Language Models Learn Rare Phenomena from Less Rare Phenomena: The Case of the Missing AANNs

Kanishka Misra[1,2]*, Kyle Mahowald[2]

** Work done as a Postdoc at the University of Texas at Austin*

[1] TTIC TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO

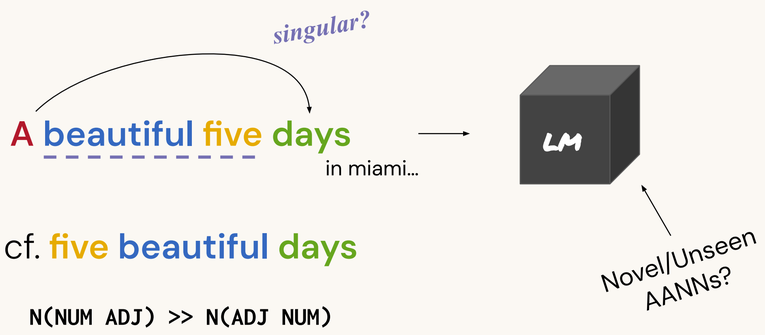
[2] TEXAS The University of Texas at Austin

NSF

Camera ready!

Camera-ready: https://bit.ly/aanns

The AANN Construction: Article + Adjective + Numeral + Noun

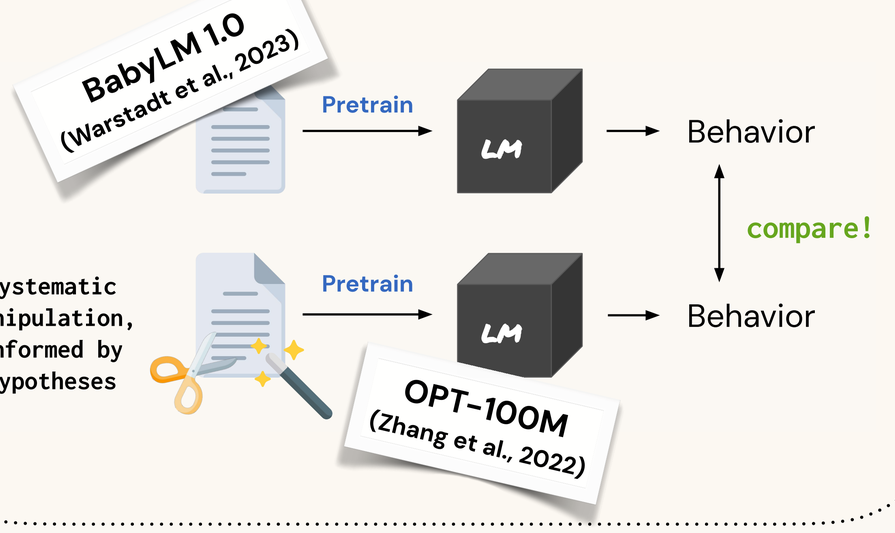


Two obvious generalizations about English that LMs might learn:

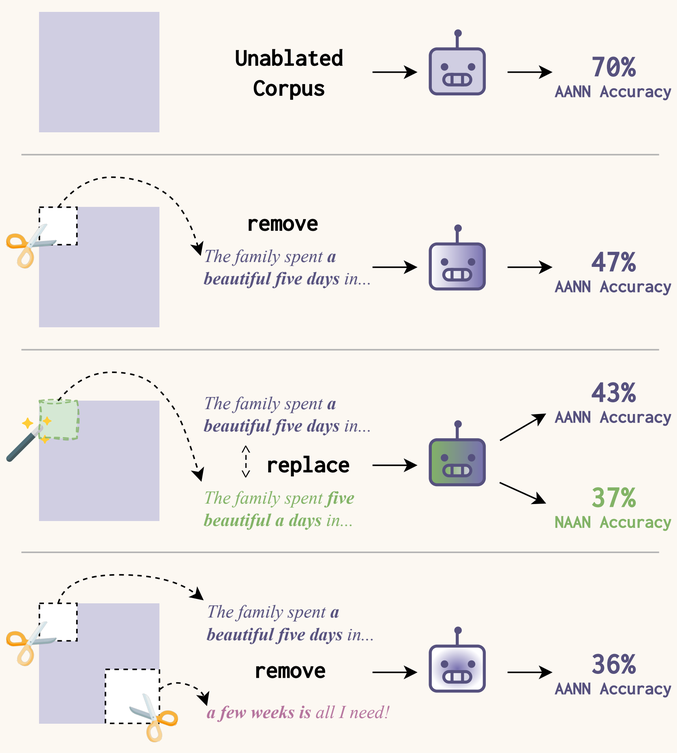
- The indefinite article "A" goes with singular nouns
- Numerals precede Adjectives!

