From Form to Function: A Constructional NLI Benchmark

Claire Bonial[1], **Taylor Pellegrin**[2], **Melissa Torgbi**[3], **Harish Tayyar Madabushi**[3],
[1]DEVCOM Army Research Laboratory, U.S.A.
[2]Oak Ridge Associated Universities, U.S.A. [3]University of Bath, U.K.
claire.n.bonial.civ@army.mil

Abstract

We present CoGS-NLI, a Natural Language Inference (NLI) evaluation benchmark testing understanding of English phrasal constructions drawn from the Construction Grammar Schematicity (CoGS) corpus. This dataset of 1,500 NLI triples facilitates assessment of constructional understanding in a downstream inference task. We present an evaluation benchmark based on the performance of two language models, where we vary the number and kinds of examples given in the prompt, with and without chain-of-thought prompting. The best-performing model and prompt combination achieves a strong overall accuracy of .94 when provided in-context learning examples with the target phrasal constructions, whereas providing additional general NLI examples hurts performance. This evidences the value of resources explicitly capturing the semantics of phrasal constructions, while our qualitative analysis suggests caveats in assuming this performance indicates a deep understanding of constructional semantics.

1 Introduction

This research addresses the challenge of how we determine what computational systems know of a language; specifically, we focus on the large portion of the English language in which meaning goes beyond the sum of lexical parts—phrasal constructions. Whereas our past NLP tools were developed and therefore grounded in some form of grammatical theory (e.g., phrase structure or dependency parsing), LLMs lack grounding in linguistic theory. Instead, their development is based on the encoder-decoder architecture, which was originally designed for sequence-to-sequence tasks, specifically translation (Bahdanau et al., 2016). This dichotomy impedes methods for evaluating LLMs, as their performance on meta-linguistic tasks, such as semantic role labeling, which previously served as benchmarks for the individual components in an NLP pipeline, are poor predictors of LLM fluency on downstream applications.

Premise	I had brushed my hair smooth.
Hypothesis	I had smooth hair because I brushed it.
Relation	Entailment

Table 1: CoGS-NLI example for a premise including the Resultative cxn; inferring the entailment relies upon recognition of the constructinoal semantics.

Although LLMs lack theoretical grounding, evaluation of language proficiency benefits from analysis through a particular theoretical lens, which enables one to hypothesize the appropriate formal units of a language and the way in which meaning is associated with those formal units. We leverage Construction Grammar (CxG) to analyze language (specifically English) as a set of constructions (cxn), or pairings of meaning and form at any structural level, including morphemes, lexemes, and phrases. As a usage-based linguistic theory, CxG provides an experimentally-validated framework for how speakers acquire language and generalize knowledge of frequently heard cxns to totally creative and novel instantiations (e.g., Tomasello (2009); Johnson and Goldberg (2013)). **CxG research demonstrates that speakers attribute meaning to special syntactic templates (phrasal cxns)—meaning that goes beyond that of the individual lexical items alone; CoGS-NLI allows us to evaluate if LLMs also attribute the appropriate meaning to phrasal cxns.**

We leverage Construction Grammar Schematicity (CoGS) corpus instances (Section 2) as the premises in the subsequent development of a comprehensive dataset of 1500 Natural Language Inference (NLI) triples (see Table 1), which serves as a downstream test of functional understanding of cxns (Section 3). We benchmark performance on

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 172–179

this task with two models (GPT-3.5-turbo, GPT-4o), and demonstrate that including examples with constructional premises in few-shot prompting boosts performance to reach a top-end accuracy of .94 (Section 4).[1] This shows first that resources exemplifying the target constructional semantics are beneficial to performance, and second that constructional premises do not pose a problem for state-of-the-art models in this task. However, there is qualitative evidence that tempers the conclusion that models must grasp constructional semantics in order to perform successfully on the task (Section 5). We close with recommendations for future steps in evaluating constructional understanding (Section 6).

