11.09.2024

The University of Texas at Austin

CONSTRUCTIONS ALL THE WAY DOWN: RETHINKING COMPOSITIONALITY IN LLMS

LEONIE WEISSWEILER
The University of Texas at Austin

What problem do we have to solve when we solve language tasks?

How can the model know what a new sentence means?

And how can it make new sentences mean something?

Is it as simple as combining words and rules?

(Rule-)Compositionality

“The meaning of an expression is a function of the meanings of its parts and the way they are syntactically combined” Partee (1984: 153)

→ A compositional model trained on the meanings of novel words: having seen *dax*, *flug*, and *flug twice*, you should be able to interpret the meaning of *dax twice* (Lake & Baroni 2018)

LLMs struggle with Rule-Compositionality

"[LLMs] STILL haven't learned that if A is the same as B, then B is the same as A" --Yann LeCun

Criticism from Computer Science, calls for different architecture

X is more Y than Z entails *Z is less Y than X* for all possible X, Y, and Z, irrespective of their specific values (Dasgupta et al. 2020)

Criticism from Linguistics/Cognitive Science, calls for more cognitive plausibility

But is language rule-compositional?

___ ***think twice*** = hesitate (≠ think a second time)

work twice ___ = work twice as hard/much (≠ work a second time)

___, ***going twice*** = last chance to buy at auction (≠ going twice somewhere)

From top dozen VERB twice examples on COCA

Constructions

- Conventional pairings of form and function
- Provide the compositional generalisations
- Various levels of schematicity:

words	*banana*
morphemes	*V-ing*
idioms	*give the devil his due*
partially filled idioms	*jog <someone’s> memory*
grammatical constructions	*Subj V Obj1 Obj2*

Construction-Compositionality

Constructions constrain how we combine units, and give us the meaning of the combinations

sneeze has no arguments and causes no movement, but:

sneeze + **caused motion construction**:

She sneezed the foam off the cappuccino.

Required Skills for Language Models

- Segmenting language into constructions
- Associating constructions with the correct meaning
- Representing distributional information about construction usage
- Combining the meaning of constructions compositionally

Can they do it?

USING DISTRIBUTIONAL INFORMATION ABOUT CONSTRUCTIONS

Counting the Bugs in ChatGPT's Wugs: A Multilingual Investigation into the Morphological Capabilities of a Large Language Model

Leonie Weissweiler*, Valentin Hofmann*, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer, David R. Mortensen

Constructions, LLMs, and Morphology

- Even children already know a lot about constructions and in which specific circumstances to use them
- Morphology is an ideal test bed for this: large groups of words (e.g. verbs) all undergoing the same change (e.g. past tense) with different outcomes (e.g. suffixes)

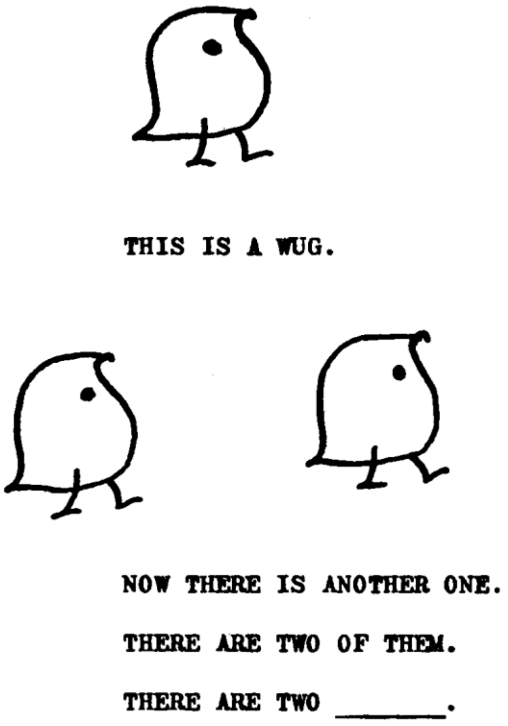
We know that **talk → talked**, but **teach → taught** and **swim → swam**

Generalising over Morphemes

We don‘t just memorise, we also form abstractions, find patterns, and generalise

This is a man who knows how to SPOW. He is SPOWING.
He did the same thing yesterday. What did he do yesterday?
Yesterday he ________.

spowed ?   spewed ?



Do LLMs form the same generalisations?