AANNs violate both these rules.
So do LMs simply memorize seen AANNs verbatim, or are they able to generalize to novel instances?

Method: Controlled Rearing



Experiment 1: How well do LMs learn about AANNs?



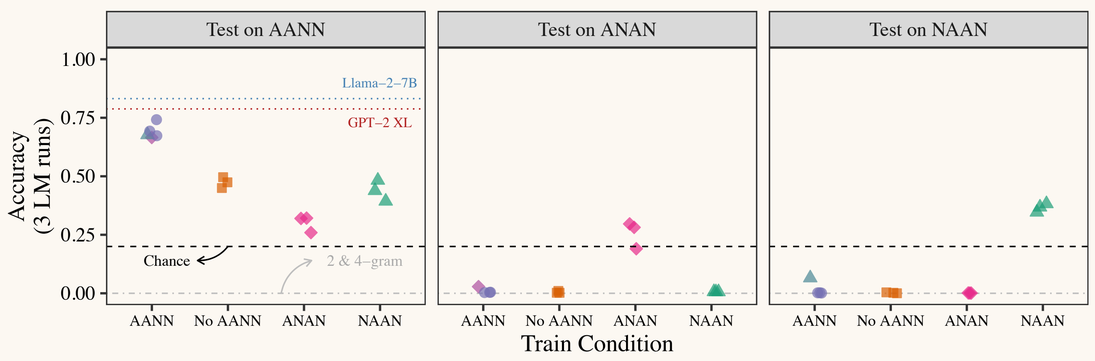
Context	AANN	ANAN	NAAN
WELL-FORMED	a whopping ninety LMs	a ninety whopping LMs	ninety whopping a LMs
	Corruptions		
ORDER-SWAP	a ninety whopping LMs	a whopping ninety LMs	whopping ninety a LMs
NO ARTICLE	whopping ninety LMs	ninety whopping LMs	ninety whopping LMs
NO MODIFIER	a ninety LMs	a ninety LMs	ninety a LMs
NO NUMERAL	a whopping LMs	a whopping LMs	whopping a LMs

$$\text{SLOR}_{\text{prefix}} = \frac{1}{|\mathcal{C}|} \log \frac{p_m(\mathcal{C} \mid \text{prefix})}{p_u(\mathcal{C})}$$

prefix: The researchers trained
$\mathcal{C}$: a whopping 90 LMs

Accuracy = Proportion of time:

$\text{SLOR}_{\text{prefix}}$(*a whopping 90 LMs*) > $\text{SLOR}_{\text{prefix}}$(*a 90 whopping LMs*)
$\text{SLOR}_{\text{prefix}}$(*a whopping 90 LMs*) > $\text{SLOR}_{\text{prefix}}$(*whopping 90 LMs*)
$\text{SLOR}_{\text{prefix}}$(*a whopping 90 LMs*) > $\text{SLOR}_{\text{prefix}}$(*a 90 LMs*)
$\text{SLOR}_{\text{prefix}}$(*a whopping 90 LMs*) > $\text{SLOR}_{\text{prefix}}$(*a whopping LMs*)



BabyLM-trained LMs learn about the AANN...
... even without encountering a single instance ...
... more strongly than they learn counterfactual variants.