2 Related Work

Related work in the area of evaluating LLMs through the lens of CxG fall broadly into two types of research: i. testing for LLM recognition and classification of certain cxns; and ii. testing for LLM functional understanding

In the first area, Tayyar Madabushi et al. (2020) demonstrated that a variety of base and fine-tuned BERT models are able to distinguish between sentences that instantiate a particular cxn and those that do not. Li et al. (2019) recreate a psycholinguistic test in which models of varying sizes are tested for their ability to group sentences by semantic similarity, where some sentences include the same cxn (e.g., Caused-motion), and others involve different cxns but semantically similar lexical verbs (e.g., sneeze, burp). The authors find that while the smallest language model with 1 million parameters, MiniBERTas (Pérez-Mayos et al., 2021), groups the sentences according to lexical semantics, the largest model with 30 billion parameters, RoBERTa (Liu et al., 2019), groups sentences according to constructional semantics. Of particular relevance to this research, Bonial and Tayyar Madabushi (2024a) develop the initial test set of corpus examples of cxns later released as the CoGS corpus, and test larger models (GPT-3 and 4) for recognition of sentences containing a cxn. The authors find a clear trend demonstrating that the models can recognize substantive cxns with some fixed words (e.g., Much-less), but have increasing difficulty recognizing cxns of increasing schematicity or variability.

[1]The evaluation data subset, prompts, and outputs can be found here: https://github.com/melissatorgbi/from-form-to-function

Overall, the research in the first area demonstrates that while models can recognize and classify some cxns, more abstract cxns present a problem for recognition. Furthermore, studies of recognition and classification do not directly demonstrate whether or not LLMs are proficient users of the cxns of a language; i.e. whether or not the models "understand" the constructional semantics.

Thus, we emphasize the importance of the second area of research, which aims to test LLM functional understanding of cxns in a downstream task. Both Weissweiler et al. (2022) and Zhou et al. (2024) set up evaluations of formal recognition of cxns as well as semantic understanding of the Comparative-correlative and Causal-excess cxns respectively. In both cases, the authors find that models are able to distinguish the cxns, but perform poorly on tests of semantic understanding in the form of downstream questions. Similarly, Scivetti et al. (2025a) finds that smaller-scale LLMs are sensitive to the formal properties of the Let-alone cxn, but reflect no sensitivity to the semantic properties, again in a set of downstream questions testing for understanding.

3 Dataset Development

NLI is a task in which a premise is presented followed by a hypothesis, and the task is to determine if the hypothesis i. must be true given the premise (entailed); ii. may or may not be true given the premise (neutral); iii. cannot be true given the premise (contradicted). We base our task guidelines on the Stanford NLI (SNLI) corpus, which was developed to test semantic representations, as the authors consider understanding entailment and contradiction to be fundamental to natural language understanding (Bowman et al., 2015). NLI has since been adopted as a relatively common test of semantic understanding with several community evaluations (e.g., Marelli et al. (2014); Lee et al. (2024)). As a result, there is widespread availability of NLI data on the web, and it is a relatively common benchmark for LLMs. This also influenced our choice—as there is abundant data on LLM performance for the NLI task, we can distinguish baseline abilities of models on this task from performance on the constructional variant (Sarlin et al., 2020; Raffel et al., 2020; Wei et al., 2022).

We draw our premises from the corpus instances of the 10 cxn types in CoGS (Bonial and Tayyar Madabushi, 2024b); there are about 50 unique corpus instances of each cxn type, giving us about

500 unique premises. The cxns in CoGS vary in *schematicity* (how many words of constructional slots are substantive/fixed or schematic/variable), which enables us to test constructional understanding for fixed-word cxns in which meaning is consistently associated with a particular form, as well as variable-word cxns, in which meaning is associated with templatic syntactic patterns (such as the DITRANSITIVE: The student [noun phrase] handed [verb] the teacher [noun phrase] a book [noun phrase]—i.e., NP V NP NP). A listing of all cxns and example NLI triples from CoGS-NLI is given in Appendix B, Table 4.

One native English speaker (and author of this paper) with an undergraduate degree in Linguistics (but no training in CxG specifically) was given a spreadsheet of the CoGS premises and asked to produce 3 NLI triples—an entailed, neutral, and contradicted hypothesis for each premise; thus, the corpus totals 1500 triples associated with 500 unique premises. We provide guidelines adapted from SNLI definitions of the relations. The NLI author selected triples to create in any order desired to prevent getting stuck on more difficult cases. Depending on the length and complexity of the premise, the hypotheses could take several minutes to process, or come to the author immediately. Overall, the development of the CoGS-NLI corpus was done over the course of a year to prevent fatigue and degraded quality.

