Constructions Are So Difficult That Even Large Language Models Get Them Right for the Wrong Reasons

Shijia Zhou[1]**, Leonie Weissweiler**[1,3]**, Taiqi He**[2]**,**
Hinrich Schütze[1,3]**, David R. Mortensen**[2]**, Lori Levin**[2]
[1]LMU Munich, [2]LTI, Carnegie Mellon University, [3]Munich Center for Machine Learning
zhou.shijia@campus.lmu.de, weissweiler@cis.lmu.de

Abstract

In this paper, we make a contribution that can be understood from two perspectives: from an NLP perspective, we introduce a small challenge dataset for NLI with large lexical overlap, which minimises the possibility of models discerning entailment solely based on token distinctions, and show that GPT-4 and Llama 2 fail it with strong bias. We then create further challenging sub-tasks in an effort to explain this failure. From a Computational Linguistics perspective, we identify a group of constructions with three classes of adjectives which cannot be distinguished by surface features. This enables us to probe for LLM's understanding of these constructions in various ways, and we find that they fail in a variety of ways to distinguish between them, suggesting that they don't adequately represent their meaning or capture the lexical properties of phrasal heads.

1. Introduction

In this paper, we test the ability of large language models (LLMs) to identify different meanings in sentences that are superficially similar. The three sentences shown in Figure 1 each contain the intensifier *so*, an adjective that heads an adjective phrase, and a *clausal complement*, a clause that is a dependent in the adjective phrase. In spite of their surface similarity, the three sentences have different semantic properties. In the first sentence, *I was so certain that I saw you*, there are no causal connections between the adjective and the clausal complement. Seeing you did not make me certain, and being certain did not make me see you. In the second sentence, *I was so happy that I was freed*, there is a causal connection. Being freed caused me to be happy. In the third sentence, *It was so big that it fell over*, there is also a causal connection, but it is the reverse, being excessively big caused it to fall over. This is an example of the *Causal Excess Construction* as studied by Kay and Sag (2012) and others.

We examine these sentences from the perspective of Construction Grammar (CxG, Goldberg, 2006; Fillmore et al., 1988; Croft, 2001), an approach to grammar in which arbitrary connections of form and meaning are not limited to the lexicon. In CxG meaning bearing units can contain multiple words, morphemes, and syntactic relations. The purpose of this paper is to test LLMs for their ability to differentiate between the three constructions in Figure 1, despite their surface similarity.

The differences between the constructions can be observed semantically in terms of the presence of causality and the direction of causality as shown in Figure 1. The Causal Excess Construction (CEC)
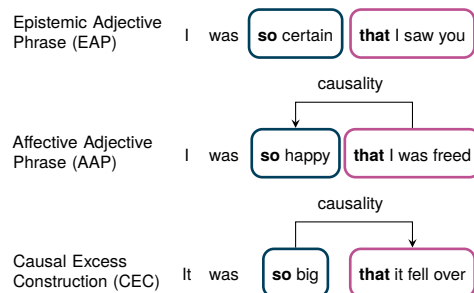

Figure 1: Examples of the clausal complement and causal excess constructions. Constructions involving intensifier, adjective, and clausal complement

also differs from the others in underlying lexico-syntactic properties. This can be seen by removing the word *so* from each sentence. *I was certain that I saw you* and *I was happy that I was freed* are grammatical sentences. In contrast, **It was big that it fell over* is not grammatical. The difference is explained in terms of *licensing*. Every phrase or clause in a sentence must be licensed by a head that selects it. A simple example of licensing is that transitive verbs license direct objects (*The student heard the lecture*) whereas intransitive verbs do not (**The student yawned the lecture*). In the examples we are dealing with here, epistemic and affective adjectives license complement clauses just as transitive verbs license direct objects. Adjectives like *big* on the other hand, do not license clausal complements. In the CEC the clausal complement is licensed by the presence of *so*, or more accurately, by the constellation of elements in the CEC construction (Kay and Sag, 2012). Therefore the clausal complement cannot be present when *so* is not present.

