Why it Matters That Babies and Language Models are the Only Known Language Learners

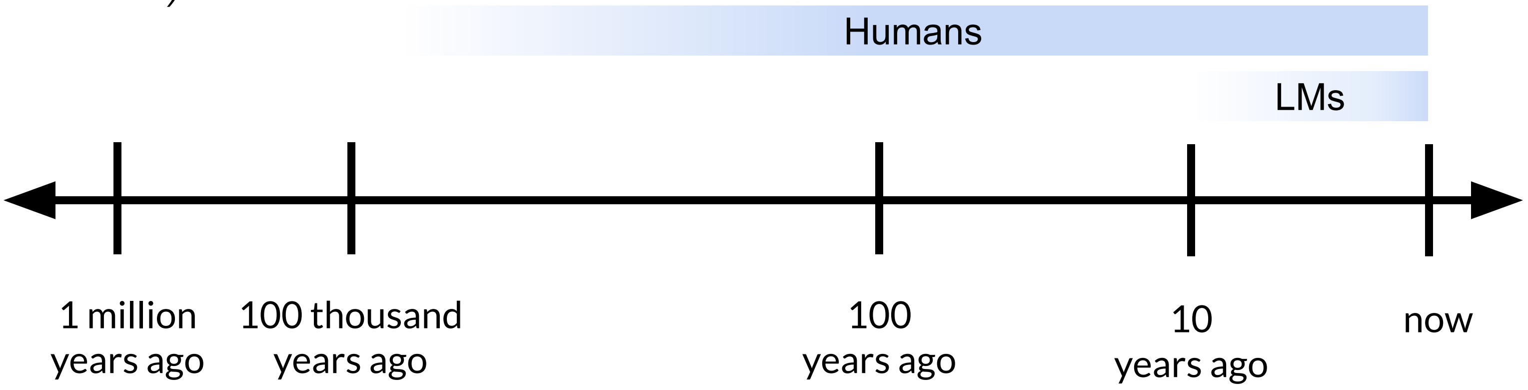
Presented by
Alex Warstadt
Assistant Professor
Halıcıoğlu Data Science Institute
Department of Linguistics

Presented at
Workshop on LLMs, Cognitive Science, Linguistics, and Neuroscience
Simons Institute
4 February 2025

For most of history, humans were the only thing in the known universe that could learn language.

In the last few years, remarkable improvements in neural language models (LMs) make us seem a little less unique.

Timeline: Things that can "learn language"
(not to scale)



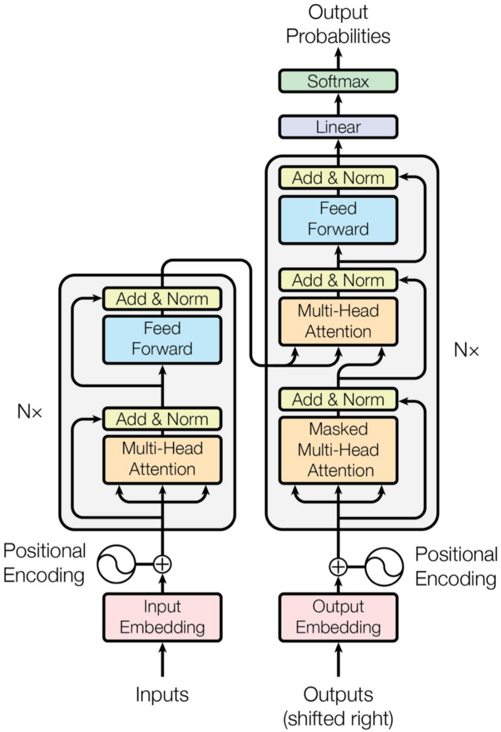
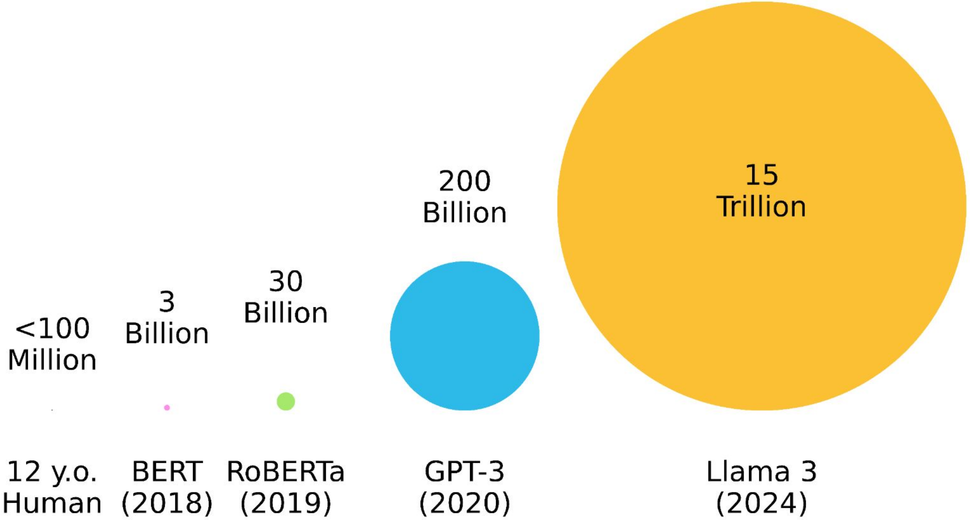


Figure 1: The Transformer - model architecture.

Figure 2: Human baby

Figure 1: The Transformer - model architecture.



# of words in learning environment

Pharaoh Psamtik
(664 – 610 BCE)

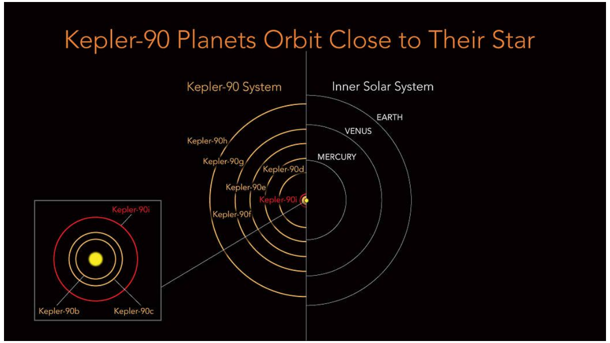
Frederick II
(1194-1250)

James IV
(1473-1513)

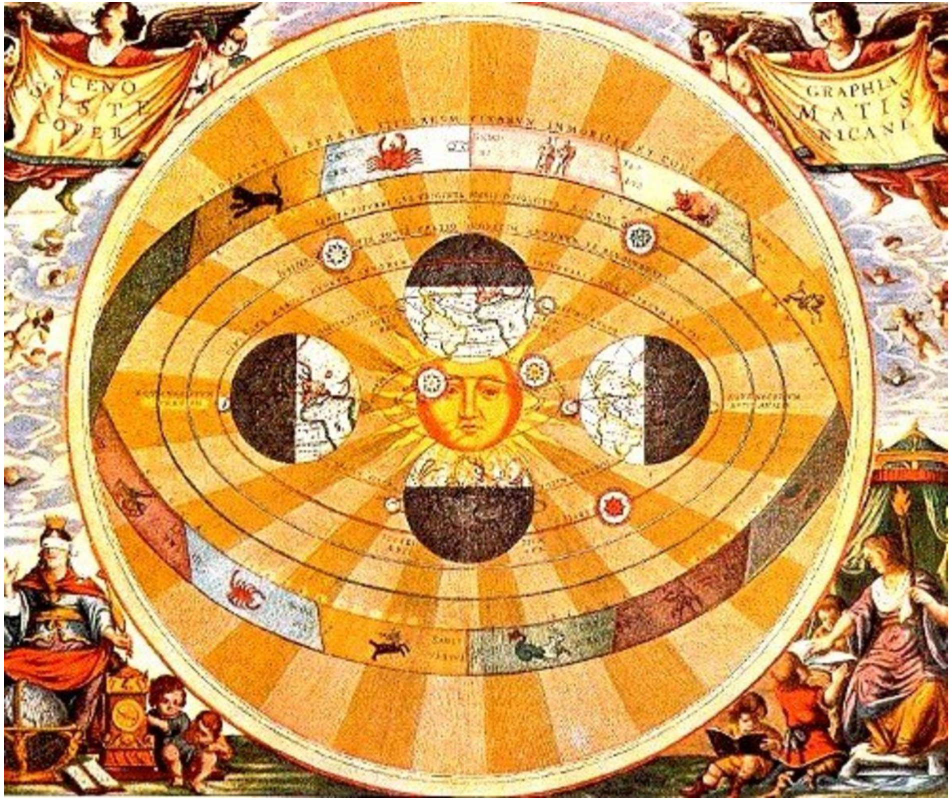
Carried out language deprivation experiments

Figure 2: Human baby

What do “Language Learners” look like in general?



By NASA/Ames Research Center/Wendy Stenzel



What can you learn with domain-general biases?

Vision

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy*,†, Lucas Beyer*, Alexander Kolesnikov*, Dirk Weissenborn*, Xiaohua Zhai*, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby*,†
*equal technical contribution, †equal advising
Google Research, Brain Team
{adosovitskiy, neilhoulsby}@google.com

ViT (>50k citations)

Protein Folding

Article

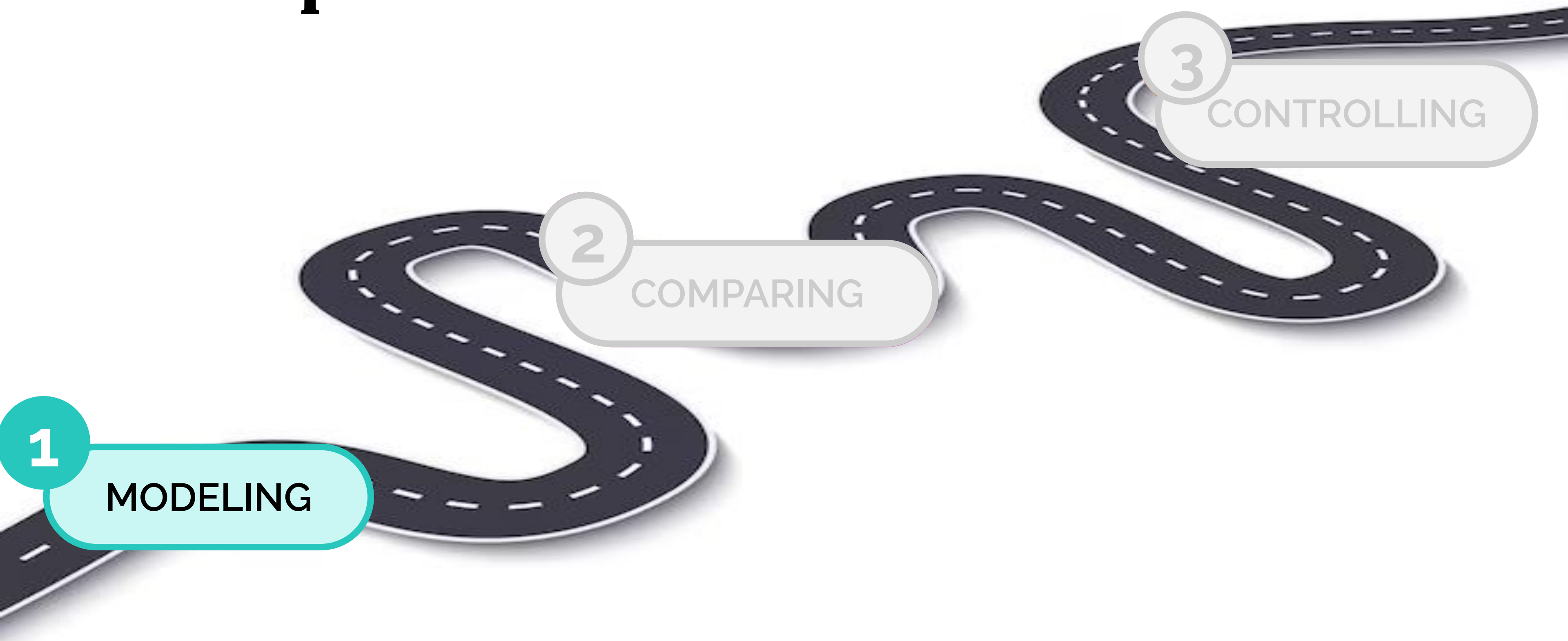
Highly accurate protein structure prediction with AlphaFold

https://doi.org/10.1038/s41586-021-03819-2
Received: 11 May 2021
Accepted: 12 July 2021
Published online: 15 July 2021
Open access
Check for updates

John Jumper[1,4]✉, Richard Evans[1,4], Alexander Pritzel[1,4], Tim Green[1,4], Michael Figurnov[1,4], Olaf Ronneberger[1,4], Kathryn Tunyasuvunakool[1,4], Russ Bates[1,4], Augustin Žídek[1,4], Anna Potapenko[1,4], Alex Bridgland[1,4], Clemens Meyer[1,4], Simon A. A. Kohl[1,4], Andrew J. Ballard[1,4], Andrew Cowie[1,4], Bernardino Romera-Paredes[1,4], Stanislav Nikolov[1,4], Rishub Jain[1,4], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[2,3], Michalina Pacholska[1], Tamas Berghammer[1], Sebastian Bodenstein[1], David Silver[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1] & Demis Hassabis[1,4]✉

AlphaFold (>30k citations)

Roadmap



The BabyLM Challenge

Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora

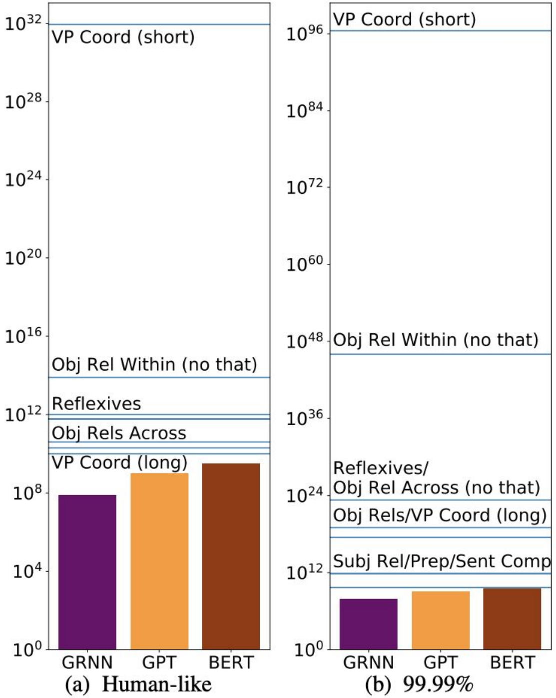
Alex Warstadt[1*] **Aaron Mueller**[2,3*] **Leshem Choshen**[4,5] **Ethan Wilcox**[1] **Chengxu Zhuang**[4]
Juan Ciro[6] **Rafael Mosquera**[6] **Bhargavi Paranjape**[8]
Adina Williams[6,7] **Tal Linzen**[9] **Ryan Cotterell**[1]
[1]ETH Zürich [2]Northeastern University [3]Technion [4]MIT
[5]IBM Research [6]MLCommons [7]Meta AI (FAIR)
[8]University of Washington [9]New York University

+ Candace Ross (FAIR)
+ Michael Hu (NYU) … in 2024

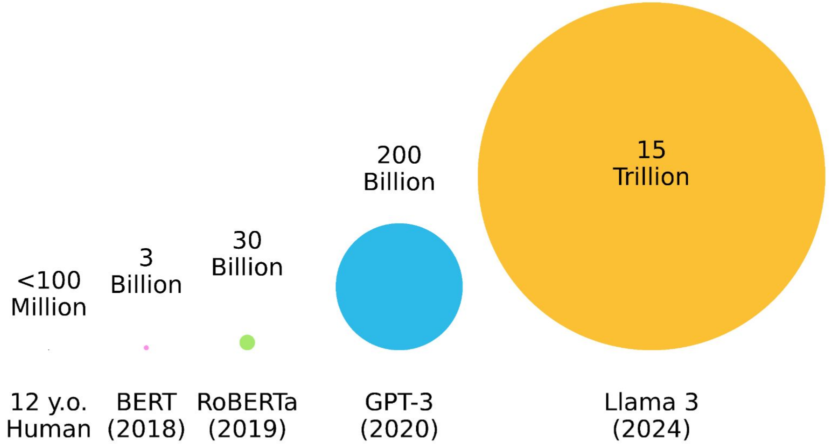
Shared task @ CoNLL 2023, 2024
Workshop @ EMNLP 2025

Humans are far better language learners than LMs in terms of data-efficiency.

