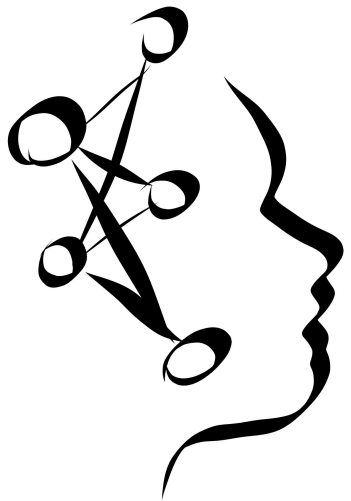
The Role of Input in Shaping Linguistic Generalization

Insights from Controlled Rearing of Language Models

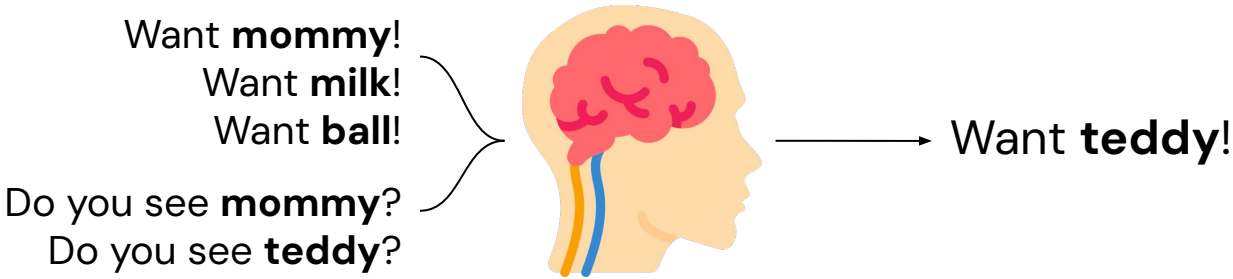
Kanishka Misra
UT Austin, Linguistics

Elias' NLP Class



*Art by **Dr. Judy Fan**, for the neural-nets-for-cognition workshop at CogSci 2022*

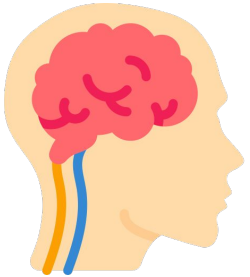
Linguistic Generalizations



Linguistic Generalizations

"The vase **broke**."

"Jessy **broke** the vase."



"The baby **giggled**" ✔

"Mommy **giggled** the baby!" ✘



Input is especially sacred to Constructivist theories of language acquisition!

Generalizations and abstractions emerge from distributional/usage-based analyses of the linguistic (and other forms of-) input.

Elman, (1993); Goldberg (2003); Rowland (2013); Ambridge and Lieven (2015); Behrens (2021); others

Direct positive evidence	Direct negative evidence
2 cute kittens (cf. *cute 2 kittens)	
give me the ball **give** the ball to me	Not systematically observed…
(Contested under PovStim for many phenomena)	

Brown and Hanlon (1970); Bowerman (1988); Pearl (2016);

Indirect positive evidence

The **wug** is **crying**

⇒

wug is animate!

An upcoming finding!

Indirect negative evidence

I **went** to the bookstore

⇏ (blocks)

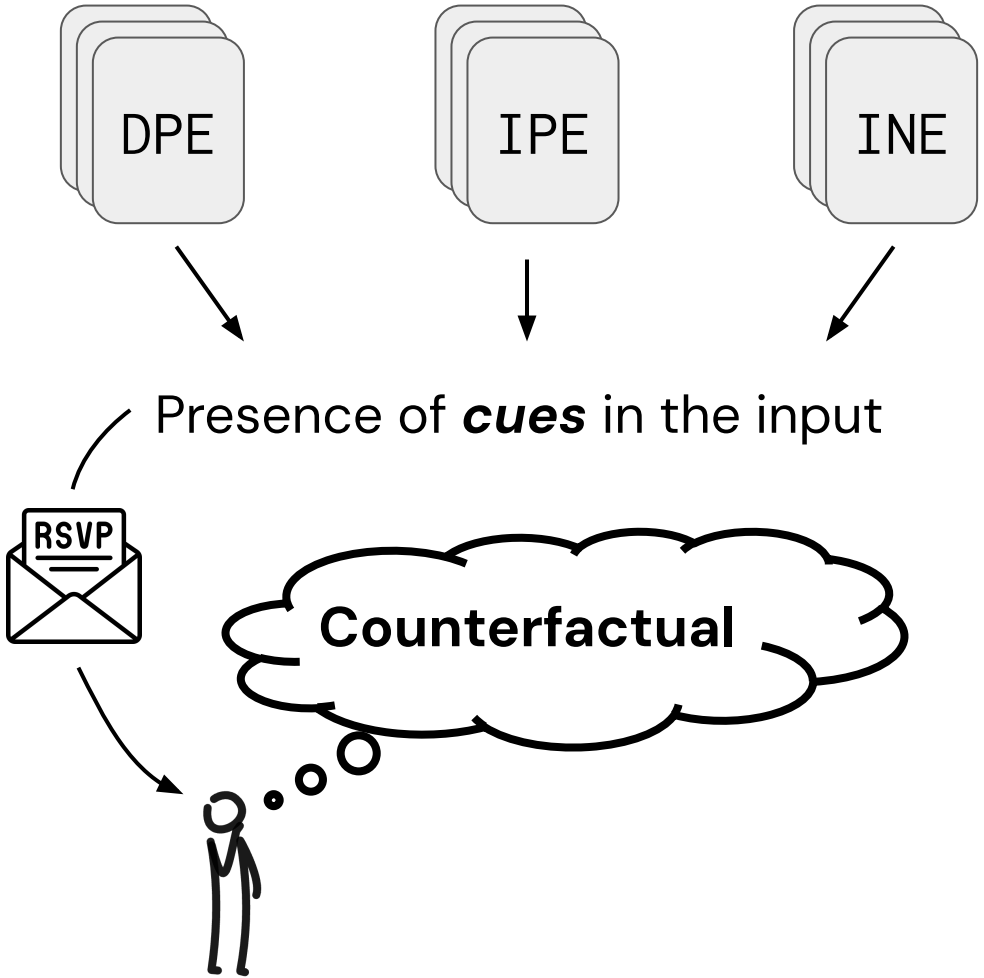
I **goed** to the bookstore

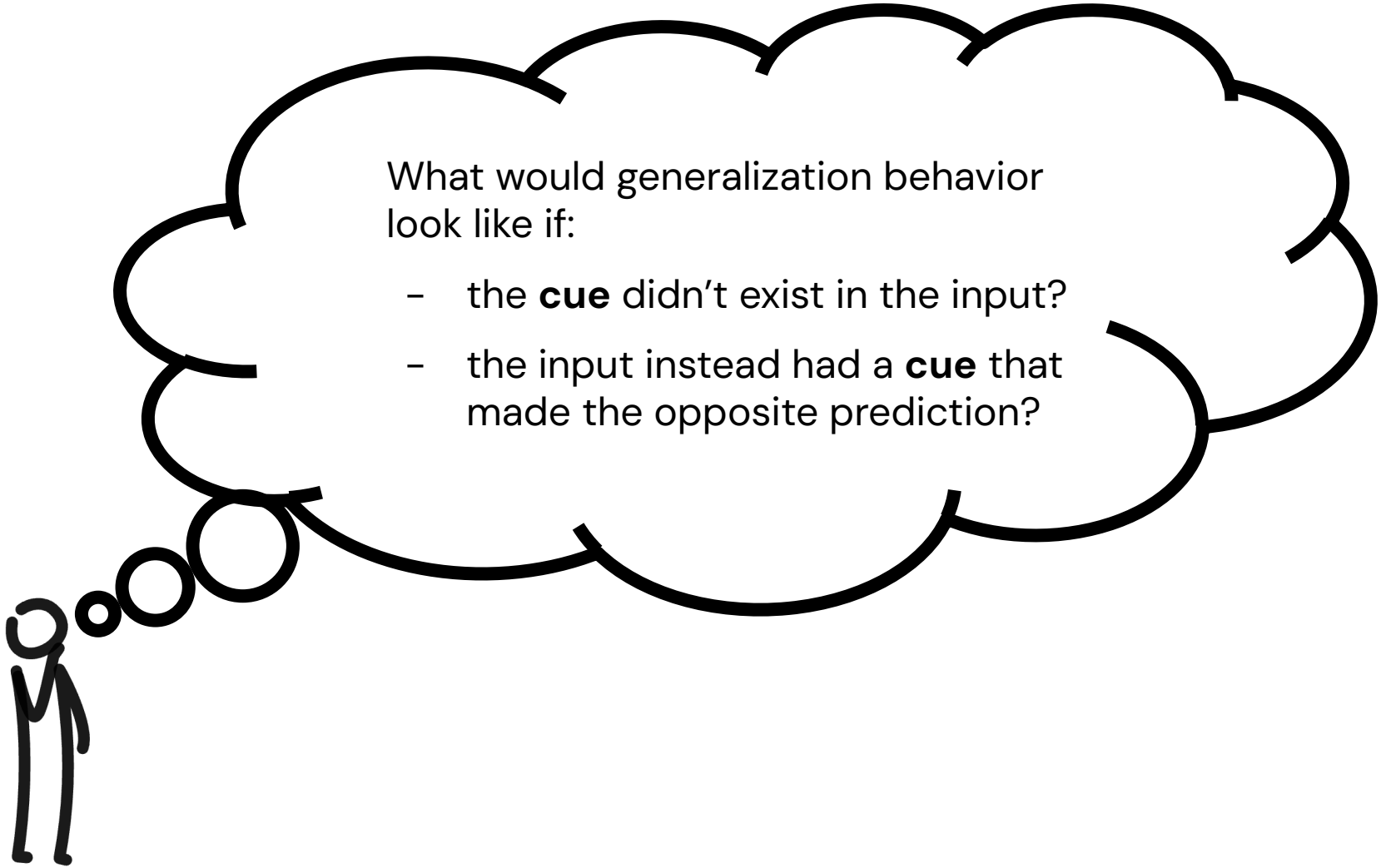
Mommy **made** the baby **giggle**!

⇏ (blocks)

Mommy **giggled** the baby!

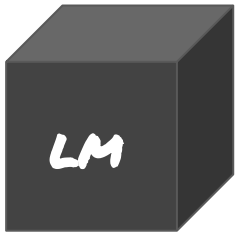
Names: Ambridge, Goldberg, Lieven, Rowland, Pearl (2016), Reali and Christiansen (2005)


DPE
IPE
INE
Presence of ***cues*** in the input
RSVP
Counterfactual

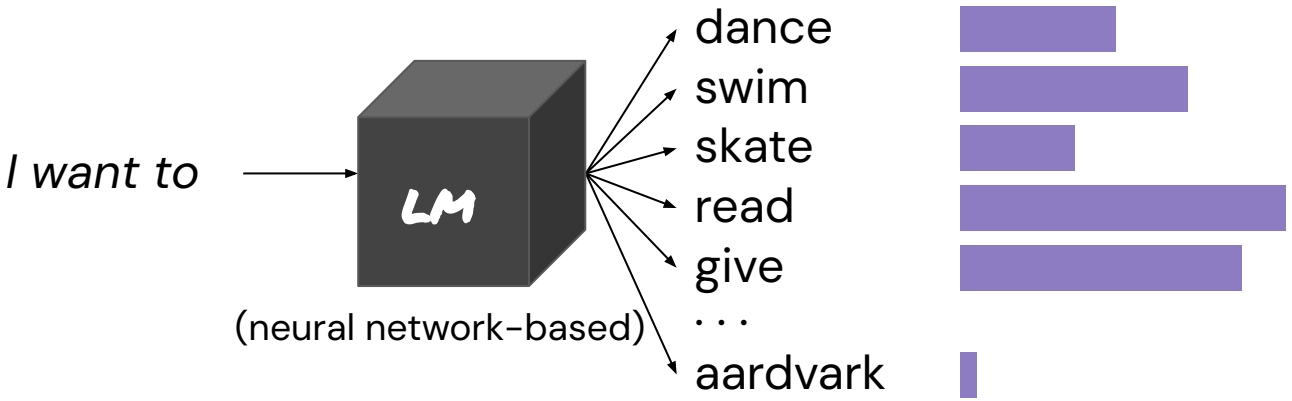

What would generalization behavior look like if:
- the **cue** didn't exist in the input?
- the input instead had a **cue** that made the opposite prediction?



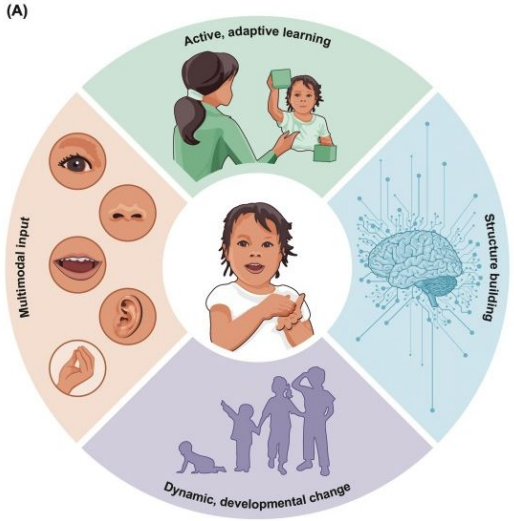
Alter learner's experience to understand the effect of a specific cue!

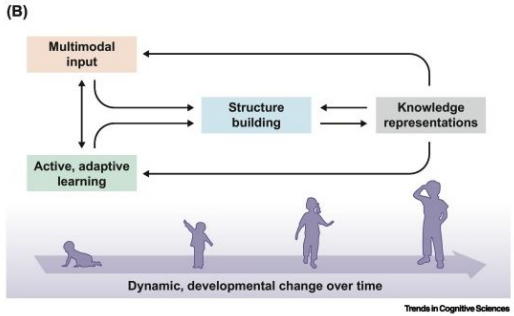


(neural network-based)

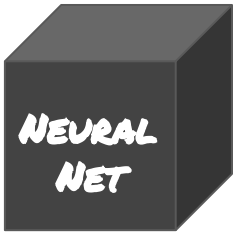


- Error-driven learning to predict the next word (Elman, 1990)
  - Learning only from positive evidence
- Trained from *scratch* on corpora we have full control over
- Abstract away effect of audio/morphological signal
- Probabilistic model of language use
  - at a level of abstraction captured by sequences of words.





