the corresponding templates.
4. Manually validate the resulting dataset.

We explain this process through the following example, beginning with the premise: *"He hammered the metal flat."* This is a RESULTATIVE Cxn, which has the meaning of an action causing a change in state. In this case, the action verb *hammered* leads to *the metal* to have a resulting state of *flat*. Regarding the syntactic form of a phrasal Cxn, we can use a constructional template to define the syntactic nature of the slots that are filled by a given Cxn. A

[2] We do not argue that the type frequencies of all Exp. 2 Cxns are low (e.g., INTRANSITIVE has high type frequency), rather that the specific sub-Cxns of Exp. 2 (e.g., INTRANSITIVE+"at"-LOCATIVE) are lower than type frequencies of argument structure Cxns of Exp. 1.

general template for the RESULTATIVE is shown in Example (1), and its application to the above sentence in Example (2).

(1) [SBJ$_1$ [V$_2$ OBJ$_3$ ADJ$_4$]$_{VP}$]$_5$

(2) [*He*$_1$ [*hammered*$_2$ *the metal*$_3$ *flat*$_4$]$_{VP}$]$_5$

Given constructional examples like those in Example (2), our goal is to produce NLI tuples that consistently target the Cxn's meaning. We do this by manipulating the slots in the Cxn templatically to produce hypotheses with consistent relations to the premise. For instance, consider the hypothesis *"The hammering did not cause the metal to become flat."* This is clearly a contradiction to the above premise and is directly in conflict with the meaning of the Cxn. We can produce this hypothesis, and similar contradicted hypotheses, by following the template in Example (3), instantiated with our current example in Example (4).

(3) [THE [V]$_2$-ING DID NOT CAUSE [OBJ]$_3$ TO BECOME [ADJ]$_4$].

(4) [THE [*hammer*]$_2$-ING DID NOT CAUSE [*the metal*]$_3$ TO BECOME [*flat*]$_4$].

Thus, for each Cxn, we create a template to generate NLI hypotheses that target constructional meaning. [3] See Appendix D for the NLI templates for each Cxn (Table 8), along with examples (Table 9).

Manual Verification We manually validate every test instance of our dataset with double or triple annotation and measure human agreement (ranging from 78-90%) to ensure the robustness of our claim that people are able interpret the specific Cxn instances that we present to the LLMs. Once the dataset was created, a second and sometimes third author annotated the relations, and hypotheses were amended until achieving an Inter-Annotator Agreement (IAA) of 90%. Thus, if we take the original author's assigned relation to be the gold standard, then native speaker accuracy on the NLI task is 90%. The final Exp. 1 "CxNLI" dataset totals 435 triples. Descriptive statistics for this dataset along with all other datasets can be found in Appendix B (Table 7).

4.2 Formalism and Task Design

We define a Cxn, C, to be a pairing of a **constructional schema** (form), $\mathcal{T}_C$, and a **semantic interpretation** (meaning), $M(C)$. For example, for the RESULTATIVE Cxn has the **schema** (**$\mathcal{T}_C$**) `NP V NP ADJ` and **Meaning** (**$M(C)$**) 'The action of the Verb causes the Object to enter the state described by the Adjective.'

Our method of evaluating models' ability to 'understand' this Cxn begins with a premise sentence, p, that is an instance of a given Cxn C ($p \in C$). We then generate a hypothesis, h, by applying a pre-defined **hypothesis template**, $\mathcal{H}_{C,L}$. This template is a function that takes the premise p as input, extracts its relevant components, and generates a new sentence h. The template is designed to probe the Cxn's core meaning, $M(C)$, in a way that produces a predictable NLI label, $L \in \{\text{Entailment, Contradiction}\}$. Therefore, the hypothesis is generated as: $h = \mathcal{H}_{C,L}(p)$. For example, to generate a contradiction (L = Contradiction) for a RESULTATIVE premise:

- **Premise (p):** "He hammered the metal flat."
- **Hypothesis (h):** $h = \mathcal{H}_{\text{RESULTATIVE,Contradiction}}(p) \Rightarrow$ "The hammering did not cause the metal to become flat."
- **Resulting Tuple:** $\langle p, h, \text{Contradiction}\rangle$.

The goal of this experiment is to assess if models can correctly classify these NLI tuples. High accuracy on this task indicates that a model has learned the fundamental association between a syntactic schema $\mathcal{T}_C$ and its meaning $M(C)$, even when instantiated with creative or atypical words.

4.3 Empirical Evaluation and Analysis

We test three OpenAI models on our constructional NLI dataset: GPT-4o-2024-05-13 and GPT-3.5-turbo-0125, as well as o1-preview-2024-09-12.[4] We also test two Llama models: Llama-3-8B-instruct and Llama-3-70B-instruct. These models were chosen for their large sizes, which make them illustrative examples of the capabilities of state-of-the-art LLMs in general. We test 3 main scenarios: *zero-shot*, *in-context learning* with examples randomly selected from Stanford NLI (SNLI, Bowman et al. 2015), and *in-context learning* with Construc-

[3] To ensure that the templates did not bias the models in some way towards the appropriate NLI relation, we produced free-form NLI triples for the same premises in a separate research effort; we found that model performance improves on the freeform hypotheses over our templated hypotheses (Bonial et al., 2025). This indicates that the templates do not seem to cue the model to the correct relation.

[4] `https://platform.openai.com/docs/models`

Setting	IC	Accuracy				
	Data	GPT			Llama 3	
		3.5	4o	o1*	8B	70B
0-shot	None	0.6	0.88	**0.90**	0.59	0.74
1-shot	CxNLI	0.69	0.9	-	0.65	0.84
3-shot	CxNLI	**0.79**	**0.96**	**0.90**	**0.73**	**0.91**
1-shot	SNLI	0.58	0.86	-	0.59	0.75
3-shot	SNLI	0.59	0.86	0.89	0.58	0.74

Table 4: Results for Exp. 1 - Evaluation on our CxNLI dataset. “IC Data" refers to the type of data used as in-context examples.*GPT-o1 is only tested in zero-shot and three-shot settings on a subset of the overall data due to API costs.

tional NLI (CxNLI, our dataset).[5] A summary of our results for Exp. 1 are reported in Table 4.