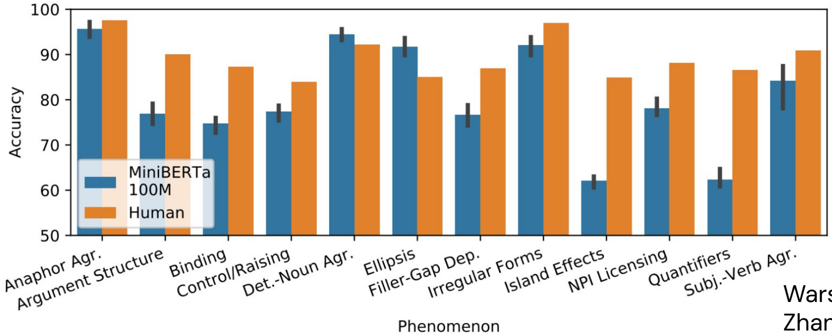
The data efficiency gap

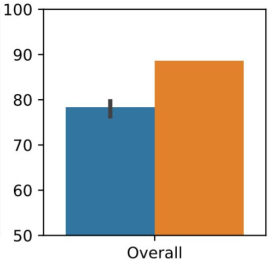


Van Schijndel, Mueller, Linzen (2019)



# of words in learning environment





Warstadt & Bowman (2022);
Zhang et al. (2020)

Motivation

1. Data efficient pretraining
2. Plausible cognitive models
3. Democratization of pretraining research

Shared Task Setup

BabyLM Tracks

BabyLM Challenge

Sample-efficient pretraining on a developmentally plausible corpus

Track 1: Strict
- 100 million words
- BabyLM dataset or BYOD
- Shared eval (grammar, generalization, NLU)

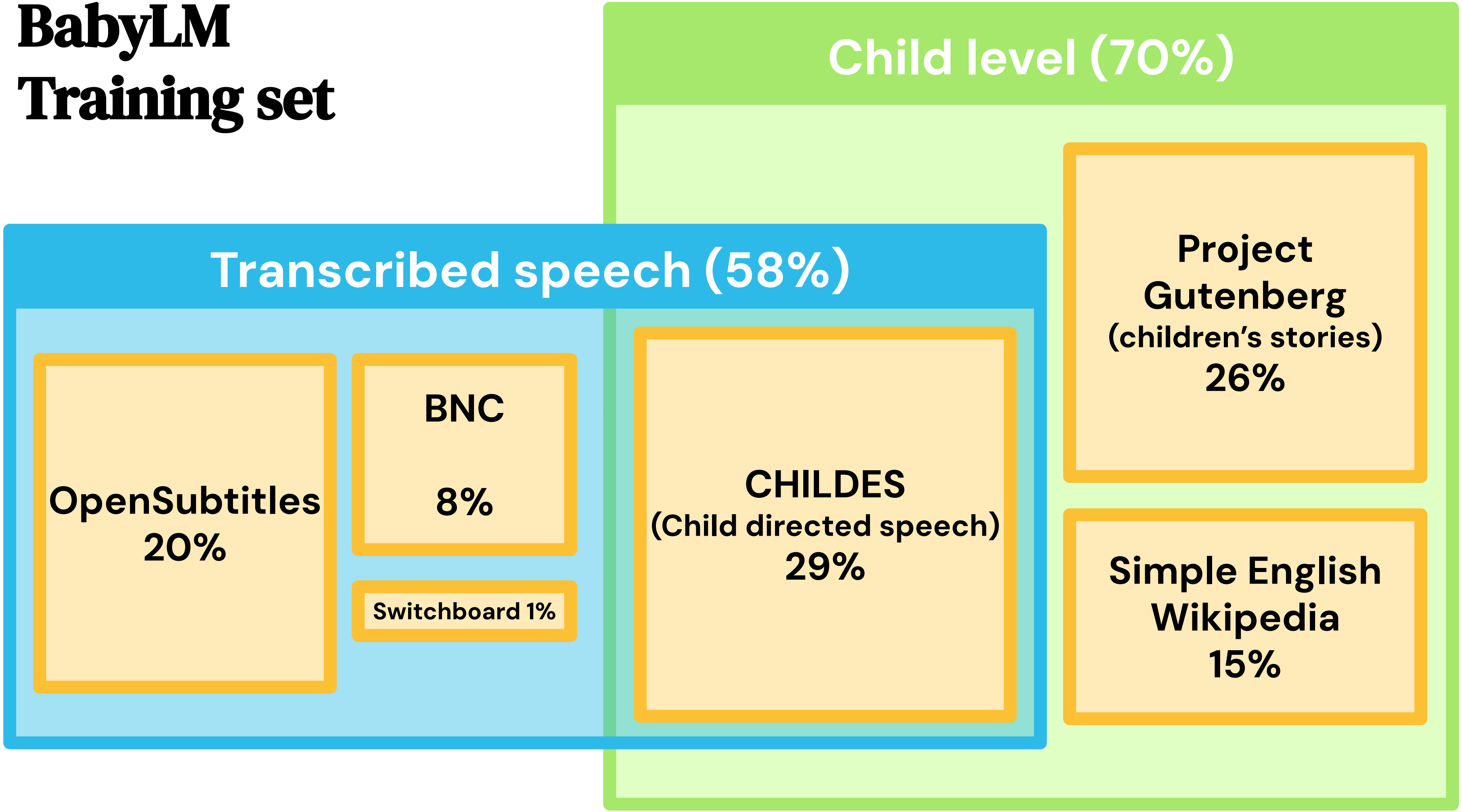
Track 2: Strict-small
- 10 million words
- BabyLM dataset or BYOD
- Shared eval (grammar, generalization, NLU)

Track 3: Multimodal
- 100 million words
- BabyLM dataset or BYOD
- Potentially unlimited vision data
- Shared eval (VQA, grounding, classification)

Track 4: Paper

Original research related to the goals of BabyLM without any competition component.

BabyLM Training set

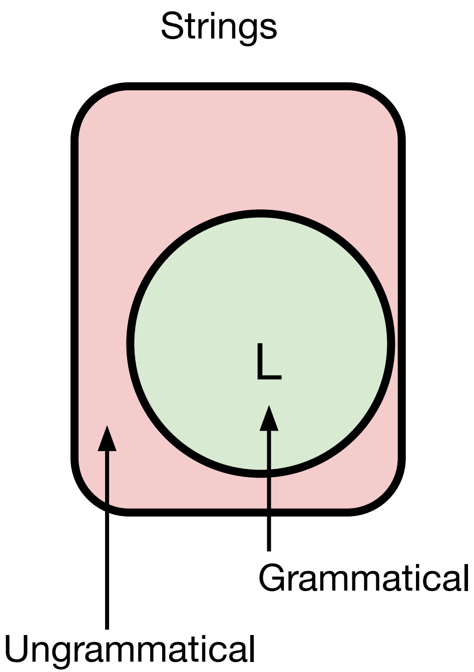


Evaluation Tasks

Hidden Tasks

BLiMP *Syntax*	(Super)GLUE *Understanding*	BLiMP Supplement *Discourse*	MSGS *Generalization*	EWOK *World Knowledge*
Subject-verb agreement	Natural language inference	Turn-taking	Syntactic construction detection	Social reasoning
Filler-gap/Islands	Question answering	Hypernyms	Syntactic category detection	Physical reasoning
Anaphora/binding	Sentiment classification	Question-answer congruence	Syntactic position detection	Spatial reasoning

Acceptability Judgments



Examples from linguistics publications

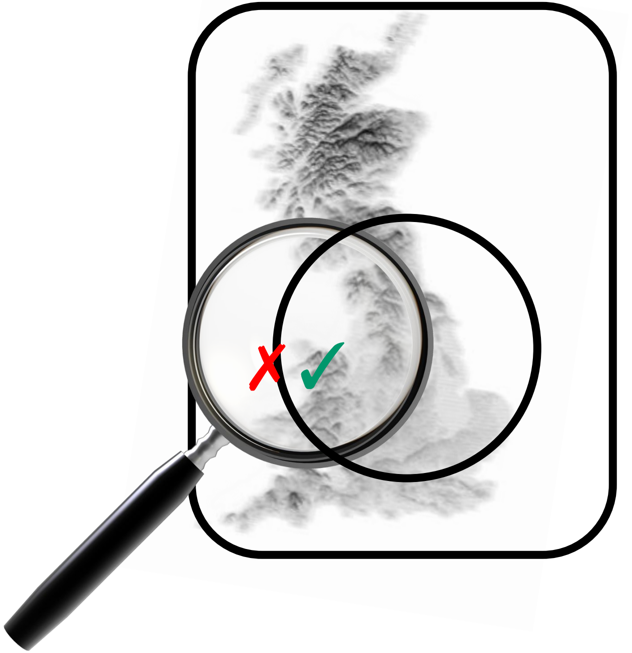
✓ Mary should know that you must go to the station.
✓ I promised that around midnight he would be there.
✓ Susan whispered the news to Rachel.
✗ When time will you be there?
✗ Patrick is likely that left.
✗ Harry coughed us into a fit.

Minimal Pairs

A pair of two nearly identical sentences which differ in acceptability.

✓ Betsy is eager to sleep.

✗ Betsy is easy to sleep.



1. Targeted
2. Reproducible
3. Unsupervised

$P_{LM}(S_{\checkmark}) > P_{LM}(S_{\times})$

The Benchmark of Linguistic Minimal Pairs (BLiMP)

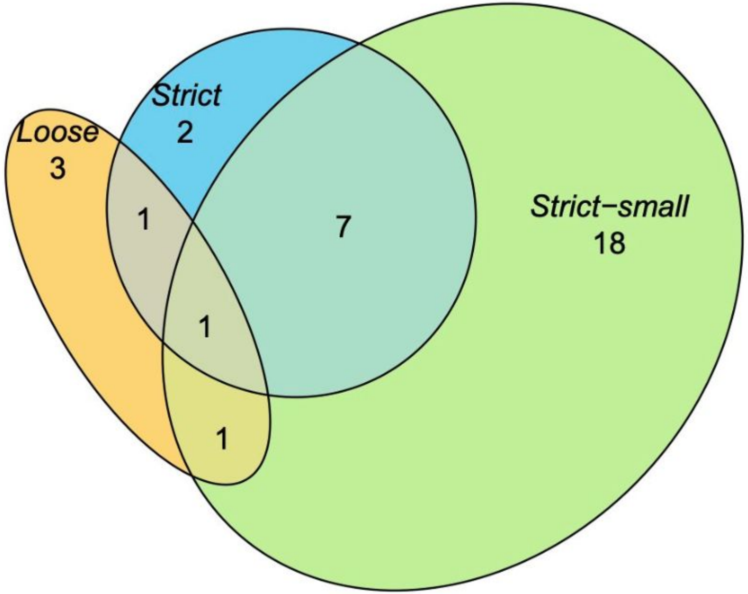
(Warstadt et al., 2020)

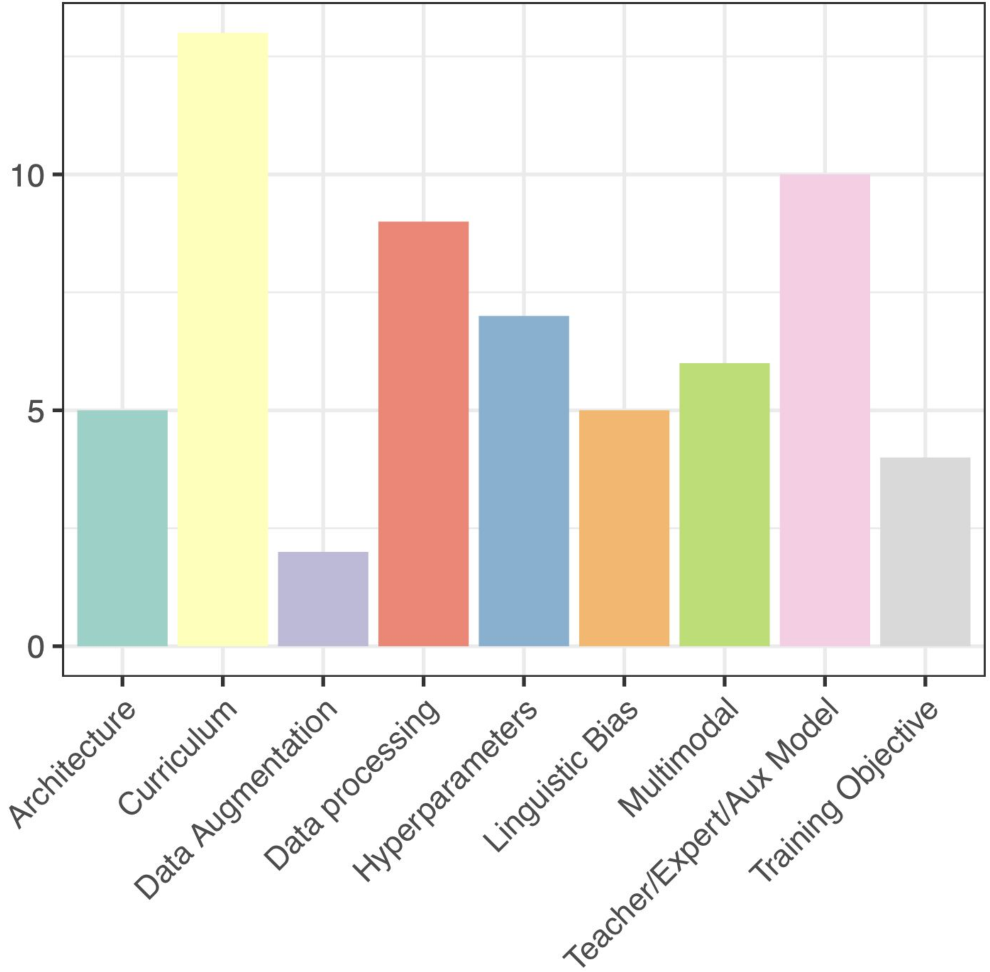
Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
ARG. STRUCTURE	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
BINDING	7	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*
CONTROL/RAISING	5	*There was bound to be a fish escaping.*	*There was unable to be a fish escaping.*
DET.-NOUN AGR.	8	*Rachelle had bought that chair.*	*Rachelle had bought that chairs.*
ELLIPSIS	2	*Anne's doctor cleans one important book and Stacey cleans a few.*	*Anne's doctor cleans one book and Stacey cleans a few important.*
FILLER-GAP	7	*Brett knew what many waiters find.*	*Brett knew that many waiters find.*
IRREGULAR FORMS	2	*Aaron broke the unicycle.*	*Aaron broken the unicycle.*
ISLAND EFFECTS	8	*Whose hat should Tonya wear?*	*Whose should Tonya wear hat?*
NPI LICENSING	7	*The truck has clearly tipped over.*	*The truck has ever tipped over.*
QUANTIFIERS	4	*No boy knew fewer than six guys.*	*No boy knew at most six guys.*
SUBJECT-VERB AGR.	6	*These casseroles disgust Kayla.*	*These casseroles disgusts Kayla.*

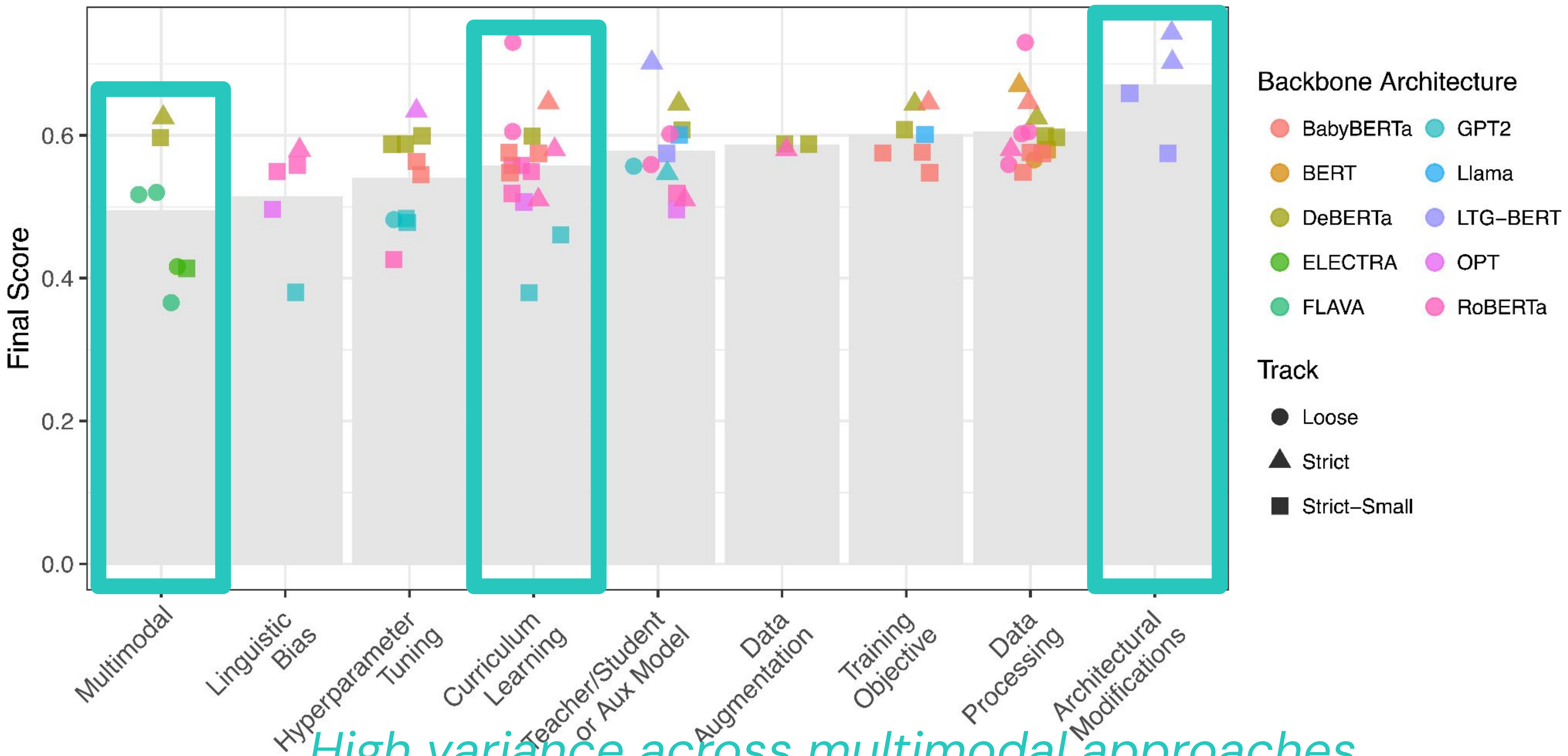
- 67 different minimal pair contrasts
- 1000 sentences each
- 12 broad categories

Results

Submissions (Year 1)

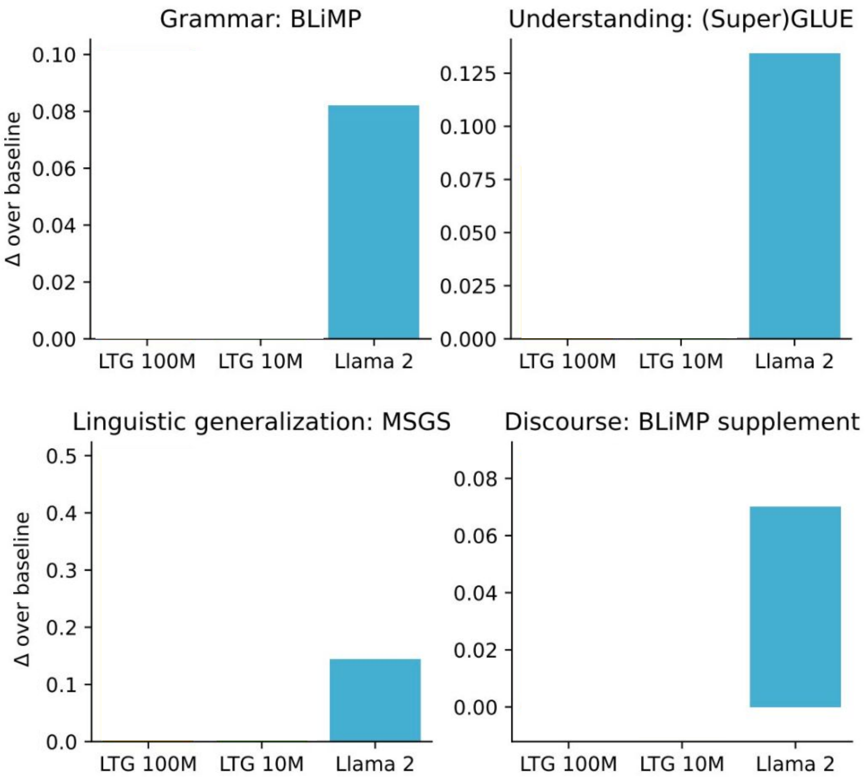


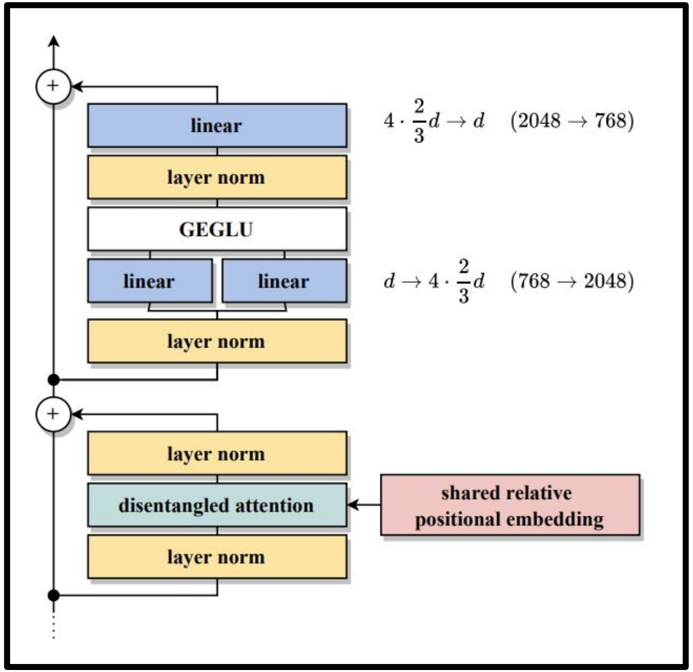




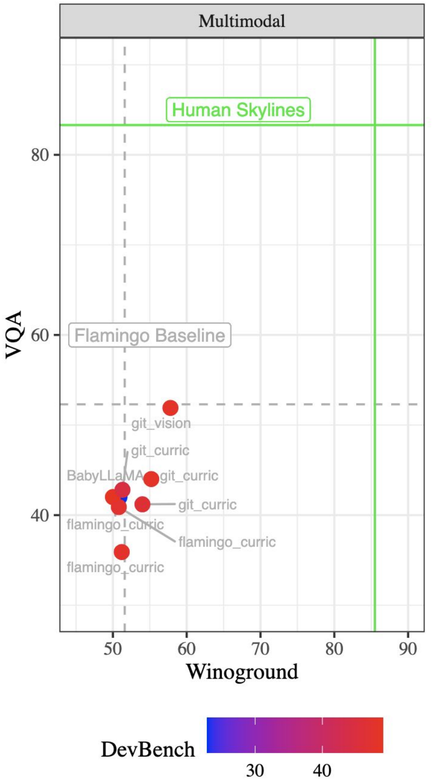
High variance across multimodal approaches
Curriculum learning is difficult
Largest gains from architectural modifications

Winning submission (year 1): LTG-BERT (Charpentier et al.)