Sabine Doebel
@sabine_doebel

ChatGPT passes the wug test! (And with the enthusiasm of of a child who knows the answer!) But only inconsistently generalizes to other novel words @JeanBerkoG

Here is a wug. Now there are two. There are two ______.

Wugs!

Here is a dax. Now there are two. There are two ______.

Here is a dax. Now there are two. There are two dax.

Remi van Trijp @RemivanTrijp · Mar 2, 2023
Wugs have almost certainly been in its training data, so it is already known by it. You have to invent a new set of promps to test it properly

1 1 160

Sabine Doebel @sabine_doebel · Mar 2, 2023
That's what I figured. I used some made up ones and it was inconsistent.

1 100

MMitchell
@mmitchell_ai

Um, has anyone succeeded in replicating something like the "wug" study with ChatGPT? I've been trying for the past 20 minutes and I already have a headache.
I think it might fail from the very start.

MA I'm going to try to teach you some new words, and then I'd like to ask you some questions.

Sure, I'm always happy to learn new words and answer your questions. Please go ahead!

MA Okay, the first new word is "glack".

I'm sorry, but "glack" doesn't seem to be a real word in the English language. Could you provide some context or a definition for it?

Experimental Setup

Long prompt

zero-shot

one-shot

few-shot

Fill in the blank with the correct past tense of the verb "bebit". Answer with one word.
They bebit all the time. In fact, they ___ just yesterday!
test : tested
teach : taught
build : built
sing : sang
hit : hit
___ :

The verb "bebit" is a made-up word, so its past tense doesn't follow standard English conjugation rules. Based on the pattern in your examples, it appears that you're asking for an irregular past tense form.

For consistency with the examples like "hit: hit," where the base and past tense forms are the same, the past tense of "bebit" could simply be "bebit."

So the answer would be:

bebit: bebit

Short prompt

zero-shot

one-shot

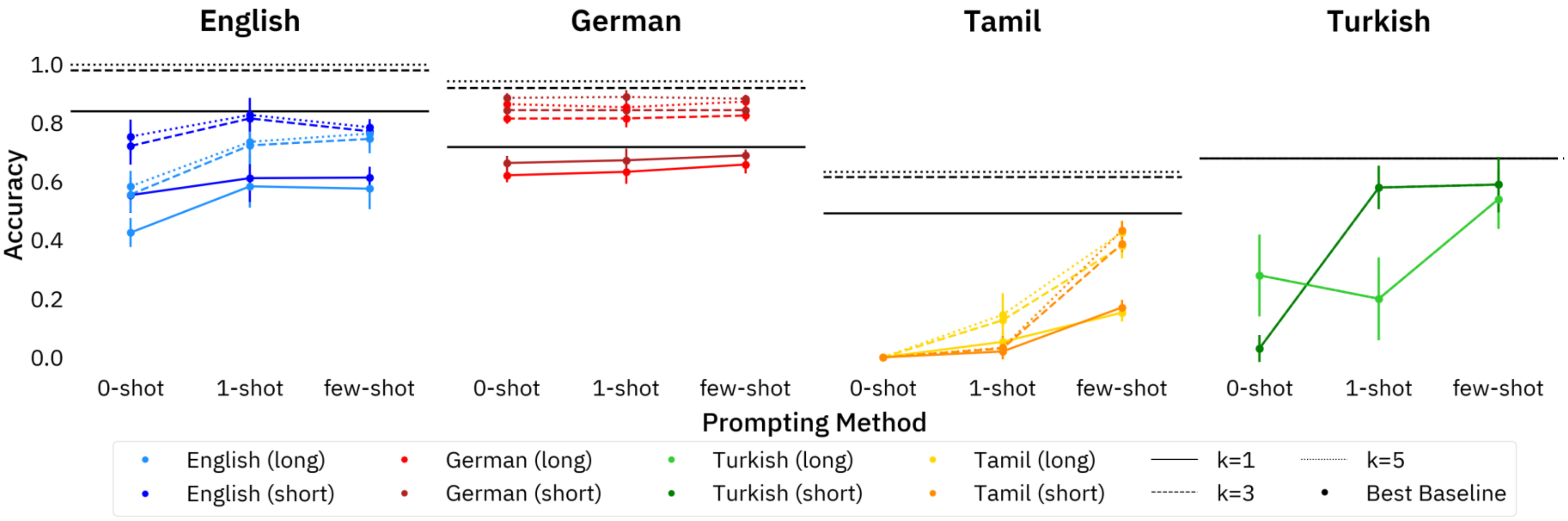
few-shot

Form the correct past tense of the verb "bebit". Answer with one word.
test: tested
teach: taught
build: built
sing: sang
hit: hit

The correct past tense of "bebit" is "bebet."