Overall, we see that performance is high even in the zero-shot setting for GPT-4o and GPT-o1. We also observe that GPT-4o and Llama 3 70B consistently perform better than their smaller model counterparts GPT-3.5 and Llama 3 8B. Adding examples of Cxns for in-context learning boosts performance, while additional SNLI examples do not boost performance. This is especially true for GPT-3.5 and Llama 3 8B, which benefit substantially more from in-context learning from CxNLI. This reliance on in-context learning indicates that our datasets test a different axis of semantic knowledge than more general datasets like SNLI.

5 Experiment 2: NLI for Distinguishing Syntactically-Identical Cxns

5.1 Dataset

In Exp. 1, we show that the models perform impressively on an NLI task that specifically targets constructional semantics. The Cxns tested are common to the English language, and the templates we generate target aspects of meaning that are highly salient for the Cxn. In Exp. 2, we test whether models can generalize the appropriate constructional semantics for syntactically identical phrasal Cxns that should be ascribed distinct semantics. This enables us to determine if models have a robust capability to attribute and understand constructional semantics, or if this capability might be limited to the more entrenched phrasal Cxns of the language that were tested in Exp. 1.

Thus, for the 5 argument structure Cxns of our 8 Cxns, we add test instances which share a surface syntax with our Cxns, but convey a different meaning.[6] Table 5 provides examples of the original Cxns used in Exp. 1 and parallel, syntactically identical Cxns tested in Exp. 2. Consider the following:

(5) He hammered the metal flat. *(resultative)*

(6) I bought the apples fresh. *(depictive)*

In the above two examples, the syntactic forms are identical: both have a subject pronoun, a verb, an object noun phrase followed by an adjective. However, the two Cxns convey different meanings: In Example (5) the adjective is the *result* of the action of the verb, whereas in Example (6), the adjective is the state of the noun *during* the action of the verb, but it is *not* the resulting state of the action. This difference in meaning is associated with two different Cxns, specifically the RESULTATIVE and the DEPICTIVE. We can tease apart this difference in meaning by leveraging our template-based hypotheses from our CxNLI dataset in 4. Consider the following templatic hypotheses:

(7) [THE [*hammer*]$_2$-ING CAUSED [*the metal*]$_3$ TO BECOME [*flat*]$_4$]. *(entailment)*

(8) [MY [buying]$_2$-ING CAUSED [*the apples*]$_3$ TO BECOME [*fresh*]$_4$]. *(contradiction)*

As we can see in Examples (7) and (8), templatically generating hypotheses for these two examples leads to different relations to the premises. The Cxns we use for this dataset are INTRANSITIVE + PP$_{\text{AT}}$, INTRANSITIVE + PP$_{\text{TO}}$, DITRANSITIVE with NP, PP complements, TRANSITIVE + PP$_{\text{WITH}}$, and the DEPICTIVE.[7]

Manual Verification After the final version of this dataset was created, a second author evaluated the dataset, achieving an IAA of 83% with the original judgments. The final Exp. 2 “CxNLI-Distinction” dataset totals 99 NLI triples.

5.2 Formalism and Task Design

Let C be the target Cxn (e.g., RESULTATIVE) and C' be the syntactically identical distractor Cxn (e.g., DEPICTIVE). The Cxns are selected such that the entrenchment of the distractor is lower than

[5] We experiment with a variety of prompt formats and report results for the best performing prompts; details can be found in Appendix E. We also perform Chain-of-Thought (CoT) experiments (Wei et al., 2022) in each of these scenarios; it does not lead to performance gains: see Appendix F.

[6] We test 5 of 8 Cxns from Exp. 1 because 3 Cxns lack syntactically identical counterparts with distinct meanings.

[7] More examples of each of these Cxns along with NLI tuples are shown in Appendix D, Table 10.

Exp. 1		Exp. 2	
Cxn	**Argument Structure Cxn**	**Syntactically-Identical New Cxn**	**Cxn**
Resultative	I brushed my hair smooth.	A famous emperor buried scholars alive.	Depictive
Conative	Marco grabbed at the ladder railing.	Other units exploded at this complex.	Intransitive+"at"
Caused-motion	We exported nickel to the United States	I introduced her to my boss.	DitransitiveV+NP+PP
Intransitive-motion	The 23 scrambled to the rear of the sub.	They're listening to the same podcast.	Intransitive+"to"
Causative-with	Samsung flooded the market with advertising.	I use a mouse with my left hand.	Transitive+"with"

Table 5: Example CxNLI premises from Exp. 1 and 2, illustrating syntactically identical, semantically distinct Cxns.

that of the target: $Entr(C') < Entr(C)$ While both Cxns are realized via the same **constructional schema**, meaning their syntactic templates (e.g., `NP V NP ADJ`) are identical ($\mathcal{T}_C = \mathcal{T}_{C'} = \mathcal{T}$), their underlying semantic interpretations are distinct, $M(C) \neq M(C')$.

Our evaluation focuses on premises that are instances of the distractor Cxn C'. As an illustration, consider one such premise, p':

> Schema: $\mathcal{T}_{C'} =$ `NP V NP ADJ`
> Example: $p' \in C' \Rightarrow$ "A famous emperor buried China's scholars alive..."

The core of the experiment is to generate **two different hypotheses** for this single premise, each designed to probe for the meaning of either the (correct) distractor ($M(C')$) or (incorrect) target Cxn ($M(C)$).

