Paper Sharings

刘柱

2025.11.06

Overviews

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

Constructions are Revealed in Word Distributions

Joshua Rozner[1], **Leonie Weissweiler**[2], **Kyle Mahowald**[3], **Cory Shain**[1]

[1]Stanford University [2]Uppsala University [3]The University of Texas at Austin
{rozner, cashain}@stanford.edu
leonie.weissweiler@lingfil.uu.se kyle@utexas.edu

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

LREC-COLING 2024

(Joint International Conference on Computational Linguistics, Language Resources and Evaluation)

short paper

Introduction

- To evaluate LLMs on understanding construction (构式) meaning
  - Different meaning in sentences that are superficially similar
  - Epistemic Adjective Phrase (EAP)
  - Affective Adjective Phrase (AAP)
  - Causal Excess Construction (CEC)

Introduction

- Same surface structure
  - so + adj. + that + clause
- Different meanings

	Causality	**Adjective**
EAP	No	Epistemic
AAP	From Clause to Adj	Affective
CEC	From Adj to Clause	Excess/Affective/ Epistemic

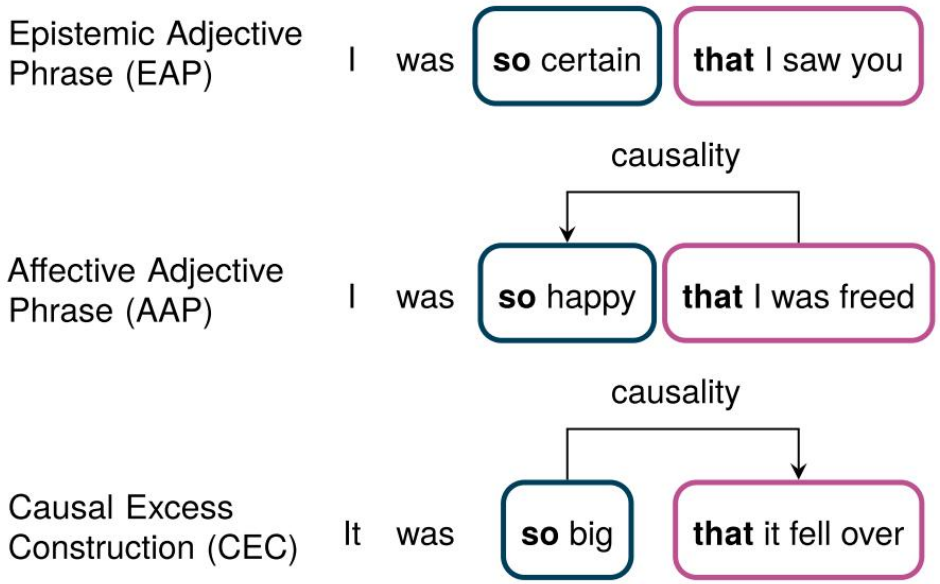

Figure 1: Examples of the clausal complement and causal excess constructions. Constructions involving intensifier, adjective, and clausal complement

Note: Whether *so* can be removed can be explained by ***licensing***

- EAP&AAP: clauses are licensed to adj.
- CEC: Clauses are licensed to *so* (So, so can't be removed)

I was so happy that I cried

I was so certain that I didn't plan for the alternative

Introduction

- Same surface structure
  - so + adj. + that + clause
- Different deep structures

	Remove *so*?	**Clausal Compose**
EAP	YES	EAP + CEC
AAP	YES	AAP + CEC
CEC	NO	NO

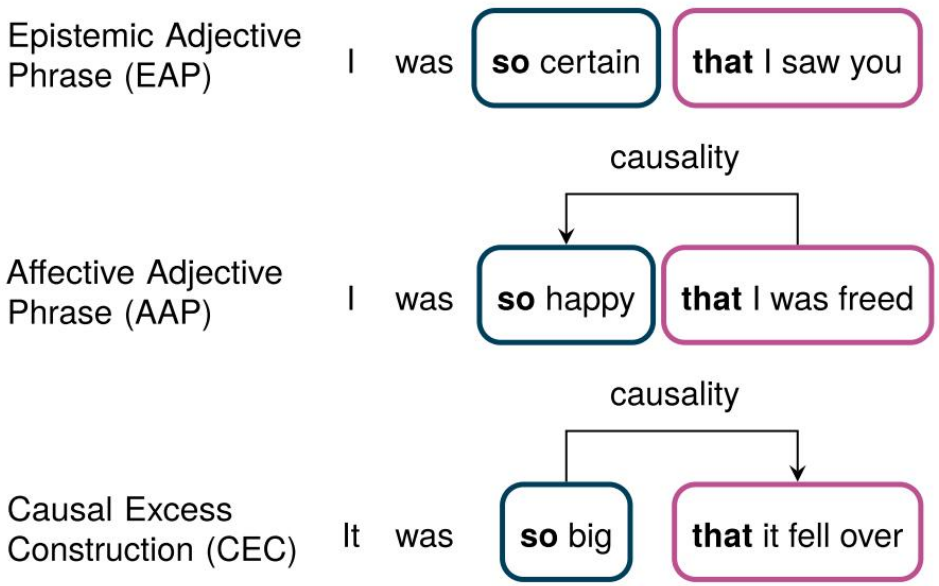

Figure 1: Examples of the clausal complement and causal excess constructions. Constructions involving intensifier, adjective, and clausal complement

For example:

- 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.