Further underlying syntactic properties of the three constructions make our work more interesting. Our experiments rely heavily on the fact that epistemic and affective adjectives can appear in the CEC as in *I was so certain that I didn't plan for the alternative* and *I was so happy that I cried*. In these sentences, the excess of certainty caused me not to plan for the alternative and the excess of happiness caused me to cry. Epistemic and Affective adjective constructions in the CEC may license two clausal complements, the first licensed by the adjective and the second licensed by the CEC: *I was so certain that I was right that I didn't plan for the alternative*, *I was so happy that I was freed that I cried*.

Importantly, for the experiments we present, any sentence without *so* cannot be CEC. So although *so happy that I cried* can be CEC, *happy that I cried* cannot be CEC—crying is interpreted as the cause of happiness in the latter example.

In this work, we ask how well LLMs differentiate between affective adjective phrases (AAP), epistemic adjective phrases (EAP) and CEC. These offer an ideal testbed for linguistic probing, since they are structurally identical, with no exploitable surface cues. Above-baseline performance would have to stem from (1) knowledge about what each adjective can license, (2) knowledge about whether causality is typically associated with the adjective, and (3) understanding the direction of causality (adjective causes clause or clause causes adjective). Specifically, we investigate GPT-3.5 (OpenAI, 2021), GPT-4 (OpenAI, 2023), OpenAI's `ada-002` and Llama 2 (Touvron et al., 2023) with various questions, for each using both a prompt and a probing classifier.[1] We observe that LLMs exhibit limited capability to effectively discriminate between these constructions and display a strong bias towards CEC, meaning LLMs tend to judge sentences containing *so... that...* as causal and the adjective being the reason for the clausal complement. Generally, Llama 2 demonstrates superior performance compared to both GPT-3.5 and GPT-4.

2. Related Work

Our work is situated in the framework of CxG (Croft, 2001; Fillmore et al., 1988; Goldberg, 1995, 2006; Hoffmann and Trousdale, 2013). In this work, we define a construction as a pairing of form and meaning. We are considering AAP, EAP and CEC to be three different constructions. They differ in meaning (affective, epistemic, and excessive states) and in form (a clause that is licensed by a lexical head and a clause that is licensed by the intensifier *so*).

Recent studies use CxG to probe the inner workings of LLMs (Weissweiler et al., 2022; Chronis et al., 2023; Mahowald, 2023), and make general observations about the compatibility of the theory of CxG with the recent successes of LLMs (Goldberg, To appear; Weissweiler et al., 2023). Most related to our work, McCoy et al. (2019) construct a challenging dataset for Natural Language Inference (NLI) that uses pairs of sentences with high lexical overlap. They show that the surface similarity of words almost always fools BERT into assuming that one sentence entails the other. Recent work (Si et al., 2022; Basmov et al., 2023) suggests that the performance of recent LLMs on the McCoy et al. (2019) data has improved, though it is still far from perfect. We are therefore motivated to provide another challenging dataset to this line of inquiry.

3. Dataset

Our data collection process takes advantage of Universal Dependency (Nivre et al., 2020) annotations, which we use for prefiltering before manual annotation. We use SPIKE (Shlain et al., 2020) to access a parsed Wikipedia corpus and a parsed Amazon Reviews corpus. We establish that the parse trees of all CEC, AAP and EAP constructions have an edge labelled `ccomp` from the adjective to the head verb of the complement clause.

We extract all sentences matching *so... that...* pattern with a structural search string `The bookshelf was so adj:[tag]tall that it verb:[tag]toppled` in SPIKE, and group the sentences by adjective. Manually, where possible, we extract a sentence pair where one is CEC and one either AAP or EAP resulting in 111 such pairs. For the adjectives that cannot license a clausal complement, we extract 101 sentences, each with a different adjective. We call this class OCE (only Causal Excess). Our set of CEC sentences excludes any OCE adjectives.

In total, we collect 323 sentences with 212 different adjectives.[2] An example of each sentence type is given in the first row of Table 1.

4. Natural Language Inference

As shown in Table 2 (Prompts 1-1 to 1-4), we design four prompt variants to test whether LLMs can detect significant changes in entailment when *so* is removed from CEC sentences (DS type in Table 1). For example, *I was so happy that I cried* does not automatically entail *I was happy that I cried*, whereas *I was so happy that I was freed* entails *I was happy that I was freed*.