Open Challenge: Multimodality

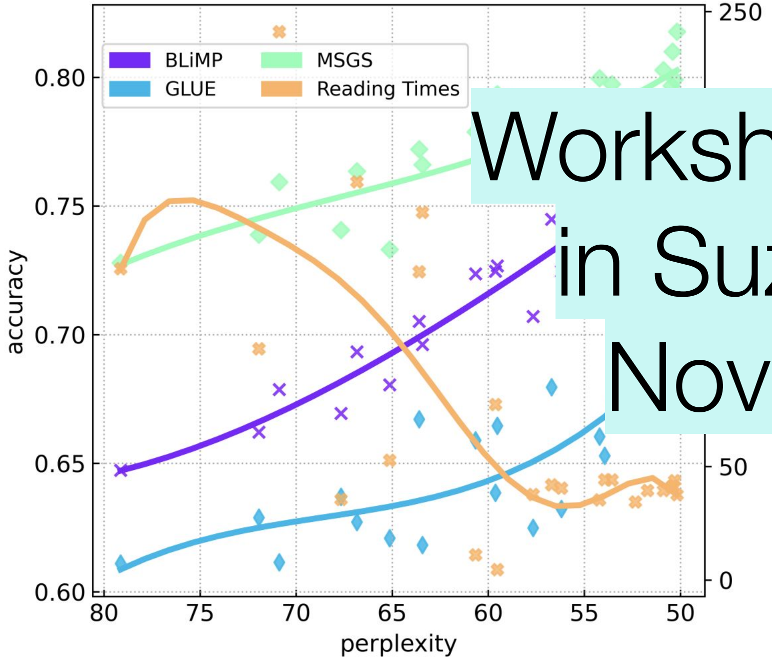


Tasks	10M Words				100M Words			
	None	40K	400K	40M	None	40K	400K	40M
(Super)GLUE (Acc., F1, MCC)	65.49	64.68	65.17	<u>65.76</u>	<u>70.42</u>	69.24	68.9	69.07
BLiMP (Acc.)	63.96	63.98	63.31	<u>64.53</u>	71.32	70.45	<u>71.9</u>	70.93
MSGS (MCC)	-12.88	-12.16	<u>-8.84</u>	-18.62	-8.66	<u>-6.18</u>	-7.41	-7.47

(Amariucai & Warstadt, 2023)

What's new for BabyLM's 3rd birthday?

Cognitive Plausibility Benchmark

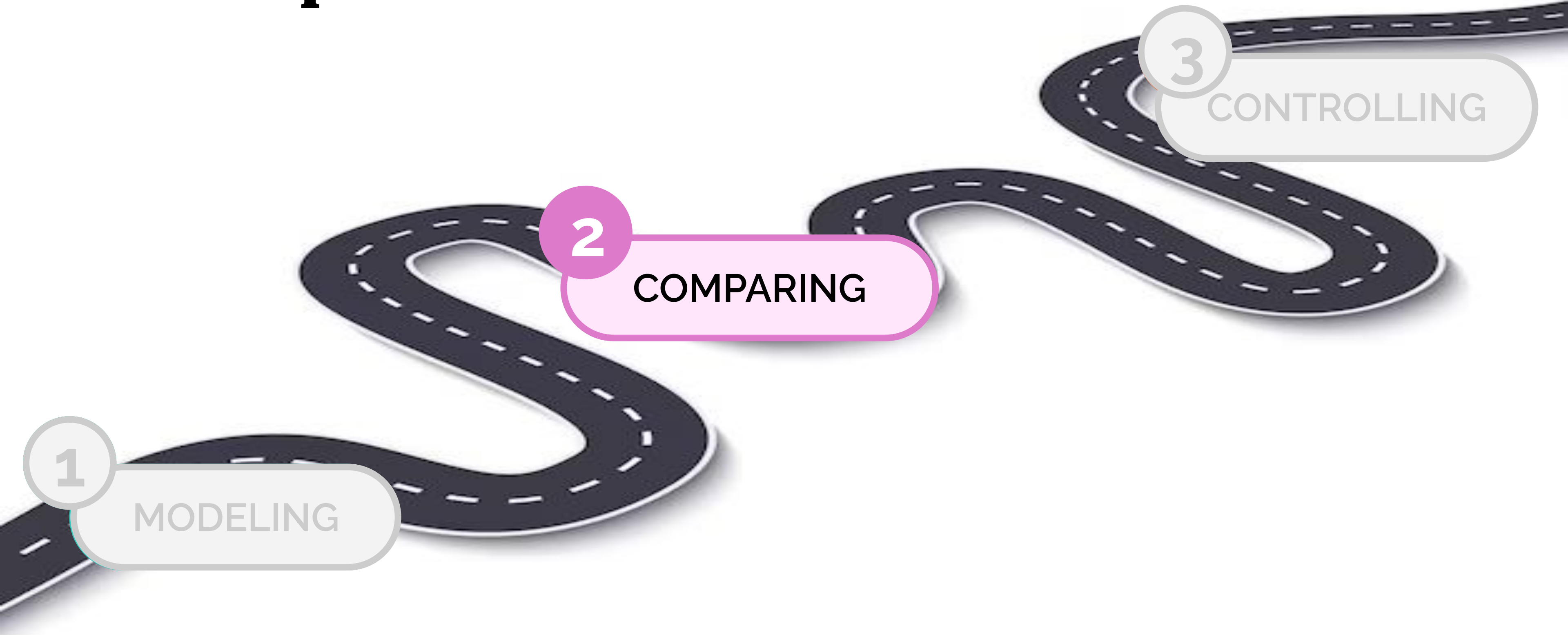


(Steuer et al., 2023)

Interaction Track

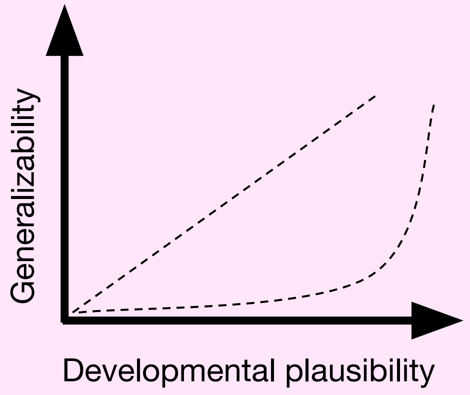
Workshop @ EMNLP
in Suzhou, China
Nov 5-9, 2025

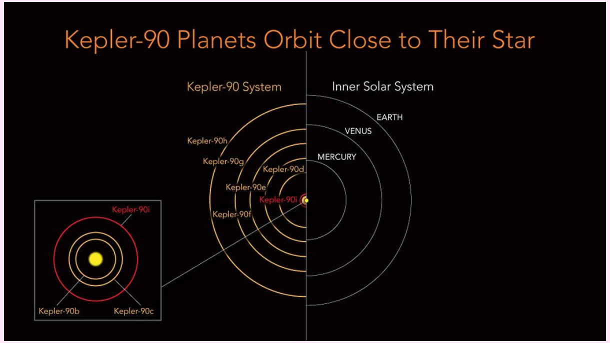
Roadmap



Why compare learning trajectories in humans and LMs?

- Establish and improve plausibility of model learners.
- Reverse-engineer sufficient components of human language learning (Dupoux, 2016).
- Determine what is idiosyncratic about humans, i.e., what is likely to be innate.





By NASA/Ames Research Center/Wendy Stenzel

Word Learning

A Distributional Perspective on Word Learning in Neural Language Models

Filippo Ficarra [1] **Ryan Cotterell** [1] **Alex Warstadt** [1]
[1]ETH Zürich
{fficarra, rcotterell, warstadt}@ethz.ch

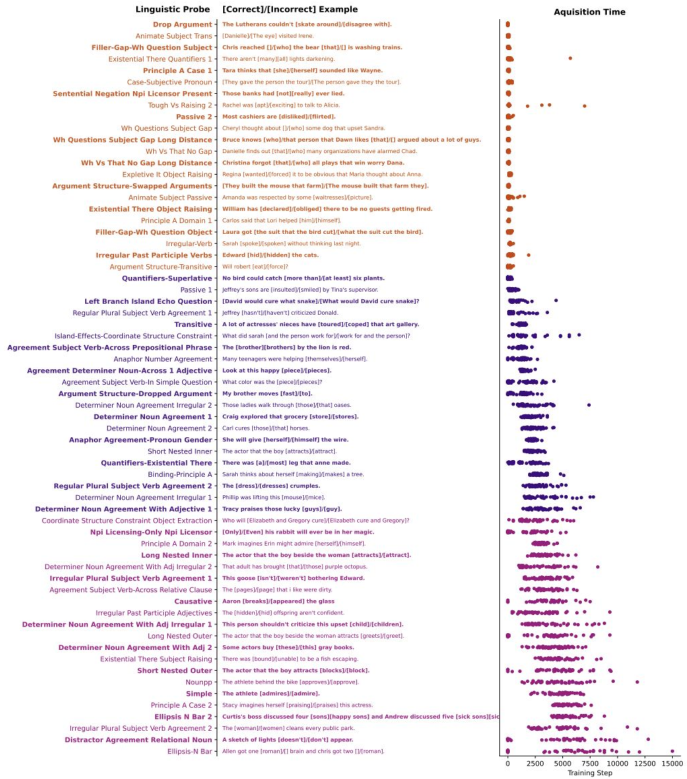
Just accepted to NAACL, 2025

Preprint Soon!

Comparing learning trajectories in LMs and humans is necessary to develop plausible model learners.

How do we compare learning trajectories in humans and BabyLMs?

Acceptability judgments?



Language acquisition: do children and language models follow similar learning stages?

Linnea Evanson
Meta AI Paris;
Laboratoire des systèmes perceptifs
École normale supérieure
PSL University
linnea.evanson8@gmail.com

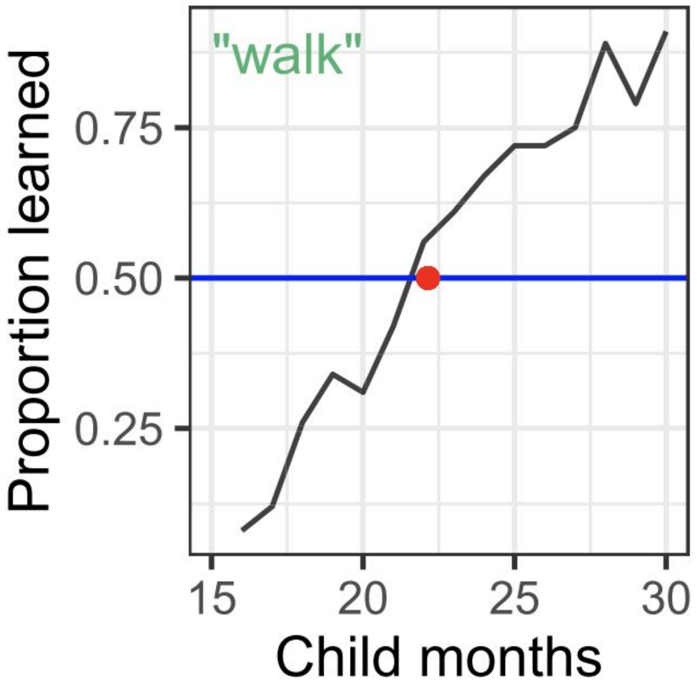
Yair Lakretz*
Cognitive Neuroimaging Unit
CEA, INSERM
Université Paris-Saclay
NeuroSpin Center
yair.lakretz@gmail.com

Jean-Rémi King*
Meta AI Paris;
Laboratoire des systèmes perceptifs
École normale supérieure
PSL University
jeanremi@meta.com

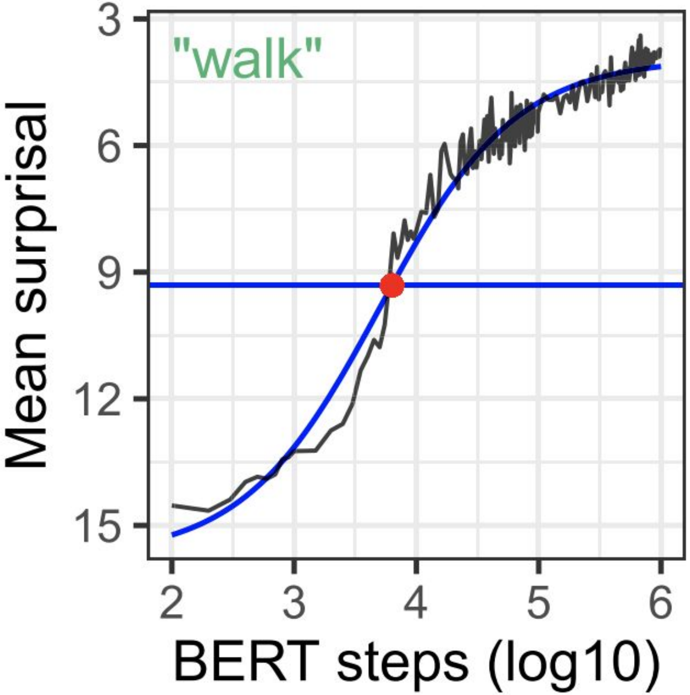
- Stage 1: Simple sentences in subject-verb (SV) order
- Stage 2: Wh Questions
- Stage 3: Relative Clauses

(Friedmann et al, 2021)

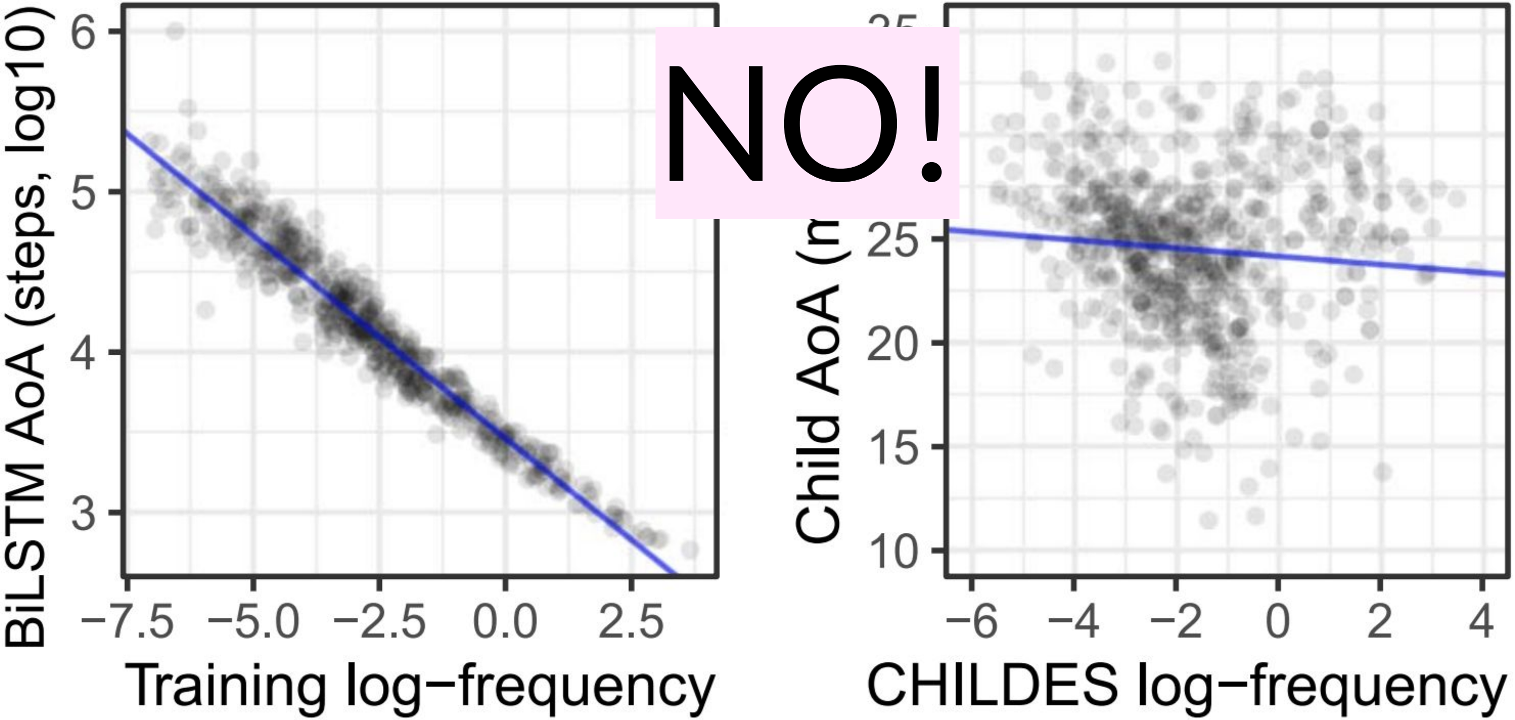
Word learning



WordBank (Frank et al., 2017)
(Braginsky et al., 2019)



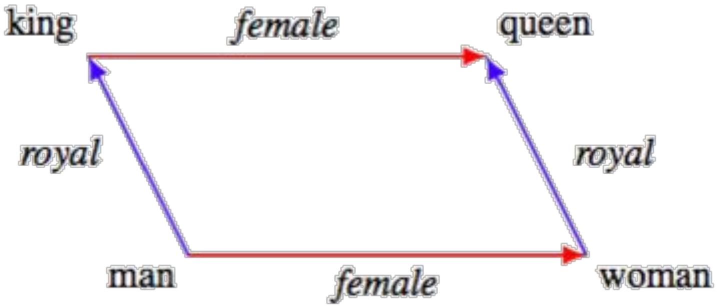
Are LM word learning trajectories human-like?