Contributions

- To evaluate how well LLMs differentiate between three types
  - what each adjective can license
  - whether causality is typically associated with the adjective
  - the direction of causality
- Methods
  - Probing
  - Prompting
- Conclusions
  - Limited capability to discriminate
  - A strong bias towards CEC (so that tends to regard as causal / adj is the reason for clausal complement)
  - Llama2 is better than GPT-3.5/4

Related Work

- Construction Grammar (CxG)
  - Form-meaning pair (conventional meaning) rather than syntax+lexicon
    - Lexical Construction (idiom, e.g., kick the bucket)
    - Syntactic Construction (What's this fly doing in my soup? What's [NP] doing [PP]?)
    - Ditransitive Construction (She gave him a book [Subj] + [V] + [Obj1] + [Obj2]) ...
  - Probing
    - Dataset [McCoy et al. (2019)] with highly overlapped sentences
      - Recent work (Si et al., 2022; Basmov et al., 2023) suggests LLM is still far from perfect
    - Probing CxG on LLMs (Gold- berg, To appear; Weissweiler et al., 2023)

Dataset

- Wikipedia corpus and Amazon Review corpus
  - Universal Dependency by SPIKE
  - all three constructions have an edge labeled *ccomp* from the adjective to the head verb of the complement clause.
  - Matching [So...that...] and group by adjective
  - Manual labeling three types

  - 111 <CEC, AAP/EAP> pair with the same adj., e.g., *happy*
  - 101 with adj which cannot license a clausal complement, e.g., *big*
  - Overall 323 sentences with 212 different adjectives

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.

Exp.1 - NLI

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.

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.

- NLI: Can premise infer hypothesis? Entailment, Neutral, contradiction
- Note: AAP and EAP are entailment, while CEC are neutral
  - I was so happy that I cried vs. I was happy that I cried

Result

- Models perform bad in CEC (most are biased into entailment)
- Two hypothesis
  - LLMs are unable to identify causality in sentences
  - LLMs do not recognize the change in the direction of causality.

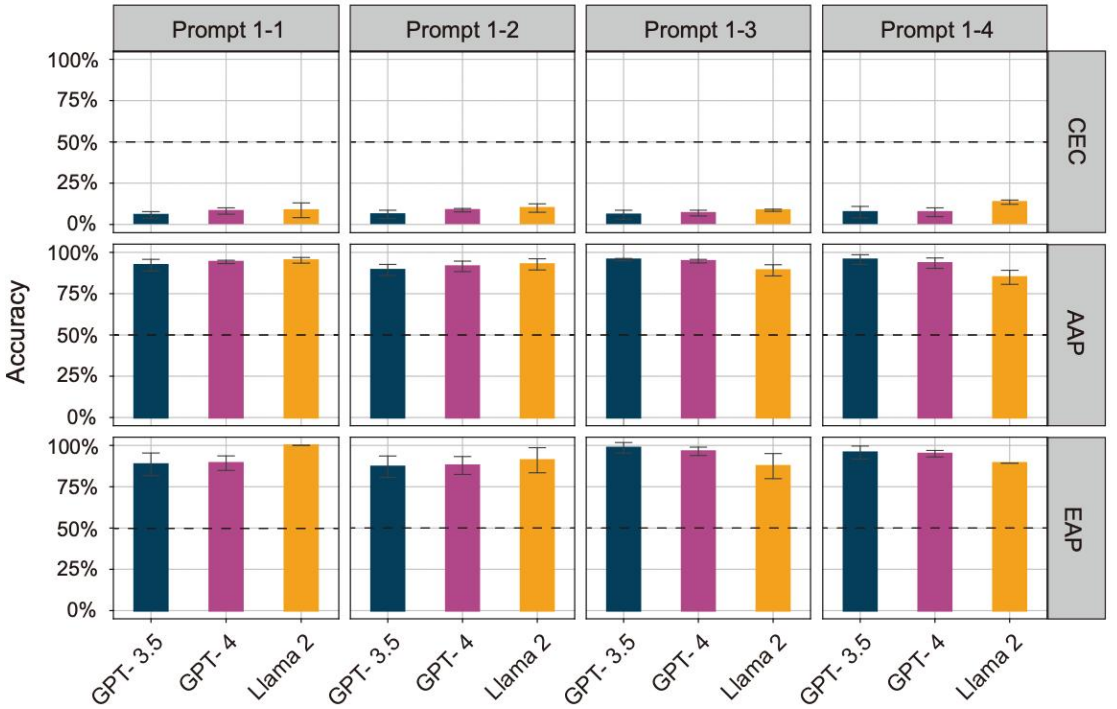

Figure 2: Performance of CEC, AAP, EAP on the central NLI task. Corresponding gold labels are no entailment, no entailment, entailment, and entailment. All models have a strong bias to answer entailment. For OCE verbs, DS (delete so) is ungrammatical, therefore, we did not prompt on the OCE type.