---

[1] See Appendix Table 7 for hyperparameters

[2] Data and code are available at https://github.com/shijiazh/Constructions-Are-So-Difficult

Type	Transformation	OCE	CEC	AAP	EAP
O	Original	It was so big that it fell over.	I was so happy that I cried.	I was so happy that I was freed.	I was so certain that I saw you.
DS	− 'so'	It was {} big that it fell over .	I was {} happy that I cried.	I was {} happy that I was freed.	I was {} certain that I saw you.
DT	− 'that'	It was so big {} it fell over .	I was so happy {} I cried.	I was so happy {} I was freed.	I was so certain {} I saw you.
DST	− 'so' & 'that'	It was {} big {} it fell over .	I was {} happy {} I cried.	I was {} happy {} I was freed.	I was {} certain {} I saw you.
AN	+ 'not'	It was **not** so big that it fell over .	I was **not** so happy that I cried.	I was **not** so happy that I was freed.	I was **not** so certain that I saw you.
P1	main clause	It was so big.	It was so happy.	I was so happy.	I was so certain.
P2	sub. clause	It fell over.	I cried.	I was freed.	I saw you.
Y-N	yes–no question	Did it fall over?	Did I cry?	Was I freed?	Did I see you?

Table 1: Transformations of the dataset with examples

The results shown in Figure 2 are striking: For CEC sentences, models achieve accuracy below 10%, while demonstrating 90% accuracy for AAP and EAP. It indicates a pronounced bias towards *entailment* in the models, which replicates a behaviour shown for BERT: "assume that a premise entails all hypotheses constructed from words in the premise" by McCoy et al. (2019). The following subsections investigate two hypotheses about why LLMs overestimate entailments: (i) LLMs are unable to identify causality in sentences (§4.2); (ii) LLMs do not recognize the change in the direction of causality (§4.3).
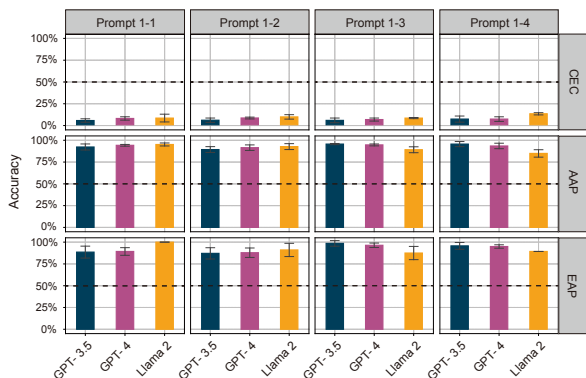

Figure 2: Performance of CEC, AAP, EAP on the central NLI task. Corresponding gold labels are no entailment, no entailment, entailment, and entailment. All models have a strong bias to answer entailment. For OCE verbs, DS (delete so) is ungrammatical, therefore, we did not prompt on the OCE type.

4.1. Experimental Setup

We conduct our experiments with two methods. The first approach involves the development of both implicit and explicit prompts. In the second approach, we extract the last-layer embeddings from LLMs and then apply perceptron-based classification to these embeddings, to assess how well the CEC, OCE, AAP and EAP categories are internally represented in the models.

We employ a perceptron classifier to test the final layer embeddings of Llama 2, and `ada-002` sentence embeddings. For Llama 2, we use the mean over token embeddings as a sentence em-

bedding. We hypothesise that the contextual representation of the adjective itself will encode the presence or direction of causality, and therefore also test its embedding separately. When the two input classes are imbalanced, we upsample the smaller class. We group sentences containing the same adjectives together and train the perceptron using cross-validation over adjectives, meaning that the adjective itself is no longer an exploitable feature. A Bag-of-Words (BoW) model serves as the baseline.