LMs could be relying on *indirect evidence* in the training data that might contribute to their learning of AANNs!

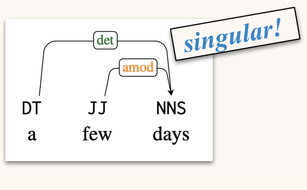
Experiment 2: What is the Key to Learning AANNs?

Hypothesis Space

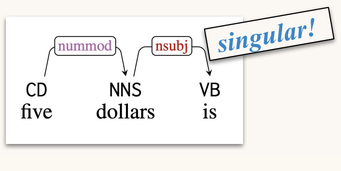
DT ANNs
(N = 15,781)

"**The** beautiful 5 days" ⇒ "**a** beautiful 5 days" ✔

Indef articles w/ PL NNS
(N = 55,373)



Measure NNS with SG verbs
(N = 7,371)



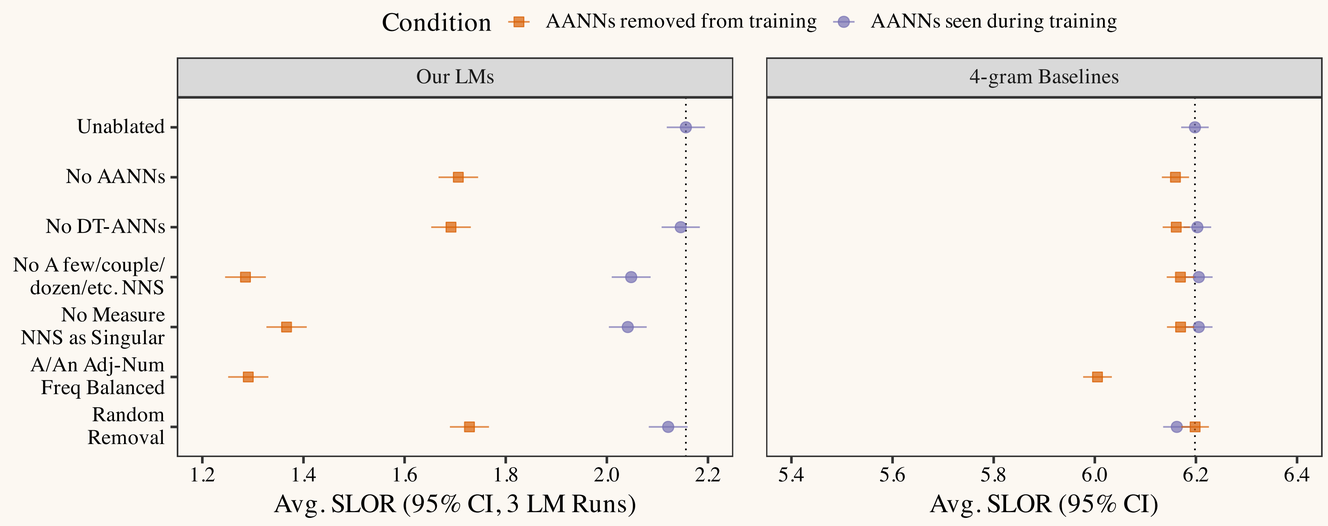
Freq. Balancing
(N = 571,874)

Freq(A/An + JJ)
> → =
Freq(A/An + NUM)

Random removal
(N = 571,874)

Remove utterances that do not conform to any hypothesis

Compare orange to orange, purple to purple!



LMs can demonstrate a completely novel phenomenon (AANN) by relying on other related—and more frequent—phenomena! E.g., by observing other instances of measure NPs with plural nouns being treated as singular units (*a few days, five dollars is plenty!*)

This cannot be explained by (1) data loss (random ablations have little effect); and (2) shallow ngram processing (4-grams do not show the same sensitivity)

Question: Is there an Experiment 3?

Yes! Check out the camera-ready for analyses on how the properties of seen AANNs affect LM generalization!

Correspondence: kanishka@ttic.edu

EMNLP 2024, Miami

https://kanishka.website