Results

bebitted	x 13
bebited	x 10
bebit	x 2
bebited	x 1
bebat	x 1



Overgeneralisation

- German few-shot results by noun ending
- Main source of error is overgeneralisation to the most productive morphemes, +*en* and +*s*

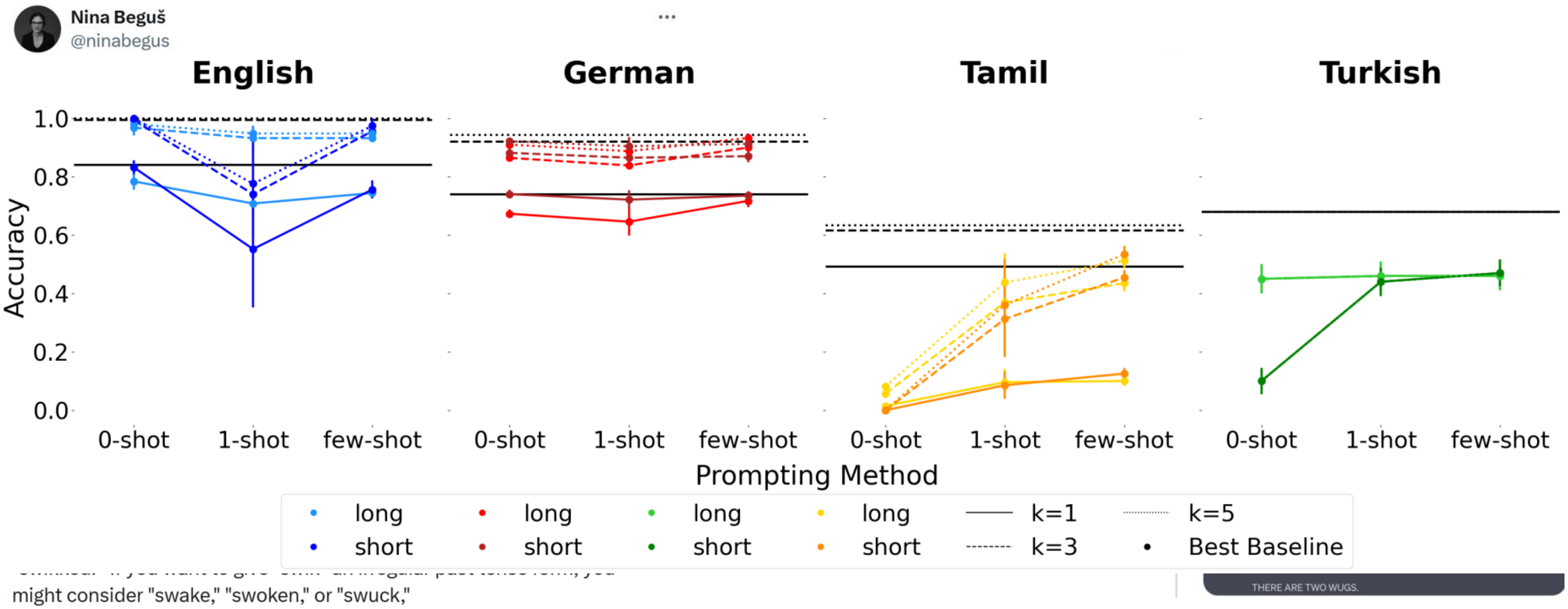
→ Amplification of training data biases

Gold \ Predicted	No change	+ e	+ en	+ er	+ s	Vowel change	Vowel change + e	Vowel change + er	Real Word	Unknown
No change	317	6	0	0	10	0	0	0	9	8
+ e	47	262	77	5	79	0	13	5	22	19
+ en	3	20	64	0	18	4	0	0	0	11
+ s	25	15	5	0	109	0	1	0	1	4
Vowel change + e	0	13	0	0	1	0	22	3	0	1
Unknown	7	3	0	0	13	0	0	0	0	25

Caveat: Instruction Following

- Prompts without original wug-context performed better
- Classic wug-question produces no parseable results
- For Tamil any reply required prompt-tuning and arguing
- Even for English: *bebit* → *drained, drank, drank, bebitted, bebit, bebit, bet, bebitted, bebit, bdrank*

Upon Popular Request: Repeat with GPT-4



ASSOCIATING THE RIGHT MEANING WITH THE RIGHT CONSTRUCTION

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

Shijia Zhou, Leonie Weissweiler, Taiqi He, Hinrich Schütze, David R. Mortensen, Lori Levin

How can we test LLMs for the edge cases?