Prospectively

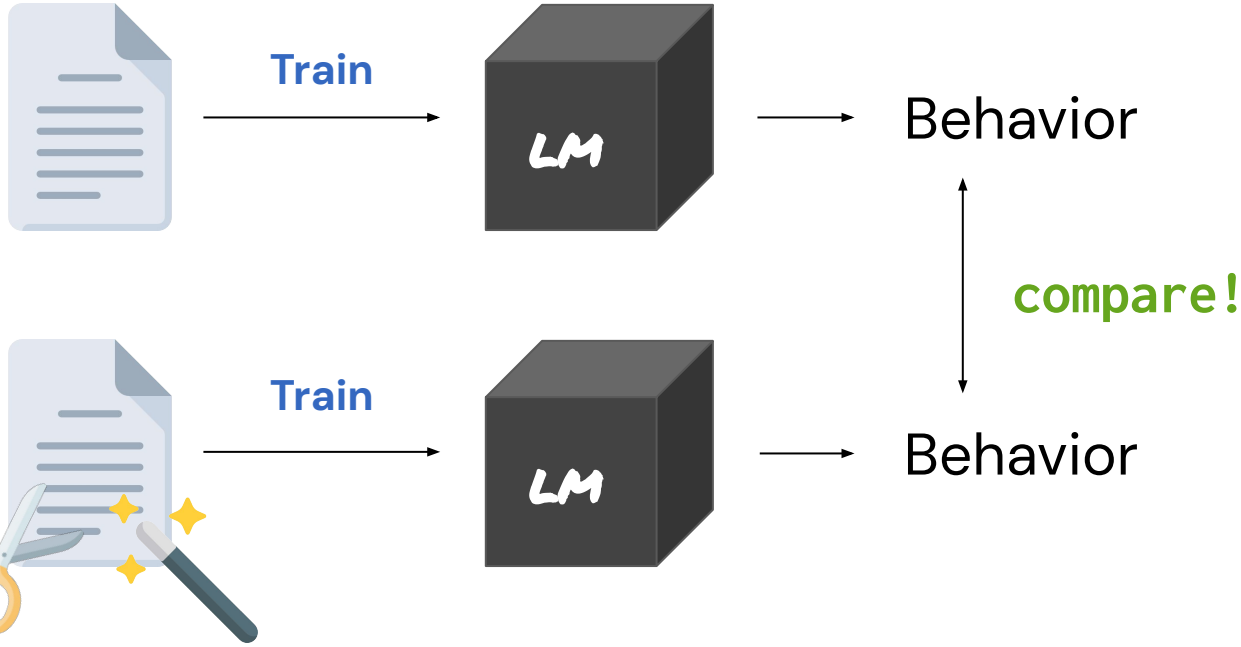


Can only raise the excitement/interestingness of the kinds of RQs one can ask wrt “what’s in the input?”

Rowland et al., 2025

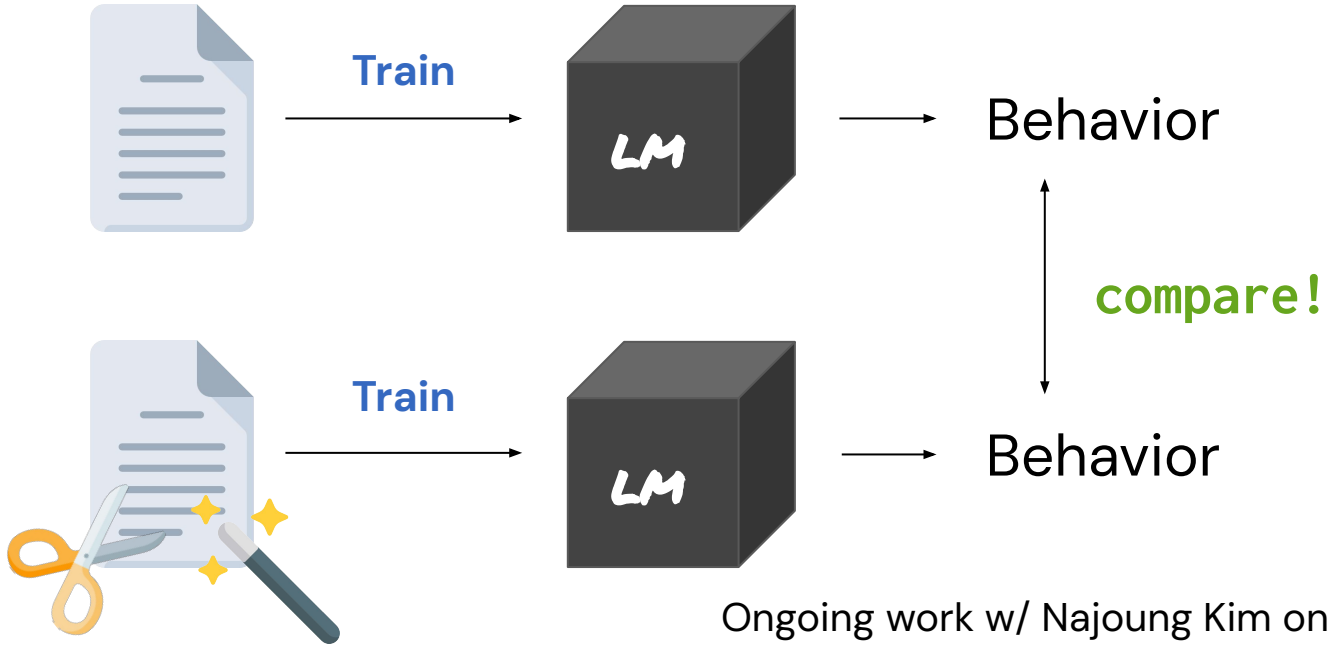
Controlled rearing of language models





(Misra & Mahowald, EMNLP 2024)

You can think of cool controlled rearing work for mildly larger LMs too!



Ongoing work w/ Najoung Kim on 1B and 7B models

People are doing it! (would've been nice if they acknowledged the comp. ling origins of the methods the')



HUBBLE: A MODEL SUITE TO ADVANCE THE STUDY OF LLM MEMORIZATION

Johnny Tian-Zheng Wei[*,1], **Ameya Godbole**[*,1], **Mohammad Aflah Khan**[*,2], **Ryan Wang**[1], **Xiaoyuan Zhu**[1], **James Flemings**[1], **Nitya Kashyap**[1], **Krishna P. Gummadi**[2], **Willie Neiswanger**[1], **Robin Jia**[1]

[1]University of Southern California [2]Max Planck Institute for Software Systems

{jtwei, ameyagod, robinjia}@usc.edu, afkhan@mpi-sws.org



Article+Adjective+Numeral+Noun = AANN



cf. five beautiful days

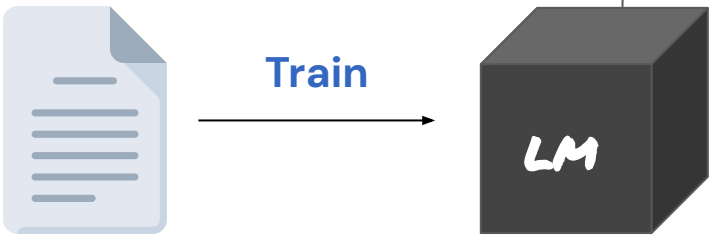
*A tall seven chairs
*A blue six pencils

"Stubbornly distributive"

(Misra & Mahowald, EMNLP 2024)

Controlled rearing for AANNs

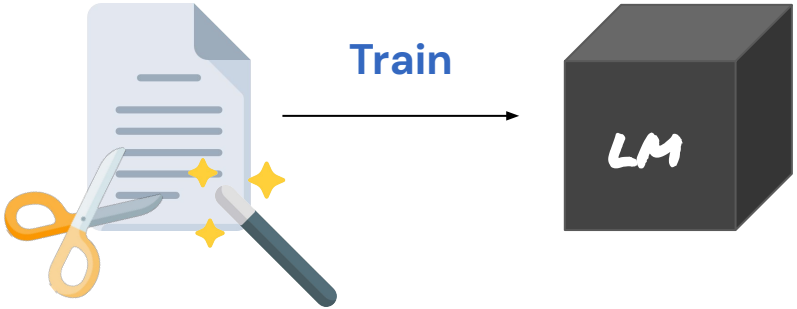
Trained from **scratch** on BabyLM (Warstadt et al., 2023) ~100M words of English!



Unseen AANNs

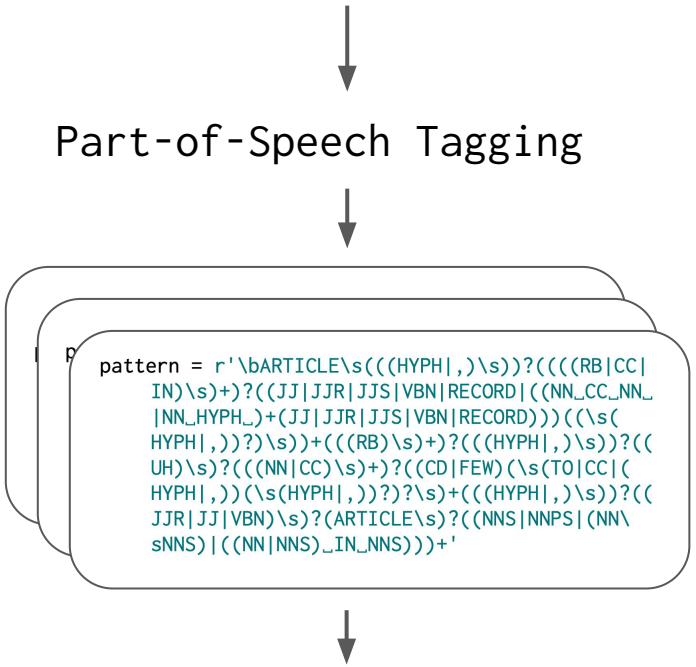
compare!

Remove ***all*** AANNs (*N* = 2,448/11M utterances)



Unseen AANNs

BabyLM Corpus (100M)



Part-of-Speech Tagging

2448 AANNs!

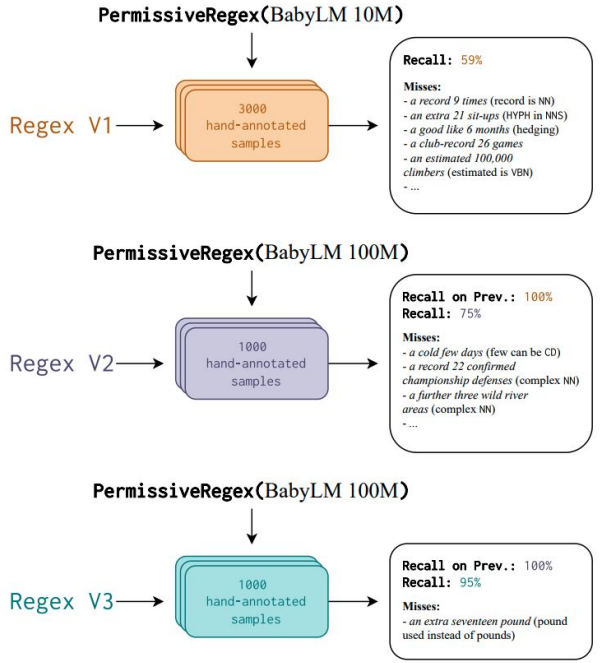


Figure 6: Pipeline to assess the recall of our AANN-detecting regex patterns, along with examples of cases missed by each regex. The recall for our final regex (`Regex v3`) is 95% (missing only one instance where there was a typo), and it is able to handle complex and sophisticated forms of the construction.

Mahowald (2023) tests for AANNs:

- 3.4k AANN containing sentences with human judgments of acceptability (1–10)
- Use 7 as threshold for “good” AANNs + **ensure that they do not occur in BabyLM ⇒ 2,027 final sentences**!