4.2. Identifying Causality

Prompting As depicted in Table 3, we devise two prompts to assess the models' capability to identify causal relationships. The first simply asks about a causal relationship between the main and the subordinate clause, while the second additionally provides the pre-segmented parts. In the EAP category, all models have accuracy below 20%, suggesting a predisposition to infer causality in sentences containing *so... that...* even when the adjective is epistemic. Llama 2 displays a stronger bias, attributing causality to over 90% of sentences in all categories. Combining the previous observation that Llama 2 tends to categorise every sentence containing *so* and *that* as grammatically correct, along with its sensitivity to the absence of *so* in CEC, this suggests a limited grasp of semantic nuances and an overreliance on simple lexical cues. GPT models both struggle about equally, with less than 50% accuracy in AAP instances. Even more perplexing, EAP sentences are classified as causal at a significantly higher rate than CEC and OCE.
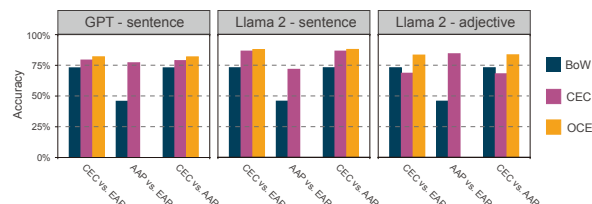

Figure 3: Accuracy of perceptrons trained with different embeddings across three tasks. In all sub-tasks involving CEC structures, we attempt to replace CEC with OCE. OCE adjectives are mutually exclusive with those in EAP and AAP.

No.		Template
1	1	premise: **O** \n hypothesis: **DS** \n Classify as entailment, no entailment, or contradiction.
	2	premise: **O** \n hypothesis: **DST** \n Classify as entailment, no entailment, or contradiction.
	3	**O** Can we infer that "**DS**"? \n Answer with yes, no or uncertain.
	4	**O** Can we infer that "**DST**"? \n Answer with yes, no or uncertain.
3	1	**O** \n Is there a causal relationship between the main clause and the subordinate clause? \n Answer with yes, no or uncertain.
	2	**O** \n Part1: **P1** \n Part2: **P2** \n Is there a causal relationship between part 1 and part 2? \n Answer with yes, no or uncertain.
4	1	premise: **O** \n hypothesis: **P2** \n Classify as entailment, no entailment, or contradiction.
	2	**AN** \n **Y-N** \n Answer with yes, no or uncertain.
	3	**O** \n Part1: **P1** \n Part2: **P2** \n Can we infer that Part1 is the cause of Part2? \n Answer with yes, no or uncertain.
	4	**O** \n Part1: **P1** \n Part2: **P2** \n Can we infer that Part2 is the cause of Part1? \n Answer with yes, no or uncertain.
	5	**O** \n This entails one of two options. \n 1) **P1** because **P2** 2) **P2** because **P1** \n Answer with the correct number.

Table 2: Prompt templates of all tasks. To test the stability of model responses, we design variants of each prompt, removing only *so* from premise as hypothesis or removing both *so* and *that*.

No.	Model	OCE	CEC	AAP	EAP	Gold Lab.
3-1	GPT-3.5	67.33	60.90	41.68	**18.57**	
	GPT-4	63.37	58.74	41.20	15.00	Y|Y|Y|N
	Llama 2	**95.05**	**98.65**	**95.18**	08.93	
3-2	GPT-3.5	54.46	64.14	49.15	06.43	
	GPT-4	57.03	65.95	46.02	04.28	Y|Y|Y|N
	Llama 2	**95.54**	**99.10**	**92.78**	08.93	

Table 3: Accuracy of the task of identifying causality with different prompts

Probing Classifier The classifiers for CEC vs EAP and AAP vs EAP serve to assess the models' representation of causality. As illustrated in Figure 3, on CEC vs EAP, perceptrons trained on sentence embeddings beat the baseline while those trained on adjective embeddings do not, suggesting that causality is encoded, but not necessarily in the adjective. Interestingly, the adjective perceptron beats the baseline on the AAP vs EAP test, even though the sets of adjectives are mutually exclusive and we perform cross-validation over them, meaning that the only source of information left would be a deeper commonality between them. This suggests that the models may have learned common features for affective and epistemic adjectives, respectively. Furthermore, the result also demonstrates that the distinction between EAP and CEC is more pronounced in the model's perspective compared to the distinction between EAP and AAP, especially for Llama 2 sentence embeddings.