1. **An NLI tuple probing the correct (DEPICTIVE) meaning:**
   - $h_1 = \mathcal{H}_{C',\text{Entailment}}(\mathbf{p}') \Rightarrow$ "China's scholars were fully alive before being buried."
   - This creates the NLI tuple: $\langle p', h_1, \text{Entailment}\rangle$.
2. **An NLI tuple probing the incorrect (RESULTATIVE) meaning:**
   - $h_2 = \mathcal{H}_{C,\text{Contradiction}}(\mathbf{p}') \Rightarrow$ "Burying caused the scholars to become alive."
   - This creates the NLI tuple: $\langle p', h_2, \text{Contradiction}\rangle$.

Crucially, this experimental design assesses if models can avoid over-generalizing the meaning of the more entrenched target Cxn, $M(C)$, to a less entrenched distractor Cxn, C', that shares the same syntactic schema ($\mathcal{T}_{C'} = \mathcal{T}_C$) but has a distinct semantic interpretation ($M(C') \neq M(C)$). By measuring overall accuracy, we test a model's ability to reject the semantics of the highly entrenched Cxn ($M(C)$) when presented with a premise from the less entrenched distractor, while simultaneously accepting the correct meaning ($M(C')$). We use performance on the more straightforward instances in Exp. 1 as a baseline; therefore, a significant drop in accuracy on this second task indicates a failure to distinguish Cxns based on their subtle semantic cues, highlighting a key difference from human generalization.

5.3 Empirical Evaluation and Analysis

We utilize this new Exp. 2 dataset as an evaluation dataset to test if LLMs can perform NLI successfully on the new phrasal Cxn examples, where performance requires distinguishing the unique semantics of the syntactically identical Cxn. As we can see in Table 6, performance is significantly lower than our results from Exp. 1 in almost every prompt setting and across all models. The difference in performance is stark. While these examples are also slightly more difficult for humans, the ceiling of human performance (IAA 83%) is well above current LLM performance, even for GPT-4o, GPT-o1 and Llama 3 70B. Again we see a large difference between the GPT-4o and the smaller models, and also see that GPT-o1 performs worse than GPT-4o.

We take these results as evidence that the ability of models to abstract over the same syntactic slots and assign the appropriate semantics is limited to the more entrenched Cxns of English, of which our original Exp. 1 dataset consisted. By shifting to phrasal Cxns that are syntactically identical but potentially less entrenched, we are essentially testing whether or not models can abstract over less data to arrive at a less statistically likely semantic interpretation. This also highlights that the abstraction process in people clearly goes beyond the syntactic character of the slots alone—people are able to balance their knowledge of lexical semantics with their knowledge of constructional semantics to arrive at the most pragmatically likely interpretation. People's lexical awareness is also imbued with physical world knowledge; thus, people are abstracting over a distinct set of information than what LLMs have access to.

Setting	IC	Accuracy				
	Data	GPT			Llama 3	
		3.5	4o	o1*	8B	70B
0-shot	None	0.26	**0.58**	**0.46**	0.38	0.52
1-shot	CxNLI	0.26	0.53	-	0.29	**0.60**
3-shot	CxNLI	**0.31**	0.57	0.45	0.35	0.50
1-shot	SNLI	0.3	0.55	-	0.37	0.54
3-shot	SNLI	0.28	0.53	**0.46**	**0.39**	0.57

Table 6: Results for Exp. 2, testing on CxNLI-Distinction. "IC Data" refers to the type of data used in the in-context examples. *GPT-o1 is only tested in zero-shot and three-shot settings due to limited resources.

5.4 Statistical Analysis

We tested whether performance on CxNLI-Distinction (Exp. 2) is worse than on CxNLI (Exp. 1) for humans and for each model. Specifically, we conducted a Bayesian A/B test to quantify the evidence for a performance drop on the more challenging CxNLI-Distinction (Exp. 2) dataset. This analysis was performed separately for human participants and three different models: LLaMA, GPT-4o, and GPT-o1. We assigned a strong prior belief of 99% that performance on the CxNLI-Distinction dataset would be equal to the CxNLI (Exp. 1) dataset and a 1% probability of lower performance.

Despite encoding strong priors favoring equal performance, our data consistently point to a performance drop. The analysis, based on 10,000 posterior samples, gives a Bayes Factor (BF) for each group. Using the standard interpretation where a Bayes Factor over 10 constitutes strong evidence, we find that for humans, the data are 4.3 times more likely under the hypothesis that performance is worse (BF = 4.31). Given that a BF of 10 constitutes strong evidence, this is only weak evidence that the performance is worse. In contrast, **for LLaMA the evidence is very strong**, with the data being over 2,600 times more likely under this hypothesis (BF_{10} = 2,684). Furthermore, **for GPT-4o and o1 the evidence is extreme** with the BFs being 3.6 x 10^8 and 8.2 x 10^{16} respectively.

Thus, the evidence overwhelmingly supports the hypothesis that the CxNLI-Distinction (Exp. 2) set is significantly more difficult for LLMs than for humans. Additionally, this demonstrates that our CxNLI-Distinction dataset size is large enough to conclude that performance is decisively worse on this dataset, which requires distinguishing between syntactically identical constructions.

6 Error Analysis

In Exp 1, we describe our CxNLI experiments, which find that GPT-4o, GPT-o1 and Llama 3 70B are extremely proficient at CxNLI while GPT-3.5 and Llama 3 8B lag behind substantially. In Exp 2, we show that all our tested models do not perform well when tested on our CxNLI-Distinction dataset with five additional, syntactically identical Cxns. For example, though GPT-4o's performance on the CxNLI RESULTATIVE is near perfect, it struggles to demonstrate understanding of the syntactically identical DEPICTIVE:

(9) **Premise:** *I bought the apples fresh.*
Hypothesis: *The apples were completely fresh before I bought them.*
Correct Response: Entailment
Model Response: Contradiction

Here, we investigate if some Cxns are harder for LLMs than others. Among Cxns from Exp. 1, LET-ALONE and COMPARATIVE-CORRELATIVE are the weakest Cxns for GPT-4o, though it is strong across the board with a minimum accuracy of 88%. GPT-3.5 is much more variable by Cxn, with a maximum accuracy of 92% for the CONATIVE and a minimum of 67% for LET-ALONE. We show an example of GPT-4o misunderstanding the scale of LET-ALONE in Example (10).