Exp 2 - Identifying Causality

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.
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.
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.			

- Prompting
  - EAP has a much lower accuracy (most are wrongly labelled causality) -> Causality bias
  - Llama 2 has a more biased trend than GPT

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

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

Exp 2 - Identifying Causality

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.

- Probing
  - Sentence vs. adjective embeddings
  - Baseline model: BoW
  - binary classification for any two types
- Results
  - Sentence embeddings can beat baselines in all classification, while only adjective is not enough (except in AAP vs. EAP or OCE vs. others)

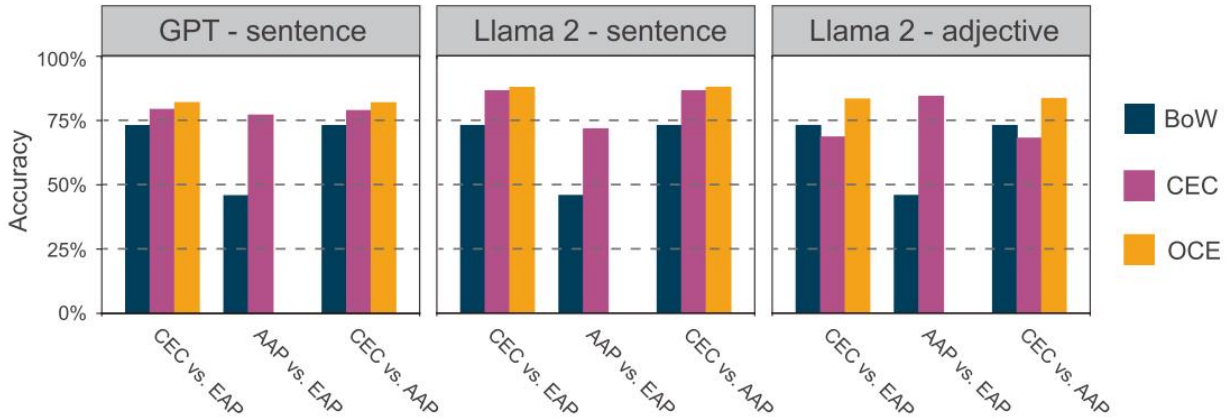

Figure 3: Accuracy of perceptrons trained with different embeddings across three tasks. In all sub-tasks involving CEC structures, we attempt to replace CEC with OCE. OCE adjectives are mutually exclusive with those in EAP and AAP.

Exp 3 - Direction of Causality

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.
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.
Y-N	yes–no question	Did it fall over?	Did I cry?	Was I freed?	Did I see you?

	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.
4	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** \n 2) **P2** because **P1** \n Answer with the correct number.

- Prompting
  - For OCE and CEC, a negation does not imply P2, while in AAP it does. (different modifier domain [not [so... that ...]] vs. [[no so...] that])

Exp 3 - Direction of

- Prompts 4-1 & 4-2:
  - → All models show a strong “Yes” bias regardless of sentence content.
- Prompts 4-3 & 4-4:
  - → Bias remains, but Llama 2’s scores correlate better with gold labels.
  - → Suggests Llama 2 analyzes P1–P2 relations rather than answering uniformly.
- Prompt 4-5:
  - → GPT models perform better, indicating this prompt suits multiple-choice formats.
- Overall conclusion:
  - → All models capture some sense of causal direction,
  - but their understanding is still imperfect.

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

Exp 3 - Direction of Causality

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.