We conducted several quality checks of the CoGS-NLI corpus by comparing agreement on the assigned relation of subsets of data (totaling 441 NLI instances) across three annotators (and authors of this paper) against the author's originally assigned relation. Percentage agreement on the initial set of triples ranged from 71-80%, or .55-.70 when measured as Cohen's κ, indicating substantial agreement. All disagreements were revisited, and a second author reworded the hypotheses. Agreement on the reworded hypotheses then reached 89%, or .84 Cohen's κ, indicating very strong agreement equal to the published agreement of individual annotators with respect to gold relation for SNLI.

4 Evaluation Experiments

4.1 Methodology

We provide a performance benchmark by testing models on the same subset of the data that was evaluated for human agreement. Specifically, we hold out 50 instances for in-context learning and use the remaining 391 instances as the test set. The in-context learning examples were randomly chosen where each example contains a single premise with a neutral hypothesis, entailment hypothesis and contradiction hypothesis. The in-context learning examples provided are paired with target phrasal cxns in the test set in order to provide clear examples of the phrasal constructional semantics within the NLI task.

Setting	IC Data	GPT-3.5	GPT-4o
0-shot	None	0.74	0.89
1-shot	CoGS-NLI	0.78	0.91
3-shot	CoGS-NLI	0.83	**0.94**
1-shot	SNLI	0.70	0.89
3-shot	SNLI	0.69	0.90

Table 2: Evaluation results, reported in accuracy, on the CoGS-NLI dataset. "IC Data" refers to the type of data used as in-context examples.

We evaluate GPT-4o-2024-05-13 and GPT-3.5-turbo-0125 models; these models were chosen as representatives of LLM capabilities due to their large size. The temperature is set to 0 to minimize randomness in the model outputs.

We compared results for six different prompt variations, with and without explicitly prompting for Chain of Thought (CoT). We report results for our best-performing prompt, provided in full in Appendix A. We also experimented with 0-shot through 3-shot learning, with two different sources of examples: held-out examples from the CoGS-NLI dataset and selected examples with full-sentence premises from the SNLI corpus. We conduct this comparison in order to determine if the constructional examples boost performance, or if general SNLI examples are sufficient. Note that the CoGS-NLI examples include the target phrasal cxns included in the evaluation, providing clear examples of how these cxns should be interpreted with respect to the NLI task. While the SNLI examples also include cxns of English, they do not include the target phrasal cxns of CoGS.

4.2 Results

Results are reported in Table 2. We see a 5-point boost in performance in the 3-shot setting with constructional examples and achieve a top-end performance of 94% accuracy from GPT-4o. We do not see an equivalent boost in GPT-4o performance in the 3-shot setting with general SNLI examples. The constructional examples are even more helpful for GPT-3.5, where 3-shot outperforms zero-shot

Premise 1	Constance squeezed her way down the platform looking for the first-class carriages.
Hypothesis	Constance waited in line for the first-class carriages.
Relation	Gold: Contradiction; GPT-4o: Neutral
Premise 2	The 23 frantically scrambled to the rear of the sub.
Hypothesis	The 23 were calm at the rear of the sub.
Relation	Gold: Contradiction; GPT-4o: Contradiction

Table 3: Premise 1 exemplifies an error for the most frequently mis-analyzed cxn (Way-manner). Premise 2 (Intransitive-motion) exemplifies a hypothesis with information outside of the constructional semantics that cues the contradiction (i.e. "frantically" vs. "calm".)

by 9 points. Notably, the 3-shot setting with SNLI examples actually *hurts* performance by 5 points.

5 Discussion

Given that the CoGS developers found that models were able to recognize and classify substantive cxns (with fixed words) with much greater accuracy than schematic cxns (with no fixed words and only variable syntactic-semantic slots) (Bonial and Tayyar Madabushi, 2024a), we also assessed if there were performance differences in CoGS-NLI for those cxns classed as fully fixed/substantive, partially fixed, or fully variable/schematic. In contrast to the earlier findings, we do not find a notable difference in performance based upon the schematicity level of the cxn in the premise (see Appendix B Table 5 for performance results separated by phrasal cxn type). However, when we analyze distinct cxns, GPT-4o achieves the highest accuracy on the fully variable Resultative cxn (see Table 1) and the lowest accuracy on the partially variable Way-manner cxn. We provide an error case in (Premise 1) of Table 3. The stronger performance that we see on schematic cxns like the Resultative in the functional understanding NLI task may relate to the frequency of the cxn—LLMs may be better at "understanding" more frequent cxns with greater representation in pretraining data, and the fully schematic argument structure cxns of CoGS are also some of the most frequent cxns of English. We begin to explore this question further in ongoing research (Scivetti et al., 2025b).