(10) **Premise:** *Beecher's reputation as a preacher, let alone as a Man of God, was not universally accepted.*
Hypothesis: *Beecher's reputation as a Man of God was easier to accept than his reputation as a preacher.*
Correct Response: Contradiction
Model Response: Entailment

In Figure 1 we report the accuracy by Cxn in our Exp. 1 NLI and Exp. 2 NLI datasets.[8] We see performance is lower for all Cxns we test in Exp. 2. This provides evidence supporting our hypothesis, outlined in §3, that the entrenchment of a Cxn contributes to model ability to understand the constructional semantics.

[8] Accuracies are from the highest performing prompt. We only visualize GPT-4o for visual clarity, though trends are similar across models.

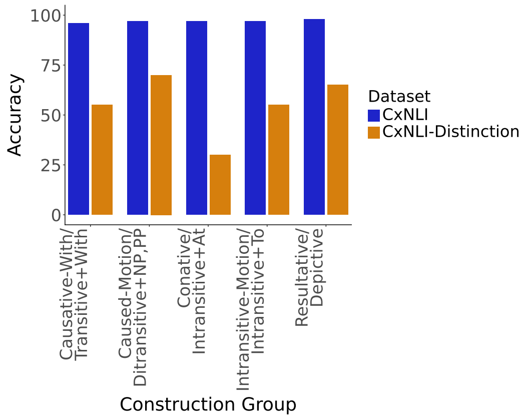

Figure 1: Constructional NLI Accuracy broken down by Cxn (for best prompts). Accuracy drops substantially for the Cxns in Exp. 2 relative to those in Exp. 1.

7 Discussion

Our results show a clear discrepancy in the ability of LLMs to process constructional meaning. While models demonstrated a surprisingly high capacity to interpret familiar constructions even with novel lexical fillers (Exp. 1), their performance dropped significantly when required to distinguish between syntactically identical constructions that carry different meanings (Exp. 2). This discrepancy highlights a failure in how models generalize from frequent patterns to more nuanced, creative uses of language and has significant implications to the development and use of LLMs

Data, Bias, and the Limits of Scale Our Exp. 2 findings challenge the prevailing "more data is better" paradigm (Kaplan et al., 2020). The models' bias towards the most frequent constructional meaning suggests that simply scaling web-text data may reinforce these errors. This implies a need for new data strategies, such as up-sampling rare-but-important structures or using targeted (Ye et al., 2025), adversarial fine-tuning (Dong et al., 2021) to correct the biases of the base model.

The Need for Diagnostic Evaluation The significant difference in performance between our two experiments shows how broad-coverage benchmarks can overestimate a model's true linguistic competence. Our work demonstrates the importance of contrastive, diagnostic benchmarks to test specific, theoretically-grounded phenomena.

A Failure of Causal Reasoning and Its Safety Implications The model's inability to distinguish a DEPICTIVE from a RESULTATIVE construction is a failure to reason about causality. *This is not a niche linguistic error;* it has direct implications for AI safety as misunderstanding the difference between a co-occurring state and a caused outcome could lead to catastrophic errors.

Architectural Limitations and Future Directions *The systematic nature of this error across models suggests the issue may be rooted in the architecture of these models.* Therefore, the path forward may benefit from the use of novel and hybrid architectures, such as the incorporation of constructional resources or the addition of long term memory (Wang et al., 2023).

8 Conclusions and Future Work

We have shown where even the latest models do not demonstrate a functional understanding of Cxns: although models can generalize the semantics of entrenched Cxns to creative instantiations, the same models cannot robustly distinguish between syntactically identical Cxns with distinct semantic interpretations. While GPT-4o, GPT-o1 and Llama 3 70B do perform quite impressively on our original constructional NLI task, they fail at the NLI scenario requiring constructional distinction, which requires generalization of the appropriate constructional semantics to syntactically-identical Cxns. Also, we see that GPT-4o substantially outperforms GPT-3.5 in all settings, and in-context learning is especially crucial for GPT-3.5. Overall, these experiments show that the constructional awareness of GPT-4o, GPT-o1 and Llama 3 70B are far more robust than that of GPT-3.5 and Llama 3 8B, but their ability to generalize constructional meaning to both novel instantiations and distinct Cxns still lags substantially behind that of humans.

Thus, our targeted series of experiments demonstrate that LLMs do process constructional semantics up to a point, yet our datasets and experiments reveal the breaking point of understanding—where speakers are able to recognize the appropriate constructional semantics despite both novel instantiations and despite the fact that there are multiple, syntactically identical Cxns that could be candidates for interpreting the phrase at hand. Overall, we find that CxG serves as a valuable theoretical lens for probing the functional language understanding of LLMs with a methodology that tests for linguistic generalization beyond memorization and dependency on pretraining priors and comparing this with human linguistic knowledge. Greater contributions to resources such as corpora of Cxns will facilitate empirical data on which constructional understanding can be evaluated with more detail.

Limitations

This work is limited in that we only evaluate our methods on English. More work is needed on the targeted evaluation of LLM performance using Cxn information in non-English settings. Our tasks are only one possible method for investigating LLM understanding of Cxns. Expanding research to include complementary methodologies will be necessary to build a complete picture of LLM knowledge in relation to CxG. This work can also be extended beyond the 8 Cxns that we use to generate our dataset, although these were selected for the extensive coverage of the English language. Also, while we consciously choose to create a smaller, more carefully curated dataset that also allows for careful expert manual evaluation, there is scope to increase the size of our dataset, which we leave to future work.