Everything is a construction

→ most constructions are easy to learn and clearly working well

Some constructions are difficult:

- Non-rule-compositional meaning
- Require world-knowledge
- Require complex distributional information
- No lexical giveaways

A Naturally Occurring Semantic Minimal Triple

causality

I was **so** happy **that** I cried → I was happy, therefore I cried

Causal Excess Construction

causality

I was **so** happy **that** I was freed → I was freed, therefore I was happy

Affective Adjective Phrase

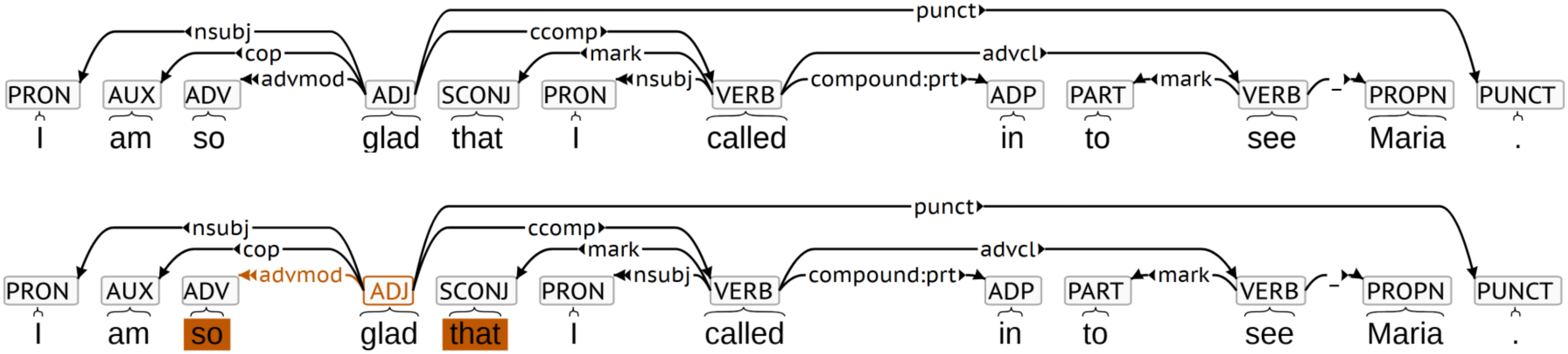
I was **so** certain **that** I saw you → ?

Epistemic Adjective Phrase

Data Sources

Annotation by hand needed!