(Chang & Bergen, 2022)

But wait…what does it mean to learn a word?

Distributional Hypothesis



John Rupert Firth

You shall know a word by the company it keeps.

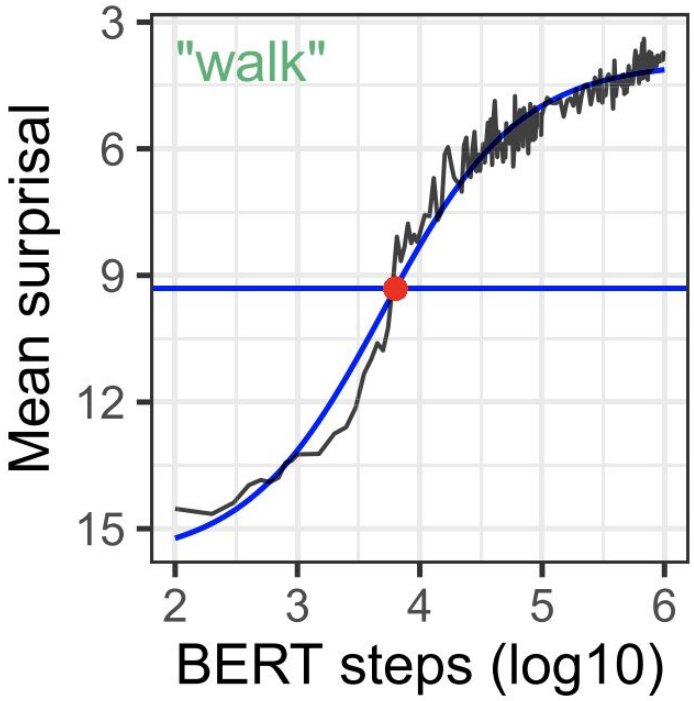
Revisiting Chang & Bergen

$$\sigma_+(\boldsymbol{w}) \stackrel{\text{def}}{=} -\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{p_\kappa}(\boldsymbol{c} \mid \boldsymbol{w}) \log \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c}).$$

Distributional signature

Weighted by context probability (Monte Carlo estimate)

LM surprisal



What about knowing where a word DOESN'T occur?

$$\sigma_{-}(\boldsymbol{w}) \stackrel{\text{def}}{=} -\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{p_\kappa}(\boldsymbol{c} \mid \neg\boldsymbol{w}) \log \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c}).$$

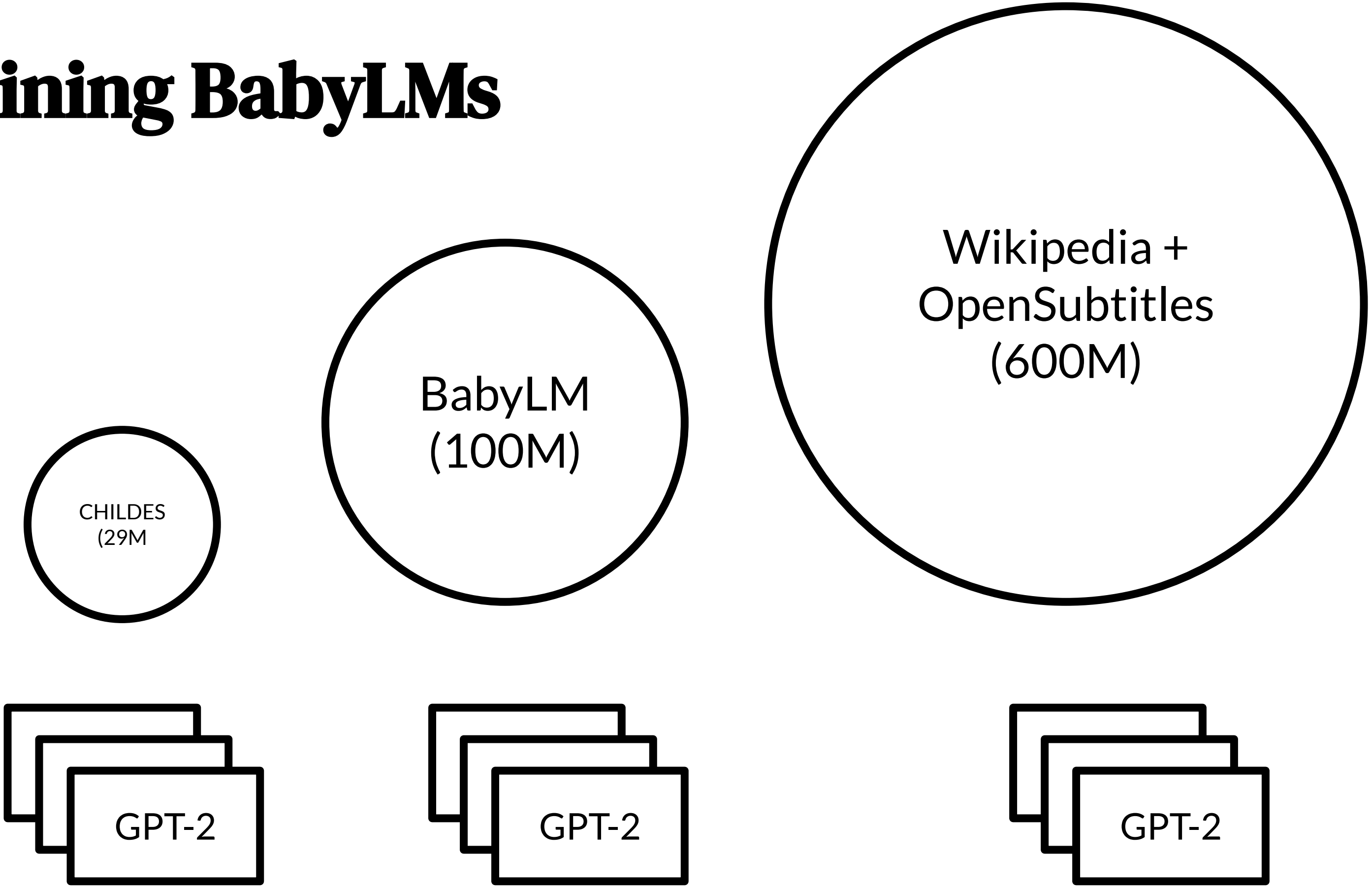
Weighted by the probability of the context given the word DIDN'T occur
(Monte Carlo estimate)

A Typology of Distributional Signatures

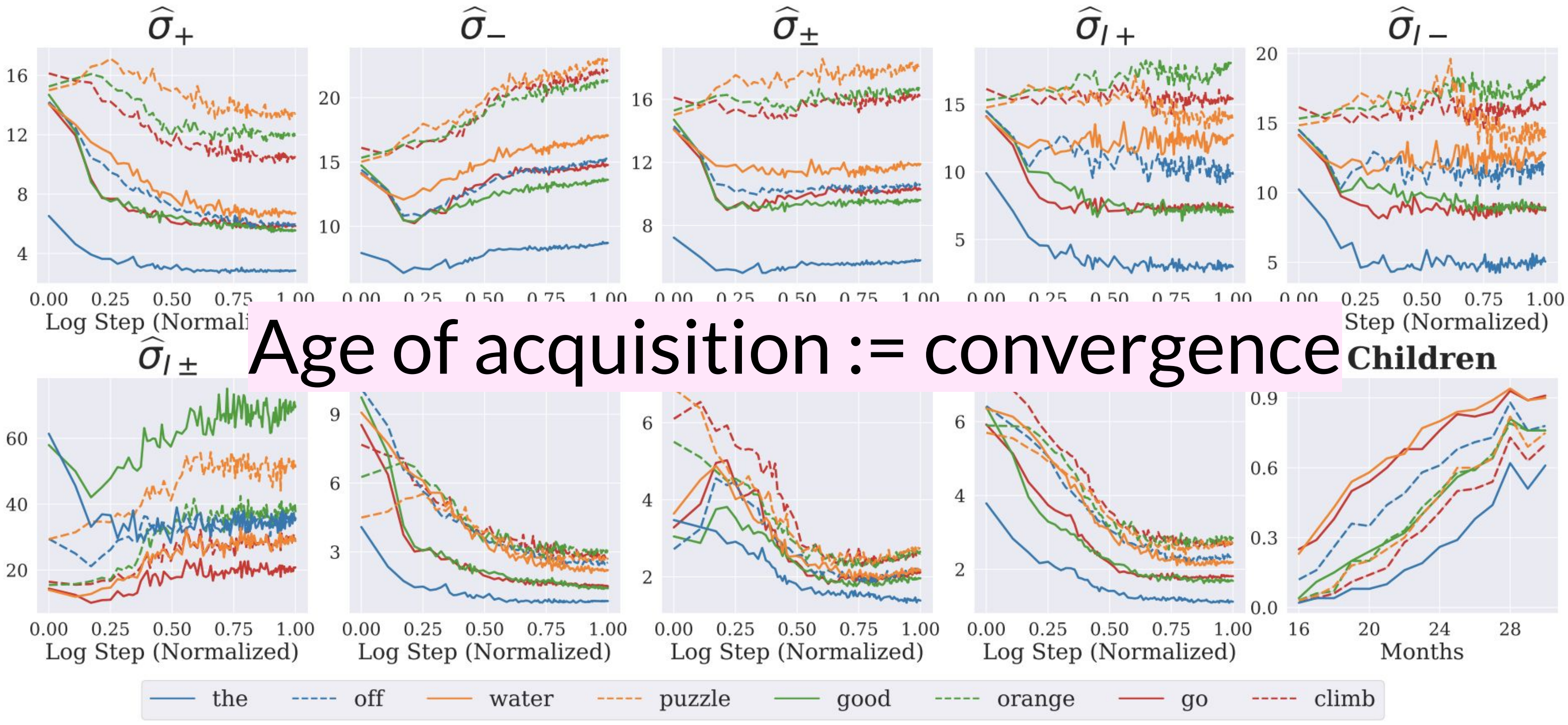
	Positive	Negative	All
True	$-\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{p_\kappa}(\boldsymbol{c} \mid \boldsymbol{w}) \log \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c})$	$-\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{p_\kappa}(\boldsymbol{c} \mid \neg\boldsymbol{w}) \log \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c})$	$-\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{p_\kappa}(\boldsymbol{c}) \log \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c})$
Intrinsic	$-\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{q_\kappa}(\boldsymbol{c} \mid \boldsymbol{w}) \log \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c})$	$-\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{q_\kappa}(\boldsymbol{c} \mid \neg\boldsymbol{w}) \log \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c})$	$-\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{q_\kappa}(\boldsymbol{c}) \log \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c})$
Reference	$\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{p_\kappa}(\boldsymbol{c} \mid \boldsymbol{w}) \left\lvert \log \frac{\overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c})}{\overrightarrow{r}(\boldsymbol{w} \mid \boldsymbol{c})} \right\rvert$	$\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{p_\kappa}(\boldsymbol{c} \mid \neg\boldsymbol{w}) \left\lvert \log \frac{\overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c})}{\overrightarrow{r}(\boldsymbol{w} \mid \boldsymbol{c})} \right\rvert$	$\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{p_\kappa}(\boldsymbol{c}) \left\lvert \log \frac{\overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c})}{\overrightarrow{r}(\boldsymbol{w} \mid \boldsymbol{c})} \right\rvert$

Experiments

Training BabyLMs

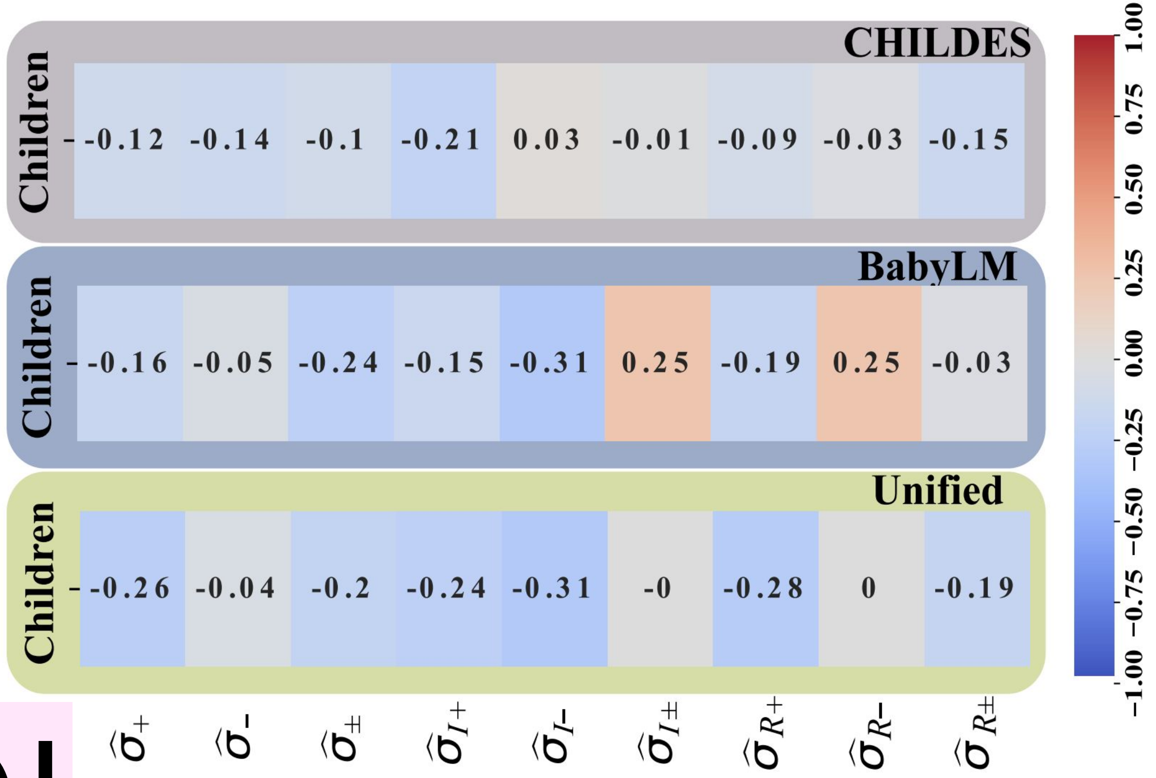


Sample learning curves (BabyLM)



Age of acquisition := convergence

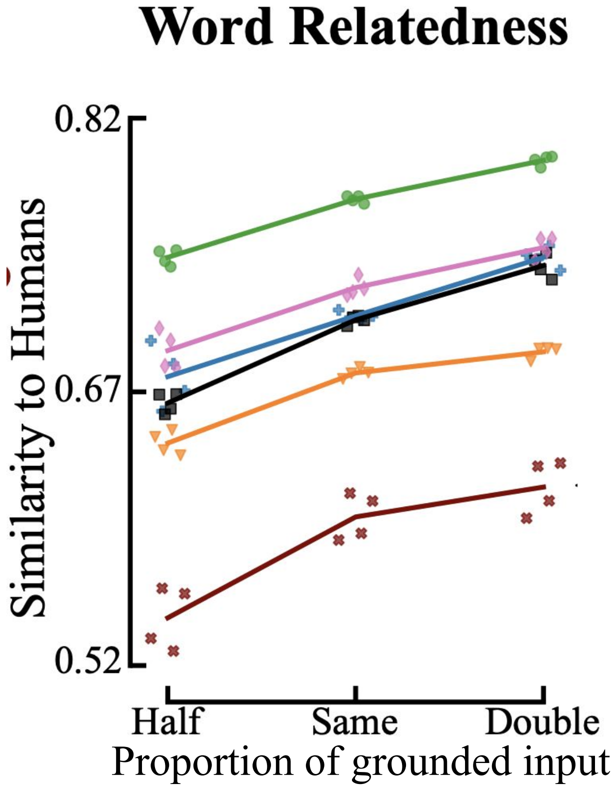
So do any of these signatures have trajectories that look human-like?



Still NO!

Why don't LMs have human-like trajectories?

- Training distribution?
- Grounding
- Interaction
- Production constraints
- Cross-entropy loss
- ...



(Zhuang et al., 2024)

Comparing learning trajectories in LMs can humans can tell us what is idiosyncratic about humans, i.e., likely innate.

It also can help us reverse-engineer the ingredients of human learning.

(Dupoux, 2016)

The Critical Period

Investigating Critical Period Effects in Language Acquisition through Neural Language Models

Ionut Constantinescu Tiago Pimentel Ryan Cotterell Alex Warstadt
ETH Zürich

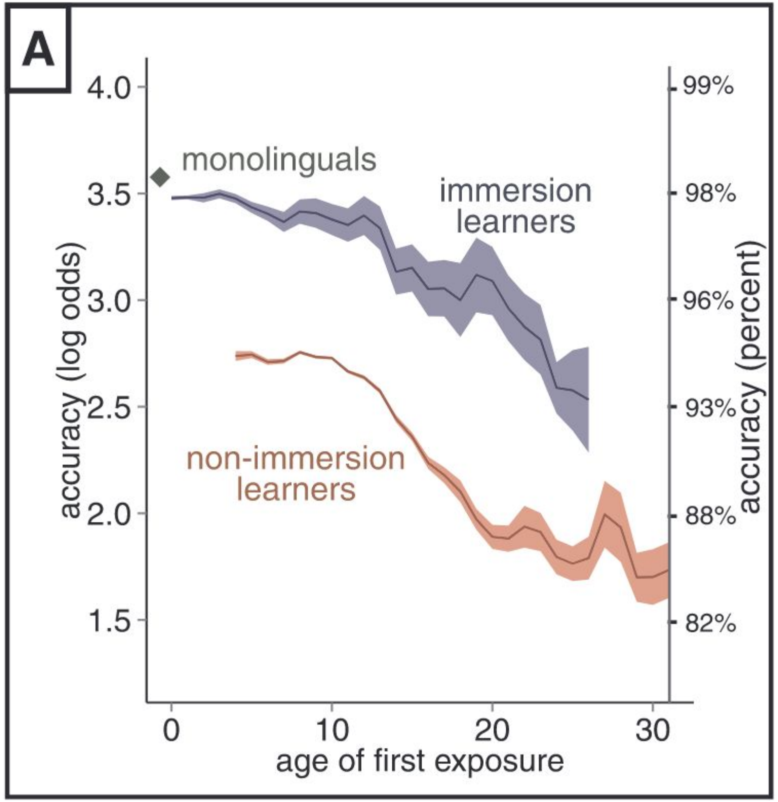
Just out in TACL, 2025

What is the critical period for language acquisition?