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

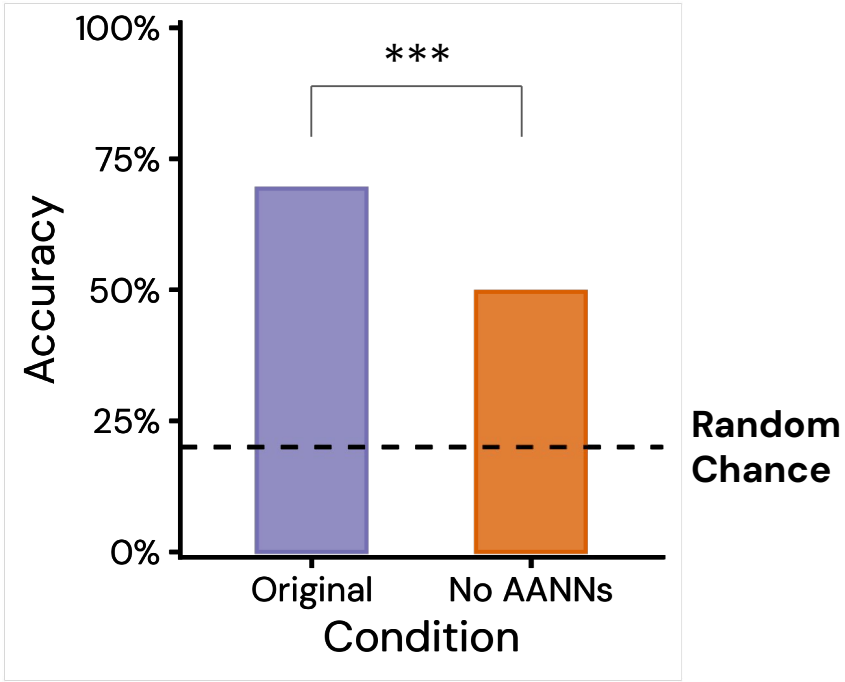
acceptable

A beautiful five days

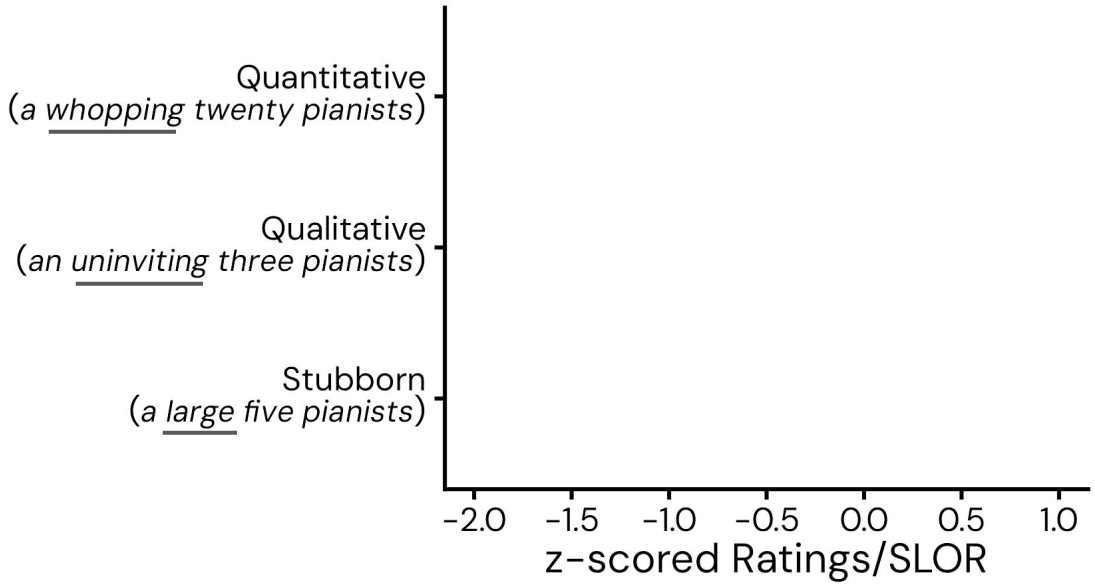
>

beautiful five days
A five days
A beautiful days
A five beautiful days

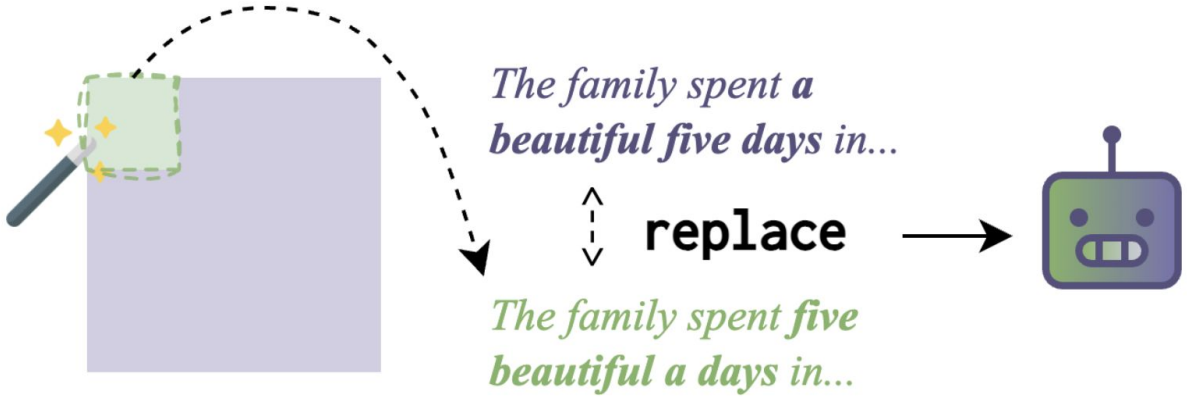
unacceptable



Constraints on the Adjective Slot



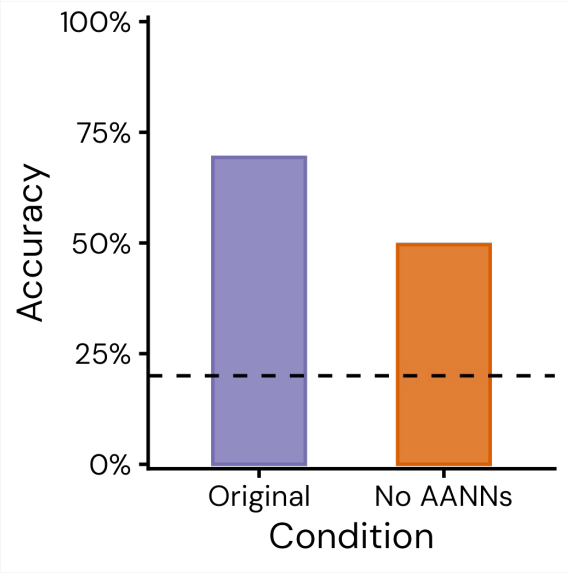
What if there was a counterfactual variant of AANN?

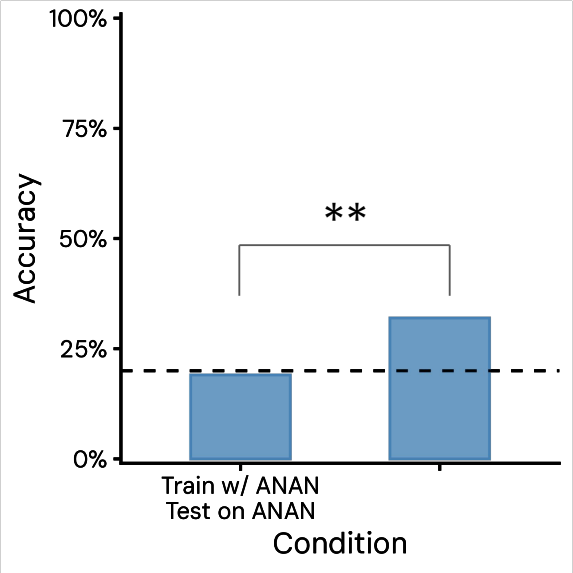


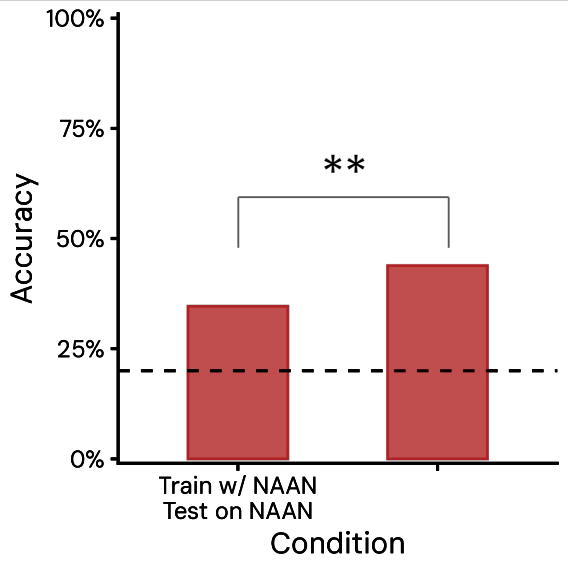
AANN ↔ NAAN

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

Results


100%
75%
50%
25%
0%
Accuracy
Original
No AANNs
Condition


100%
75%
50%
25%
0%
Accuracy
**
Train w/ ANAN
Test on ANAN
Condition

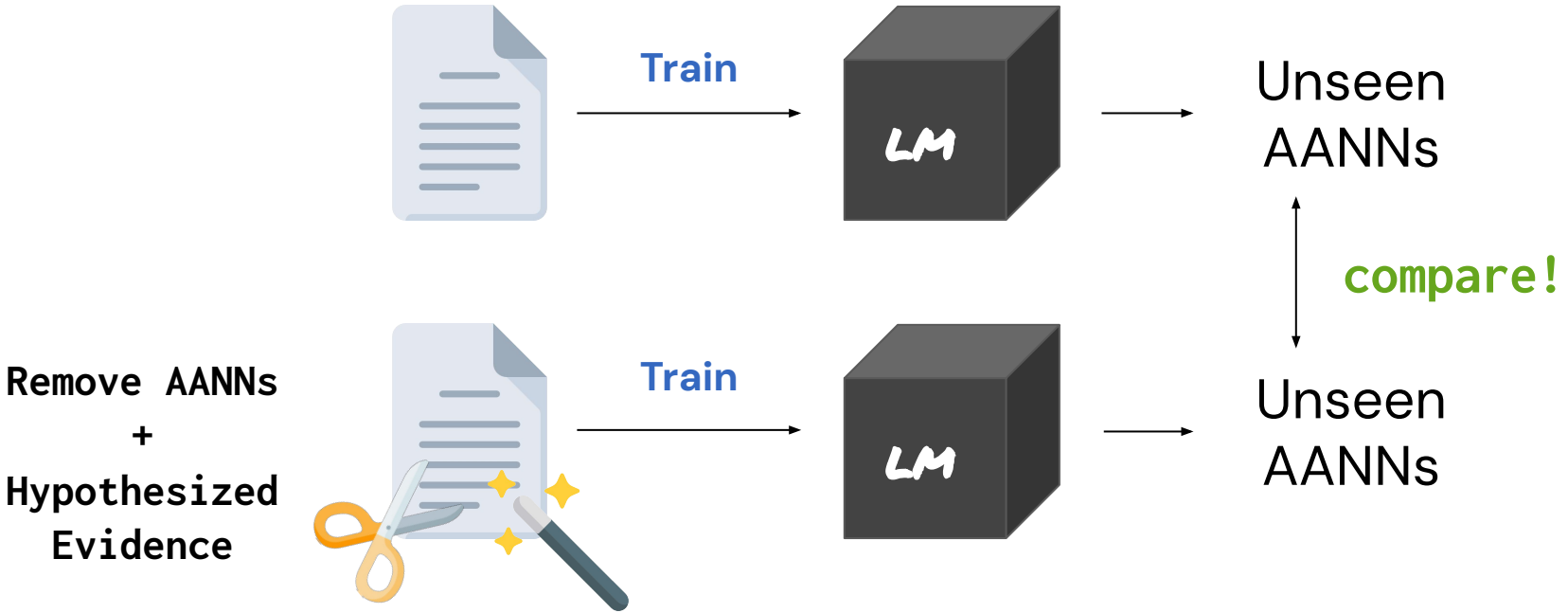

100%
75%
50%
25%
0%
Accuracy
**
Train w/ NAAN
Test on NAAN
Condition

Interim Takeaways:

- LMs learn about the AANN…
- …even without encountering a single instance…
- …more strongly than they learn its counterfactual variants…
- …in a way that extends to lexical semantic constraints!

There is *indirect* evidence in the training data that can be picked up by our target learners (LMs)

Indirect Evidence for AANNs

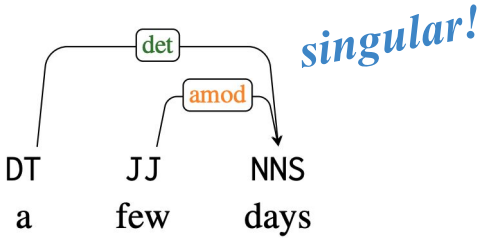


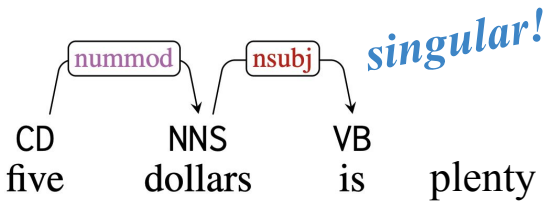
Indirect Evidence for AANNs

Determiner generalization

“**The** beautiful 5 days” ⇒ “**a** beautiful 5 days” ✔

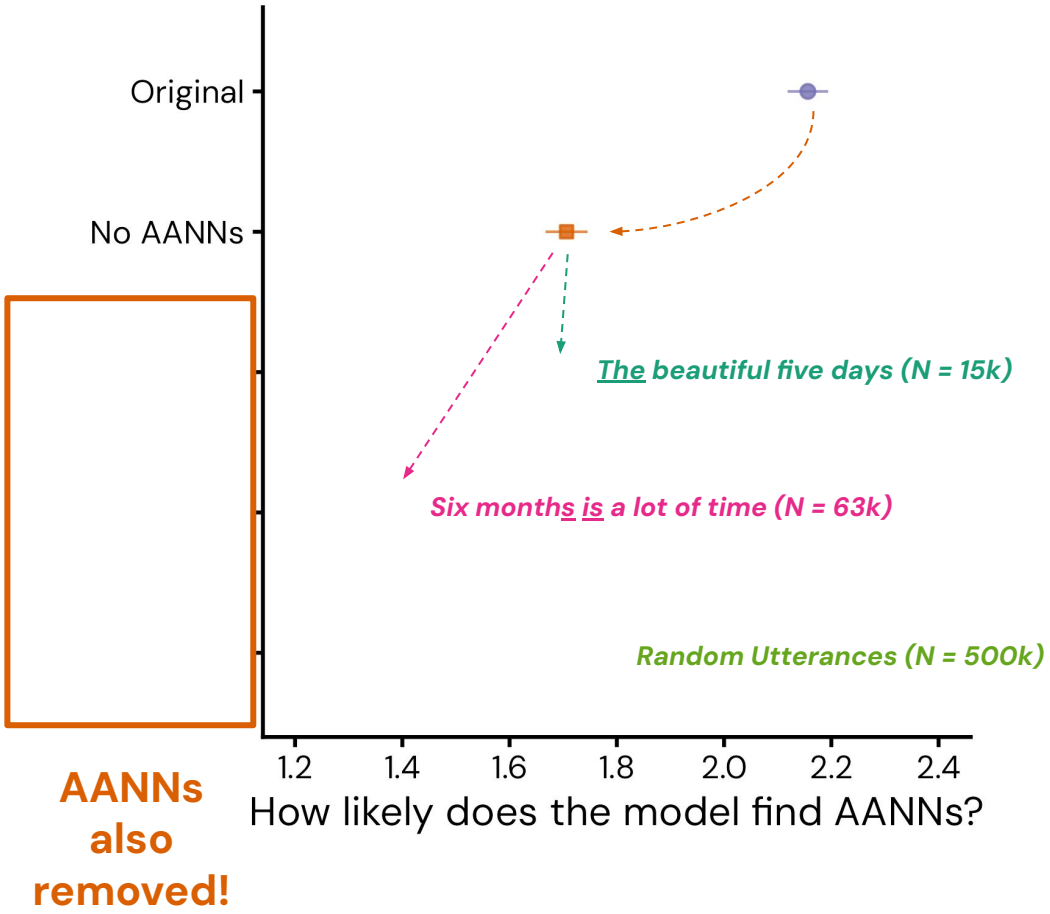
When plural things are construed as singular





Control: Random removal

Remove utterances that do not conform to these hypotheses



AANNs also removed!

AANNs can be learned, **even without direct evidence,** from other instances of measure NPs with **plural nouns treated as singular units**!

N-gram (4) LMs don't show this!

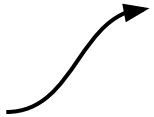
Constructions: a new theoretical approach to language

Adele E. Goldberg

Linguistics Department, University of Illinois, Urbana, IL 61801-0168, USA

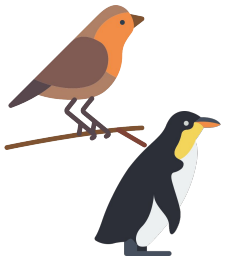
Box 1. Questions for Future Research

- Does learning one construction facilitate learning other related constructions?



It can!

Slot variability (*Premise Diversity* in Induction)



Robins can dax.
Penguins can dax.

All **birds** can dax.

$>$

Robins can dax.
Sparrows can dax.

All **birds** can dax.



High Variability

A **colossal** 127 pounds.
An **accumulative** 44 years.
A **balmy** 8 degrees!
A **dawdling** 5 hours..
A **glorified** 700 people!