But: finding sentences is easy because of ***so*** and ***that***

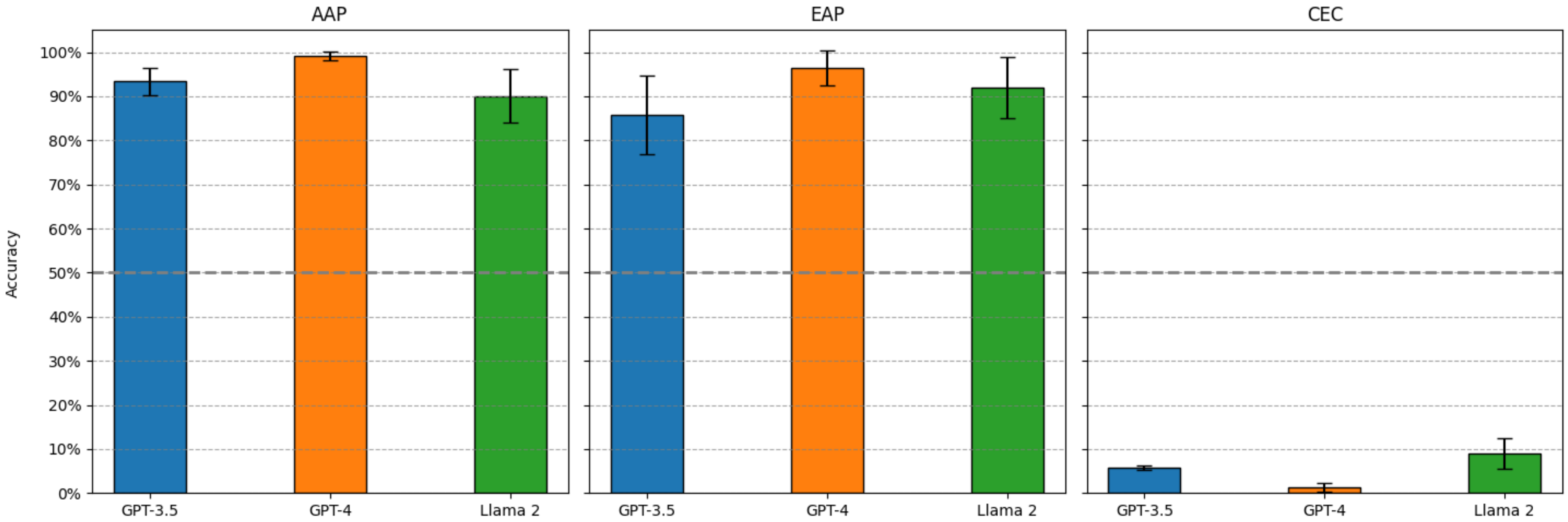


An Inference Task With Minimal Changes

- ✅ I was **so** certain that I saw you → I was certain that I saw you
- ✅ I was **so** happy that I was freed → I was happy that I was freed
- ❌ I was **so** happy that I cried → I was happy that I cried
- ❌ I was **so** confident that I won the title → I was confident that I won the title

Results

Entailment ✅ Entailment ✅ Entailment ❌



TESTING AN ARGUMENT-STRUCTURE CONSTRUCTION WITH NO FIXED SLOTS

Hybrid Human-LLM Corpus Construction and LLM Evaluation for Rare Linguistic Phenomena

Leonie Weissweiler, Abdullatif Köksal, Hinrich Schütze

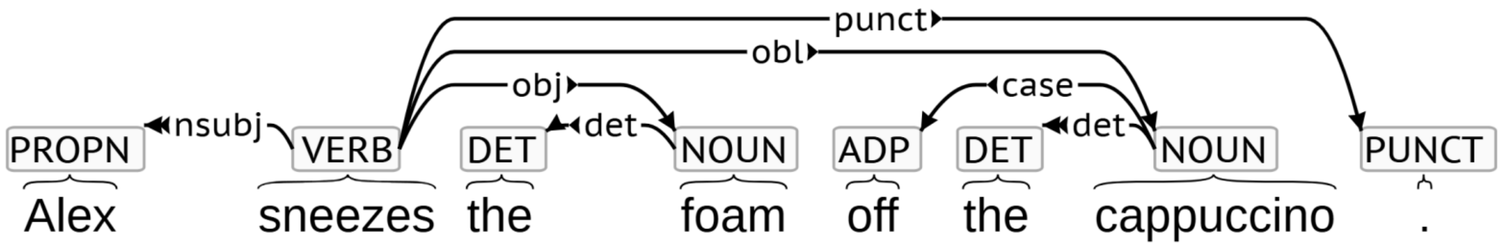
Construction properties that could be difficult for LLMS

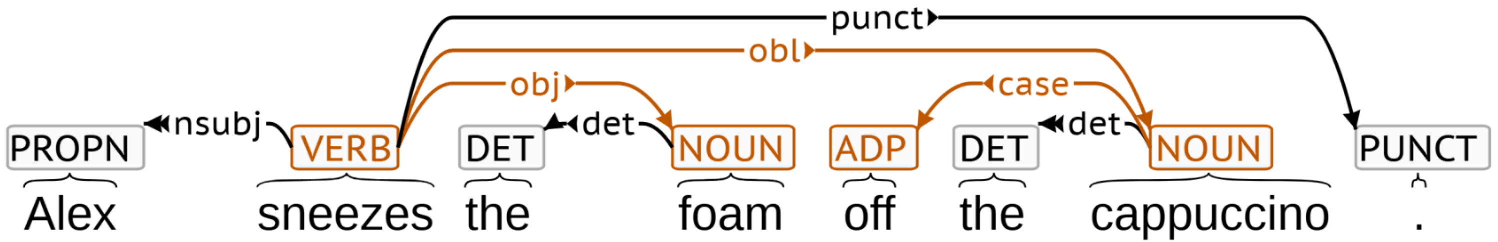
- Sparse
- Violate some „rule“ of the language
- Interpretation requires world knowledge