The performance of both models on the CoGS-NLI dataset is comparable to performance on SNLI (Ye et al., 2023; OpenAI et al., 2024); thus, we can conclude that including constructional premises does not pose a significant challenge in this task. We note two intertwined limitations in drawing the strong conclusion that these models therefore have a functional understanding of the semantics of the cxn. First, in the NLI task generally, models may rely on spurious features (e.g., the number of tokens) of the premise and hypothesis to solve the task without actually understanding the constructional semantics (Gururangan et al., 2018). Second, the hypotheses may probe other aspects of meaning of the premise outside of the constructional semantics. Premise 2 in Table 3 provides an example where the hypothesis includes the modifier "calm" which contradicts the modifier "frantically" in the premise, but bears no relation to an understanding of the Intransitive-motion constructional semantics. Taken together, these limitations mean that NLI task solvability generally, including that of CoGS-NLI, may be correlated with features outside of a deep semantic understanding.

Thus, on the whole, our results demonstrate that while constructional resources are needed for boosting performance on downstream tasks in which language includes phrasal cxns (note that this is not rare—argument structure cxns are some of the most common phrasal cxns of English), more precise probing evaluations are needed for assessing constructional understanding.

We take steps to craft hypotheses that more precisely target the constructional semantics in Scivetti et al. (2025b). This research also leverages a subset of the CoGS corpus for premises in setting up an NLI evaluation of constructional understanding; however, unlike the present research, we semi-automatically generate the NLI triples by leveraging templates for the neutral, contradicted and entailed hypotheses across all instances of a given cxn type. We note that the templatically generated NLI triples may inadvertently simplify the task by consistently patterning different types of hypotheses. In contrast, CoGS-NLI enables testing understanding through NLI while leveraging free-form, human-authored triples. Together, CoGS-NLI and the templatically-generated NLI dataset of Scivetti et al. (2025b) provide complementary evaluation resources.

6 Conclusions & Future Work

The evaluation of constructional information encoded in LLMs has been approached in several ways. A significant limitation of early methods, such as probing internal model weights, is that discovering the presence of constructional information encoded in weights does not guarantee it is functionally utilized. While prompting allows us to observe how models interact with cxns, metalinguistic tasks that test an LLM's ability to identify sentences as instances of the same cxn measure a classificatory skill, not whether the model can make use of that cxn's meaning to solve a problem.

Our work directly addresses this gap by focusing on the functional application of constructional knowledge. To this end, we created an NLI dataset where premises are carefully selected to feature specific cxns. Our results show that including examples with constructional premises does boost performance, indicating a value to constructional resources like CoGS-NLI. While our results suggest that current models can often correctly solve this task, we recognize that the NLI task does not always isolate the exact semantic meaning carried by the cxn itself. Therefore, in our ongoing and future work we are developing more targeted evaluations to verify that an LLM's reasoning is guided by the precise meaning conveyed by a grammatical cxn (Scivetti et al., 2025b).

Furthermore, any claim about an LLM's *understanding* must contend with recent findings that their performance relies on "context-directed extrapolating from training data priors" (Tayyar Madabushi et al., 2025). Therefore, to genuinely test a model's reasoning capabilities, it is not enough to evaluate it on problems for which priors readily exist in model training data. A systematic evaluation must present novel scenarios with minimal or nonexistent priors, forcing the model to demonstrate *inherent* 'reasoning' or 'understanding' rather than relying on statistical shortcuts. We will continue to leverage CxG as a formalism for targeting language that is creative and novel, but readily understandable by people in order to support such systematic evaluation.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

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.

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.

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.

Lung-Hao Lee, Chen-Ya Chiou, and Tzu-Mi Lin. 2024. Nycu-nlp at semeval-2024 task 2: Aggregating large language models in biomedical natural language inference for clinical trials. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1455–1462.

Hao Li, Wei Lu, Pengjun Xie, and Linlin Li. 2019. Neural Chinese address parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3421–3431, Minneapolis, Minnesota. Association for Computational Linguistics.

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*.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, et al. 2024. Gpt-4o system card.

Laura Pérez-Mayos, Miguel Ballesteros, and Leo Wanner. 2021. How much pretraining data do language models need to learn syntax? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1571–1582.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. SuperGlue: Learning Feature Matching With Graph Neural Networks . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4937–4946, Los Alamitos, CA, USA. IEEE Computer Society.