L2 Critical Period

age of exposure to L2

proficiency and **learning** in L2



(Hartshorne et al., 2018)

L1 Attrition

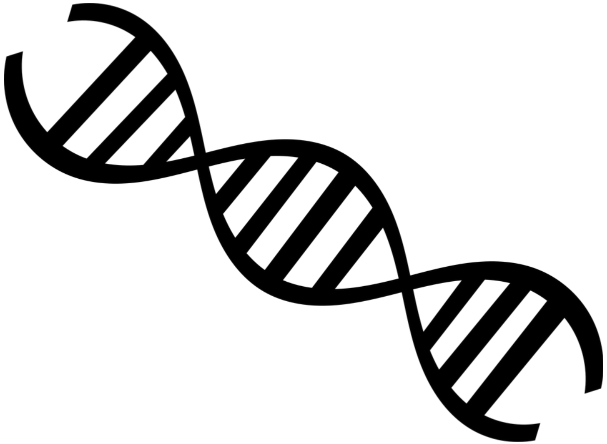
age of ceasing exposure to L1

attrition (forgetting) of L1

(Pallier et al., 2003)

Why is there a critical period?

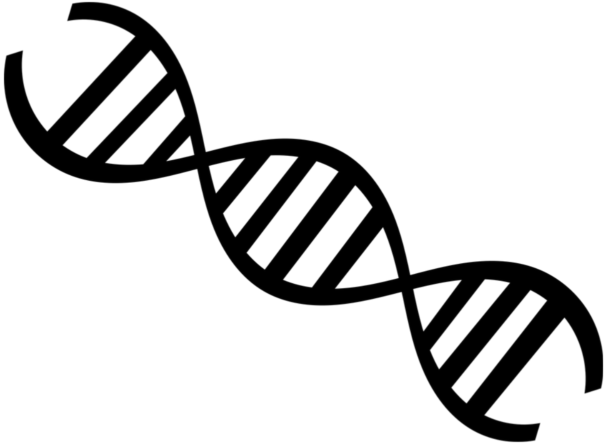
Nature

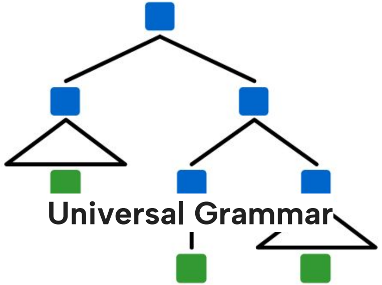


vs.

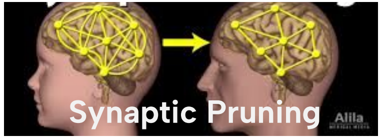
Nurture

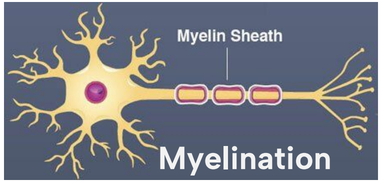
Nature

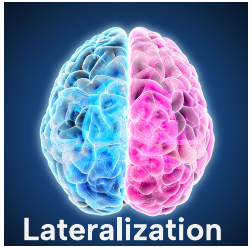




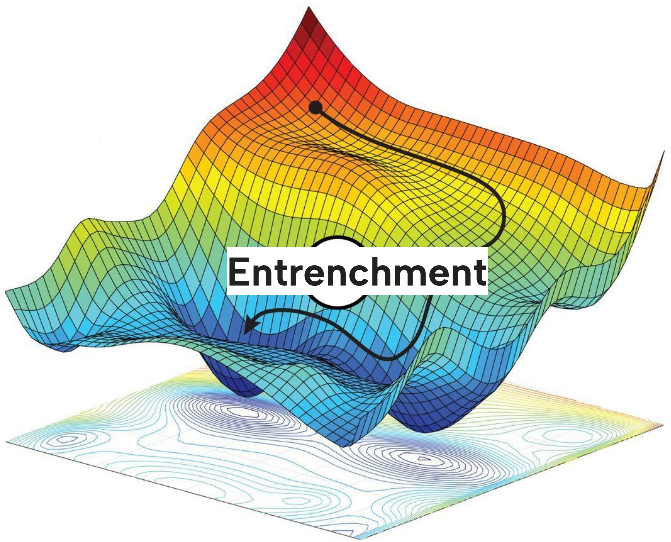
Chomsky, 1965
Newport, 1990







Huttenlocher, 1979
DeBot, 2006
Lenneberg, 1967



Munro, 1986
Elman et al., 1996
Zevin and Seidenberg, 2002
Achille et al., 2019

Nurture

How can BabyLMs tell us something about the critical period in humans?

Nature **vs.** **Nurture**

Strong Innate Hypothesis

Innate learning constraints are <u>necessary</u> to explain critical period effects.

Strong Experiential Hypothesis

Critical period effects are a <u>necessary</u> consequence of successful statistical learning.

If LMs <u>do</u> show critical period effects.

Nature vs. Nurture

Strong Innate Hypothesis

Innate learning constraints are <u>necessary</u> to explain critical period effects.

Strong Experiential Hypothesis

Critical period effects are a <u>necessary</u> consequence of successful statistical learning.

If LMs <u>don't</u> show critical period effects.

Space of effective learners

Weak Innate Hypothesis

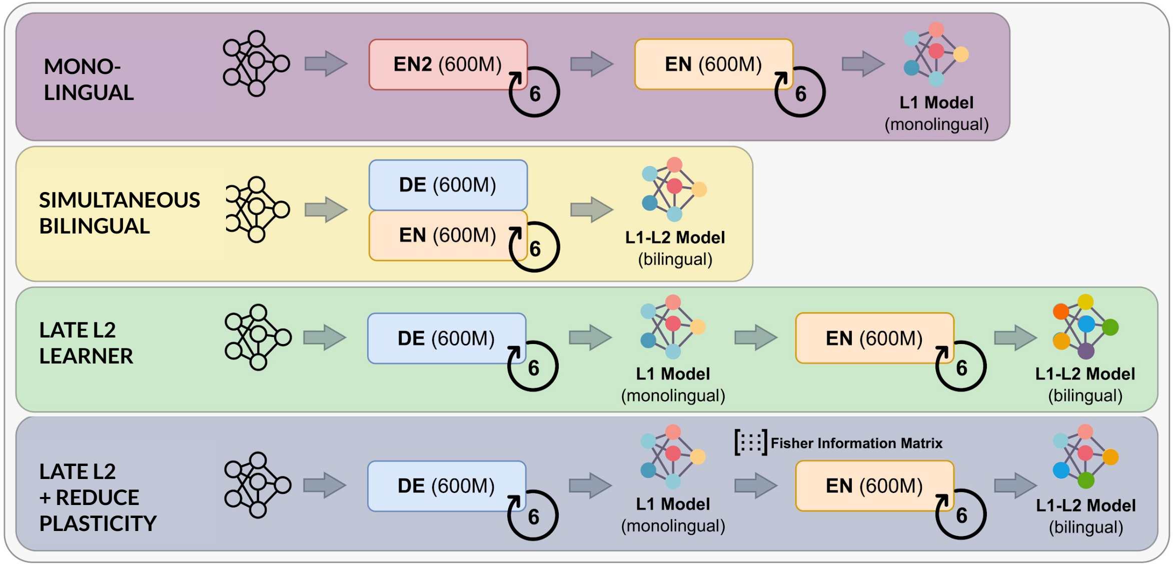
Innate learning constraints are the main driver of critical period effects in arbitrary learners.

Weak Experiential Hypothesis

Statistical learning is the main driver of critical period effects in arbitrary learners.

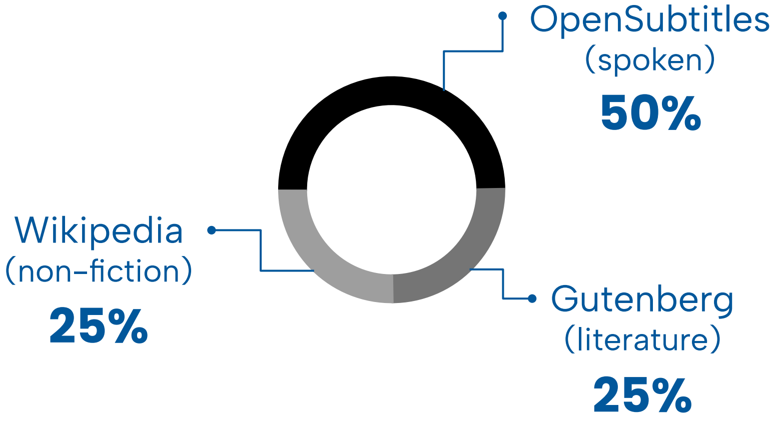
Experiments

Training Conditions



Data

Domains



Languages

Main language: English

Other languages:

- Germanic: German, Dutch
- IE, L: Spanish, Polish
- IE, non-L: Greek, Russian
- non-IE, L: Finnish, Turkish
- non-IE, non-L: Arabic, Korean
- Programming language: Java

IE: Indo-European
L: Latin

Training

Architectures

- GPT2 (autoregressive decoder)
- RoBERTa (masked encoder)

Tokenization

- Train bilingual BPE tokenizers
- Dataset is split into fixed-size blocks of 512 tokens

Hyperparameters

- We do a hyperparameter sweep (W&B) for each model and choose 3 best configs

Training schedule

- Linear learning rate decay
- Restart optimizer between sequential stages

Evaluation

All evaluations are done on English only for fair comparisons!

We evaluate models at every epoch (except for GLUE)

1. BLiMP
2. Perplexity
3. GLUE

Results

Do LMs show human-like <u>L2</u> critical period effects?

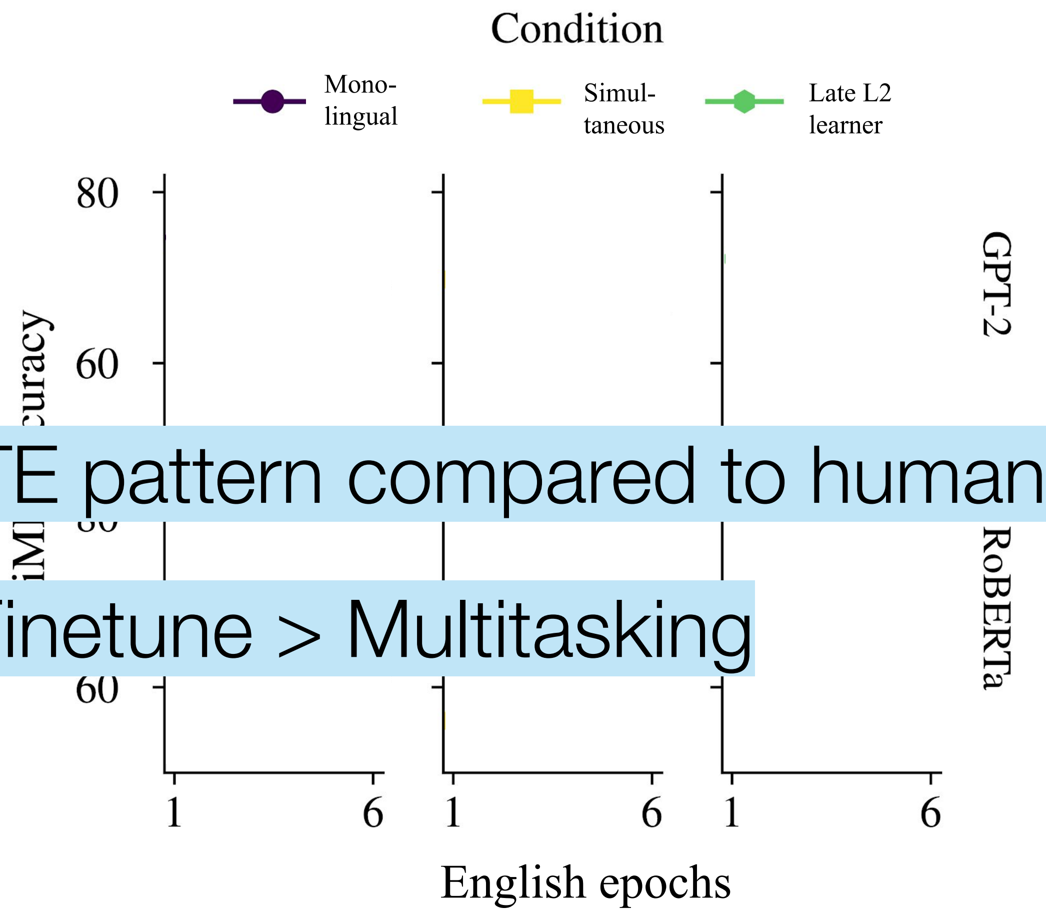
LMs show OPPOSITE pattern compared to humans!

age of exposure to L2

Pretrain + Finetune > Multitasking

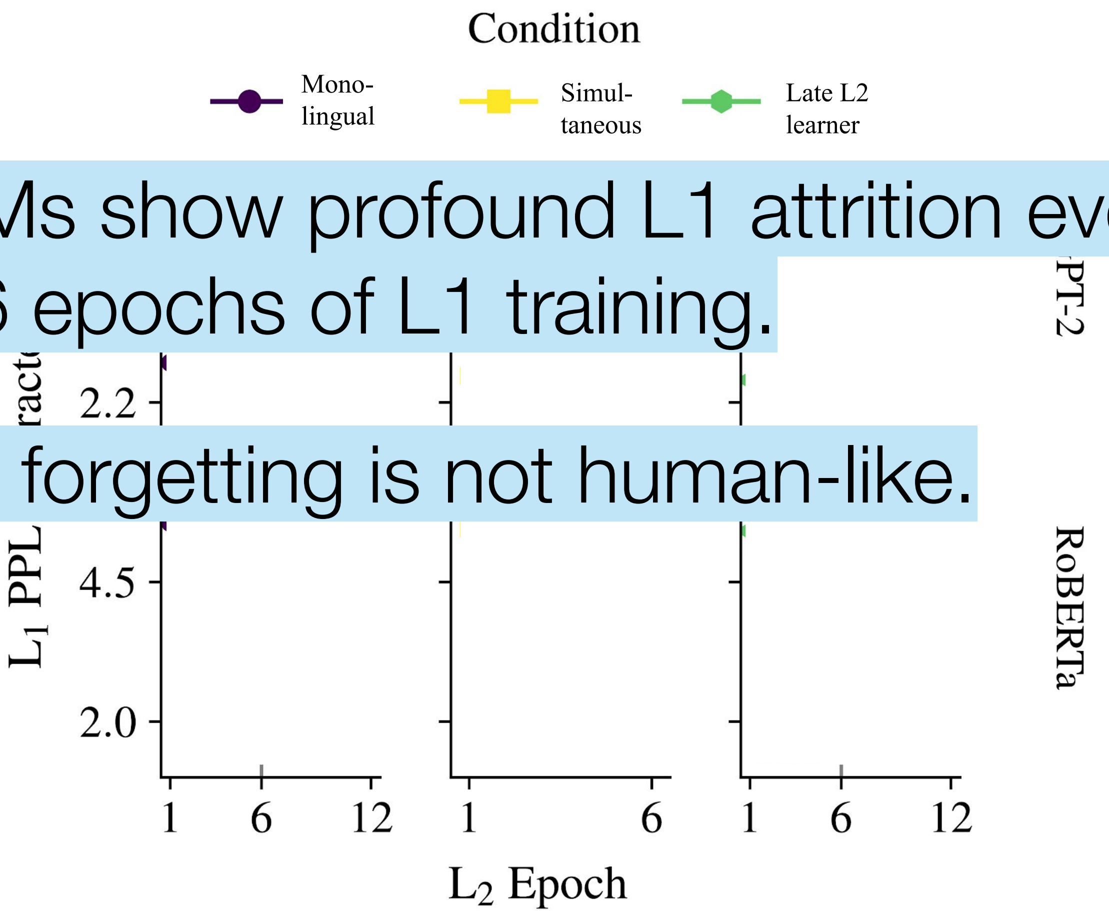
proficiency a

learning in L2



Do LMs show human-like L1 critical pe effects?

Unlike humans, LMs show profound L1 attrition even after 6 epochs of L1 training.