The Caused-Motion Construction

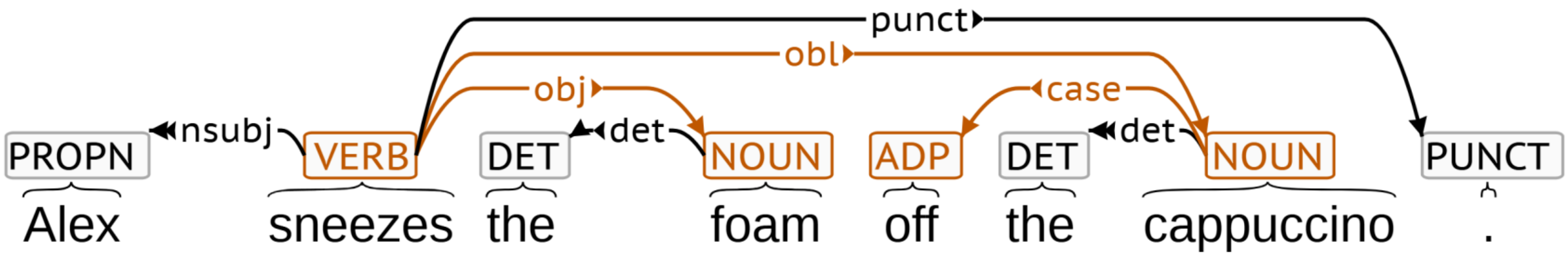
- I can pop my shoulder out of my socket.
- Alex sneezes the foam off the cappuccino.
- They laughed him off the stage.
- She had sauce on her lip so he kissed it off of her.

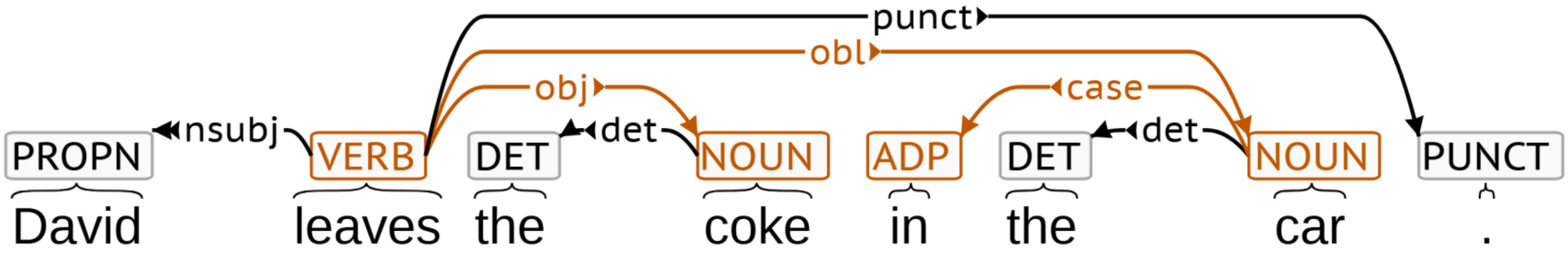
Reducing the Haystack with Dependencies





Syntactic Structure is not Enough





Can GPT4 solve all our problems?

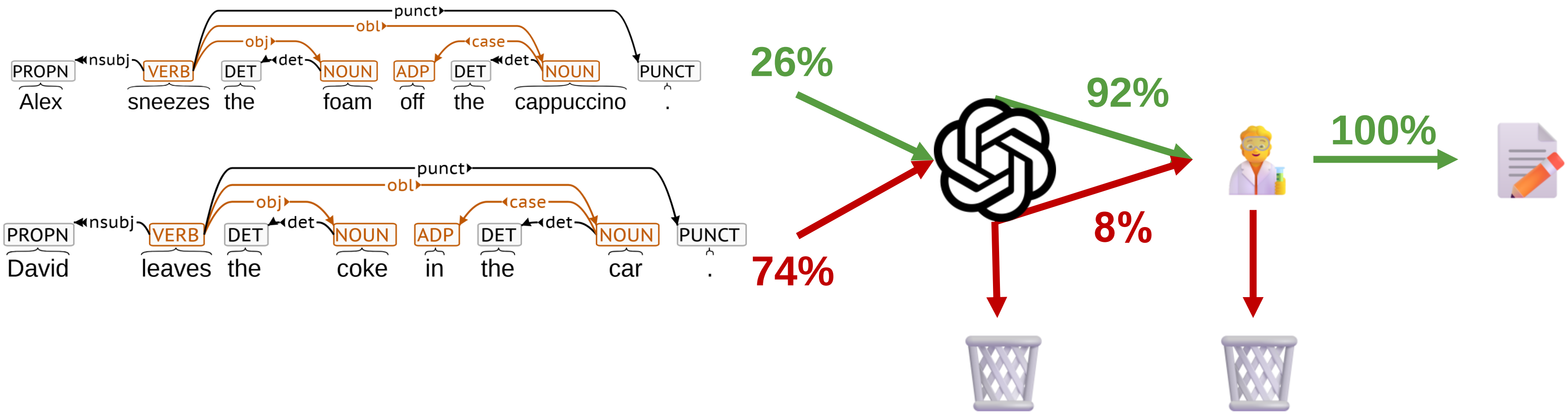
- Passing all sentences with matching dependencies on to GPT4
- Binary classification task with few shots