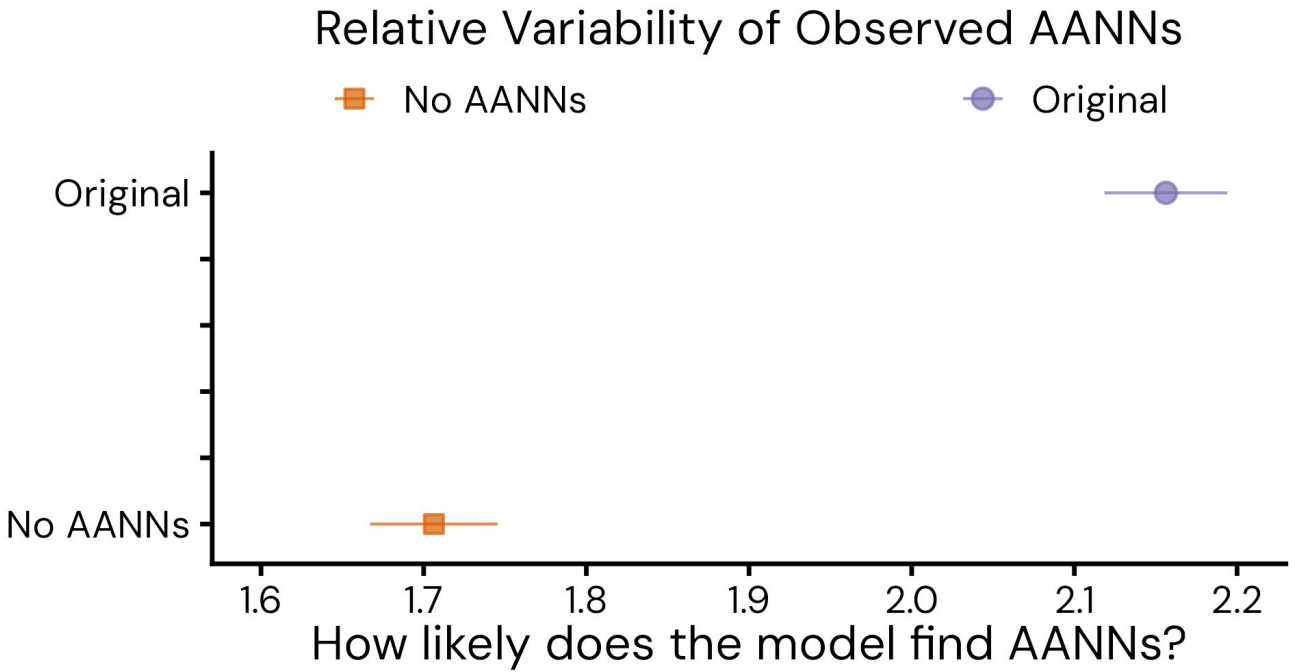
$>?$

Low Variability

A **mere** hundred yards.
A **mere** 2 cents.
A **good** five days.
A **good** 10 years.
A **good** 50 pounds.

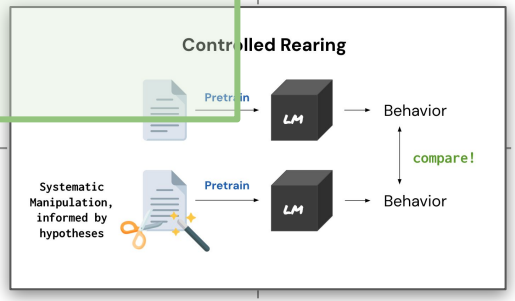
Osherson et al., 1990; Suttle and Goldberg, 2011

Evidence of productivity promotes generalization!



Controlled Rearing = Fruitful Toolkit!

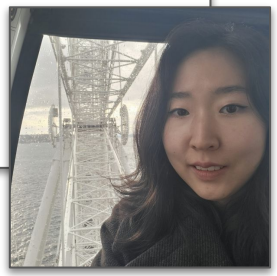
- NPIs – Jumelet et al., 2021
- A host of syntactic structures (w/ sentences with shared dependency structure as indirect evidence) - Patil et al., 2024
- Entrenchment vs. affectedness in English passivization - Leong and Linzen, 2024/6
- Paired focus Constructions – Scivetti et al., 2025
- Role of function words in syntax learning – Yang et al., 2026
- NPN Constructions (day to day, face to face) – Scivetti et al., 2026, *forthcoming*
- Word order preferences (det num adj noun) – Popadich and Steinert-Threlkeld, 2026
- Unlike Coordination – Luo and Steinert-Threlkeld, 2026
- Differential argument marking - Deng et al., 2026


Controlled Rearing
Pretrain
LM
Behavior
compare!
Systematic Manipulation, informed by hypotheses
Pretrain
LM
Behavior

A systematic framework for generating novel experimental hypotheses from language models

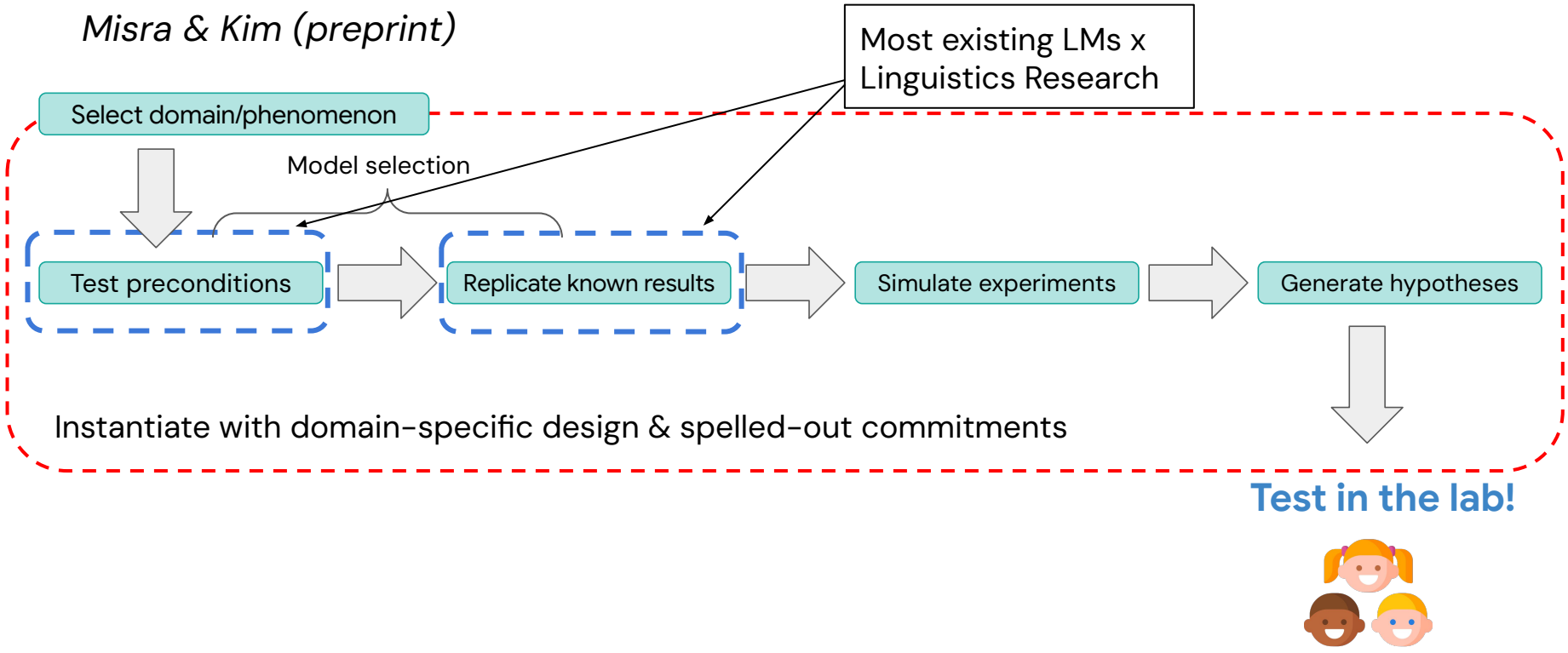
Kanishka Misra[*,1]
Department of Linguistics
The University of Texas at Austin
kmisra@utexas.edu

Najoung Kim[*,2]
Department of Linguistics
Boston University
najoung@bu.edu



V1 (2024) → V2 (2026)

A fleshed out hypothesis generation framework for neural networks

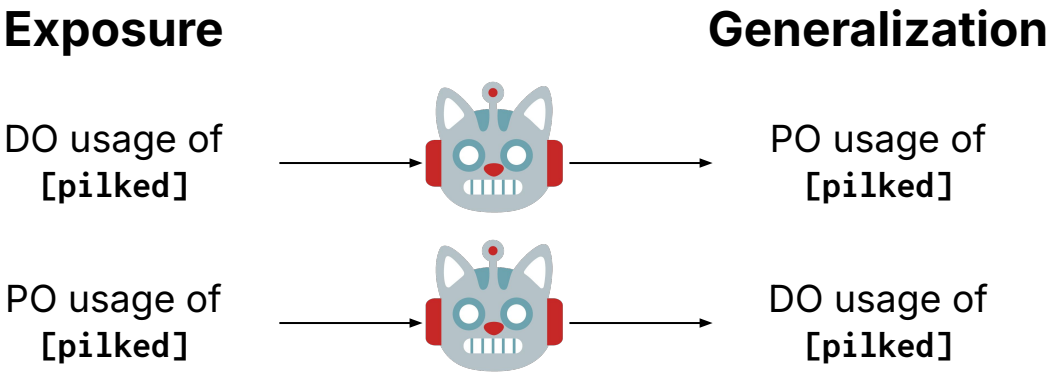


Slide due to Najoung Kim (slightly edited by me)

English Cross-dative Generalization

Double Object (DO): she **[pilked]** me the ball.

Prepositional (PO): she **[pilked]** the ball to me.



Starts at 3;0 – 4;0 for English speaking children!

Conwell and Demuth, 2007, Arunachalam, 2017

If the learner is exposed to ***pilked*** in **construction A**, then what properties of this exposure **facilitate** or **constrain** the learner's usage of ***pilked*** in **construction B**?

Large hypothesis space (Information Structure)

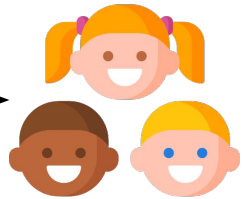
Theme/Recipient
Feature Space

Pronominality
Animacy
Length
Definiteness
Givenness

Prior Experimental Work

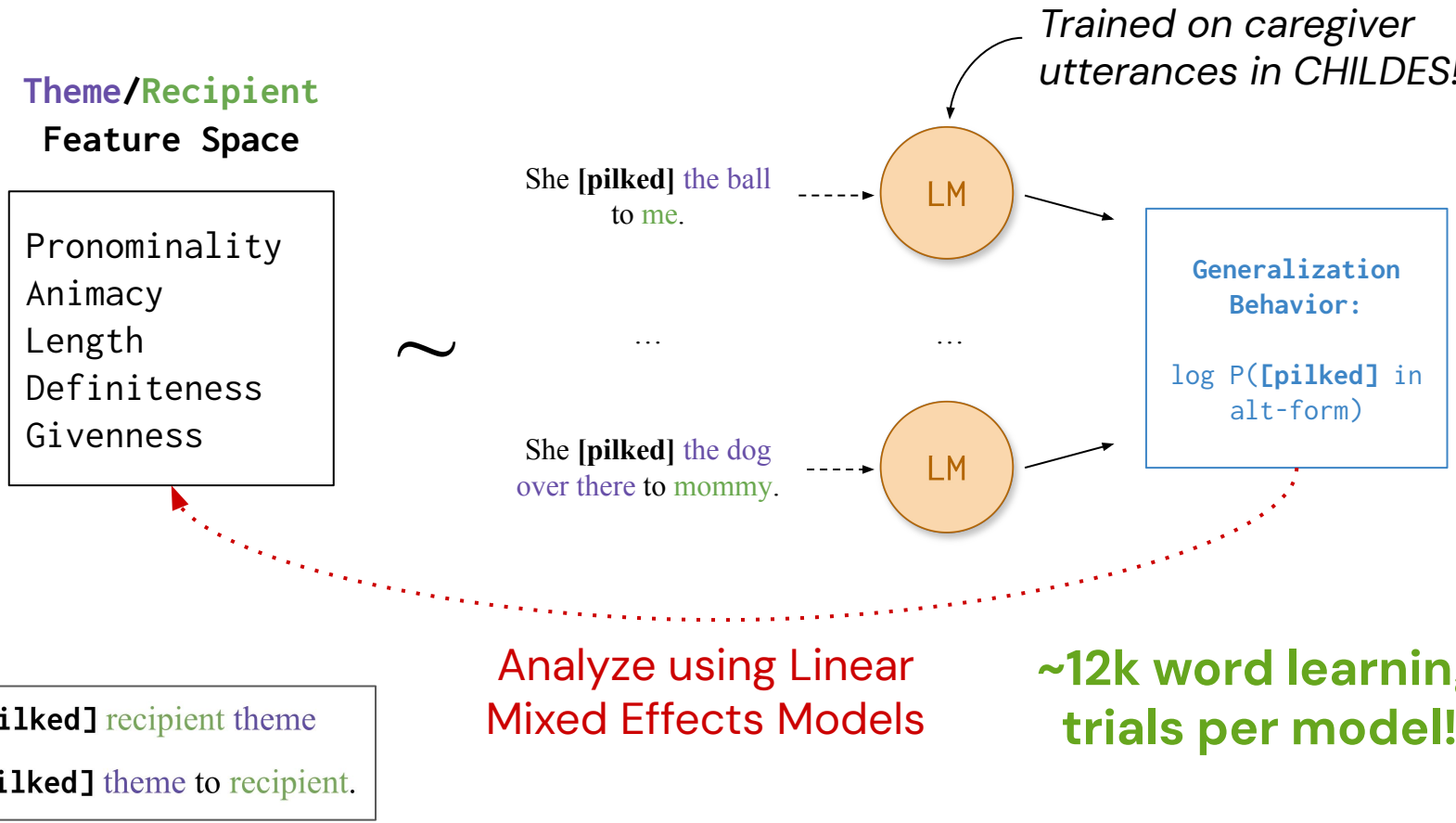
(Conwell and Demuth, 2007; Arunachalam, 2017)

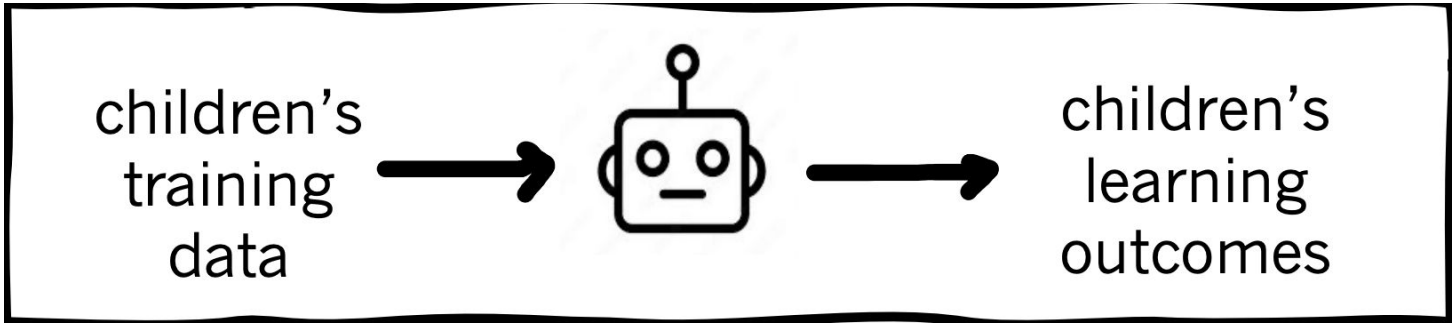
Do you see the ball?
toby is **[pilking]** the ball to me



Recipient: pronoun, animate, short, definite, given
Theme: nominal, inanimate, short, definite, given

Thompson et al., 1995; Bresnan et al., 2007; de Marneffe et al., 2012


Theme/Recipient
Feature Space
Pronominality
Animacy
Length
Definiteness
Givenness
~
Trained on caregiver utterances in CHILDES!
She [pilked] the ball to me.
She [pilked] the dog over there to mommy.
LM
LM
Generalization Behavior:
log P([pilked] in alt-form)
Analyze using Linear Mixed Effects Models
~12k word learning trials per model!
DO: [pilked] recipient theme
PO: [pilked] theme to recipient.