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. *arXiv preprint arXiv:2506.04408*.

Wesley Scivetti, Melissa Torgbi, Austin Blodgett, Mollie Shichman, Taylor Hudson, Claire Bonial, and Harish Tayyar Madabushi. 2025b. Assessing language comprehension in large language models using construction grammar.

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*, pages 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.

Michael Tomasello. 2009. The usage-based theory of language acquisition. In *The Cambridge handbook of child language*, pages 69–87. Cambridge Univ. Press.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

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.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models.

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.

A Prompts

The following prompt was the best performing variation, achieving .94 accuracy with gpt-4o.

Prompt 1:

"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."

B CoGS-NLI Constructions: Examples & Results by Construction

We provide a listing of all ten cxns included in CoGS-NLI, along with example NLI triples, in Table 4. We then provide performance results across individual cxn types in Table 5.

Construction	Premise, Hypothesis	Relation
Much-less	P: When my dad catches swarms sometimes he doesn't even wear a veil, much less a bee suit. H: When my dad handles swarms, he sometimes wears a veil.	Entailment
Let-alone	P: None of these arguments is notably strong, let alone conclusive. H: All of the given arguments are strong and conclusive.	Contradiction
Way-manner	P: As she felt her way forward, suddenly a knight on horseback galloped past her. H: She was moving forward when a knight on horseback almost ran her over.	Neutral
Comparative-correlative	P: The fewer things we make the more sustainable we are. H: We are more sustainable if we make fewer things.	Entailment
Causative-with	P: The waiter filled her glass with white wine. H: She ordered the white wine in a glass.	Neutral
Conative	P: He nibbled at the filet, then ate ravenously. H: He took big bites of the filet, then slowed down.	Contradiction
Ditransitive	P: They threw me a surprise party. H: They forgot to give me a surprise party.	Contradiction
Caused-motion	P: The MiG-25 fired an AAM at the Predator. H: The MiG-25 tried to hit the Predator.	Neutral
Intransitive-motion	P: Armed troops marched to the substations and turned the power back on. H: The power was turned back on by armed troops that marched to the substations.	Entailment
Resultative	P: He ate himself sick. H: He felt wick from eating.	Entailment

Table 4: One example for each of the ten phrasal cxns included in CoGS-NLI. Note that premises are drawn directly from CoGS, and CoGS-NLI contributes three hypotheses for each premise: one entailed, one contradicted, and one neutral.

Setting	IC Data	Construction	GPT-3.5	GPT-4o
0-shot	None	Let-alone	0.79	0.92
		Way-manner	0.58	0.79
		Comparative- correlative	0.60	0.70
		Causative- with	0.83	0.94
		Conative	0.69	0.88
		Caused- motion	0.78	0.92
		Intransitive- motion	0.78	0.91
		Resultative	0.80	0.94
1-shot	CoGS-NLI	Let-alone	0.83	0.92
		Way-manner	0.64	0.79
		Comparative- correlative	0.57	0.73
		Causative- with	0.85	0.93
		Conative	0.88	0.91
		Caused- motion	0.81	0.97
		Intransitive- motion	0.74	0.94
		Resultative	0.79	0.94
3-shot	CoGS-NLI	Let-alone	0.67	0.92
		Way-manner	0.79	0.85
		Comparative- correlative	0.67	0.87
		Causative- with	0.89	0.94
		Conative	0.90	0.94
		Caused- motion	0.83	0.97
		Intransitive- motion	0.86	0.97
		Resultative	0.80	0.98
1-shot	SNLI	Let-alone	0.79	0.92
		Way-manner	0.52	0.88
		Comparative- correlative	0.60	0.73
		Causative- with	0.78	0.93
		Conative	0.69	0.86
		Caused- motion	0.69	0.92
		Intransitive- motion	0.78	0.93
		Resultative	0.67	0.91
3-shot	SNLI	Let-alone	0.62	0.92
		Way-manner	0.58	0.85
		Comparative- correlative	0.63	0.67
		Causative- with	0.78	0.93
		Conative	0.71	0.91
		Caused- motion	0.67	0.92
		Intransitive- motion	0.72	0.96
		Resultative	0.68	0.92

Table 5: Evaluation results, reported in accuracy, on the CoGS-NLI dataset for the best performing prompt for each individual construction. “IC Data" refers to the type of data used as in-context examples.