- Probing
  - CEC vs. AAP

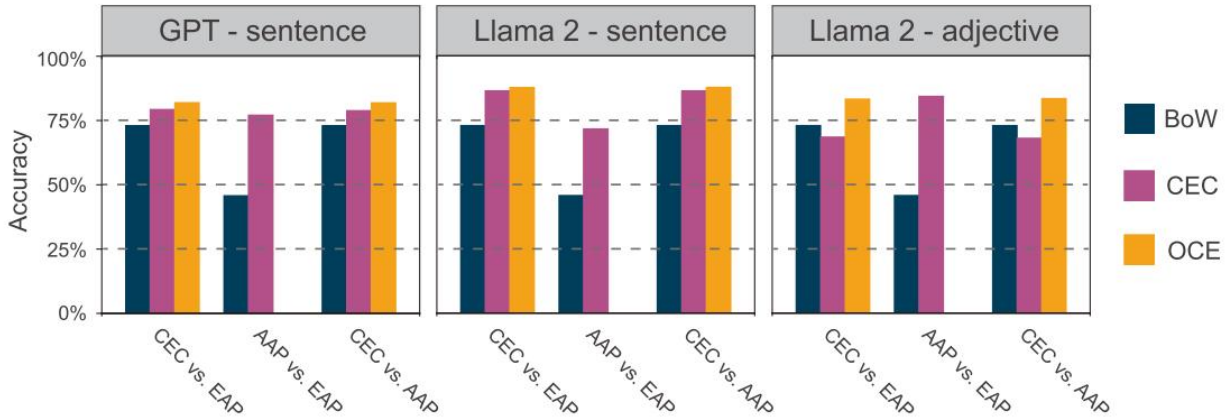

Figure 3: Accuracy of perceptrons trained with different embeddings across three tasks. In all sub-tasks involving CEC structures, we attempt to replace CEC with OCE. OCE adjectives are mutually exclusive with those in EAP and AAP.

Grammatical Acceptability

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.		

- Prompting
  - Llama tends to regard a ungrammatical sentence as Good than GPT
  - For GPT, CEC and OCE can be distinguished in DS

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

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.

Conclusion

- No LLM performed adequately on NLI task
- Llama2 generally performed better than GPT (GPT-4 does not significantly perform better than GPT-3.5)
- Prompting results are often consistently below random and probing classifier results only slightly above baseline.
- Bias: positive answer (grammatical, entailment, casual ...)

Constructions are Revealed in Word Distributions

Joshua Rozner[1], **Leonie Weissweiler**[2], **Kyle Mahowald**[3], **Cory Shain**[1]

[1]Stanford University [2]Uppsala University [3]The University of Texas at Austin
{rozner, cashain}@stanford.edu
leonie.weissweiler@lingfil.uu.se kyle@utexas.edu

EMNLP 2025 Long

Introduction

- Construction grammar (CxG)
  - Form-meaning pair
  - acquired through experience with language (distribuitional learning)
  - kick the bucket vs. *I was so happy that I cried and I was so happy that I was freed*. (similar surface form)
- Distrbution over strings
  - Unavailable for children learning
  - Corpus-based methods, like collostructional analysis (Stefanowitsch and Gries, 2003, 2005; Hilpert, 2014) -> only static, not causal
  - PLMs can approximate the distribution as an *interactive* simulator

Introduction

- Related work on PLMs as a tool to analyze CxG
  - simulations of the learner vs. simulations of the distribution
  - model's behaviour (prompting, probing) vs. Context-sensitive distribution
  - current evidence on the learnability of constructions by PLMs is mixed
- This paper
  - Collostructional analysis (statistical affinities)
  - + intervention methods (effect changes by altering input)
  - -> Pertubed masking to develop affinity methods
  - Inspired by idioms (semi-fixed MWD): meaning activated by constraining

Contributions

- **Extension** of prior work (perturbed masking) as *affinity* methods that reveal constructions as patterns of statistical interaction (§3)
- **Resolution** of previously reported challenges using the methods (§4)
- **Generalization** of the methods to a wide range of other construction types (§5, §6)
- **Qualitative analysis** to characterize method behavior and inform the limits of purely distributional approaches (§7)

Method

- Model
  - Bidirectional PLMs, RoBERTa
- Global affinity
  - interaction between a single word and the entire context
- Local affinity
  - pairwise interactions between words