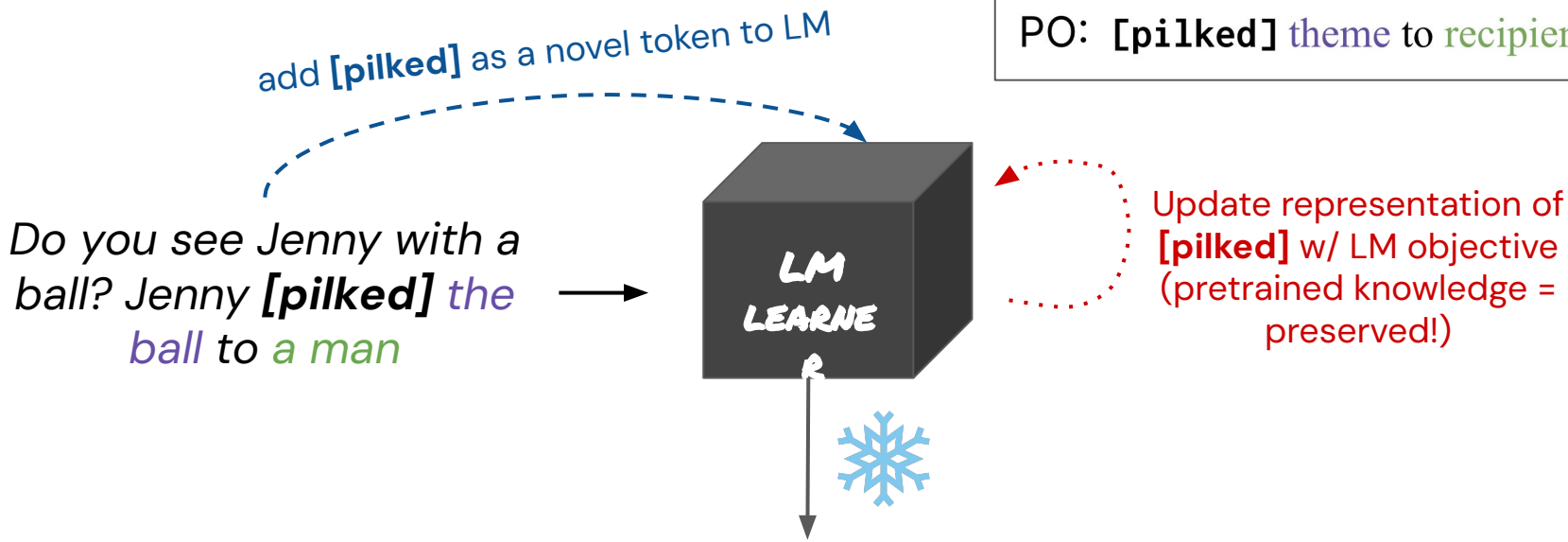


– Michael Frank's *CoNLL* 2026 keynote

Novel verb learning simulations

DO: **[pilked]** recipient theme

PO: **[pilked]** theme to recipient.



$$log\ P(\ \text{they } \textbf{[pilked]} \text{ him a book.}\ /\ \text{he } \textbf{[pilked]} \text{ the cat some milk.}\ /\ \text{she } \textbf{[pilked]} \text{ mommy the orange.}\ /\ \text{daddy } \textbf{[pilked]} \text{ him the block.}\ )$$

Rumelhart and Todd (1993); Rogers and McClelland, 2004; Kim and Smolensky, 2021

Preconditions for generating hyps. from LMs

DO: [pilked] recipient theme

PO: [pilked] theme to recipient.

Learn alternation preferences for <u>known</u> verbs...

She **kicked** the ball to me
She **kicked** me the ball

She **explained** the book to me
*She **explained** me the book

Models only see these!

...without direct evidence!

Replicate known cross-dative gen. behavior

DO → PO › PO → DO
Asymmetric Cross-dative Gen
(Conwell and Demuth, 2007)

PO → DO › DO → DO
Advantage of Cross-structure training
(Arunachalam, 2017; *in tandem with* Rowland and Noble 2010)

Prereq: Dative Constructions and Information Structure

Recipient: me
Theme: a big red ball with stars on it

DO: She gave me a big red ball with stars on it.

PO: She gave a big red ball with stars on it to me.

Recipient: a man in the hawaiian shirt
Theme: the ball

DO: She gave a man in the hawaiian shirt the ball.

<

PO: She gave the ball to a man in the hawaiian shirt.

Bresnan et al., 2007; de Marneffe et al., 2012

Prereq: Dative Constructions and Information Structure

Recipient: me
Theme: a big red ball with stars on it

Recipient: a man in the hawaiian shirt
Theme: the ball

Both the DO and PO are associated with specific Information structure properties!

both want their first argument to be:

DO: She gave me a big red ball with stars on it.

DO: She gave a man in the hawaiian shirt the ball.

```
short > long
definite > indefinite
given > new
pronoun > nominal
...
```

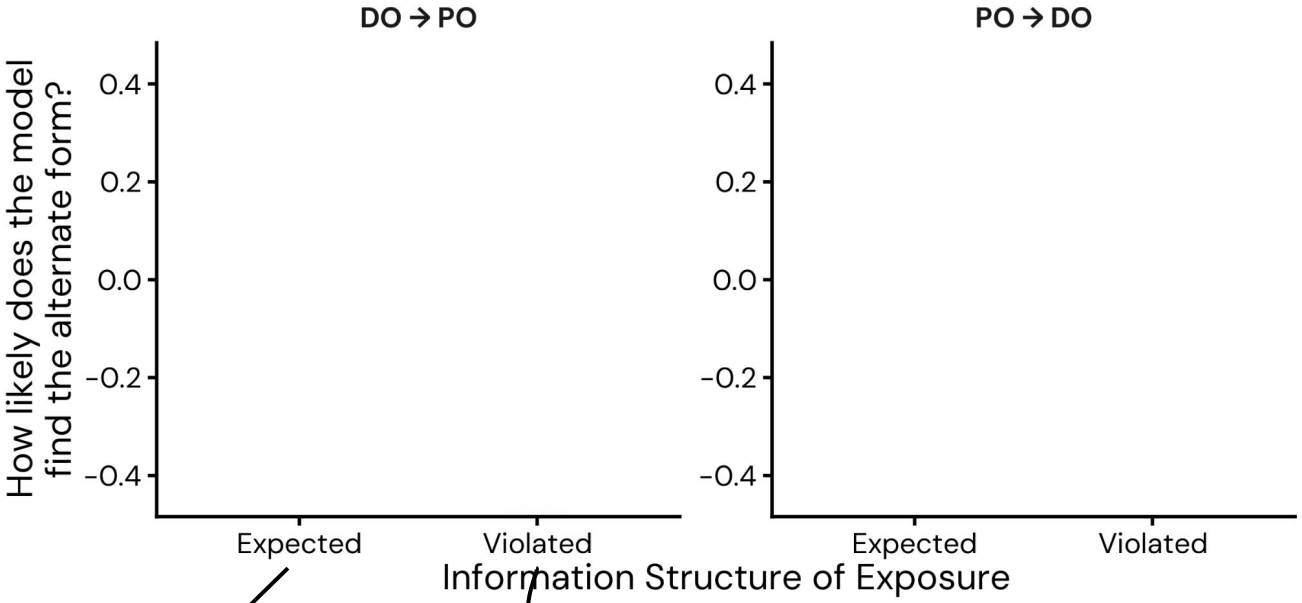
<

PO: She gave a big red ba stars on it to me.

e gave the ball to a man in the hawaiian shirt.

Bresnan et al., 2007; de Marneffe et al., 2012

A snapshot of (critical) results

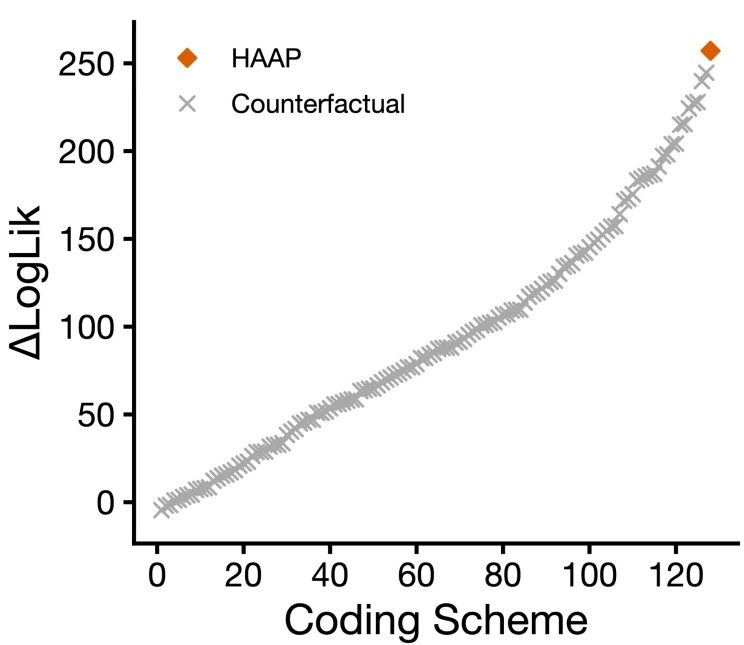


short > long
definite > indefinite
given > new
...

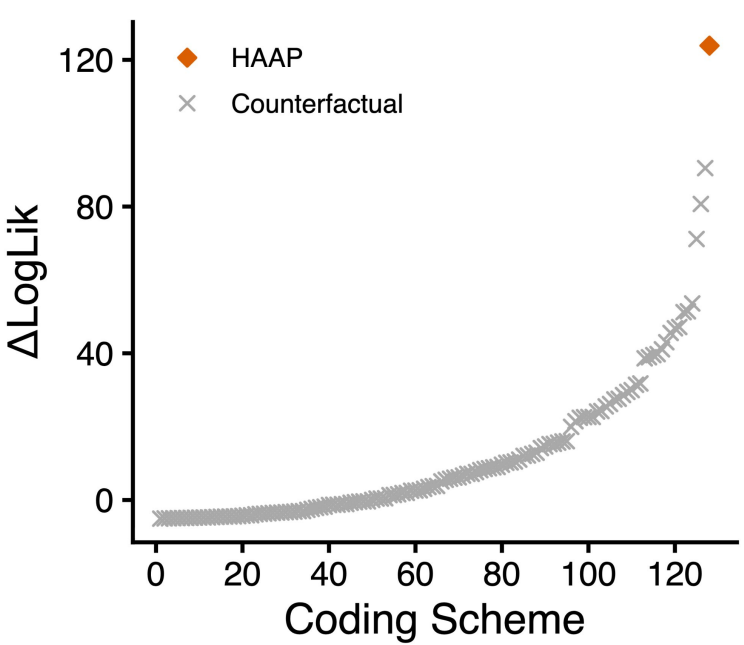
long > short
indefinite > definite
new > given
...

Considering all possible ways to encode information structure

DO → PO


HAAP
Counterfactual
ΔLogLik
250
200
150
100
50
0
0
20
40
60
80
100
120
Coding Scheme

PO → DO


HAAP
Counterfactual
ΔLogLik
120
80
40
0
0
20
40
60
80
100
120
Coding Scheme

Intuition behind Generated Hypothesis:

Recipient: the baby
Theme: a yummy chocolate cookie

Observed: PO
She ***[pilked]*** a yummy chocolate cookie to the baby.

[pilked] less likely to be used in DO

Expected: DO
(recipient shorter than theme; recipient definite theme indefinite, etc.)

*"This is the **DO** info-structure, but I repeatedly see **[pilked]** being used in **PO**; therefore, **DO** usage must be unlikely!"*

Test in lab





duckling
LAB

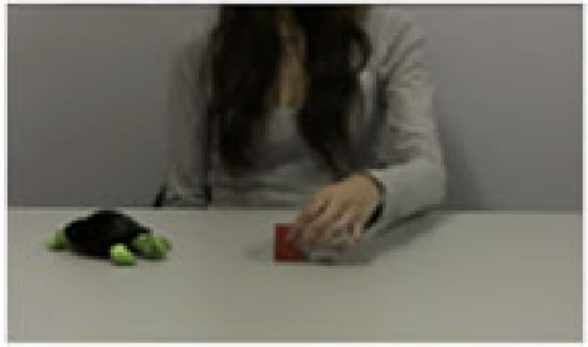
Anthony Yacovone (BU)

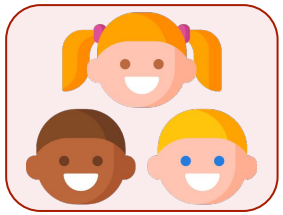
Repeatedly observing ***pilked*** used in **construction A** in a context where **construction B** was expected, is evidence *against* the usage of ***pilked*** in **construction B**

"Preemption from violation of information structure"

Goldberg 1995; Brooks and Zizak, 2002; Boyd and Goldberg 2011; Samara et al., 2025

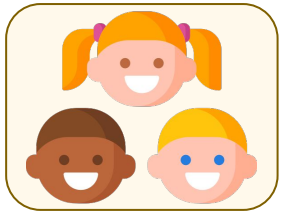
Familiarization/ Training





She is **pilking** a very cute turtle the block.