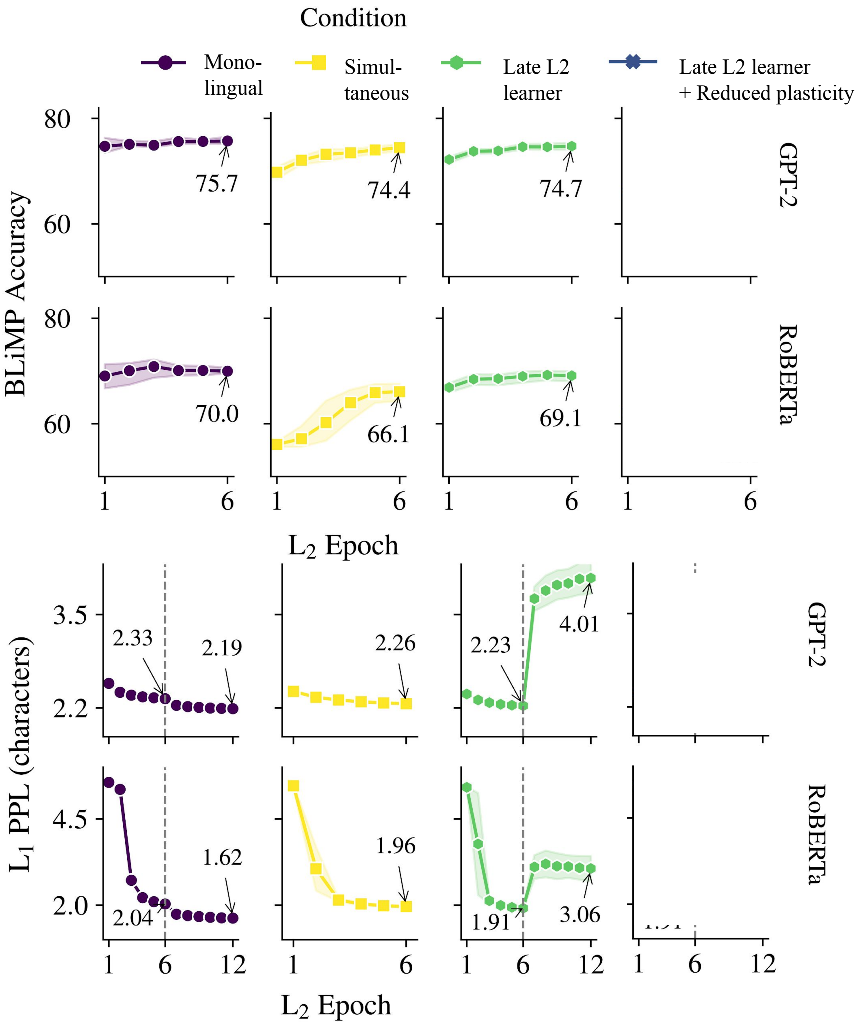


Catastrophic forgetting is not human-like.

age of ceasing exposure to L1

attrition (forgetting) of L1

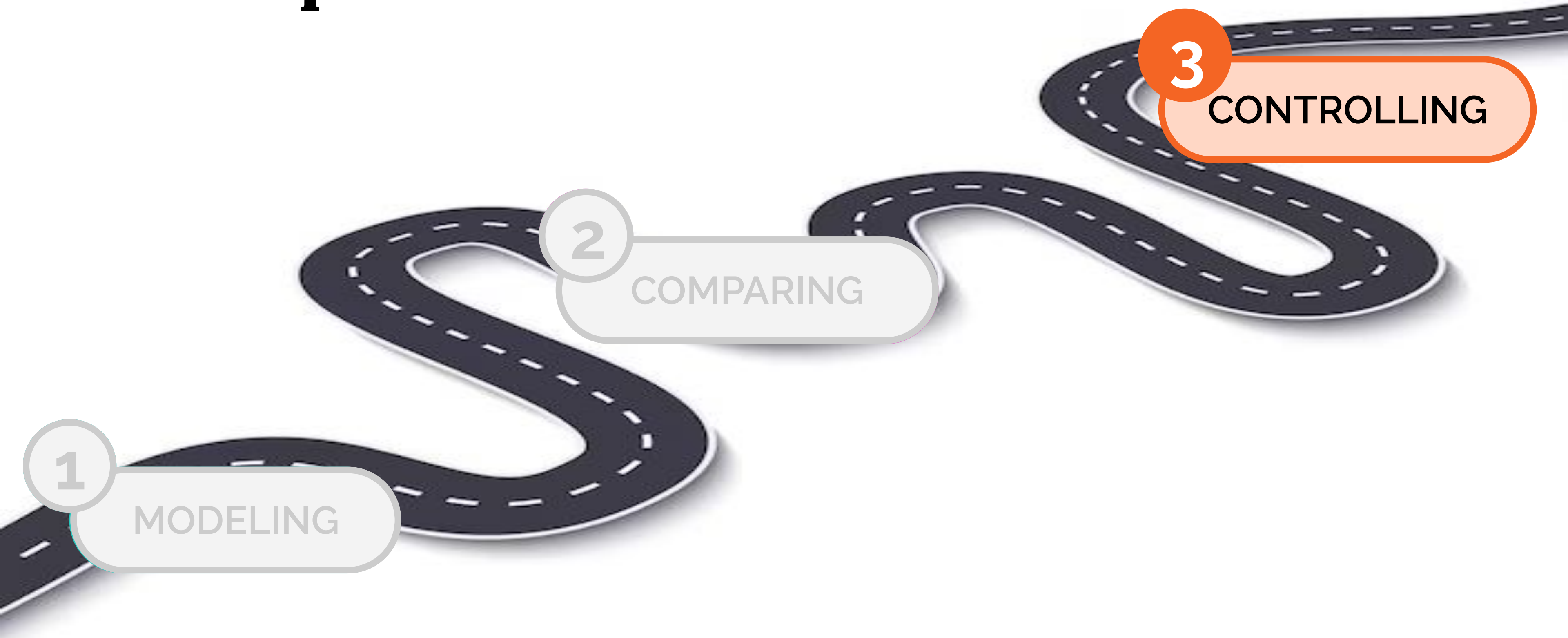
Can we reverse-engineer critical period effects?



L2 Performance

L1 Performance

Roadmap



1. Controlled experiments on LMs enable causal inferences about the impact of environment on learning.

2. LMs can also help to determine whether learning biases are domain-general or language-specific.

Typological Correlations

Can Language Models Learn Typologically Implausible Languages?

Tianyang Xu[a,*], **Tatsuki Kuribayashi**[c], **Yohei Oseki**[d], **Ryan Cotterell**[b], **Alex Warstadt**[e,*]
[a]Toyota Technical Institute at Chicago, [b]ETH Zürich, [c]MBZUAI,
[d]The University of Tokyo, [e]University of California San Diego,
*work conducted partially at ETH Zürich
sallyxu@ttic.edu, tatsuki.kuribayashi@mbzuai.ac.ae,
oseki@g.ecc.u-tokyo.ac.jp, rcotterell@inf.ethz.ch, awarstadt@ucsd.edu

Under review
Preprint soon!

Why are some types of grammars common across the world's languages while others are not?

Greenberg's Universals (1963)

Universal 2. In languages with prepositions, the genitive almost always follows the governing noun, while in languages with postpositions it almost always precedes.

Turning once more to the data of Table I, it is a striking evidence of lawful relationships among the variables that of the 12 possibilities 5, or almost half, are not exemplified in the sample. All of these types are either rare or non-existent.[7] For type I, we see that all 6 languages of the sample are Pr/N. This holds with extremely few exceptions on a world-wide basis. There are, however, a few valid examples of I/Pr/A, the mirror image, so to speak, of the fairly frequent III/Po/N. On the other hand, there are, as far as I know, no examples of either I/Po/A or I/Po/N. Hence we may formulate the following universal:

Universal 3. Languages with dominant VSO order are always prepositional.

Languages of type III are, as has been seen, the polar opposites of type I. Just as there are no postpositional languages in type I, we expect that there will be no prepositional languages in type III. This is overwhelmingly true, but I am aware of several exceptions.[8] Since, as has been seen, genitive position correlates highly with Pr/Po, we will expect that languages of type III normally have GN order. To this there are some few exceptions. However, whenever genitive order deviates, so does adjective order, whereas the corresponding statement does not hold for Pr/Po.[9] We therefore have the following universals:

Universal 4. With overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional.

Universal 5. If a language has dominant SOV order and the genitive follows the governing noun, then the adjective likewise follows the noun.

An important difference may be noted between languages of types I and III. In regard to verb-modifying adverbs and phrases

Dryer's Update (1993)

VERB PATTERNER	OBJECT PATTERNER	EXAMPLE
verb	object	*ate + the sandwich*
verb	subject	*(there) entered + a tall man*
adposition	NP	*on + the table*
copula verb	predicate	*is + a teacher*
'want'	VP	*wants + to see Mary*
tense/aspect auxiliary verb	VP	*has + eaten dinner*
negative auxiliary	VP	cf. 7 in §4.2
complementizer	S	*that + John is sick*
question particle	S	cf. 8 in §4.4.
adverbial subordinator	S	*because + Bob has left*
article	N′	*the + tall man*
plural word	N′	cf. 9 in §4.7
noun	genitive	*father + of John*
noun	relative clause	*movies + that we saw*
adjective	standard of comparison	*taller + than Bob*
verb	PP	*slept + on the floor*
verb	manner adverb	*ran + slowly*

TABLE 39. Complete list of correlation pairs.

Learnability as an Explanation

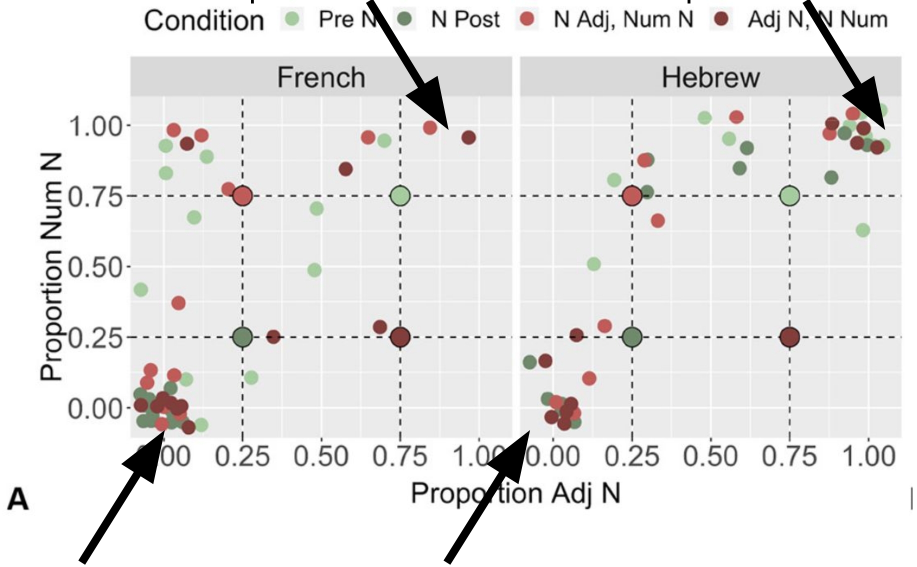
Chapter 1

The Theory of Principles and Parameters
with Howard Lasnik

(Chomsky & Lasnik, 1993)

Typologically
plausible

Typologically
plausible



Typologically
plausible

Typologically
plausible

(Culbertson et al., 2020)

So where do BabyLMs help?

Problem 1: Do humans really have a learning bias?

Table 3
IPA transcriptions (and meanings for adjectives and numerals) of French and Hebrew artificial language lexicon. Note that adjectives and numerals are pseudo-nonce (real word equivalents and IPA transcriptions in the respective languages are given in parentheses).

French				
Nouns	Adjectives		Numerals	
[bogi]	[bly]	'blue' (cf. *blu* [blø])	[doks]	'two' (cf. *deux* [dø])
[sefi]	[taʃu]	'spotted' (cf. *tacheté* [taʃte])	[tʁa]	'three' (cf. *trois* [tʁwa])
[voli]	[pølu]	'furry' (cf. *poilu* [pwaly])	[kitʁ]	'four' (cf. *quatre* [kætʁ])
[kani]				

(Culbertson et al., 2020)

Space of effective learners

Solution: Controlled Experiments at Scale on LMs

Pharaoh Psamtik
(664 – 610 BCE)

Frederick II
(1194-1250)

James IV
(1473-1513)

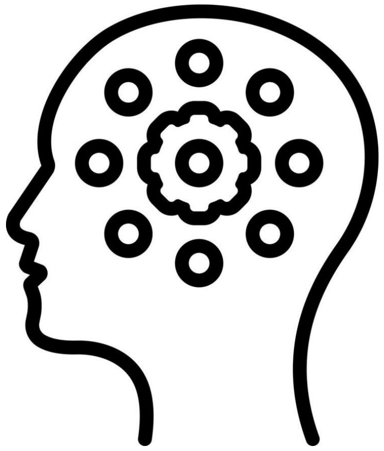
Carried out language deprivation experiments

Problem 2: Whence bias?

Language-Specific

vs.

Domain-General



Solution: Domain Generality of Transformers

Vision

AN IMAGE IS WORTH 16X16 WORDS:
TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy[*,†], Lucas Beyer[*], Alexander Kolesnikov[*], Dirk Weissenborn[*], Xiaohua Zhai[*], Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby[*,†]
[*]equal technical contribution, [†]equal advising
Google Research, Brain Team
{adosovitskiy, neilhoulsby}@google.com

ViT (>50k citations)

Protein Folding

Article

Highly accurate protein structure prediction with AlphaFold

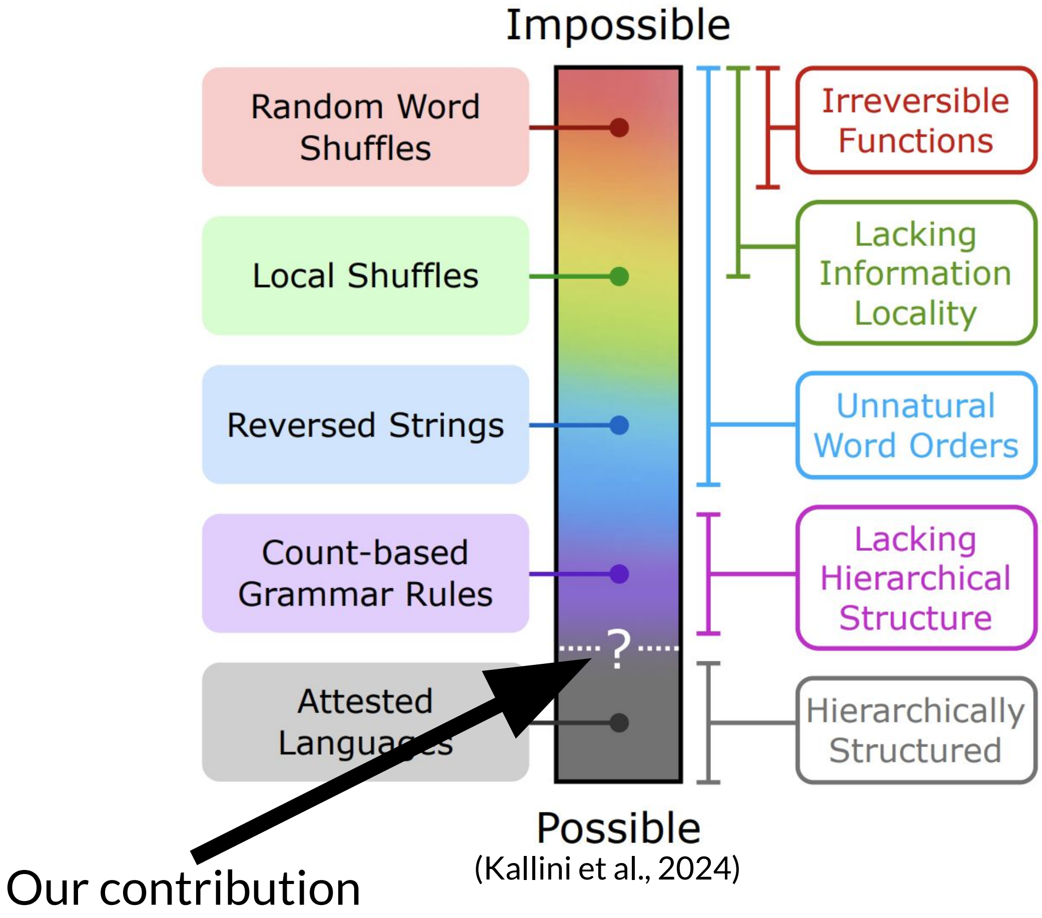
https://doi.org/10.1038/s41586-021-03819-2
Received: 11 May 2021
Accepted: 12 July 2021
Published online: 15 July 2021
Open access
Check for updates

John Jumper[1,4]✉, Richard Evans[1,4], Alexander Pritzel[1,4], Tim Green[1,4], Michael Figurnov[1,4], Olaf Ronneberger[1,4], Kathryn Tunyasuvunakool[1,4], Russ Bates[1,4], Augustin Žídek[1,4], Anna Potapenko[1,4], Alex Bridgland[1,4], Clemens Meyer[1,4], Simon A. A. Kohl[1,4], Andrew J. Ballard[1,4], Andrew Cowie[1,4], Bernardino Romera-Paredes[1,4], Stanislav Nikolov[1,4], Rishub Jain[1,4], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[2,3], Michalina Pacholska[1], Tamas Berghammer[1], Sebastian Bodenstein[1], David Silver[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1] & Demis Hassabis[1,4]✉

AlphaFold (>30k citations)

Corpus Editing & Counterfactual Language Learning

(Jumelet and Hupkes, 2018; Warstadt, 2022; Patil et al., 2024; Misra and Mahowald, 2024)



(Kallini et al., 2024)

Experiments

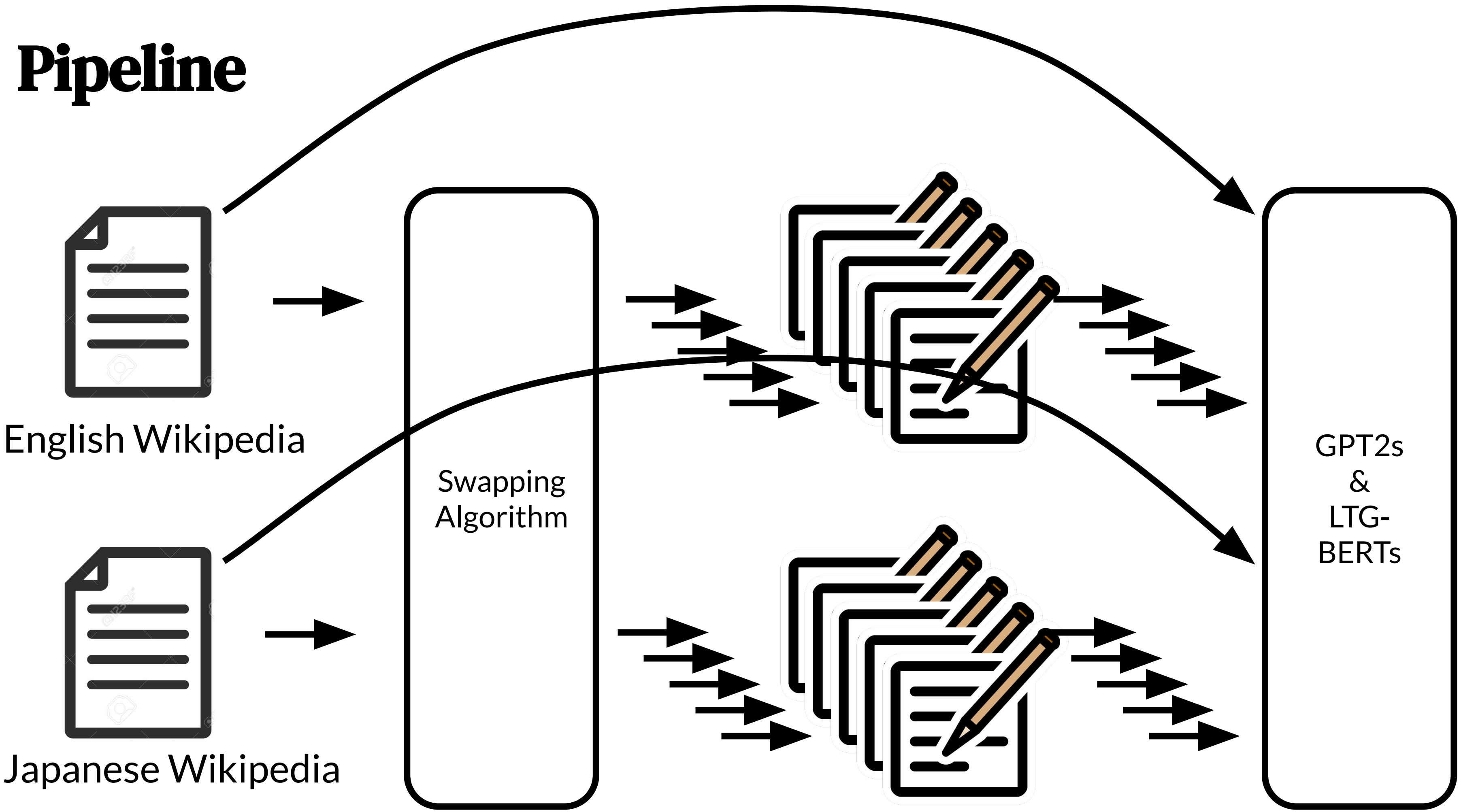
Counterfactual Corpora

Correlation Pair	Example
Original	DET NOUN AUX SCONJ DET NOUN ADP NOUN AUX VERB ADP PROPN ADP PROPN The fact is that the season of strawberries is running from July to August.
<V, O>	DET NOUN AUX SCONJ DET NOUN ADP NOUN ADP PROPN ADP PROPN AUX VERB The fact is that the season of strawberries to August from July is running.
<Adp, NP>	DET NOUN AUX SCONJ DET NOUN NOUN ADP AUX VERB PROPN ADP PROPN ADP The fact is that the season strawberries of is running July from August to.
<Cop, Pred>	DET NOUN SCONJ DET NOUN ADP NOUN AUX VERB ADP PROPN ADP PROPN AUX The fact that the season of strawberries is running from July to August is.
<Aux, V>	DET NOUN AUX SCONJ DET NOUN ADP NOUN VERB ADP PROPN ADP PROPN AUX The fact is that the season of strawberries running from July to August is.
<Noun, Genitive>	DET NOUN AUX SCONJ DET ADP NOUN NOUN VERB ADP PROPN ADP PROPN AUX The fact is that the of strawberries season running from July to August is.