Ethics

LLMs are extremely expensive to train and run. The compute costs associated with LLMs have a nontrivial environmental impact which should not be ignored. Furthermore, due to their large-scale training data, they can reflect and propagate harmful social biases in their responses if they are not properly aligned and moderated. Furthermore, there is risk of LLMs having a negative societal impact if their widespread deployment is done without proper consideration for the lives of people. While there are risks in the use and proliferation of LLMs in general, we do not believe this work incurs any specific additional risks. Despite the overall risks and dangers, we believe this research is worthwhile in order to better understand the language systems of LLMs and compare and contrast LLM language understanding with that of humans. We honor the code of ethics.

Acknowledgments

Melissa Torgbi is supported by the UKRI Centre for Doctoral Training in Accountable, Responsible, and Transparent AI [EP/S023437/1] of the University of Bath.

References

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Claire Bonial, Taylor Pellegrin, Melissa Torgbi, and Harish Tayyar Madabushi. 2025. From form to function: A constructional nli benchmark. In *Proceedings of the Second International Workshop on Construction Grammars & NLP (CxG+NLP 2025), co-located with IWCS 2025*.

Claire Bonial and Harish Tayyar Madabushi. 2024a. Constructing Understanding: on the Constructional Information Encoded in Large Language Models. *Language Resources and Evaluation*, pages 1–40.

Claire Bonial and Harish Tayyar Madabushi. 2024b. 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*. 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.

Joan Bybee. 2010. *Language, usage and cognition*. Cambridge University Press.

Noam Chomsky. 2014a. *The minimalist program*. MIT press.

Noam Chomsky. 2014b. *The Minimalist Program*.

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)*, page 242–261, Toronto, Canada. Association for Computational Linguistics.

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

Mark Davies. 2010. The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, 25(4):447–464.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, and et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Xinshuai Dong, Anh Tuan Luu, Min Lin, Shuicheng Yan, and Hanwang Zhang. 2021. How should pre-trained language models be fine-tuned towards adversarial robustness? In *Advances in Neural Information Processing Systems*, volume 34, pages 4356–4369. Curran Associates, Inc.

Shahriar Golchin and Mihai Surdeanu. 2024. Time travel in llms: Tracing data contamination in large language models. *Preprint*, arXiv:2308.08493.

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

Stefan Th Gries and Anatol Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspective onalternations'. *International journal of corpus linguistics*, 9(1):97–129.

Thomas Hoffmann. 2022. *Construction Grammar*. Cambridge University Press.

Matt A Johnson and Adele E Goldberg. 2013. Evidence for automatic accessing of constructional meaning: Jabberwocky sentences prime associated verbs. *Language and Cognitive Processes*, 28(10):1439–1452.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Martha Lewis and Melanie Mitchell. 2024. Using Counterfactual Tasks to Evaluate the Generality of Analogical Reasoning in Large Language Models. *arXiv preprint arXiv:2402.08955*.

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.

Yingcong Li, M. Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. 2023. Transformers as algorithms: generalization and stability in in-context learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

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. *arXiv preprint arXiv:1907.11692*.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2024. Are Emergent Abilities in Large Language Models just In-Context Learning? *Preprint*, arXiv:2309.01809.

Bill MacCartney and Christopher D. Manning. 2008. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 521–528, Manchester, UK. Coling 2008 Organizing Committee.

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.

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

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.

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.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023a. NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, and Eneko Agirre. 2023b. Did ChatGPT cheat on your test?

Roland Schäfer. 2015. Processing and querying large web corpora with the COW14 architecture. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora*, page 35–41, Lancaster.

Roland Schäfer and Felix Bildhauer. 2012. Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, page 486–493, Istanbul, Turkey. European Language Resources Association (ELRA).

Wesley Scivetti, Tatsuya Aoyama, Ethan Wilcox, and Nathan Schneider. 2025. 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 27491–27502, 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.

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.

Harish Tayyar Madabushi, Melissa Torgbi, and Claire Bonial. 2025. Neither stochastic parroting nor agi: Llms solve tasks through context-directed extrapolation from training data priors. *Preprint*, arXiv:2505.23323.

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

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.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023. Augmenting language models with long-term memory. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Leonie Weissweiler, Nina Böbel, Kirian Guiller, Santiago Herrera, Wesley Scivetti, Arthur Lorenzi, Nurit Melnik, Archna Bhatia, Hinrich Schütze, Lori Levin, Amir Zeldes, Joakim Nivre, William Croft, and Nathan Schneider. 2024. UCxn: Typologically informed annotation of constructions atop Universal Dependencies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16919–16932, Torino, Italia. ELRA and ICCL.

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.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862, Mexico City, Mexico. Association for Computational Linguistics.

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. 2025. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. In *The Thirteenth International Conference on Learning Representations*.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. 2023. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. *Preprint*, arXiv:2305.19420.

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)*, page 3804–3811, Torino, Italia. ELRA and ICCL.

Xin Zhou, Martin Weyssow, Ratnadira Widyasari, Ting Zhang, Junda He, Yunbo Lyu, Jianming Chang, Beiqi Zhang, Dan Huang, and David Lo. 2025. LessLeak-Bench: A First Investigation of Data Leakage in LLMs Across 83 Software Engineering Benchmarks. *Preprint*, arXiv:2502.06215.

A Construction Grammars

CxG has particular explanatory power with respect to phrasal Cxns, such as the RESULTATIVE Cxn: "The jackhammer pounded us deaf." Generative linguistic theories (e.g., Chomsky (2014b)) would generally analyze a transitive sentence with one verbal head ("pounded") that licenses the arguments of the sentence. "Pounded" generally licenses an agent subject (here, "jackhammer") and potentially a patient direct object. Generative approaches argue that this information about the verb is memorized and stored in the lexicon, while combinatory rules of how to put lexical items together are stored in a separate syntax module of language processing. However, unless a special grammatical rule or sense of the verb is postulated, there is nothing to explain why "us" is not the thing pounded here, or what licenses the adjective "deaf." Nonetheless, native speakers have no problem recognizing the special formal and semantic properties of this Cxn, namely that it entails a pounding event that causes a change in state of "us" resulting in the state of "deaf."