(Violated)

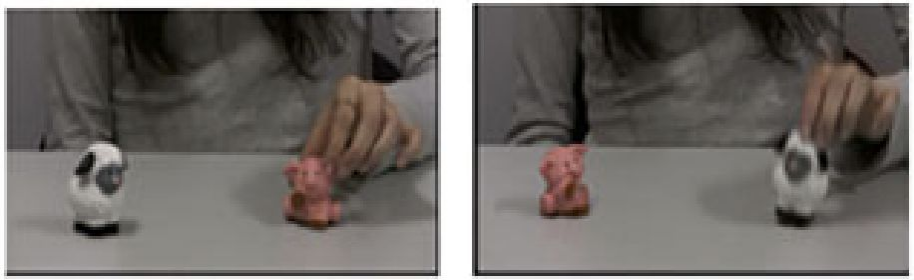


She is **pilking** him a block.

(Expected)

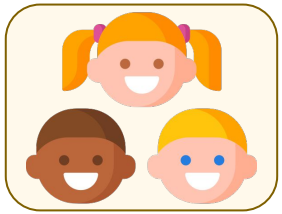
Pictures from Arunachalam, 2017; stimuli ours

Testing



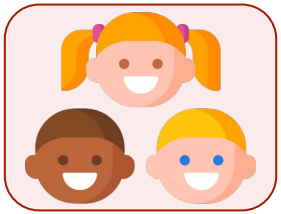
Let's see some more **pilking**!
Where is she **pilking** the pig to the sheep?

Prediction:



(expected)

>



(violated)

Pictures from Arunachalam, 2017; stimuli ours

A promise (point?) of connectionism!

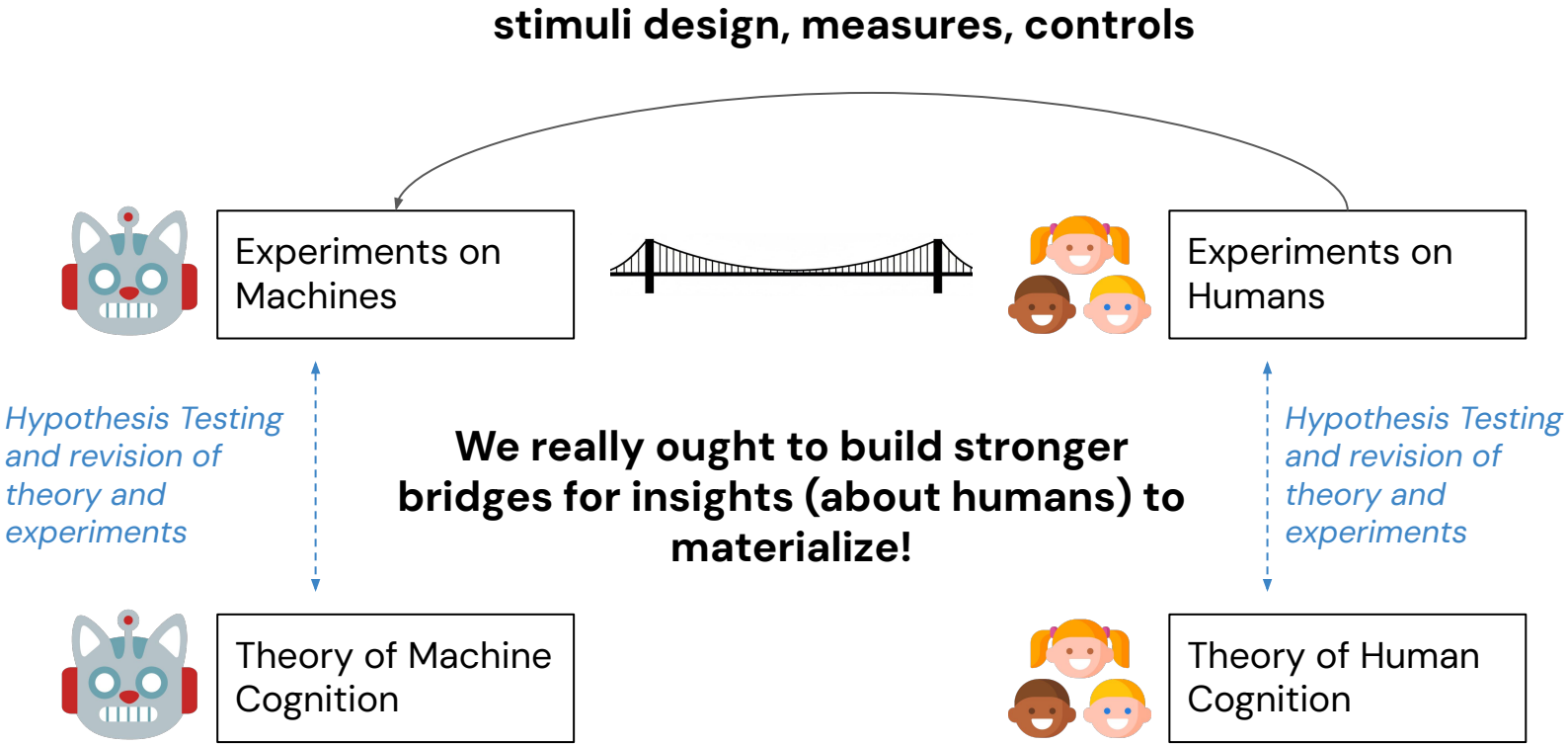


They then took this machine and they "turned the handle on the crank" and said, "let's see what else it will do", and it turned out to generate a bunch of behaviors that also were not obvious, but which they then proceeded to test with human subjects...and lo and behold the human subjects, to everybody's surprise, acted just like the machine.

Jeff Elman on Rumelhart and McClelland (1982) during an Interview with Roger Bingham at CogSci 2010



Lesson?


stimuli design, measures, controls
Experiments on Machines
Experiments on Humans
Hypothesis Testing and revision of theory and experiments
We really ought to build stronger bridges for insights (about humans) to materialize!
Hypothesis Testing and revision of theory and experiments
Theory of Machine Cognition
Theory of Human Cognition

Food for thought for:

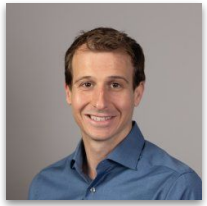
Developmentalists: What kinds of statistical learners are children (and is there a monolithic learner?); how do we account for developmental stages when theorizing about direct/indirect evidence? What are the phenomena that people really care about when talking about the input?

- Anaphoric *one*; Question formation; determiner + noun generalization (ongoing work with Afra Alishahi, Susan Goldin Meadow, and others)

Analytical folks: A theory of how indirect evidence is picked up by neural networks? Are there ways to explain the learning dynamics w/ and w/o indirect evidence?

I am not sure if this counts as analytical: Influence functions x Controlled rearing?

You!





Thanks!

Looking for PhD students!

KOALAB
Kanishka's Origin of Abstractions Lab

The University of Texas at Austin
Donald D. Harrington Fellows Program

Backup

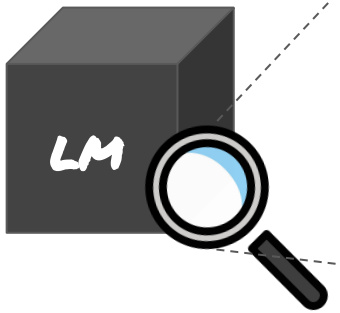
Noun!

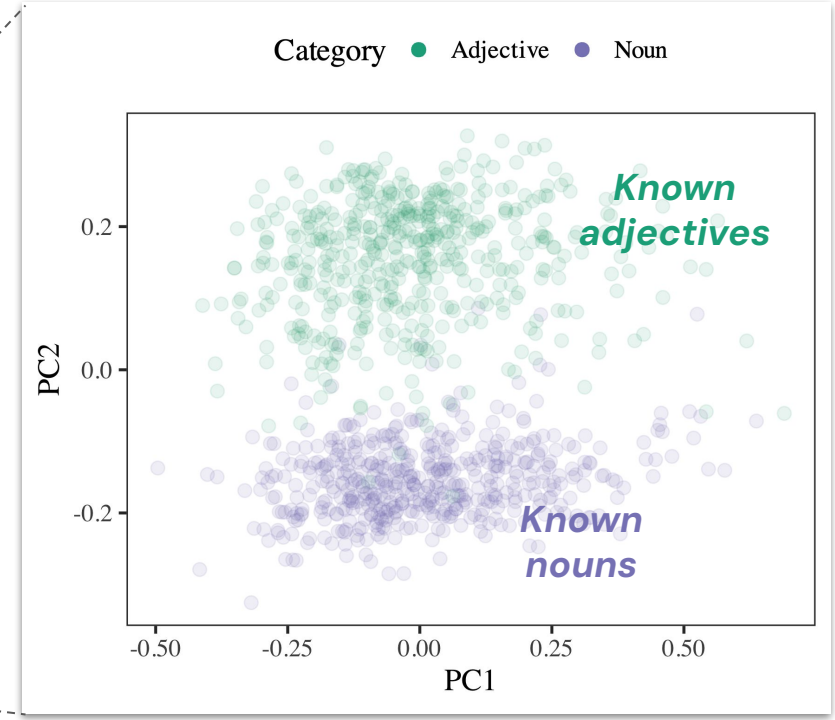
I saw the *wug* run away.

The book was very *dax*.

adjective!

repeated exposure





A simplified view of the LM parameters

(Misra & Kim, BUCLD 2023, Position Piece, forthcoming)

Main Experiment Analyses

DO: [pilked] recipient theme

PO: [pilked] theme to recipient.

Discourse Prominence Scales:

given > new pronoun > non pronoun definite > indefinite short > long

Positional Prominence Scale:

Earlier > Later

Multiple scales are in play;
We need a scoring scheme!

Harmonically Aligned | **Not/Less Harmonically Aligned**

I [pilked] ***her*** ***some chocolate cookies***

pronoun > not pronoun

Earlier > Later

😊

I [pilked] ***some chocolate cookies*** *to* ***her***

pronoun > not pronoun

Earlier > Later

😖

Aissen, 1999, 2003

What about animacy?

- In general harmonic alignment and production observations (Aissen, Macdonald, Bock, others), animates tend to precede inanimates, but this was not as empirically borne out in corpus analyses as other features:
    - Only for Recipients for DO and only in adults (Bresnan et al., 2007, de Marneffe et al., 2012)
- Animacy also interacts with the semantics of the event (Goldberg 1995, Beavers, 2011):
    - Inanimates make poor recipients in possession transfer events (Beavers, 2011).
- Prototypically, recipients should be more animate in general than themes.

Harmonic Alignment and Animacy Prototypicality Scores (HAAP)

Measures compatibility of the exposure stimulus with the harmonic alignment of argument features as well as prototypical animacy assignments.

Three components, conditional on what the exposure dative is:

1. Discrete theme features (pronominality, animacy, definiteness, givenness)
2. Discrete recipient features (same)
3. Relative Length differences (second PostVerbal arg - first PostVerbal arg)

HAAP Scoring for discrete features

DO: **[pilked]** recipient theme

PO: **[pilked]** theme to recipient.

For PO exposure:

- +1 each if theme = pronominal, definite, given, inanimate (max 4)
- +1 each if recipient = nominal, indefinite, new, animate (max 4)
- 0 otherwise (min = 0)

For DO exposure:

- +1 each if recipient = pronominal, definite, given, animate
- +1 each if theme = nominal, indefinite, new, inanimate
- 0 otherwise

Length Scoring

DO: **[pilked]** recipient theme

PO: **[pilked]** theme to recipient.

$$\Delta\text{length} = \begin{cases} \log(l_2 - l_1) + 1 & l_2 > l_1 \\ -\log(l_2 - l_1) + 1 & l_2 < l_1 \\ 0 & l_2 = l_1 \end{cases}$$

Example

DO: **[pilked]** recipient theme

PO: **[pilked]** theme to recipient.

Do you see Lucy and John? Lucy [verbed] him a chocolate cookie.

Recip Score: pron (+1), anim (+1), def (+1), given (+1)
Theme Score: np (+1), inanim (+1), indef (+1), new (+1)
Length diff Score: log(2)+1 = 1.69
HAAP: 9.69

Is HAAP the best coding scheme?

We have made theoretical commitments to the fact that both postverbal arguments should be harmonically aligned and follow animacy prototypicality.

How well does this reflect the reality of the model behavior?

We consider all possible coding schemes: 2^{8-1}

- 8 binary features: 2^8 → halved due to symmetry
- rating something as 00000000 is the same as 11111111 for the regression model

HAAP and its counterfactuals

Look! It's Laura and **Eve**. Laura ***pilked*** **her a box with blocks**!

Theme: [−pron,−animate,−given,−def] **Recipient:** [+pron,+animate,+given, +def]

HAAP (DO → PO)

Theme: [1,1,1,1] = 4
Recipient: [1,1,1,1] = 4
Length: log(3)+1 = 2.09
Score: 10.09

Counterfactual HAAP
(indefinite recipients → 1, new recipients → 1, given theme → 1; definite theme → 1, animate theme → 1):

Theme: [1,0,0,0] = 1
Recipient: [1,1,0,0] = 2
Length: log(3)+1 = 2.09
Score: 5.09

Is HAAP the best coding scheme?