Counterfactual Corpus (Japanese)

Correlation Pair	Example
Original	NOUN ADP NOUN ADP NOUN ADP NOUN ADP VERB AUX NOUN ADP NOUN AUX Ichigo no kisetsu ga shichigatsu kara hachigatsu made tsudui teiru koto wa jijitsu dearu. Strawberry of season NOM July from August to runinng is that TOP fact is.
<V, O>	NOUN ADP NOUN ADP VERB AUX NOUN ADP NOUN ADP NOUN ADP NOUN AUX Ichigo no kisetsu ga tsudui teiru hachigatsu made shichigatsu kara koto wa jijitsu dearu. Strawberry of season NOM running is August to July from that TOP fact is.
<Adp, NP>	ADP NOUN ADP NOUN ADP NOUN ADP NOUN VERB AUX ADP NOUN NOUN AUX No ichigo ga kisetsu kara shichigatsu made hachigatsu tsudui teiru wa koto jijitsu dearu. Of strawberry NOM season from July to August runinng is TOP that fact is.
<Cop, Pred>	NOUN ADP NOUN ADP NOUN ADP NOUN ADP VERB AUX NOUN ADP AUX NOUN Ichigo no kisetsu ga shichigatsu kara hachigatsu made tsudui teiru koto wa dearu jijitsu. Strawberry of season NOM July from August to runinng is that TOP is fact.
<Aux, V>	NOUN ADP NOUN ADP NOUN ADP NOUN ADP AUX VERB NOUN ADP NOUN AUX Ichigo no kisetsu ga shichigatsu kara hachigatsu made teiru tsudui koto wa jijitsu dearu. Strawberry of season NOM July from August to is running that TOP fact is.
<Noun, Genitive>	NOUN NOUN ADP ADP NOUN ADP NOUN ADP VERB AUX NOUN ADP NOUN AUX Kisetsu ichigo no ga shichigatsu kara hachigatsu made tsudui teiru koto wa jijitsu dearu. Season strawberry of NOM July from August to runinng is that TOP fact is.

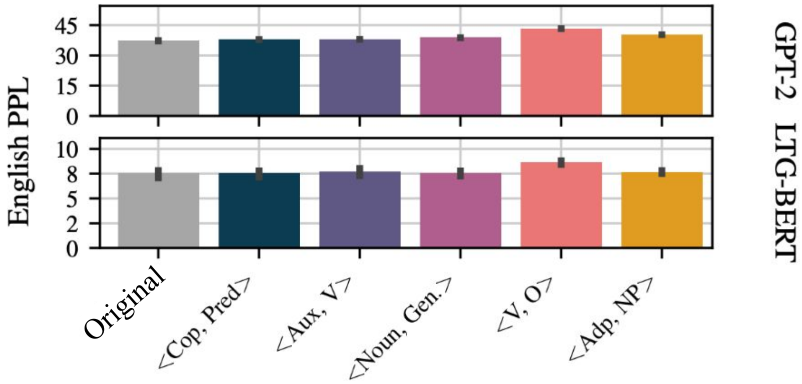
Pipeline

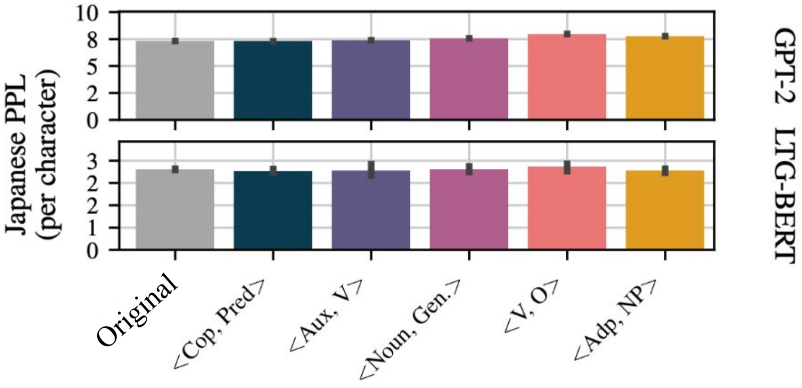

English Wikipedia
Japanese Wikipedia
Swapping Algorithm
GPT2s & LTG-BERTs

Data validation

	Train Data (En)			Train Data (Ja)		
Pair	Prec	Rec	Val	Prec	Rec	Val
<Cop, P.>	59.1	54.2	4.4	55.0	55.0	4.8
<Aux, V>	95.8	95.8	5.0	72.7	83.3	4.5
<N., Gen.>	80.0	80.0	4.8	81.0	81.0	4.8
<V, O>	74.4	73.4	4.3	85.9	81.6	4.2
<Adp, NP>	78.9	81.8	4.7	85.8	89.0	4.6

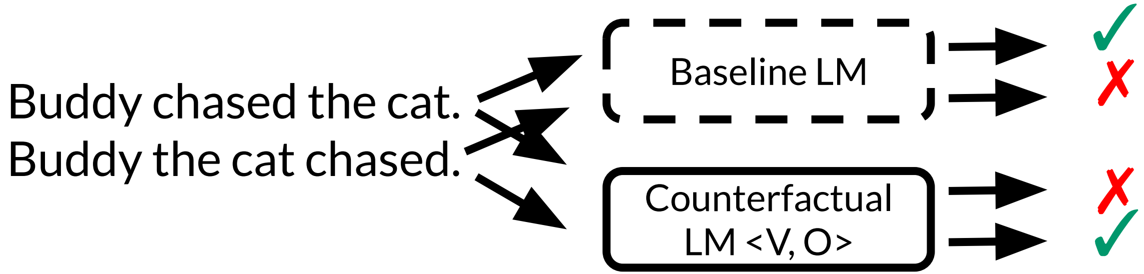
Evaluation: Perplexity

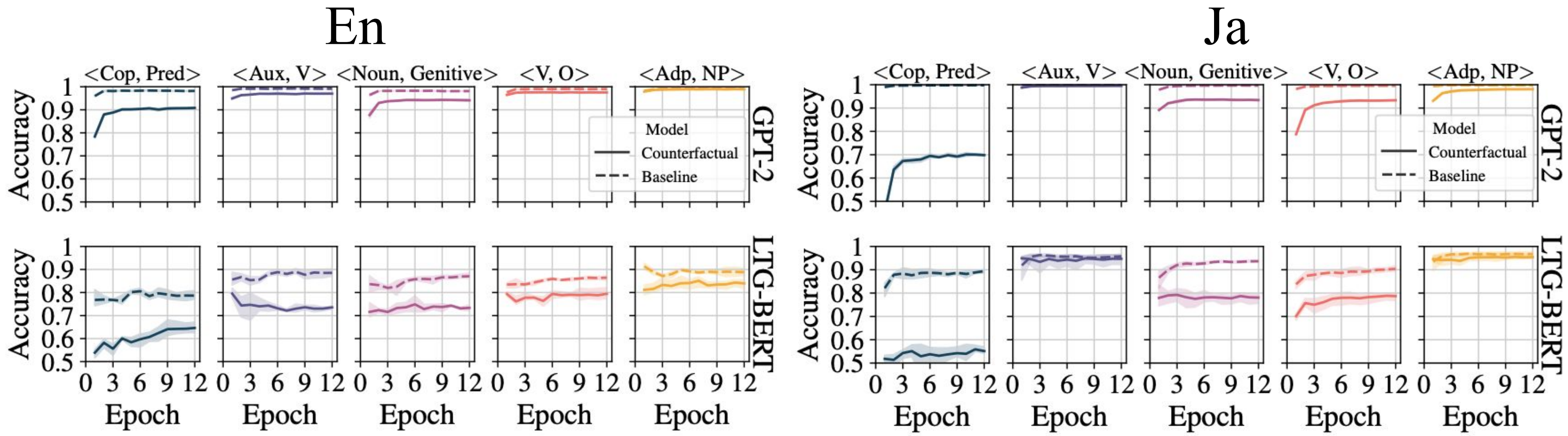




The differences between original and counterfactual are mostly NOT significant.
… BUT if we consider this comparison

Evaluation: Targeted Minimal Pairs





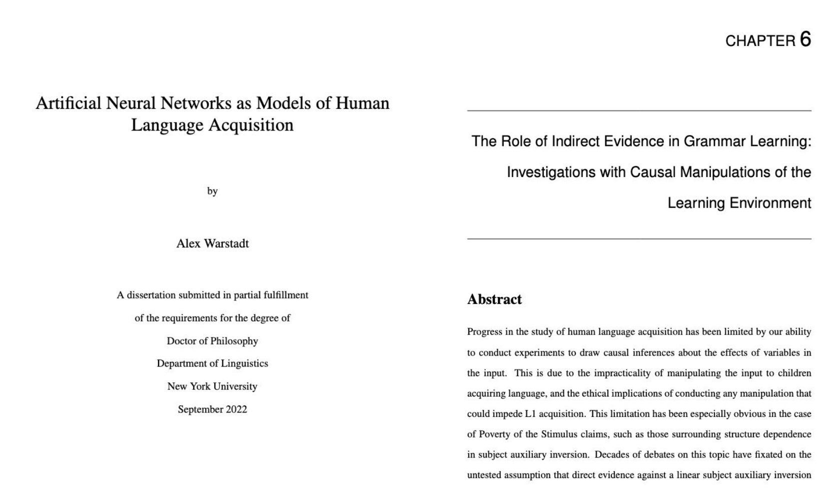
Counterfactual accuracy significantly less than baseline for <u>all settings</u>.

Typologically implausible languages seem to be somewhat harder for LMs to learn.

Implications:

1. Converging evidence with human artificial language learning experiments.
2. Evidence for domain generality of word order bias.

Indirect Evidence and the Poverty of the Stimulus



Artificial Neural Networks as Models of Human Language Acquisition

by

Alex Warstadt

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Linguistics

New York University

September 2022

CHAPTER 6

The Role of Indirect Evidence in Grammar Learning:

Investigations with Causal Manipulations of the

Learning Environment

Abstract

Progress in the study of human language acquisition has been limited by our ability to conduct experiments to draw causal inferences about the effects of variables in the input. This is due to the impracticality of manipulating the input to children acquiring language, and the ethical implications of conducting any manipulation that could impede L1 acquisition. This limitation has been especially obvious in the case of Poverty of the Stimulus claims, such as those surrounding structure dependence in subject auxiliary inversion. Decades of debates on this topic have fixated on the untested assumption that direct evidence against a linear subject auxiliary inversion

Artificial Neural Networks as Models of Human Language Acquisition

by

Alex Warstadt

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Linguistics

New York University

September 2022

CHAPTER 6

The Role of Indirect Evidence in Grammar Learning: Investigations with Causal Manipulations of the Learning Environment

Abstract

Progress in the study of human language acquisition has been limited by our ability to conduct experiments to draw causal inferences about the effects of variables in the input. This is due to the impracticality of manipulating the input to children acquiring language, and the ethical implications of conducting any manipulation that could impede L1 acquisition. This limitation has been especially obvious in the case of Poverty of the Stimulus claims, such as those surrounding structure dependence in subject auxiliary inversion. Decades of debates on this topic have fixated on the untested assumption that direct evidence against a linear subject auxiliary inversion

How does the distribution of syntactic phenomena in the input affect grammatical generalization?

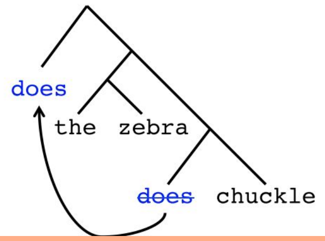
Subject Auxiliary Inversion

The zebra **does** chuckle. → **Does** the zebra chuckle?

Surface Generalization:
Move the first auxiliary to the front.

???

Linguistic Generalization:
Move the structurally highest auxiliary to the front.



Adults always acquire the linguistic generalization… Children never even entertain the surface generalization.

Example: M (Crain and Nakayama, 1987)

Poverty of the stimulus → Innate bias?

"Surely, if children hear enough [disambiguating examples], then they could reject the [linear] hypothesis. But if such evidence is virtually absent from the linguistic data, one cannot but conclude that children do not entertain the [linear] hypothesis, because the knowledge of structure dependency is innate."

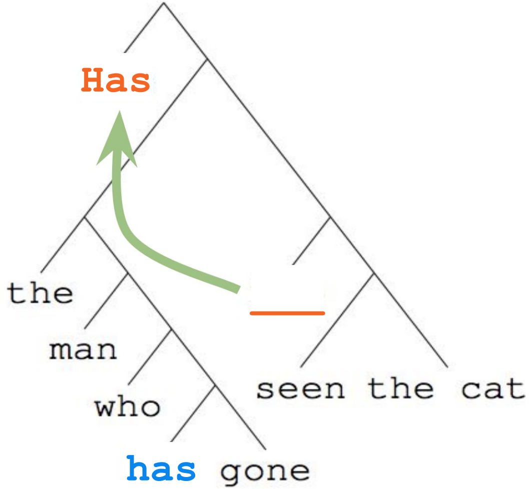
(Legate & Yang, 2001)

The man who **has** gone **has** seen the cat.

Surface Generalization:
Move the first auxiliary to the front.

Has the man who ___ gone **has** seen the cat?

Linguistic Generalization:
Move the structurally highest auxiliary to the front.



The Indirect Evidence Hypothesis

While a child may not receive direct evidence about the correctness of a particular hierarchical phrase structure rule…, there is vast indirect evidence for the general superiority of syntax with that structure throughout language. A learner who adopts a hierarchical phrase structure framework for describing the syntax of English will arrive at a much simpler, more explanatory account of her observations than a learner who adopts a linear framework.

(Perfors, Tenenbaum, Regier, 2011)

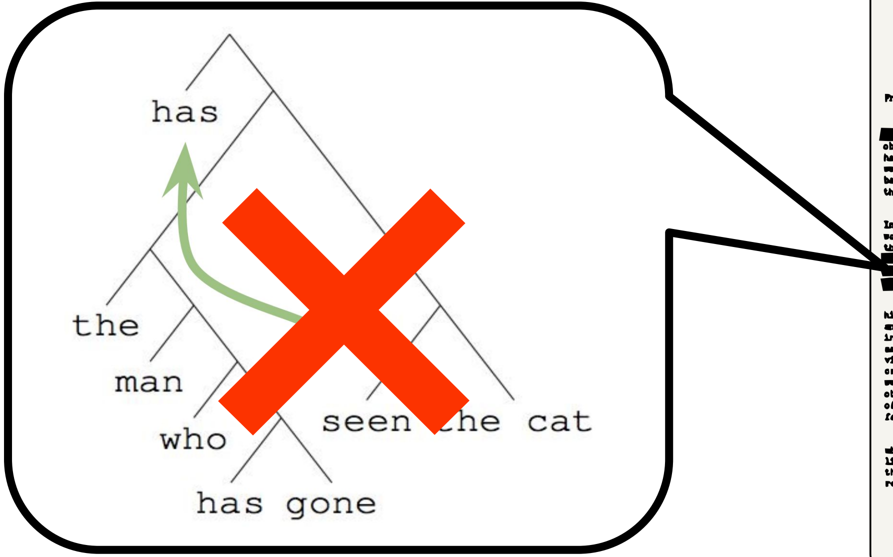
LMs and Subject Auxiliary Inversion

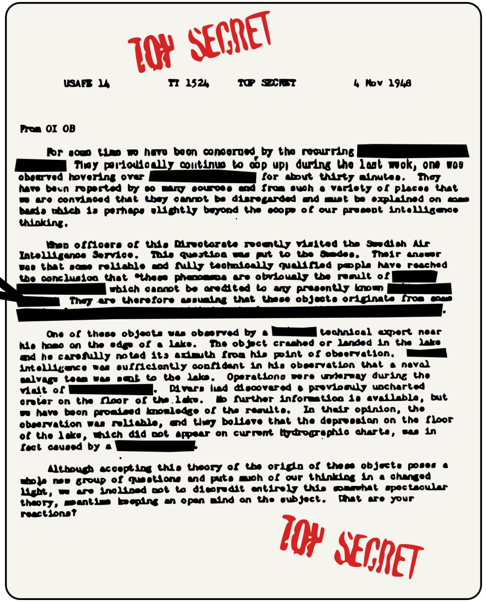
Earlier findings:

- LMs **trained from scratch** on ambiguous data usually adopt the **surface generalization.** (McCoy, Frank, and Linzen, 2018, 2020; Petty and Frank, 2022)
- **Pretrained** LMs fine-tuned on ambiguous data usually adopt the **linguistic generalization.** (Warstadt and Bowman, 2020; Mueller et al. 2020)

Confound: Pretraining data contains some direct evidence.

Language Deprivation Experiment





TOP SECRET

USAFE 14 TT 1524 TOP SECRET 4 Nov 1948

From OI OB

For some time we have been concerned by the recurring [redacted] They periodically continue to pop up; during the last week, one was observed hovering over [redacted] for about thirty minutes. They have been reported by so many sources and from such a variety of places that we are convinced that they cannot be disregarded and must be explained on some basis which is perhaps slightly beyond the scope of our present intelligence thinking.

When officers of this Directorate recently visited the Swedish Air Intelligence Service. This question was put to the Swedes. Their answer was that some reliable and fully technically qualified people have reached the conclusion that "these phenomena are obviously the result of [redacted] which cannot be credited to any presently known [redacted] They are therefore assuming that these objects originate from some [redacted].

