

LLMs Learn Constructions That Humans Do Not Know

Jonathan Dunn and Mai Mohamed Eida
University of Illinois Urbana Champaign

Are LLMs aware of syntactic constructions?

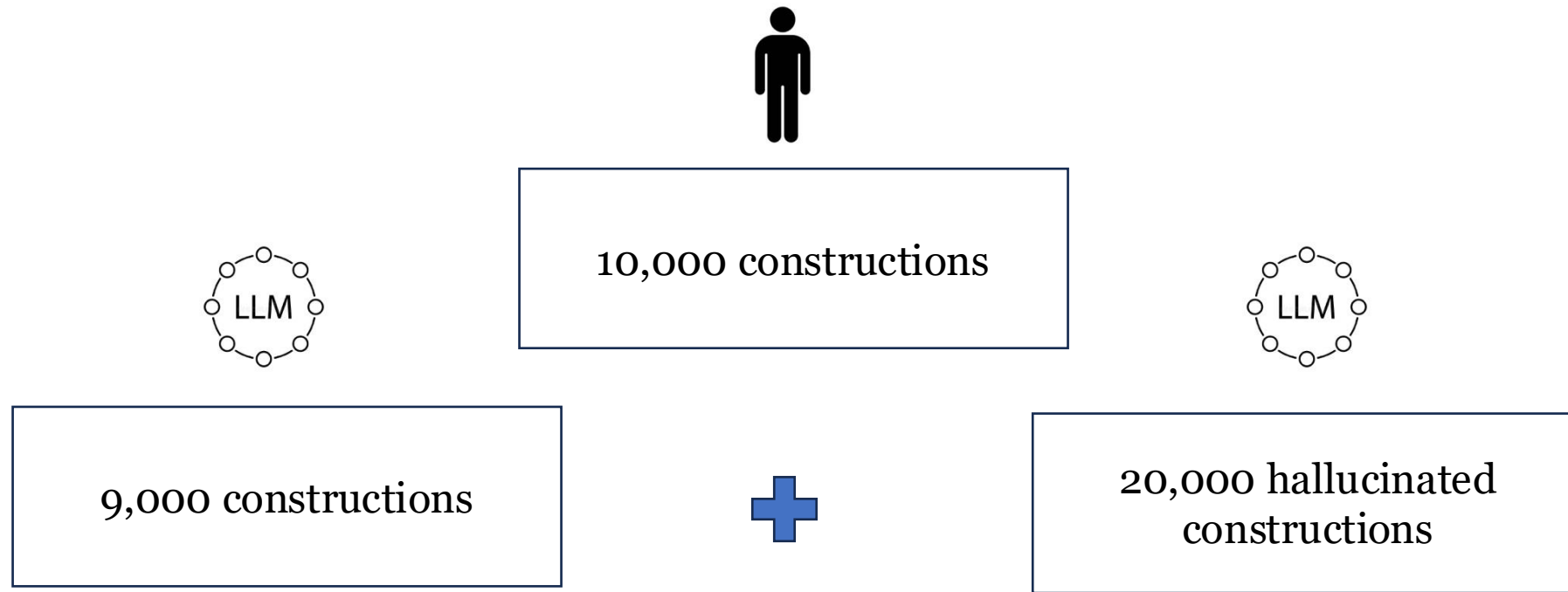
- (a) five terrible weeks
- (b) a terrible five weeks

LLMs Successful

Evidence of learnability
(Misra and Mahowald, 2024).

LLMs Unsuccessful

Model is incorrect
(Weissweiler et al., 2022).



Do LLMs make errors by over-learning constructions which do not actually exist for human speakers?

Experiment 1: GPT-4 Meta-Linguistic Prompts (Bonial and Tayyar Madabushi, 2024)

Intransitive

Transitive (NP)

Transitive (C)

Passive

Double Object

Pristine
Exemplar

Reverted Focus

Alternate
Application

Normalized
Attribution

Entrenched
Objective

From amongst the following sentences, extract the three sentences which are instances of the **Reverted Focus construction**, as exemplified by the following sentence: “**Luckily they caught the crooks before they did one on us**”. Output only the three sentences in three separate lines: **[Followed by six examples to sort from the same construction.]**

From amongst the following sentences, extract the three sentences which are instances of the **Reverted Focus construction**, as exemplified by the following sentence: “**Luckily they caught the crooks before they did one on us**”. Output only the three sentences in three separate lines: **[Followed by six examples to sort from the same construction.]**

A high stability across the sentence clustering indicate that neither the name nor the example influence which patterns are discovered.

	<i>Intransitive</i>	<i>Transitive (NP)</i>	<i>Transitive (C)</i>	<i>Passive</i>	<i>Double Object</i>
Alternate Application	92.69%	92.04%	92.84%	93.88%	92.74%
Entrenched Objective	93.93%	92.53%	91.70%	94.36%	91.25%
Normalized Attribution	94.53%	92.76%	92.20%	93.34%	93.79%
Pristine Exemplar	93.01%	93.26%	90.33%	93.18%	91.41%
Reverted Focus	93.39%	93.34%	93.64%	93.98%	93.81%

Evidence of Confirmation Bias in Methodology:

If these nonce constructions were real, we would think the models agrees with us!

Are these distinctions **false positives hallucination** or
meaningful constructions humans missed?

Category	Cluster Accuracy
<i>Intransitive</i>	74.8%
<i>Transitive (NP)</i>	71.1%
<i>Transitive (C)</i>	77.0%
<i>Passive</i>	80.5%
<i>Double Object</i>	74.2%

The resulting pairwise clusters make a stronger case for hallucinated constructions.

Are there new model-driven constructions which are plausible?

Based on linguist introspection analysis of model clustering patterns, this is unlikely.

Experiment 2: Pytha 1.4 Behavioural Probes

(Li et al., 2022)

Intransitive

Transitive (NP)

Transitive (C)

Passive

Double Object

Do embeddings correctly distinguish between the five true positive constructions?

	<i>Intransitive</i>	<i>Transitive (NP)</i>	<i>Transitive (C)</i>	<i>Passive</i>	<i>Double Object</i>
Direct Embeddings	0.92	0.88	0.92	0.91	0.93
Grammar-Focused	0.85	0.79	0.87	0.85	0.85

Intransitive

Transitive (NP)

Transitive (C)

Passive

Double Object

Puppet Hypothesis: Do LLMs distinguish between two false positive constructions that belong to one construction category?

Clustering

Classification

Puppet Hypothesis: Do LLMs distinguish between two false positive constructions that belong to one construction category?

	<i>Direct Embeddings</i>	<i>Grammar- Focused</i>
Intransitive	0.99	0.99
Transitive (NP)	0.94	0.93
Transitive (C)	0.96	0.99
Passive	NA	0.99
Double Object	0.97	0.97

Puppet Hypothesis: Does the model consistently make distinctions between hallucinated constructions that do not exist for humans?

Yes, it does. The high accuracy is true for both the true positives and the false positives.

From a computational perspective, this means that we still do not know the full linguistic knowledge within LLMs because we have only looked for what we expected to find.

Thank you! Questions?

Minimal pairs cannot be formulated for unknown false positives.

Entrenchment and exposure are specific to individuals.

Can computational models ever tell linguists something they did not know?