- **Full model:** `outcome ~ haap + (1|seed) + (1|template)`
- **Null Model:** `outcome ~ 1 + (1|seed) + (1|template)`
- `Compare ΔLogLik`

Outcomes:

DO exposures: log-prob of verb used in PO (generalization set)

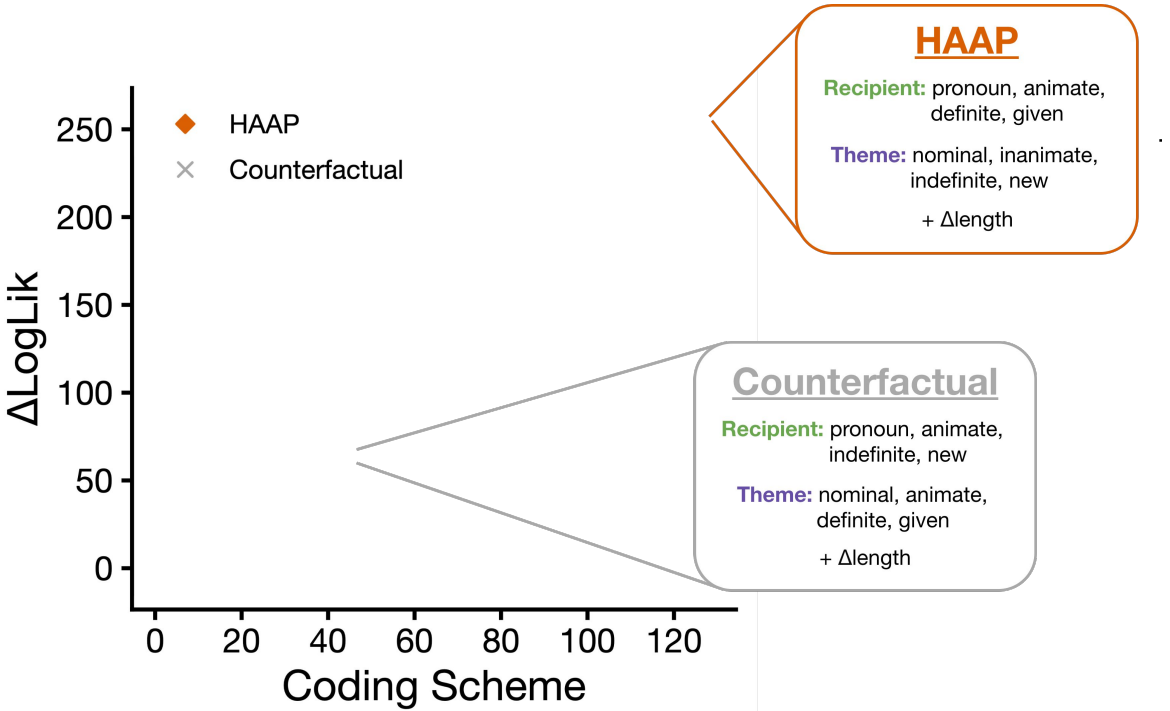
PO exposures: log-prob of verb used in DO (generalization set)

Results!

DO → PO

DO: **[pilked]** recipient theme

PO: **[pilked]** theme to recipient.



HAAP

Recipient: pronoun, animate, definite, given

Theme: nominal, inanimate, indefinite, new

+ Δlength

Counterfactual

Recipient: pronoun, animate, indefinite, new

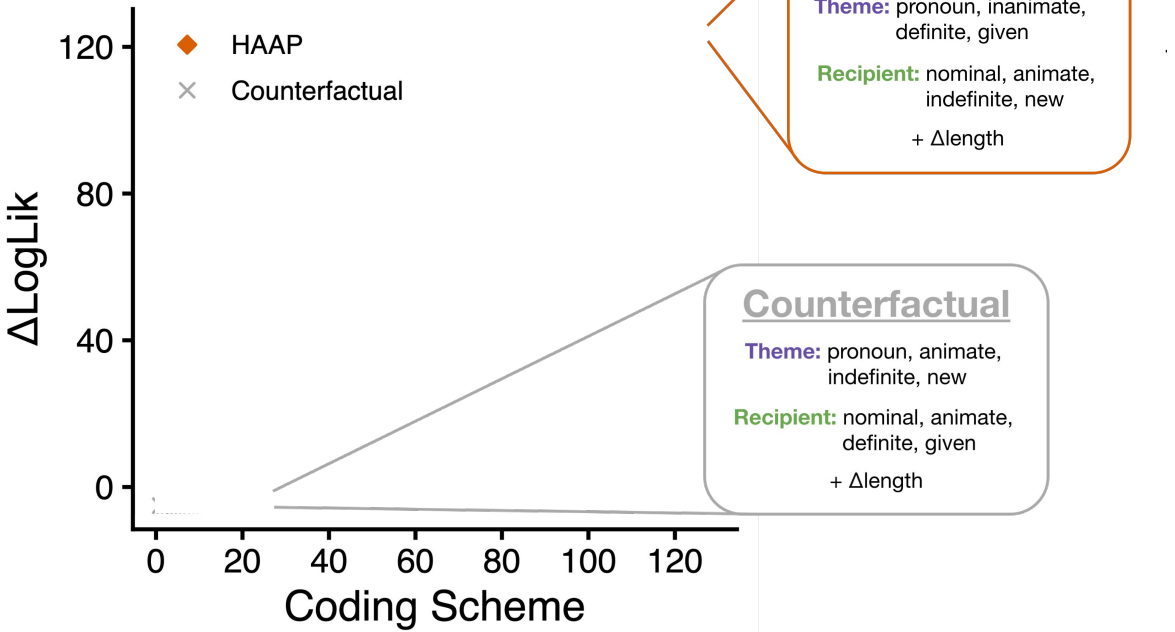
Theme: nominal, animate, definite, given

+ Δlength

PO → DO

DO: **[pilked]** recipient theme

PO: **[pilked]** theme to recipient.



HAAP

Theme: pronoun, inanimate, definite, given

Recipient: nominal, animate, indefinite, new

+ Δlength

Counterfactual

Theme: pronoun, animate, indefinite, new

Recipient: nominal, animate, definite, given

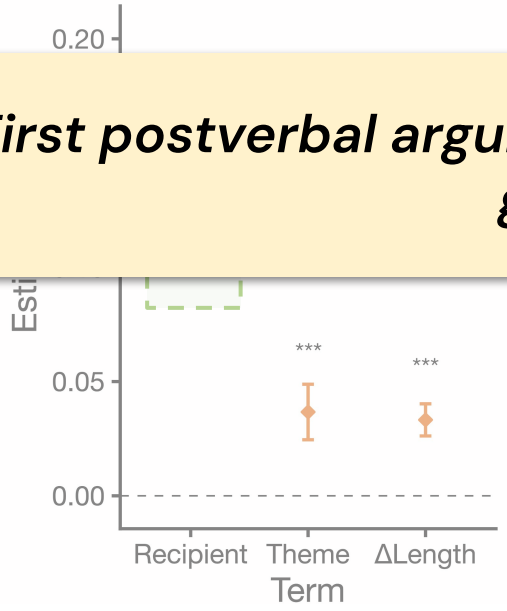
+ Δlength

haap = haap$_{recipient}$ + haap$_{theme}$ + Δlength

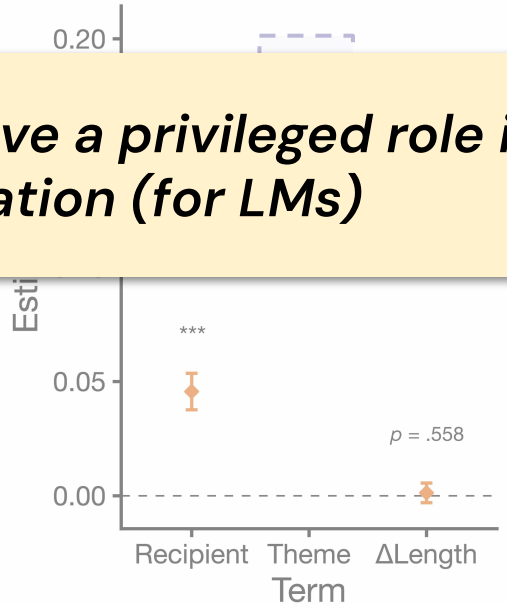
DO: [pilked] recipient theme

PO: [pilked] theme to recipient.

DO → PO



PO → DO



First postverbal arguments have a privileged role in cross-dative generalization (for LMs)

Quantitative Tests:

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

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

Syntactic
log-**o**dds
ratio

Joint probability under a unigram model trained on the same corpus.

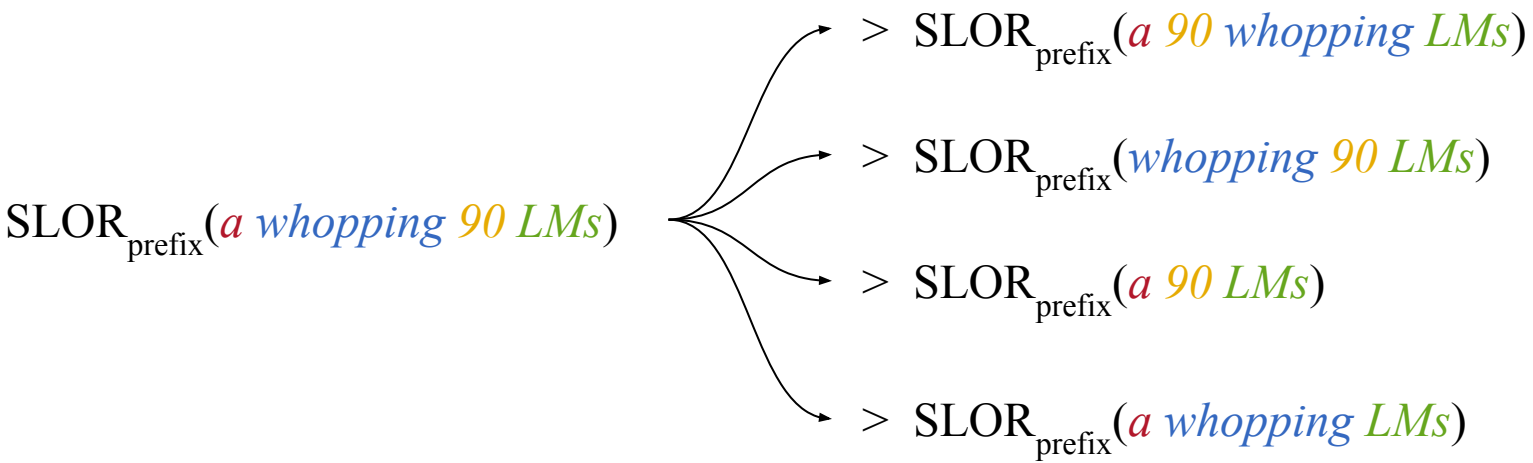
Important if we are manipulating corpora – to account for differences in frequencies!

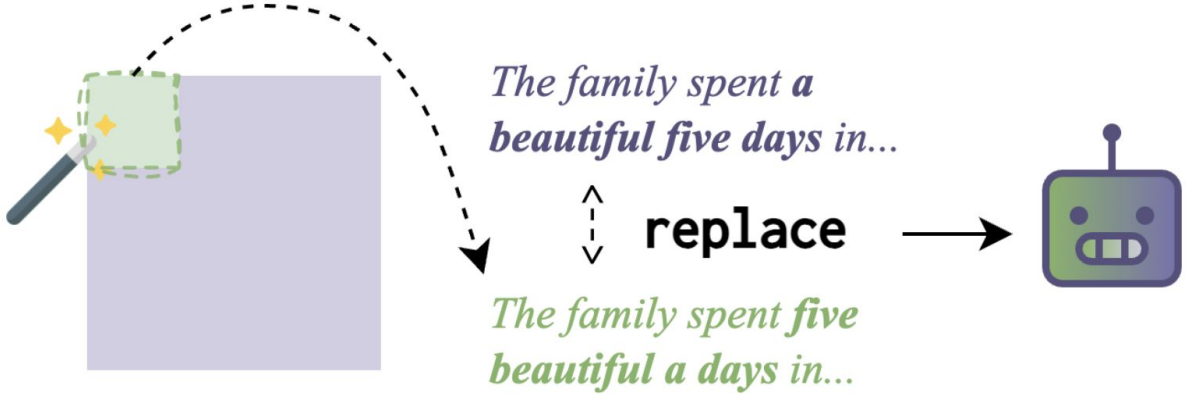
Pauls and Klein (2012), Lau et al. (2017)

Quantitative Tests:

Chance: ⅕ = 20%

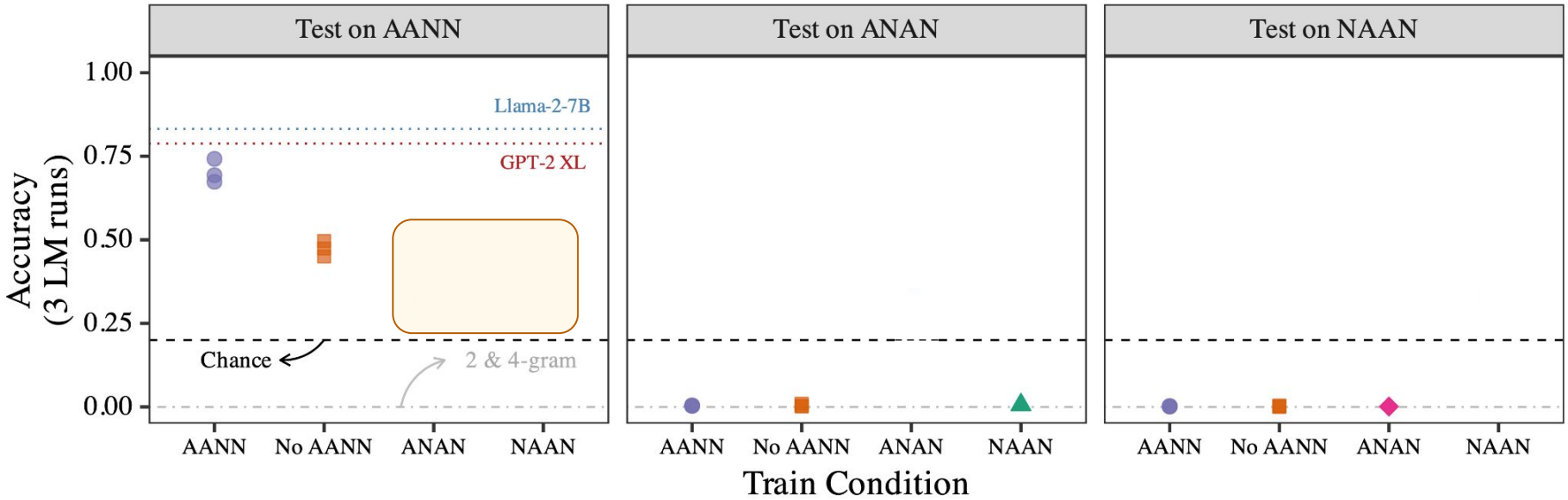
Accuracy = Proportion of time:




The family spent a beautiful five days in...
replace
The family spent five beautiful a days in...

AANN ↔ NAAN

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


Test on AANN
Test on ANAN
Test on NAAN
Accuracy
(3 LM runs)
1.00
0.75
0.50
0.25
0.00
Llama-2-7B
GPT-2 XL
Chance
2 & 4-gram
AANN
No AANN
ANAN
NAAN
Train Condition

Interim Takeaways:

- LMs learn about the AANN…
- …even without encountering a single instance…
- …more strongly than they learn its counterfactual variants…
- …in a way that extends to lexical semantic constraints!