4.3. Direction of Causality

Prompting The negator *not* before *so* influences the truth value of the subclause for CEC but has no influence in AAP. For instance, *He was not so big that he fell* (CEC) does not imply that he fell, while *He was not so happy that he went.* (AAP) suggests that he went. Asking the model to distinguish between these is equivalent to distinguishing the direction of causality.

Therefore, we devise an explicit NLI prompt 4-1 (Webson and Pavlick, 2022) and an implicit prompt 4-2 (AN + Y-N) to explore these effects. Accu-

Type	Model	OCE	CEC	AAP	Gold Label
4-1	GPT-3.5	28.71	29.19	62.41	
	GPT-4	26.93	29.01	**62.65**	N|N|Y
	Llama 2	**39.60**	**49.10**	53.62	
4-2	GPT-3.5	4.55	2.52	60.24	
	GPT-4	4.95	2.16	**60.72**	N|N|Y
	Llama 2	**16.34**	**18.47**	40.97	
4-3	GPT-3.5	74.46	82.88	13.25	
	GPT-4	77.03	83.96	8.91	Y|Y|N
	Llama 2	**87.13**	**93.69**	46.99	
4-4	GPT-3.5	50.69	42.70	44.09	
	GPT-4	48.31	40.18	47.95	N|N|Y
	Llama 2	**77.23**	**71.17**	**81.92**	
4-5	GPT-3.5	**61.19**	**60.72**	77.59	
	GPT-4	60.80	54.78	**79.27**	2)|2)|1)
	Llama 2	51.98	45.49	78.31	

Table 4: Accuracy of direction of causality task with different prompts. Y: yes/entailment, N: no/contradiction.

rate answers depend on the models' precise understanding of the causal direction. We also introduce prompt 4-3, where P1 and P2 are provided, and the question is whether P1 is the cause of P2. Prompt 4-4 maintains the same structure but now inquires whether P2 is the cause of P1. Prompt 4-5 is structured as a multiple-choice question between the two directions.

As can be seen in Table 4, results for prompts 4-1 and 4-2 suggest that they bias all models towards answering *yes* for any sentence. By contrast, prompts 4-3 and 4-4 show a clearer picture, still biased, but Llama 2 scores clearly correlate with the gold label. By comparing these two sets of prompts, we can discern that Llama 2's responses are grounded in an analysis of P1 and P2, rather than simply providing uniform *yes* or *no* answers. Interestingly, prompt 4-5 elicit better performance from the GPT models in contrast, suggesting that it might be more suited to the multiple-choice answer format. We conclude that all models have some representation of the direction of causality, but it is far from perfect.

Probing Classifier The classifier for CEC vs AAP serve to assess the models' capability to differentiate the direction of causality. The results in Figure 3 are similar to those for identifying causality in Sec-

No.	Model	OCE	CEC	AAP	EAP	Gold Label
O	GPT-3.5	89.31	92.43	80.96	73.57	
	GPT-4	89.31	92.97	81.93	73.57	G|G|G|G
	Llama 2	**100.00**	**100.00**	**100.00**	**100.00**	
DS	GPT-3.5	36.83	**84.69**	87.23	79.28	
	GPT-4	36.43	84.32	**88.92**	78.57	B|G|G|G
	Llama 2	**39.58**	76.30	80.62	**83.34**	
DT	GPT-3.5	80.40	89.55	72.29	60.00	
	GPT-4	80.40	88.83	70.60	56.43	G|G|G|G
	Llama 2	**100.00**	**100.00**	**100.00**	**100.00**	
DST	GPT-3.5	**57.83**	**67.39**	67.95	65.71	
	GPT-4	**57.83**	66.67	**68.92**	65.72	B|G|G|G
	Llama 2	49.77	49.97	60.50	**76.46**	
AN	GPT-3.5	83.76	90.63	74.46	69.29	
	GPT-4	84.36	90.45	76.14	74.29	G|G|G|G
	Llama 2	**100.00**	**100.00**	**100.00**	**100.00**	

Table 5: Accuracy of the grammaticality task. Bold font indicates the models with the highest accuracy for a type and transformation. G: good, B: bad.

tion 4.2, with the notable exception of the OCE test set, which is easiest for the adjective classifiers with no obvious explanation.