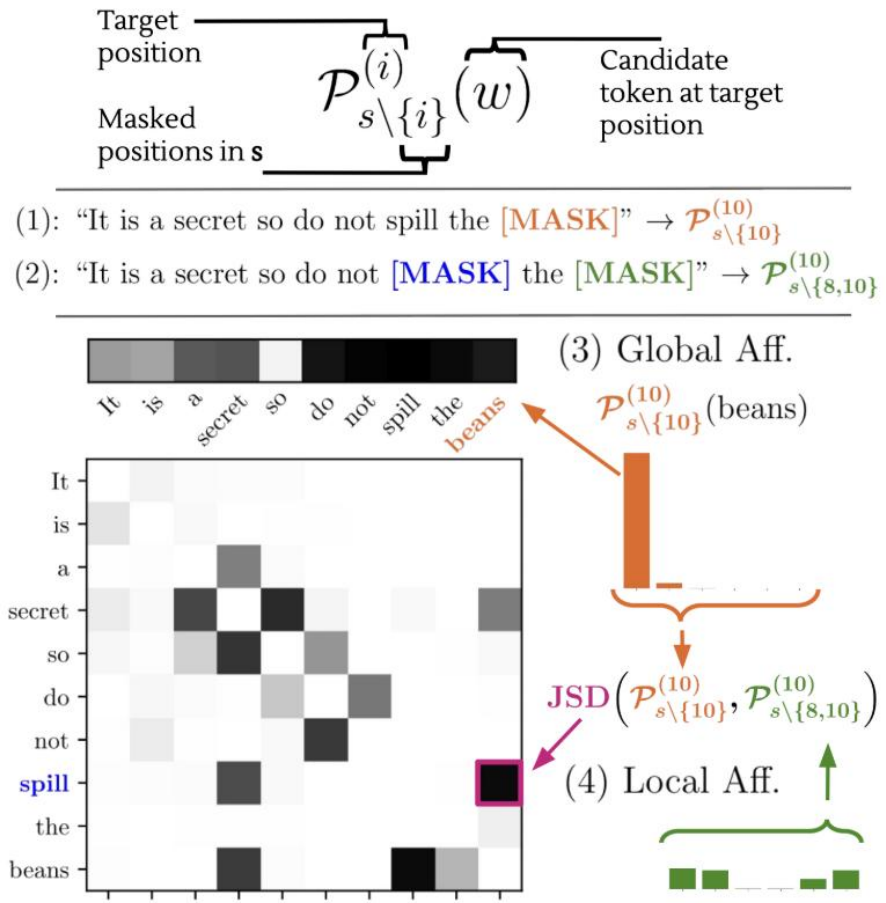

Figure 1: In s = "*It is a secret so do not spill the* ***beans***", masking *beans* (1) gives a constrained distribution, where $\mathcal{P}^{(10)}_{s\setminus\{10\}}(beans)$ is high, so *beans* has high *global affinity* (3). By also masking, e.g., *spill* (2), we get $\mathcal{P}^{(10)}_{s\setminus\{8,10\}}$, compute JSD, and find the words that constrain *beans* and thus have high *local affinity* (4).

a Challenging Case

- CEC vs. EAP and AAP
- The last paper reports the incapability of LLMs

Epistemic Adjective Phrase (EAP)

I was so certain that I saw you.

Affective Adjective Phrase (AAP)

I was so happy that I was freed.

Causal Excess Construction (CEC)

It was so big that it fell over.

$$[\ [\text{NP}]\ [\text{V}]\ \text{so}\ [\text{ADJ}]\ ]_1\ \text{that}\ [\text{S}]_2$$

This paper found...

- Models distinguish CEC from EAP and AAP
  - Global affinity of “so”
    - Accuracy 98.2% (thr: 0.78)
    - A clear margin from the distributions

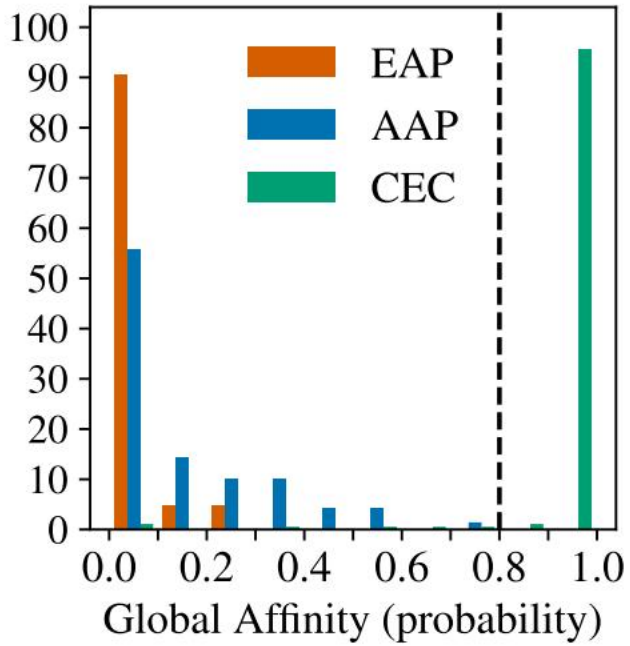

Figure 2: Percent of examples with *so* having given global affinity. The CEC is seen to be well-separated from the EAP and AAP.

This paper found...

- Models distinguish CEC from EAP and AAP
  - Global affinity of “so”
    - Accuracy 98.2% (thr: 0.78)
    - A clear margin from the distributions
    - Better than the last paper ...

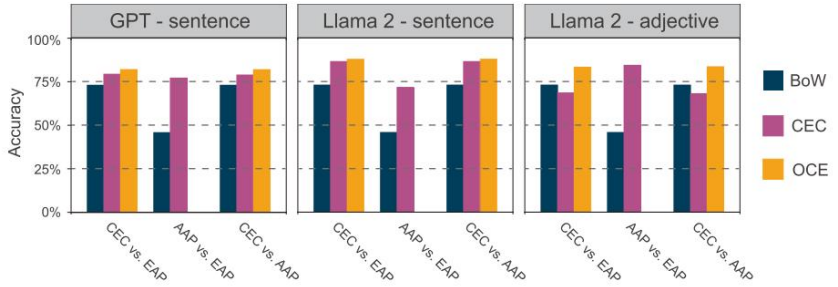

Figure 3: Accuracy of perceptrons trained with different embeddings across three tasks. In all sub-tasks involving CEC structures, we attempt to replace CEC with OCE. OCE adjectives are mutually exclusive with those in EAP and AAP.