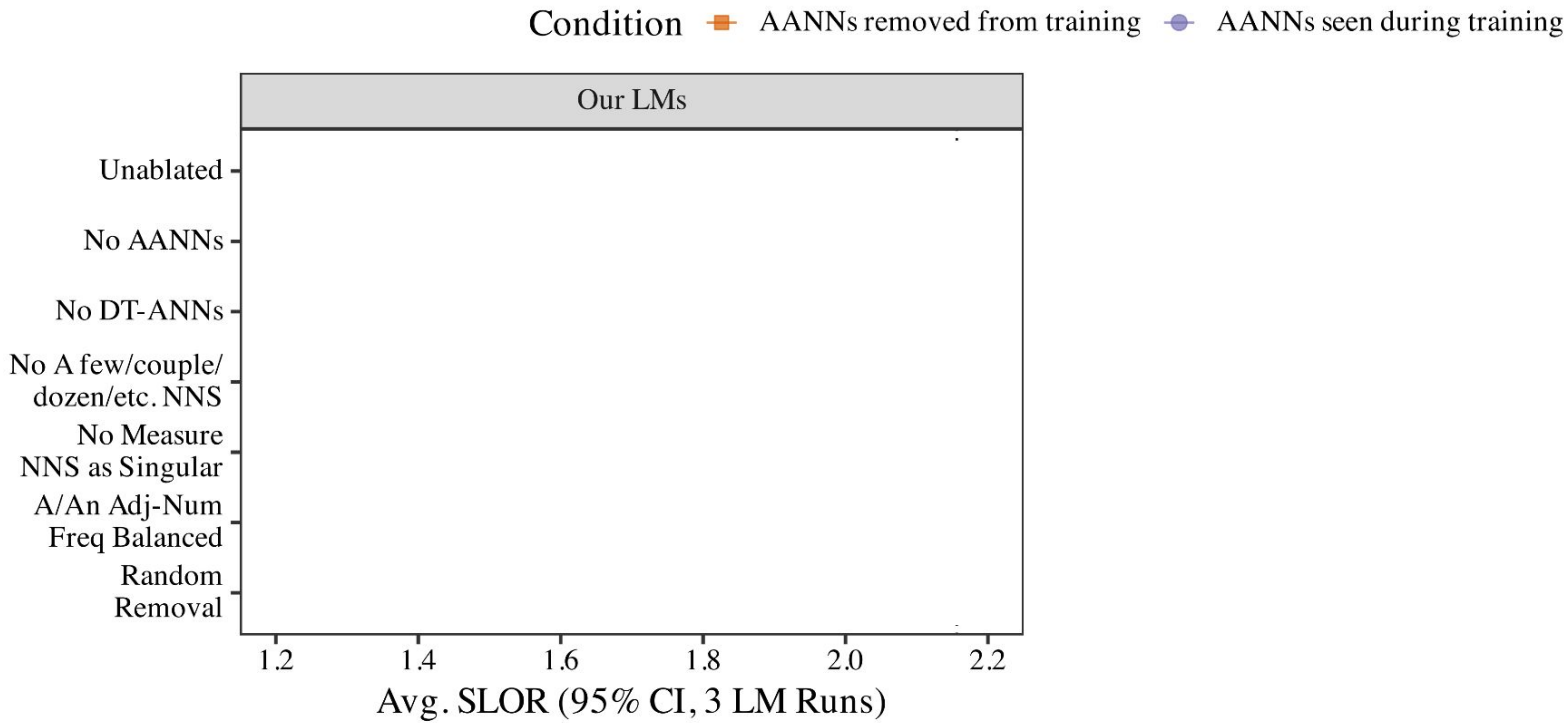
LMs could be relying on *indirect* evidence in the training data that might contribute to learning of this construction!

Phenomenon/Manipulation	**Example/`Desc.`**	**Freq.**
AANN	**a fine eighteen months**	2,448
DT ANN	**the usual forty dollars fine**	15,781
A few/couple/dozen/etc. NNS	**a few plums**	55,373
Measure NNS with SG verbs and/or indef. articles	**6 months is a long time**	62,744
A/An + ADJ/NUM balancing	`enforce freq. balance`	571,874
Random removal (**control**)	`randomized ablation`	571,874

Table 2: Manipulated Phenomena, their examples/descriptions, and their frequency in the BabyLM corpus.

Compare orange to orange, purple to purple!

Same tokenizer and training data



Cross-modal Generalization

During Training:



Visual Features

\+ Is there a **crow** in this image?

Does not "see" what a "**bird**" looks like



All it knows about the word ***bird*** is from "language exposure"

LM → **Yes**

During *Testing*:



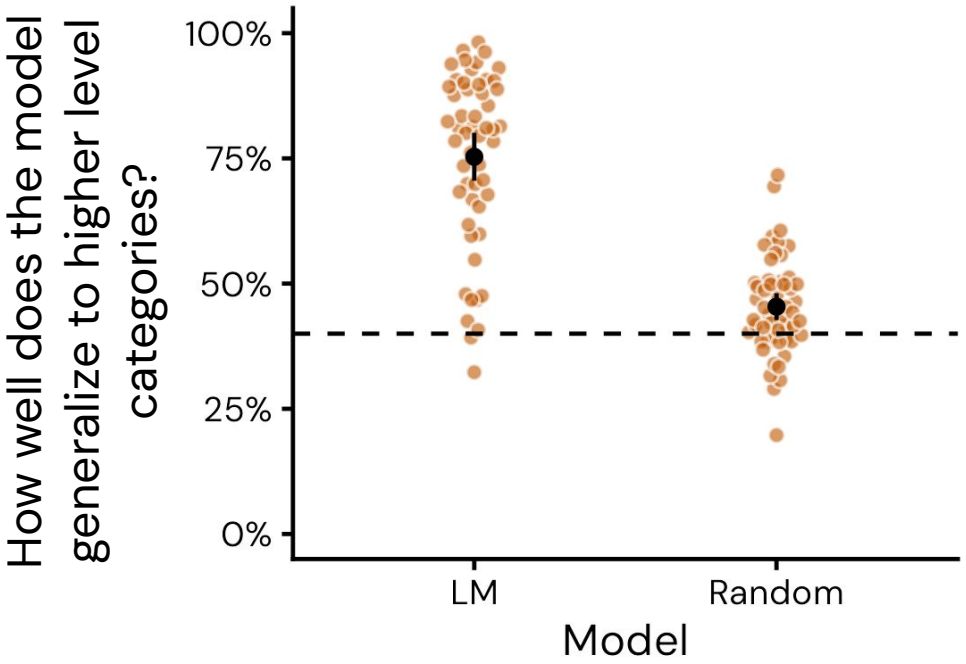
Visual Features

\+ Is there a **bird** in this image?

Crows are birds!

LM → **?**

Results



Higher order categories for nouns can be recovered from statistical patterns in language alone....

.... in a manner that generalizes to a different modality (vision!)

Dative-alternation preference tests for known verbs

- Focuses on verbs that are **not observed to alternate** in the training set.
- Alternation status decided by whatever Beth Levin says (Levin, 1993).
- NABA (Not Alternating in training But actually Alternating)
  - 6 DOs (e.g., guarantee, traded, etc.)
  - 6 POs (e.g., delivered, kicked, etc.)
- NANA (Not Alternating in training and Not Alternating)
  - 2 DOs (e.g., cost, wished)
  - 12 POs (e.g., addressed, explained, etc.)

Does the model predict that NABA verbs are more likely to alternate than NANA verbs?

Difference in DO-PO preference should be greater in NABA verbs than for NANA verbs:

"Kick is more alternatable than explain"

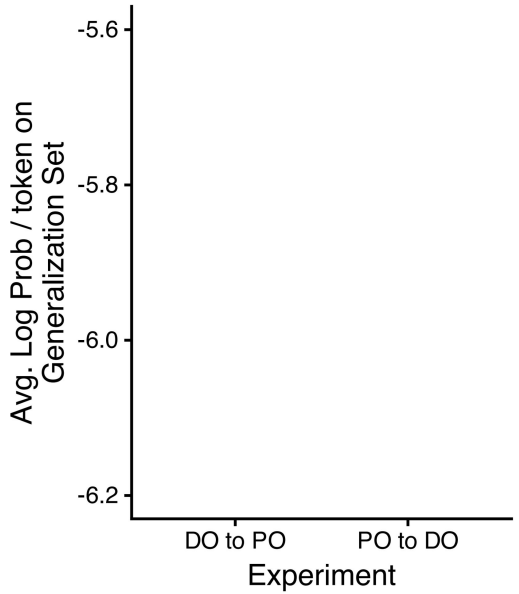
I **kicked** him the ball – I **kicked** the ball to him	>	?I **explained** him the problem – I **explained** the problem to him
NABA		NANA

Blue: attested in corpus (in terms of structure, not exact sequence)

Testing for replication of known findings about cross-dative generalization

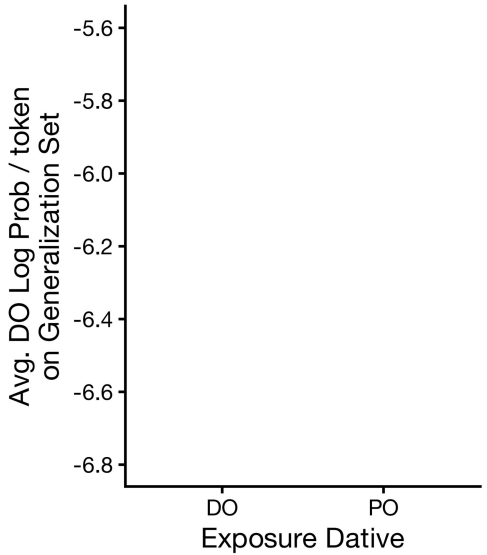
Asymmetric Cross-dative Generalization (Conwell and Demuth, 2007)

Observation: generalization from DO → PO > PO → DO



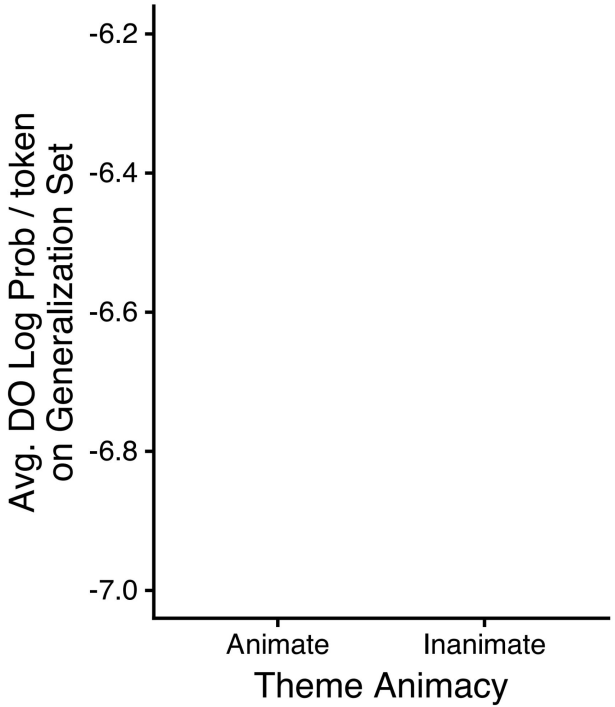
Advantage of Cross-structure Training (Arunachalam, 2017 w/ Rowland and Noble 2010)

Observation: children learn to comprehend novel verb in DO better from exposure in PO than from exposure in DO, likely due to DO being syntactically difficult (Fisher, 2002)



Note: we used a larger set of stimuli, since the feature combination used by Arunachalam only worked for 16 total sentences in our entire dataset.

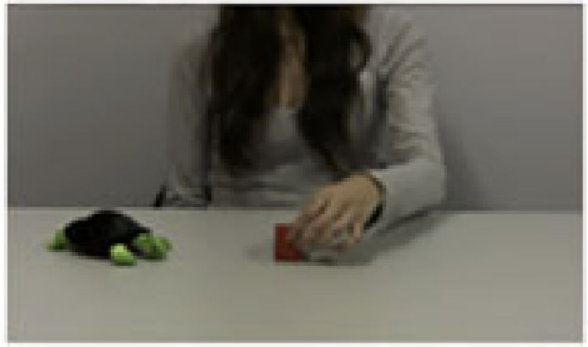
Bonus from Arunachalam: no effect of theme animacy for DO → DO

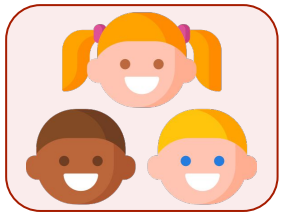


Where might these come from?

- Prevalent everywhere in children's input (De Marneffe et al., 2012; Bresnan et al., 2007 for adults)
- LMs pick up on it (Hawkins et al., 2020)
    - LMs prefer short > long (Yao et al., 2025), and this comes from indirect evidence as well (English is predominantly short-first)
    - Even in priming (Jumelet et al., 2024)

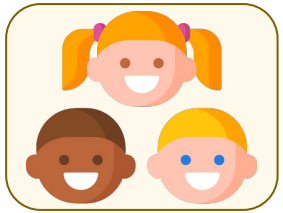
Familiarization/ Training





She is **pilking** a very cute turtle the block.

(Violated)

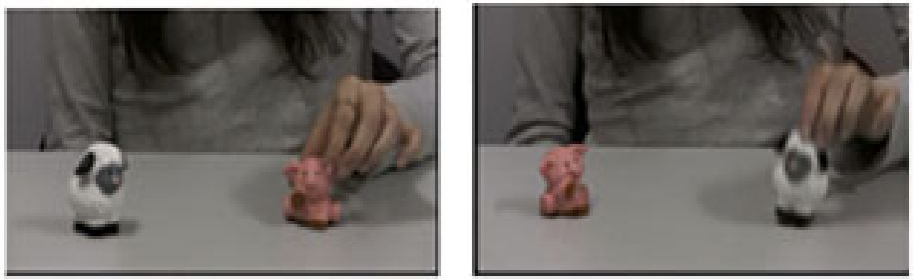


She is **pilking** him a block.

(Expected)

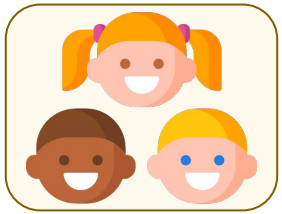
Pictures from Arunachalam, 2017; stimuli ours

Testing



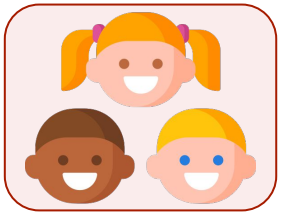
Let's see some more **pilking**!
Where is she **pilking** the pig to the sheep?

Prediction:



(expected)

>



(violated)

Pictures from Arunachalam, 2017; stimuli ours