90% Precision, 80% Recall

→ **No!**

You are a linguistic expert specializing in syntax, specifically the caused-motion construction in English sentences. Your task is to analyze given sentences and classify whether they exhibit this construction or not. Remember to carefully consider the structure and meaning of each sentence to make the most accurate determination. The task is to classify whether the sentences are instances of the caused motion construction as first introduced by Goldberg (1992) or not. A caused-motion construction is a linguistic phenomenon where a verb describes an action that results in a change of location or motion for a specific object. Your task will be to understand what is going on in the sentence and determine if the verb describes an action that results in a change of location or motion for a specific object. Keep in mind that the caused-motion construction is rare, and label the sentences accordingly.
Here are 10 examples with examples with explanations and ground truth labels: : { "id": id,"sentence": sentence, "explanation": explanation, "label": label }. Classify the following sentences: { "id": id,"sentence": sentence }.
Respond with a jsonl codeblock (wrapped in three backticks). Each object should include an "id", "sentence", "explanation", and finally a "label" field with either "true" or "false". Label all 50 sentences.

Filtering the caused-motion sentences



Back to the original question

Let‘s ask the LLMS:

In the sentence `sentence`, is `direct_object` moving, yes or no?

Example

In the sentence ‚`She sneezed the foam off the cappuccino`‘, is `the foam` moving, yes or no?

Control question

In the sentence ‚`She threw the foam off the cappuccino`‘, is the `foam` moving, yes or no?

Still a Challenge

Y→Y: correct answer

N→Y: correct answer to control but not CMC

X→N: incorrect answer to control

→ Models generally struggle to give a valid answer

→ Even when they do fairly well on that, they're often wrong about the CMC

Family	Model	IT	Y→Y	N→Y	X→N
GPT	3.5	+	43.20	10.70	46.10
	4	+	57.07	11.23	31.70
Gemini	Pro	+	43.43	12.70	43.87
Llama2	7b	–	9.54	1.09	89.37
		+	21.93	1.77	76.29
	13b	–	53.00	8.72	38.28
		+	5.59	1.23	93.19
	$70b_Q$	–	36.65	7.36	55.99
		+	37.87	5.59	56.54
Mistral	7b	–	34.20	4.50	61.31
		+	68.12	8.45	23.43
	8x7b	–	35.29	9.95	54.77
		+	69.75	12.13	18.12

Table 3: LLM evaluation results. IT=instruction-tuned. Q=quantised.

NEW HOPE FOR DATA

UCxn: Typologically Informed Annotation of Constructions Atop Universal Dependencies

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, Nathan Schneider

Scaling Up

Rule-compositionality is easier to investigate at scale than construction-compositionality