In contrast to Generative linguistic theories (e.g., Chomsky (2014b)), CxG posits that speakers acquire and store Cxns, which notably account for not only the semantic properties of the unit but also the formal syntactic properties. Cxns at all levels of language are learned through language usage; thus, in lieu of grammatical rules accounting for the grammaticality of a particular structure, frequency plays an important role in what 'sounds right' or is grammatical to a speaker. In the CxG account, speakers acquire single-word Cxns that they are frequently exposed to (e.g., "milk"), but then generalize from that to recognize how an acquired holophrastic Cxn falls into slots of larger, more complex Cxns (e.g., "want milk," "baby want milk," "want up," "I want to go"). In this process of generalization, speakers build up a taxonomically organized set of the related Cxns of their language, or a *constructicon*. Speakers use domain-general cognitive processes to incrementally generalize over such frequent Cxns to novel usages, arriving at the ability to interpret even rare and previously unseen instances of a Cxn.

B Scientific Artifacts and Descriptive Statistics for All Datasets

We use the following scientific artifacts: COCA (Davies, 2010), EnCOW (Schäfer and Bildhauer, 2012; Schäfer, 2015), the CoGS dataset (Bonial and Tayyar Madabushi, 2024b), and the OpenAI API, in addition to our created datasets. The COCA corpus contains 8 genres: Academic, Blog, Fiction, Magazine, News, Spoken, TV, and Web. It is intended to capture American English, though there is no guarantee that it does not also include some other varieties. Demographic information about the creators of the texts in the corpus is not always available given its scale. EnCOW is a large scale corpus of English text from the web. As such, the demographic information that it captures is not completely clear. All datasets besides COCA and EnCOW are open-source under a Creative Commons license. The institutions of the authors have valid licenses for COCA and EnCOW, permitting their use in academic settings. We use the artifacts as intended. Overall, we construct and experiment on 2 datasets: CxNLI (Exp. 1) and CxNLI-Distinction (Exp. 2). The sizes and IAA agreement for the final datasets is reported in Table 7. All of our datasets are exclusively in English.

Dataset	N	Tokens	IAA
CxNLI (Exp. 1)	435	15,144	90%
CxNLI-Distinction (Exp. 2)	99	3,202	83%

Table 7: Descriptive statistics for each dataset. N is the number of unique examples.

B.1 CxNLI (Exp. 1) Data Sources

The templatic dataset is constructed with real-world corpus data of Cxns, primarily coming from the CoGS dataset (Bonial and Tayyar Madabushi, 2024b), with supplementary data coming from Corpus of Contemporary American English (COCA, Davies 2010), and the English Corpus from the Web, or EnCOW (Schäfer and Bildhauer, 2012; Schäfer, 2015). Within this dataset, each premise includes one of 8 total Cxns: the COMPARATIVE-CORRELATIVE Cxn, the LET-ALONE Cxn, the

WAY-MANNER Cxn, the CAUSATIVE-WITH Cxn, the CONATIVE Cxn, the RESULTATIVE Cxn, the CAUSED-MOTION Cxn, and the INTRANSITIVE-MOTION Cxn. We include a roughly balanced sample of each Cxn, with all premises taken from corpus data. These Cxns cover a wide range of *schematicity* meaning that they have different levels of lexicalization/abstractness.

C Annotator Information

All datasets are annotated by co-authors of this paper. 2 annotators identify as men, 4 annotators identify as women. Of our annotators, 3 have a graduate degree in linguistics while 3 do not. At least one linguistic expert and one non-expert annotate each dataset. Because the source of our datasets is web corpora, there is some risk of offensive or hateful content. During annotation, annotators were asked to remove any hateful or offensive content as well as personal identifying information. The annotation task was given to annotators in a spreadsheet. The instructions provided are detailed in Appendix H.

D Example NLI Tuples

Our templates for generating hypotheses across Cxn types is shown in Table 8. In Table 9, we show examples of our constructional NLI datasets from Exp. 1. We show examples of the distinction-requiring NLI examples from Exp. 2, alongside examples of our new constructions for Exp. 2, in Table 10.

E Prompt Variation Experiments

For each experiment, we choose one setting to observe the impact of prompt variations on performance.[9] These prompt variations were explored in two sequential stages; changing the content of the prompt and changing the phrasing of the prompt. Changing the content of the prompt involved giving more details about the task, giving fewer details about the task and experimenting with the specific instructions. After selecting the best-performing prompt content, the prompt was rephrased using an LLM and variations of the prompt with new wording were used to check the robustness of the model output by observing how small variations in the prompts affect performance. We report an example of each of our prompts for variations in content in Table 11. Results for our prompt variation experiments are shared in Tables 12 and 13.

F Chain-of-Thought Results for all Experiments

We replicate each of our main experiments with the addition of Chain-of-Thought (CoT) to observe changes in performance and inspect any errors in the model's reasoning steps. For CoT thought prompting we do not provide examples of reasoning in the prompt, instead we simply ask the model to explain "step by step". The results for each experiment have been shared in tables 14 and 15.

G Model Parameters and Hyperparameters

We use the default hyperparameters for all models including GPT-4o (unreported parameter size), GPT-3.5 (175B parameters), GPT-o1 (unreported parameter size), LLaMA 3 8B (8B parameters), LLaMA 3 70B (70B parameters) though to ensure maximal reproducibility of our work, we set the temperature of model responses to 0 apart from GPT-o1 which only acceptd a temperature of 1. Our GPT experiments were done through the OpenAI API. The total cost of our experiments was approximately $250 USD. Our LLaMA experiments were run through the replicate API and the experiments cost approximately $25 USD.

H Constructional Natural Language Inference Annotation Guidelines

Here we provide the exact instructions given to people to annotate the NLI datasets.