Figure 3 displays that on CEC vs AAP, similar to CEC vs EAP, models trained with sentence embeddings outperform the baseline, while those trained with adjective embeddings slightly lag behind. Additionally, the perceptron attains the highest accuracy on the OCE test set.

As adjectives in OCE can only appear in CEC, while adjectives in CEC can also occur in AAP or EAP, this can be interpreted as OCE's adjective embeddings being more easily identified as belonging to the CEC structure than those of CEC.

5. Grammatical Acceptability

We perform experiments on grammatical acceptability to see whether LLMs are sensitive to the difference between embedded clauses that are lexically licensed by an adjective and embedded clauses that are licensed by the CEC construction. Following Mahowald (2023), we prompt LLMs for grammaticality judgements by first presenting eight pairs of sentences from the CoLA corpus (Warstadt et al., 2019), which consists of minimal pairs of grammatical and ungrammatical sentences extracted from linguistics literature. Then the target sentence is inserted, and the model is asked to generate one token, *good* or *bad*.

We test the original sentences (O) and four transformations DS, DT, DST, and AN, as mentioned in Table 1. Deleting *that* (DT) or adding *not* (AN) will not affect the grammaticality of CEC, EAP and AAP, whereas removing *so* (DS) from OCE makes them ungrammatical.

As can be seen in Table 5, compared to CEC OCE is more likely to be rated as *bad* by both GPT-3.5 and GPT-4, regardless of the transformation.

It demonstrates that GPT models indeed detect the distinction between OCE and CEC especially regarding their reliance on the grammaticality with *so*. Notably, for GPT models, removing *that* from AAP and EAP sentences results in more *good*, whereas removing *that* from DS sentences tends to yield more *bad* ratings even though it has no effect on grammaticality.

Llama 2's answers deviate from that of GPT models. It rates all O, DT, and AN sentences as *good*, which is exactly the gold label, signifying its robust inclination to not only consider sentences featuring *so... that...* as acceptable, but also acknowledge the possibility of omitting *that* in such contexts. However, its performance on DS and DST is far from perfect.

Additionally, GPT4 performs better on the DS variants of AAP.[3]

6. Conclusion

Overall, our most striking result remains that no LLM performed adequately on our NLI task, and that this result is not sufficiently explained by the mediocre but better-than-baseline performance on the sub-tasks. Llama 2 performed better in those than the GPT models, but generally, prompting results are often consistently below random and probing classifier results only slightly above baseline. Interestingly, GPT-4 does not perform significantly better than GPT-3.5 at any task.

Both in the central NLI task, and the sub-tasks, all LLMs show bias to offer positive answers. Llama 2 demonstrates a more comprehensive understanding of the grammatical structures within CEC, an enhanced ability to identify causality within sentences, and a greater proficiency in ascertaining the direction of causality compared to GPT models. These findings align with our initial observations in the central NLI task.

We exclude the following from the current work: (1) Extraposition from clausal subject (*That I left was so bad*/*It was so bad that I left*) (2) CEC meanings with other intensifiers such as *enough* and non-finite clauses (*big enough to fall over*) (3) CEC headed by nouns (*so many people that the police came*). We have also not investigated the CEC in sentences other than copula sentences, or when the CEC adjective is part of a noun phrase (*I met many people so short that they couldn't reach the counter*).

---

[3]A probing classifier for this prompting task would not be well-defined.

Bibliographical References

Victoria Basmov, Yoav Goldberg, and Reut Tsarfaty. 2023. ChatGPT and simple linguistic inferences: Blind spots and blinds.

Gabriella Chronis, Kyle Mahowald, and Katrin Erk. 2023. A method for studying semantic construal in grammatical constructions with interpretable contextual embedding spaces. *arXiv preprint arXiv:2305.18598*.

William Croft. 2001. *Radical Construction Grammar: Syntactic theory in typological perspective*. Oxford University Press, Oxford, UK.