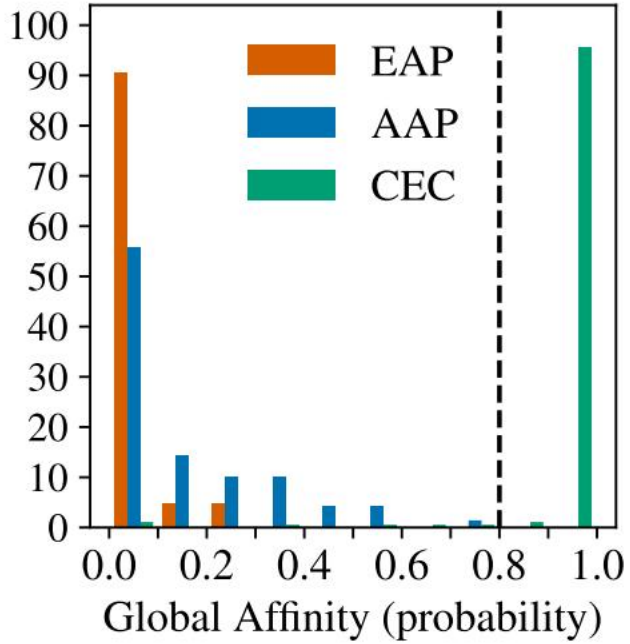

Figure 2: Percent of examples with *so* having given global affinity. The CEC is seen to be well-separated from the EAP and AAP.

This paper found...

- Models distinguish CEC from EAP and AAP
  - Global affinity of “so”
    - Accuracy 98.2% (thr: 0.78)
    - A clear margin from the distributions
    - 11 misclassified examples are actually mislabeled or invalid
      - “This was so funny that I had to buy another copy and read it to my better half,” was originally labeled AAP (should be CEC)

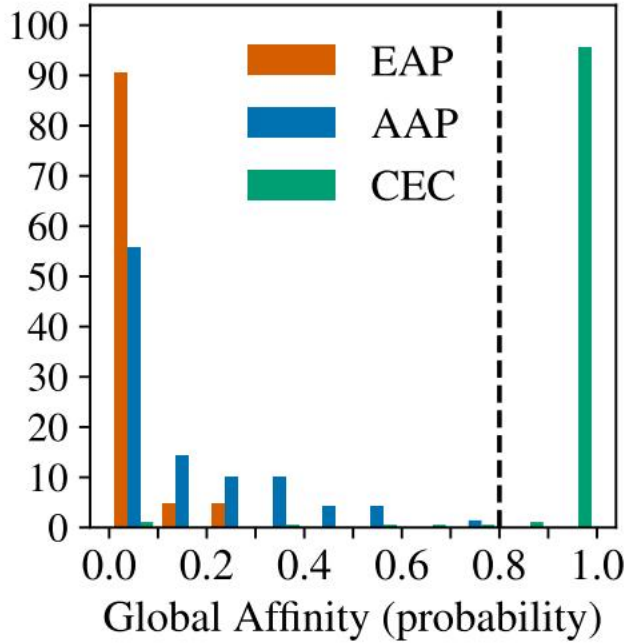

Figure 2: Percent of examples with *so* having given global affinity. The CEC is seen to be well-separated from the EAP and AAP.

This paper found...

- Models capture causal relations in the CEC
  - Local affinity
    - so-that pair
    - that1 -> excited (AAP)
    - that2 -> so (CEC)

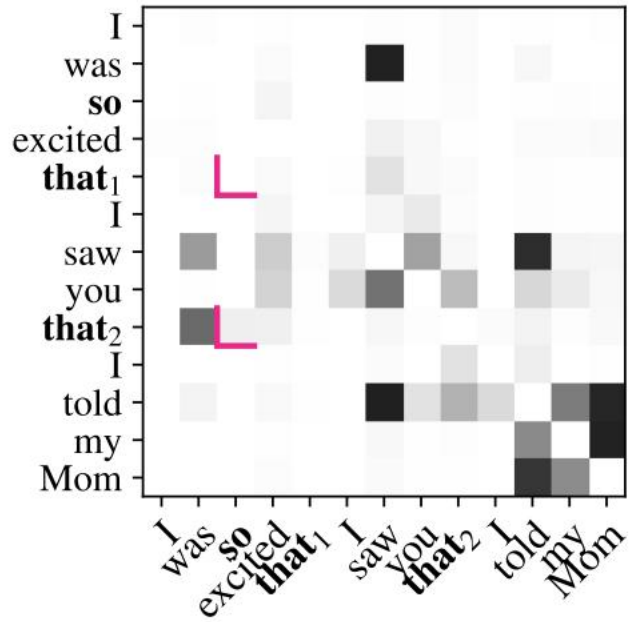

Figure 3: Local Affinity plot. A *column* shows how much the distribution for that word is affected by other words in the context. *so* is more affected by $that_2$ (CEC) than by $that_1$.