We have developed a dataset of sentences featuring different linguistic “constructions”—pairings of form and meaning. The constructions exemplified in this dataset range from purely substantive (the words filling the constructional slots are fixed), such as the *Much-less* construction, e.g., “He won't eat shrimp, much less squid;” to purely schematic (the words filling the constructional slots can vary, but fulfill some general semantic and syntactic requirements), such as the *Caused-Motion* construction, e.g., “She blinked the snow off her eyelashes.” It's okay if you aren't familiar with this terminology or the idea of these constructions!

Task Overview

Your job is to read the sentences from this dataset, which are presented as the **Premise** in a set of

[9] For NLI we choose the 3-shot CxNLI setting.

Cxn Name	**Cxn Template**	**NLI Hypothesis Template**
Causative-With	[SBJ$_1$ [V$_2$ OBJ$_3$ *with*-PP$_4$]$_{\text{VP}}$]$_5$	SBJ$_1$ *did not cause* OBJ$_3$ *to contain* OBJ-of-PP$_4$
Caused-Motion	[SBJ$_1$ [V$_2$ OBJ$_3$ PP$_4$]$_{\text{VP}}$]$_5$	OBJ$_3$ *did not change locations.*
Comparative-Correlative	[[*the*$_1$ [Comparative Phrase]$_2$ REST-CLAUSE$_3$]$_{\text{C1}}$ [[*the*$_4$ [Comparative Phrase]$_5$ REST-CLAUSE$_6$]$_{\text{C2}}$]$_7$	*The amount* [Comparative Phrase]$_2$ *is positively/negatively correlated with the amount* [Comparative Phrase]$_5$
Conative	[SBJ$_1$ [V$_2$ *at*-PP$_3$]$_{\text{VP}}$]$_4$	OBJ-of-PP$_3$ *was not the target of the* V$_2$*-ing motion.*
Intransitive-Motion	[SBJ$_1$ [V$_2$ PP$_3$]$_{\text{VP}}$]$_4$	SBJ$_1$ V$_2$ *in a static location.*
Let-Alone	XP$_1$ CONJ$_{2\text{-}3}$ XP$_{4\text{-}5}$	*If* XP$_1$ *then not* XP$_{4\text{-}5}$
Resultative	[SBJ$_1$ [V$_2$ OBJ$_3$ AP$_4$]$_{\text{VP}}$]$_5$	*The* V$_2$*-ing did not cause* OBJ$_3$ *to become* AP$_4$.
Way-Manner	[SBJ$_1$ [V$_2$ [PRON$_{3=1}$] *way*$_4$]$_{\text{OBJ5}}$ PP$_6$]$_{\text{VP}}$]$_7$	SBJ$_1$ *traveled* PP$_6$ *without* V$_2$*-ing.*

Table 8: Example templates for Cxns and templatic constructional NLI hypotheses. All examples provided here have the contradiction relation. In these templates OBJ stands for a bare object, and OBL stands for an oblique, which is a prepositional phrase that introduces a recipient, goal, or result of the verb. AP, PP, and VP stand for adjective phrase, prepositional phrase, and verb phrase respectively.

triples for the Natural Language Inference (NLI) task. Also known as Recognizing Textual Entailment (RTE), NLI is the task of determining the inference relation between two (short, ordered) texts: entailment, contradiction, or neutral (MacCartney and Manning, 2008).

- **Premise:** A man inspects the uniform of a figure in some East Asian country.
- **Hypothesis:** The man is sleeping.

Then, you will fill in the **Relation** between the Premise and the Hypothesis, which indicates the kind of entailment between the two sentences. We are using numerical coding, listed below and in your annotation spreadsheet:

0 – **entailment** – The hypothesis must be true given the premise.

1 – **neutral** – The hypothesis may or may not be true given the premise.

2 – **contradiction** – The hypothesis must not be true given the premise.

So for the example above, the correct answer would be:

> **2 – contradiction** – If the man is inspecting a uniform, then it must not be true that the man is sleeping.

The two sentences describe the *same scenario.* Entities mentioned in both premise and hypothesis refer to the same thing; e.g., "the man" refers to the same individual. The hypothesis does not describe a different time.

If you encounter unfamiliar words, you may consult a dictionary. However, it is not expected or encouraged that you would have to "do research" into a topic in order to determine a relation between a premise and hypothesis. Instead, you should rely on common sense and your understanding of the words.

You will be completing these annotations in a spreadsheet like what is shown below, where there is a relation space available below a given premise/hypothesis pair. In the cell to the right of "relation", you will provide the appropriate relation number. This space should include nothing except for the numbers 0, 1, or 2.

If you would like to note instances that are problematic, please add a Notes column to the right of Annotation Target and make your note as relevant to the right of the premise, hypothesis or relation.

There is also a space in column L where you can note the start and end times of each annotation session, or "sitting." Please kindly track about how many judgments you are able to do in each sitting, so we can get a sense of how long the annotation task takes.

There are about 100 distinct NLI judgments or triples per annotation spreadsheet. Once you have completed all the relation annotations, please save and send the spreadsheet back to me: Claire.n.bonial.civ@army.mil.

Get your inference hat on. Happy annotating!

Cxn Name		**CxNLI (Exp. 1)**
Causative-With	Premise	*Freshly ground coffee beans filled the room with a seductive, earthy aroma.*
	Hypothesis Relation	*The room did not contain a seductive, earthy aroma.* Contradiction
Caused-Motion	Premise	*I threw the stone across the river.*
	Hypothesis Relation	*I caused the stone to move across the river by throwing it.* Entailment
Comparative-Correlative	Premise	*The more they work, the more I will pay them.*
	Hypothesis Relation	*Increasing the amount they work will increase the amount I pay them.* Entailment
Conative	Premise	*I sipped at the Heineken.*
	Hypothesis Relation	*The Heineken was not the target of my sipping.* Contradiction
Intransitive-Motion	Premise	*I ran around the track.*
	Hypothesis Relation	*I ran, staying in one place.* Contradiction
Let-Alone	Premise	*It's unsurprising that such an attitude failed to produce competent screenwriters, let alone exciting ones.*
	Hypothesis Relation	*An attitude that produces exciting screenwriters can also produce competent ones.* Entailment
Resultative	Premise	*The jackhammer pounded us deaf.*
	Hypothesis Relation	*We were completely deaf before the jackhammer pounded.* Contradiction
Way-Manner	Premise	*I yawned my way back to the Narrow Neck.*
	Hypothesis Relation	*I traveled back to Narrow Neck without yawning.* Contradiction

Table 9: Examples of templatically generated hypotheses for and resulting NLI tuples for Exp. 1 CxNLI.