Charles J. Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: the case of 'let alone'. *Language*, 64(3):501–538.

Adele E. Goldberg. 1995. *Constructions: a construction grammar approach to argument structure*. University of Chicago Press, Chicago.

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

Adele E. Goldberg. To appear. A chat about constructionist approaches and LLMs. *Constructions and Frames*.

Thomas Hoffmann and Graeme Trousdale. 2013. *The Oxford handbook of construction grammar*. Oxford University Press.

Paul Kay and Ivan A Sag. 2012. Cleaning up the big mess: Discontinuous dependencies and complex determiners. In *Sign-based construction grammar*, chapter 5, pages 229–256. Citeseer.

Kyle Mahowald. 2023. A discerning several thousand judgments: GPT-3 rates the article + adjective + numeral + noun construction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 265–273, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

OpenAI. 2021. GPT-3.5: The Third-and-a-Half-Generation Language Model. https://api.openai.com/docs/gpt-3.5.

OpenAI. 2023. GPT-4 Technical Report.

Micah Shlain, Hillel Taub-Tabib, Shoval Sadde, and Yoav Goldberg. 2020. Syntactic search by example. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 17–23, Online. Association for Computational Linguistics.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, and Lijuan Wang. 2022. Prompting gpt-3 to be reliable. In *The Eleventh International Conference on Learning Representations*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

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

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of

their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Leonie Weissweiler, Taiqi He, Naoki Otani, David R. Mortensen, Lori Levin, and Hinrich Schütze. 2023. Construction grammar provides unique insight into neural language models. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, pages 85–95, Washington, D.C. Association for Computational Linguistics.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The better your syntax, the better your semantics? probing pretrained language models for the English comparative correlative. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Appendix

Now we are going to say which sentences are acceptable (i.e., grammatical) and which are not.

Sentence: Flosa has often seen Marn.
Answer: good

Sentence: Chardon sees often Kuru.
Answer: bad

Sentence: Bob walk.
Answer: bad

Sentence: Malevolent floral candy is delicious.
Answer: good

Sentence: The bone chewed the dog.
Answer: good

Sentence: The bone dog the chewed.
Answer: bad

Sentence: I wonder you ate how much.
Answer: bad

Sentence: The fragrant orangutan sings loudest at Easter.
Answer: good

Sentence: [TEST SENTENCE GOES HERE]
Answer:

Table 6: Few-shot CoLA prompts template created by Mahowald (2023). We tested 5 types of sentence: O, DS, DT, DST and AN.

	GPT-3.5/GPT-4	Llama 2
temperature	1	0.7
top_p	1	0.95
top_k	-	40
max_tokens	null	512
frequency_penalty	0	-
presence_penalty	0	-

Table 7: Hyperparameters for GPT-3.5, GPT-4, OpenAI's ada-002 and Llama 2. For each prompt, we repeat the mean over six runs of the experiment.

Adjective	Type	Sentence
frustrated	CEC	In one XFM show , he became so frustrated that he left the room before Karl finished the segment .
	AAP	I am so frustrated that a $ 500 purchase brought such short lived joy .
proud	CEC	Mandhata had dominated the whole planet and he became so proud that he wanted to rule heaven also .
	AAP	My dad was so proud that his son made " aliyah " .
afraid	CEC	One man was so afraid that he camped in the middle of his flock , hoping to evade patrolling cowboys .
	EAP	He was so afraid that rival loyalist inmates wished to kill him inside the prison .
optimistic	CEC	Like Napoleon , Hitler was so optimistic that he falsely believed he 'd make it to Moscow before Winter .
	EAP	I am so optimistic that I made the best choice .
abrupt	OCE	The growth was so abrupt that a village sprang .
beautiful	OCE	The palace was so beautiful that the king of Mengwi heard of Tan Cin Jin .

Table 8: Examples from the collected database. CEC represents causal excess construction, where the adjective is interpreted as the cause of the complement. AAP stands for affective adjective phrases, which usually trigger an inference that the complement caused the feeling expressed by the adjective. EAP stands for epistemic adjective phrases, which lexically liscence non-causal complement.