This paper found...

- Affinity patterns distinguish the EAP and AAP

  - EAP and AAP are harder to distinguish just by so
  - The previous work has a relative high accuracy (77.1, 71.7, 84.3 for GPT, Llama, and Llama-adj)
    - But they can't distinguish whether the adj itself or the construction to help the classification
  - This work use 5 highest pair-wise affinities to do the classification

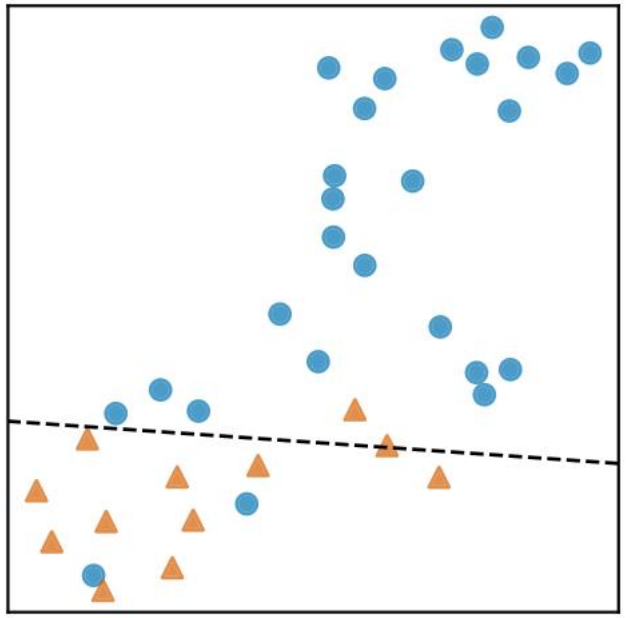

Figure 4: UMAP projection (EAP: orange, AAP: blue) using 5 pair-wise affinities. Separability with SVM (dashed) suggests interaction patterns differ for EAP and AAP.

Other Substantive Constructions

- Data set
  - CoGS (Bonial and Tayyar Madabushi 2024)
  - 50 examples for each of 10 types
  - 6 are substantive

Causative-with: She loaded the truck *with* books.
Comparative correlative: *The* more *the* merrier. (In our analysis the two *the* words are considered as a single class.)
Conative: He kicked *at* the ball.
Let-alone: None of these arguments is particularly strong, *let alone* conclusive.
Much-less: He has not been put on trial, *much less* found guilty.
Way-manner: We made our *way* home.

尝试/意向：他朝球踢

Other Substantive Constructions

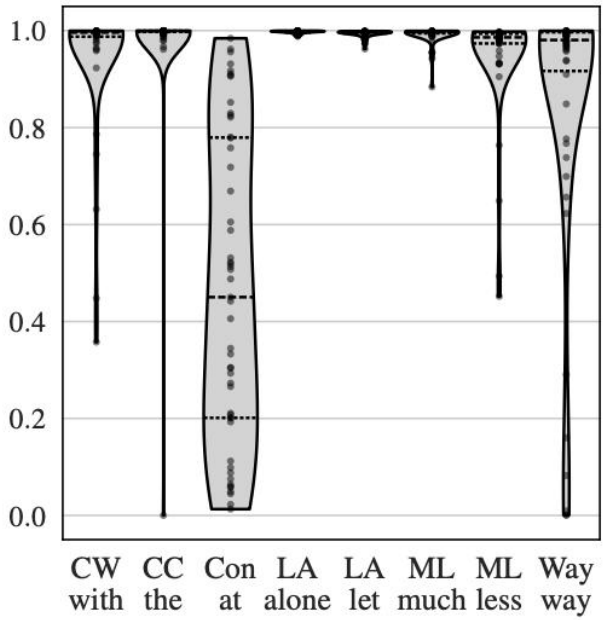

Figure 5: Global affinity for CoGS. CW: Causative-with; CC: Comp-correlative; Con: Conative; LA: Let-alone; ML: Much-less; Way: Way-manner

Causative-with: She loaded the truck *with* books.
Comparative correlative: *The* more *the* merrier. (In our analysis the two *the* words are considered as a single class.)
Conative: He kicked *at* the ball.
Let-alone: None of these arguments is particularly strong, *let alone* conclusive.
Much-less: He has not been put on trial, *much less* found guilty.
Way-manner: We made our *way* home.