CxN Name		**Example NLI Triple**
Intransitive+At	Premise Hypothesis Relation	I watch the women, their legs crossed at the ankles, try to look as if they don't sweat at all. Their legs made a crossing motion towards their ankles. Contradiction
Transitive+with	Hypothesis Premise Relation	He hit the lamp with his head. He caused the lamp to contain his head. Contradiction
Intransitive+To	Hypothesis Premise Relation	He spoke to the workers on the street corner. He changed locations by speaking. Contradiction
Ditransitive+NP,PP	Hypothesis Premise Relation	I introduced her to my boss. She remained in the same place. Neutral
Depictive	Hypothesis Premise Relation	A famous emperor buried China's scholars alive with their books. Burying caused the scholars to become alive. Contradiction

Table 10: Example Exp. 2 NLI triples corresponding to each of the 5 constructions evaluated in the Exp. 2 Cxn NLI dataset. 5 of the 8 Exp. 1 Cxns are included in this dataset; limited to those Exp. 1 constructions with syntactically identical counterpart Cxns that have different meanings.

Prompt Type	Prompt
Variation 1	You are the world's best annotator. Your task is to read sentences from a dataset, presented as the Premise in a set of triples for the Natural Language Inference (NLI) task. Also known as Recognizing Textual Entailment (RTE), NLI involves determining the inference relation between two short, ordered texts: entailment, contradiction, or neutral. Next, you will identify the Relation between the Premise and the Hypothesis, which indicates the type of entailment between the two sentences. We use numerical coding, also listed in your annotation spreadsheet as a reminder: 0 – entailment – The hypothesis must be true given the premise 1 – neutral – The hypothesis may or may not be true given the premise 2 – contradiction – The hypothesis must not be true given the premise Output a single numerical value between 0, 1, or 2, corresponding to the associated relation. Output a single number only and nothing else.
Variation 2	You are the world's best annotator. You are tasked with annotating a triple for Natural Language Inference. You must determine the inference relation between the Premise and the Hypothesis by selecting one of three numerical codes that reflect the relationship: 0 – Entailment: The Hypothesis is definitely true given the Premise. 1 – Neutral: The Hypothesis may or may not be true given the Premise. 2 – Contradiction: The Hypothesis cannot be true given the Premise. Output a single numerical value between 0 and 2 inclusive, corresponding to the associated relation.
Variation 3	You are the best at understanding language inference based on Cxn grammar. You are tasked with annotating a triple for Natural Language Inference. You must determine the inference relation between the premise and the hypothesis by selecting one of three numerical codes that reflect the relationship: 0 – entailment – The hypothesis must be true given the premise 1 – neutral – The hypothesis may or may not be true given the premise 2 – contradiction – The hypothesis must not be true given the premise Output a single numerical value between 0, 1, or 2, corresponding to the associated relation. Output a single number only and nothing else.
Variation 4	You are the world's best annotator. Your task is to read sentences from a dataset, provided as the Premise in a set of triples for the Natural Language Inference (NLI) task. Also called Recognizing Textual Entailment (RTE), NLI requires determining the inference relation between two short, ordered texts: entailment, contradiction, or neutral. Your next step is to identify the Relation between the Premise and the Hypothesis, specifying the type of entailment between the two sentences. We use the following numerical coding: 0 – entailment – The hypothesis must be true given the premise 1 – neutral – The hypothesis may or may not be true given the premise 2 – contradiction – The hypothesis must not be true given the premise Output a single numerical value between 0 and 2 inclusive, corresponding to the associated relation.

Table 11: Prompt variations for Exp. 1 - CxNLI and Exp. 2 - CxNLI-Distinction

Prompt Type	Accuracy	
	GPT-3.5	GPT-4o
Variation 1	0.74	0.95
Variation 2	0.79	0.92
Variation 3	0.76	0.94
Variation 4	0.74	0.93
Best Variation Rephrase 1	0.79	0.95
Best Variation Rephrase 2	0.67	0.95
Best Variation Rephrase 3	0.69	0.96

Table 12: Prompt Variation Results for Exp. 1 - CxNLI in the three-shot setting.

Prompt Type	Accuracy	
	GPT-3.5	GPT-4o
Variation 1	0.24	0.53
Variation 2	0.28	0.46
Variation 3	0.31	0.57
Variation 4	0.26	0.54
Best Variation Rephrase 1	0.31	0.57
Best Variation Rephrase 2	0.26	0.55
Best Variation Rephrase 3	0.23	0.54

Table 13: Prompt Variation Results for Exp. 2 CxNLI-Distinction in the three-shot setting.

Setting	IC Data	Accuracy	
		GPT-3.5	GPT-4o
Zero-shot	None	0.60	0.89
One-shot	CxNLI	0.66	0.89
Three-shot	CxNLI	0.71	0.92
One-shot	SNLI	0.63	0.89
Three-shot	SNLI	0.66	0.91

Table 14: Results for Exp. 1 - Cxn NLI with Chain-of-Thought, "IC Data" refers to the type of data used as in-context examples.

Setting	IC Data	Accuracy	
		GPT-3.5	GPT-4o
Zero-shot	None	0.29	0.43
One-shot	CxNLI	0.29	0.43
Three-shot	CxNLI	0.35	0.46
One-shot	SNLI	0.28	0.43
Three-shot	SNLI	0.31	0.45

Table 15: Results for Exp. 2 with the CxNLI-Distinction data as the test and Chain-of-Thought, "IC Data" refers to the type of data used as in-context examples.