Too complex to generate automatically

Little to no annotated data

→ No annotation models

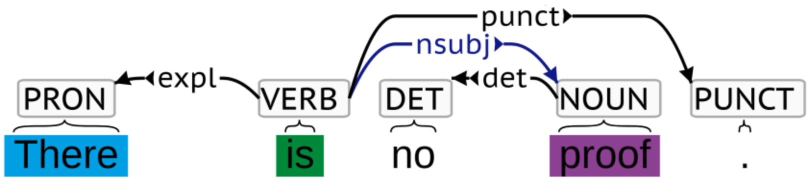
Piggybacking off Universal Dependencies

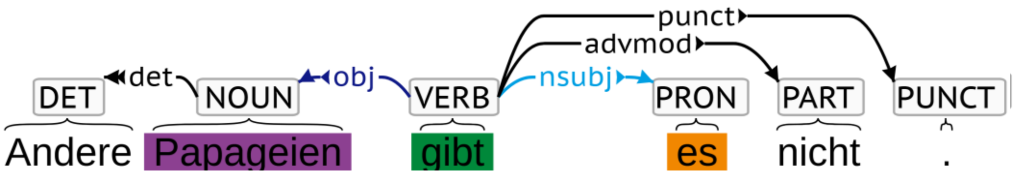
What if Universal Dependencies looked like this?

```
Who let the dogs out ?
1    Who    …    2    nsubj  …    CxnElt=2:Interrogative.WHWord
2    let    …    0    root   …
Cxn=Interrogative,Resultative|CxnElt=2:Interrogative.Clause,2:Resultative.Event
3    the    …    4    det    …    _
4    dogs   …    2    obj    …    _
5    out    …    2    xcomp  …    CxnElt=2:Resultative.ResultState
6    ?      …    2    punct  …    _
```

But where will the necessary information come from?

Using Dependencies to Annotate Constructions





There are no other parrots.

```
_anchor_ [lemma="be"];
_anchor_ -[re".subj"]-> P;
_anchor_ -> Y;
Y [xpos="EX"];
```

```
E[lemma="es"];
_anchor_[lemma="geben"];
_anchor_-[re".subj"]-> E;
_anchor_-[obj]-> P;
```

1 **There**
2 **Is** Cxn=**Existential**-CopPred-ThereExpl
3 no
4 **proof** CxnElt=3:**Existential**-CopPred-ThereExpl.**Pivot**
5 .

1 Andere
2 **Papageien** CxnElt=3:**Existential**-GivePred-ItExpl.**Pivot**
3 **gibt** Cxn=**Existential**-GivePred-ItExpl
4 **es**
5 nicht
6 .

UCxn V1: A New Resource

Lang.	Interrogative (§4)	Existential (§5)	Conditional (§6)	Resultative (§7)	NPN (§8)	total sent.	total tokens
EN	1117; 769	472; 319 (f)	762; 375 (D)	H, D	21; 12	17k; 11k	254k; 187k
DE	5483 (H)	3392 (H)	3291 (A,H)	D	40	190k	3.5m
SV	276	235	310 (H)	D	7	6k	96k
FR	368	114 (F)	213 (F)	D	12	16k	400k
ES	580	160 (F)	502 (F)	D	37	18k	567k
PT	337 (A)	340 (F)	106	D	7	9k	227k
HI	285	2058 (F)	350 (A)	D	?	16k	351k
ZH	146	58 (F)	31	78 (D)	?	1k	9k
HE	236; 22	113; 60	192; 56	D	9; 11	6k; 5k	160k; 140k
COP	150	80	185	D	2	2k	55k

Table 4: Counts of identified construction instances by treebank, along with qualifications: definitional issues (D), UD annotation errors (A), occasional false positives (f), frequent false positives (F), unattested strategies (H). ? means that the existence of the productive construction is doubtful (see Fn. 6). The two numbers for EN and HE represent the two treebanks for each (see Table 5 in the Appendix).

The State of Construction (Work) in LLMS

- Choosi ctions to use based
- Associ uctions 

- Modifyi onstructions 

- More d

Conclusion

- Language is compositional, but based on constructions, not words and rules
- Language models seem like they‘re bad at rule-compositionality but they‘re good (but not perfect) at construction-compositionality
- Some constructions are still challenging

Future Directions

- **Let‘s not climb the wrong compositionality hill**, because some of what we‘ve been criticising in LLMs is precisely what makes them so good at language
- To improve, **let‘s perform targeted evaluation** to find out if it‘s model size, data size, architecture or something else

THANK YOU

To my collaborators

- On the ideas presented in this talk: Adele Goldberg & Kyle Mahowald
- On the practical work presented in this talk:

David Mortensen, Nathan Schneider, Bill Croft, Joakim Nivre, Hinrich Schütze (to name but a few)

FOR LISTENING!

The State of Construction (Work) in LLMS

- Choosing the correct morpheme constructions to use based on context 

- Associating the right meaning with constructions 

- Modifying argument structure based on constructions 

- More data for all of the above