Global affinity distinguishes fixed slots in numerous constructions

- except for conative, since at can be replaced with out, at, over

Other Substantive Constructions

- Global affinity helps distinguish literal from figurative usages
  - kick the bucket
  - spill the beans
- 114k words that are part of a PIE in 45k sentences (10k literal, 34k figurative
- affinity as a classification probability for figurative usage
- Result of AUC: 0.71

Generalizing to Schematic Constructions

- Models generalize the NPN's covarying noun-noun slots
- NPN construction
  - day after day
  - 400 by sampling 100 nouns and prompt GPT-4 to produce NPN sentences with *by, after, upon* and *to*
- Result shows a high global affinity for global affinity for nouns in NPN

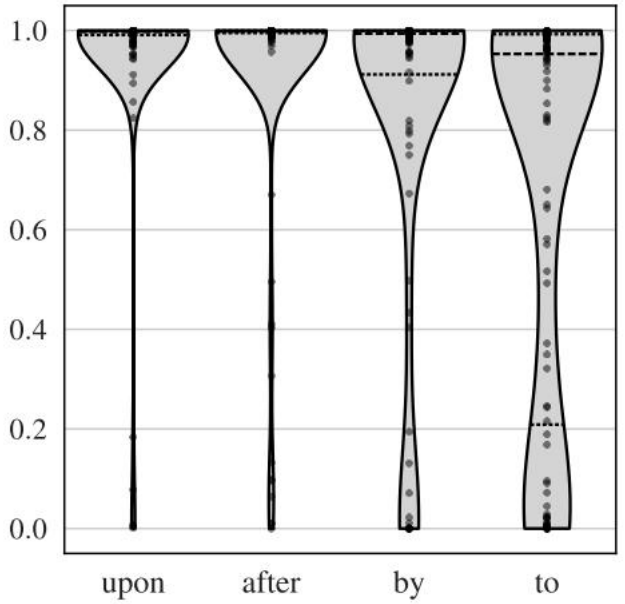

Figure 6: Global affinity for nouns in the NPN construction, grouped by preposition, for sentences with acceptability ≥ 4.

Generalizing to Schematic Constructions

- Models generalize the comparative correlative's category constraint
- Tested if model encodes the comparative adjective/adverb slot in the Comparative Construction ("The Xer, the Yer").

- Masked the comparative word and checked model predictions.(global affinity)

- Result: 98/99 cases correctly filled with comparatives → model captures the abstract syntactic constraint of the CC nearly perfectly.

Limits of Distributional Analysis

- High affinity doesn't always mean a construction: Some strong connections come from context or meaning, not the construction itself.

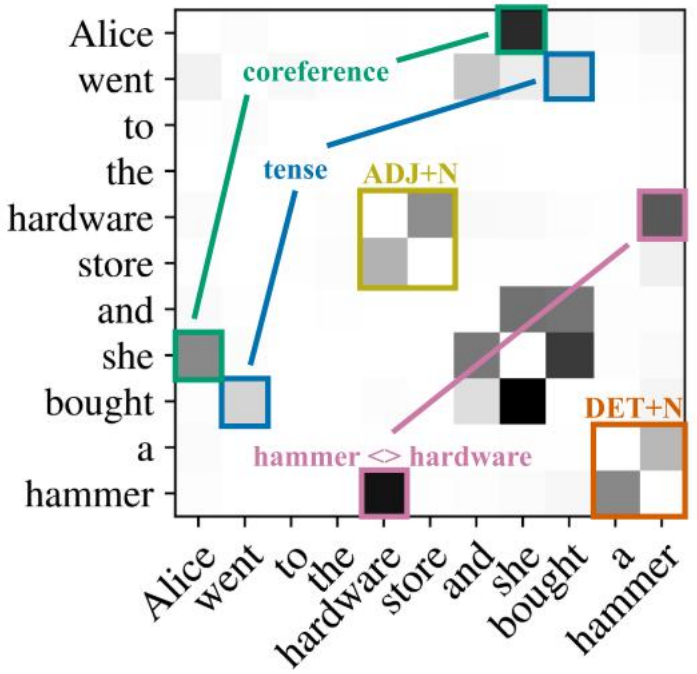

Figure 7: The local affinity matrix encodes diverse types of interactions, including both constructional and non-constructional interactions.

Limits of Distributional Analysis

- High affinity doesn’t always mean a construction: Some strong connections come from context or meaning, not the construction itself.

- Context matters: Even fixed expressions (like “Green Day”) may appear weak if the surrounding words don’t trigger them.

- Masking issues: Hiding words to test affinities can miss constructions, so affinities aren’t a perfect measure.

Conclusion

- PLM as a simulator of distribution, as the hypothesis of CxG is distributional learning hypothesis
- constructional information is reliably reflected in the causal interactions between words and their surrounding context.
- Collostructional analysis vs. counterfactual inquiry
  - 替换法（已存在的现象哪些关联性强） vs. 拓展法（干预之后能不能说）
- Probing, Prompting and Causual

清華大學
Tsinghua University

Q & A

谢谢

Note