One of these objects was observed by a [redacted] technical expert near his home on the edge of a lake. The object crashed or landed in the lake and he carefully noted its azimuth from his point of observation. [redacted] intelligence was sufficiently confident in his observation that a naval salvage team was sent to the lake. Operations were underway during the visit of [redacted]. Divers had discovered a previously uncharted crater on the floor of the lake. No further information is available, but we have been promised knowledge of the results. In their opinion, the observation was reliable, and they believe that the depression on the floor of the lake, which did not appear on current hydrographic charts, was in fact caused by a [redacted].

Although accepting this theory of the origin of these objects poses a whole new group of questions and puts much of our thinking in a changed light, we are inclined not to discredit entirely this somewhat spectacular theory, meantime keeping an open mind on the subject. What are your reactions?

TOP SECRET

Questions:

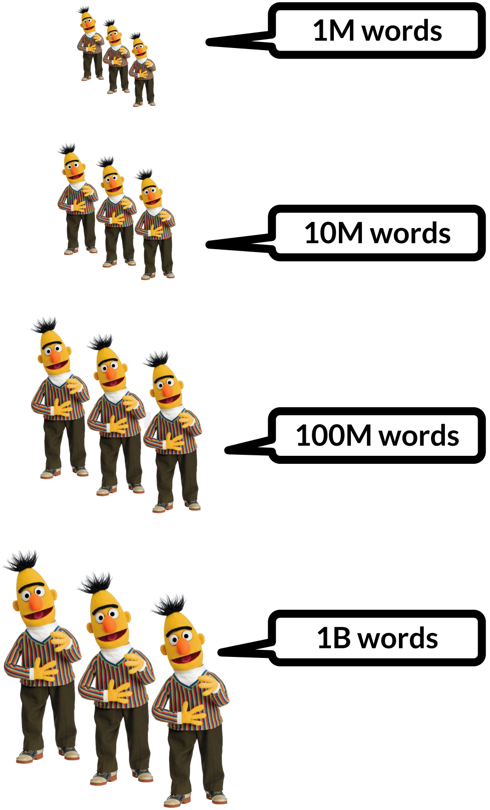
1. Does direct evidence have a causal impact on generalization?
2. Is indirect evidence sufficient to learn the linguistic generalization?

Models

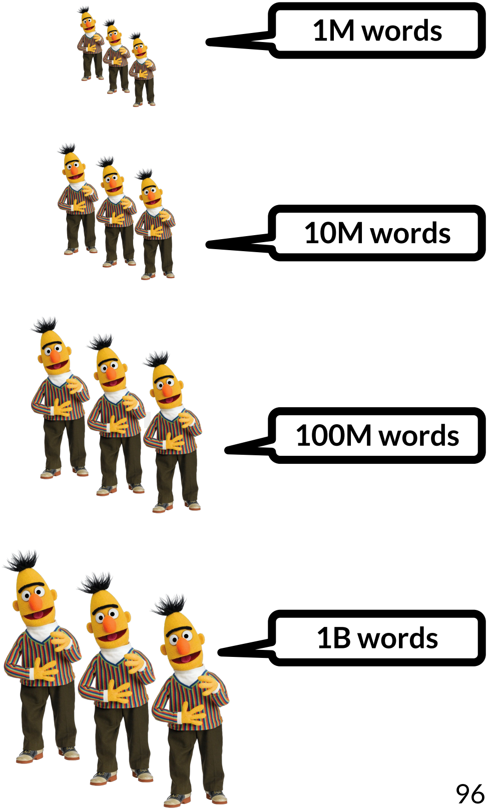
48 RoBERTa models pretrained from scratch

- 2 main conditions
- 4 sizes
- 3 runs (failed runs discarded)
- 2 domains (written, spoken)

Filtered Condition



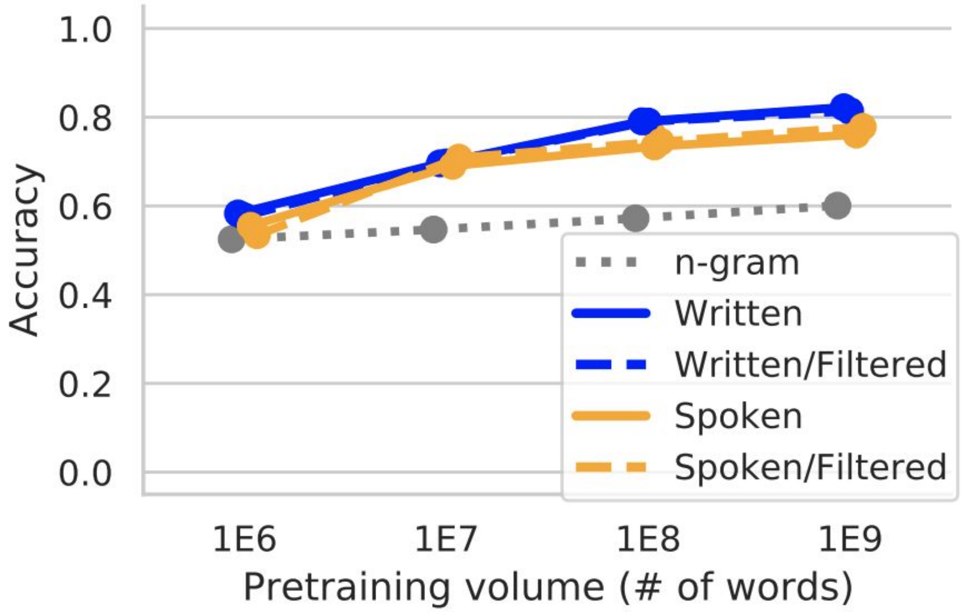
Unfiltered Condition (control)



Results: General acceptability judgments on BLiMP

Question: Did the removal of direct evidence have effects on unrelated phenomena?

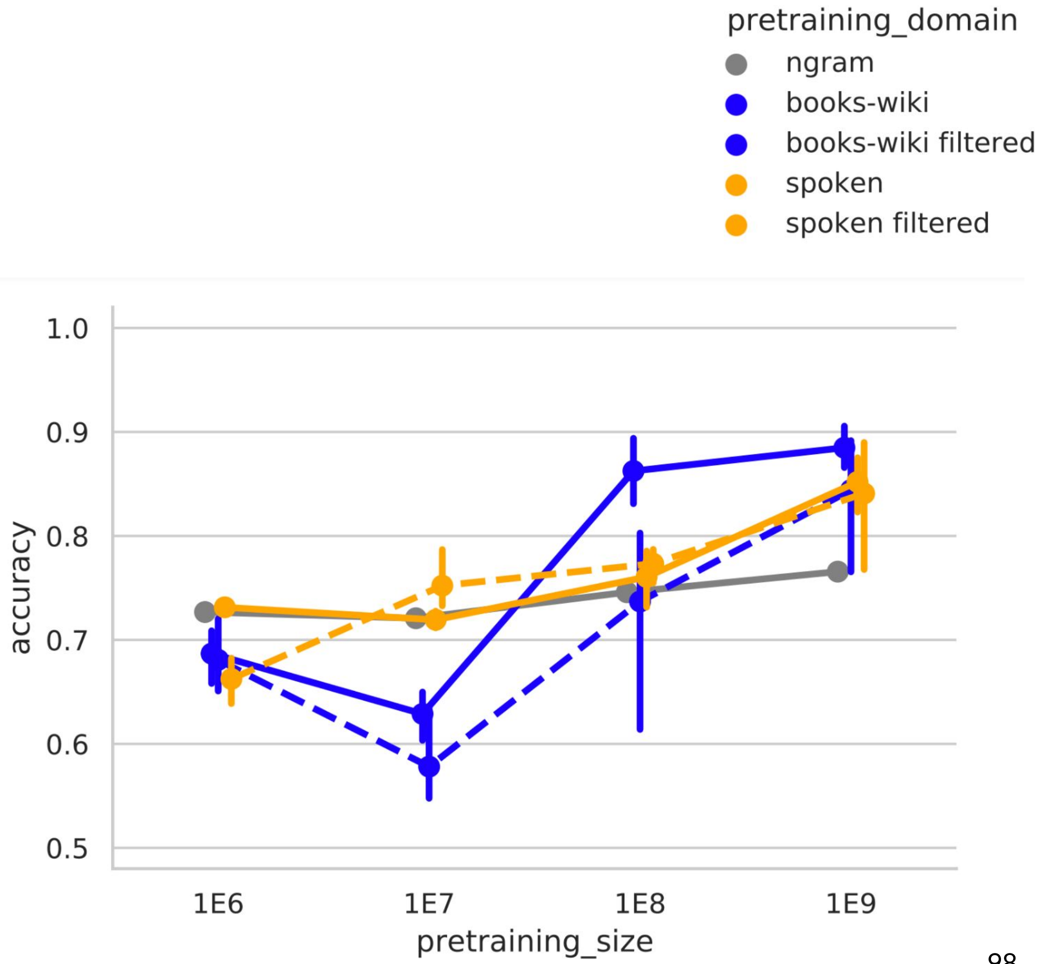
Answer: No



Results: Subject Aux Inversion

Question: Did the removal of direct evidence affect generalization on subject auxiliary inversion?

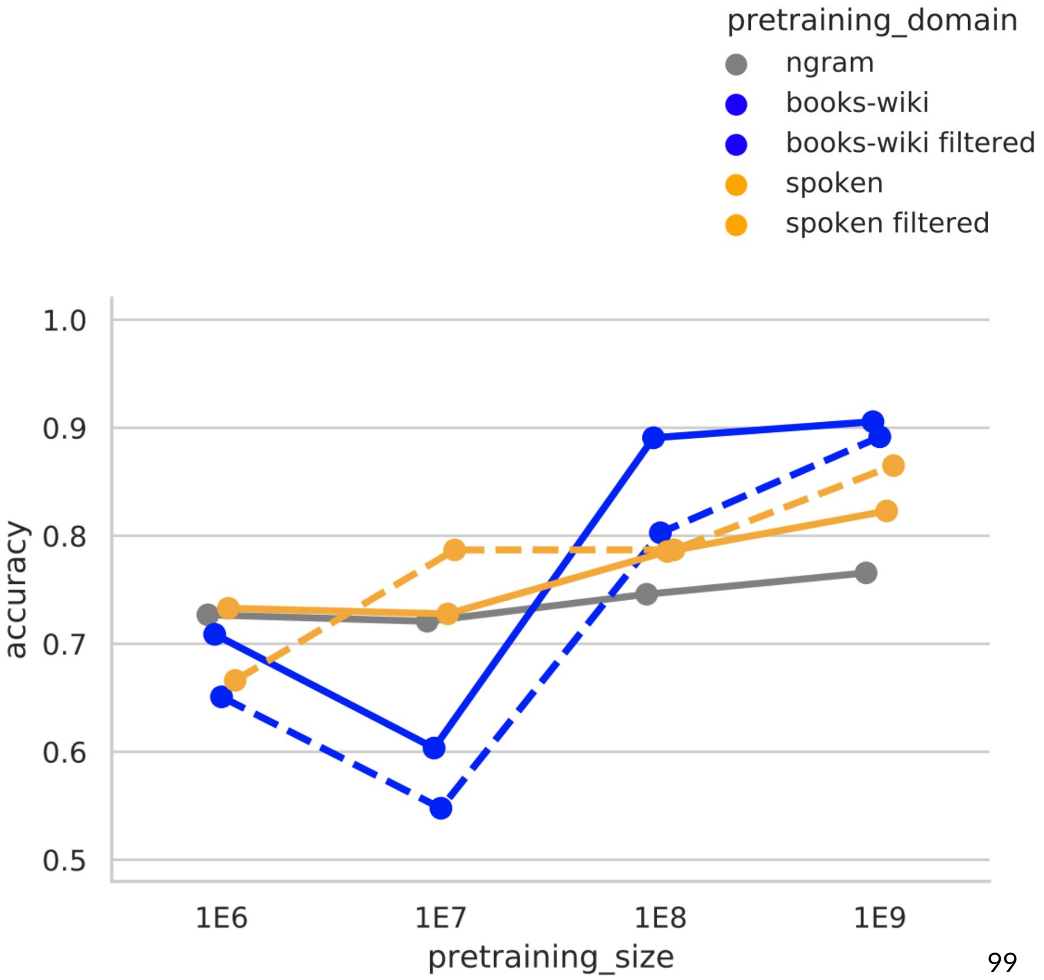
Answer: Slightly, only in the written domain.



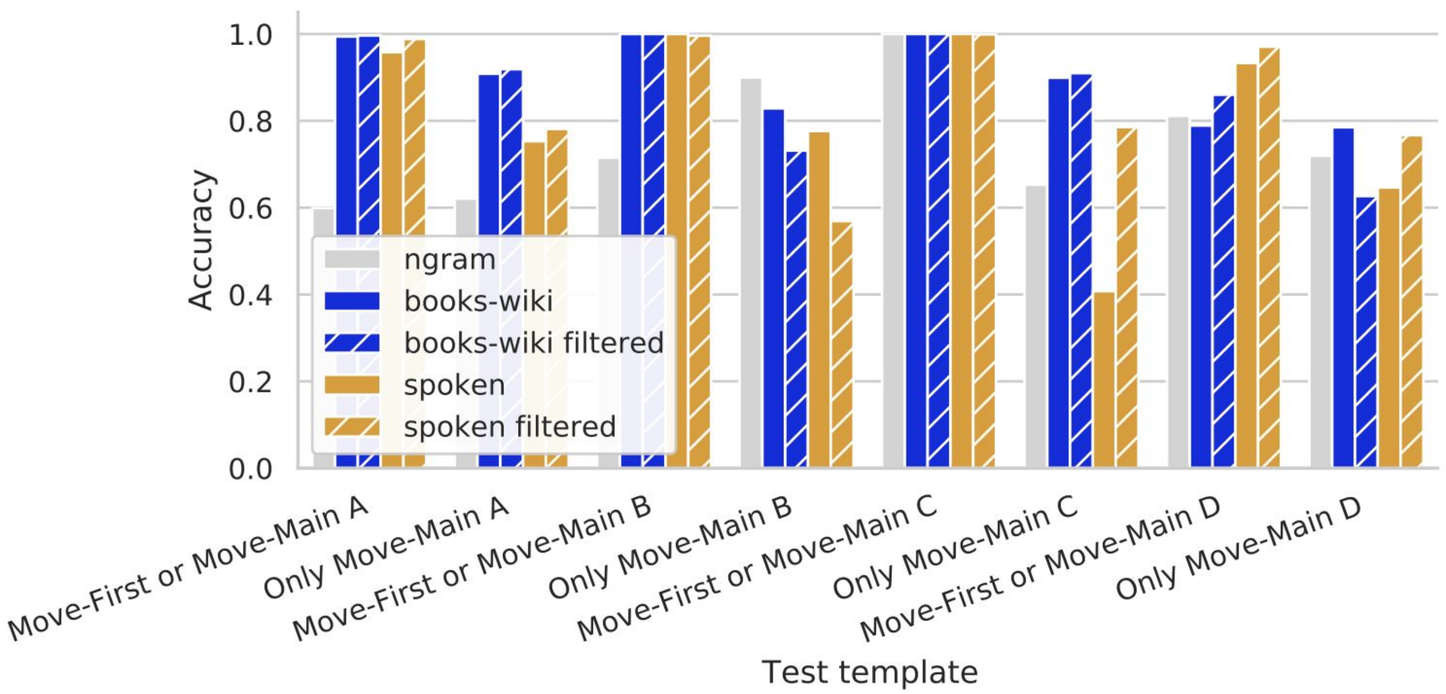
Results: Subject Aux Inversion

Question: Is indirect evidence sufficient to acquire the linguistic generalization?

Answer: Yes, but only in the best case.



Results: Subject Aux Inversion (BEST CASE)



Why it Matters That Babies and Language Models are the Only Known Language Learners

1. Improve data efficiency in LMs
2. Reverse engineer to determine the sufficient conditions for human-like acquisition
3. Map out the space of competent language learners
4. Suggest what is idiosyncratic and likely innate about human learning
5. Determine which learning biases are innate vs. domain-general
6. Establish causal relations between environmental variables and outcomes
7. …

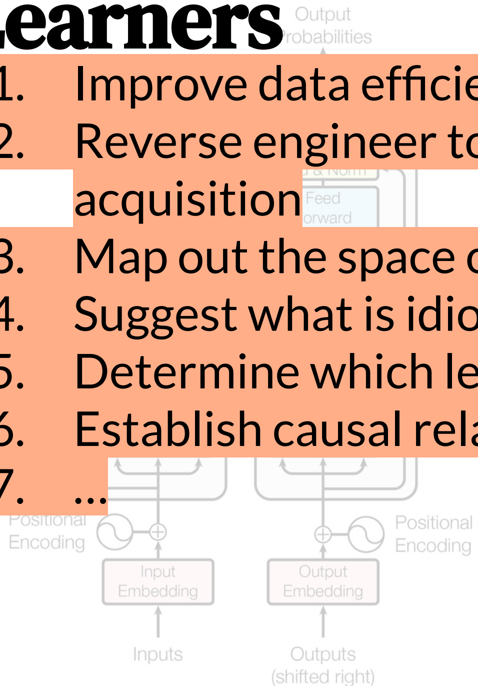


Figure 1: The Transformer - model architecture.

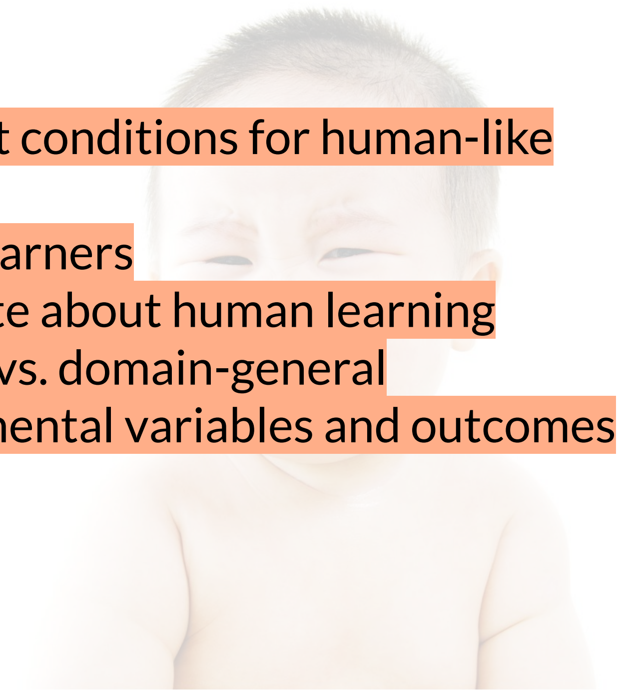


Figure 2: Human baby

Thank you

⭐ My stellar coauthors ⭐

Leshem Choshen, Juan Ciro, Ionut Constantinescu, Ryan Cottrell, Filippo Ficarra, Michael Hu, Tatsuki Kuribayashi, Tal Linzen, Rafael Mosquera, Aaron Mueller, Yohei Oseki, Tiago Pimentel, Candace Ross, Ethan Wilcox, Adina Williams, Tianyang Xu, Chengxu Zhuang