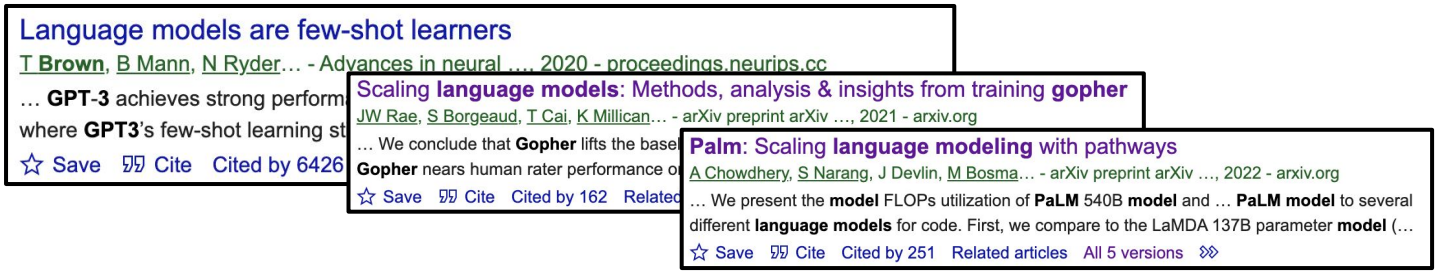
What Neural Networks Can Teach Us About Language Acquisition

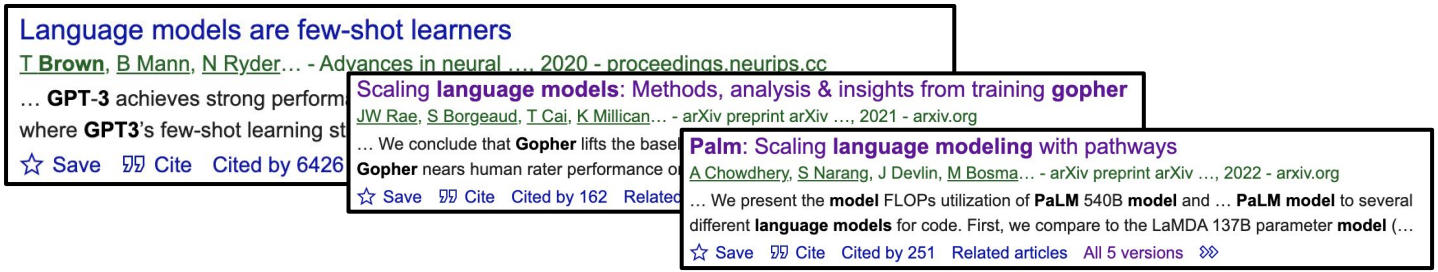
Alex Warstadt
ETH Zurich

Presented at Aix-Marseille University
1 December 2022

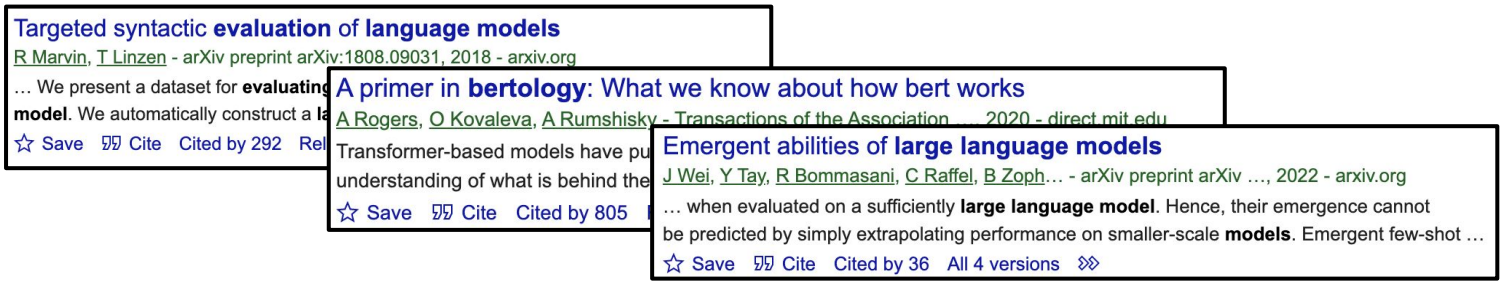
In NLP, there are bigger and more powerful language models (LMs) all the time.



In NLP, there are bigger and more powerful language models (LMs) all the time.



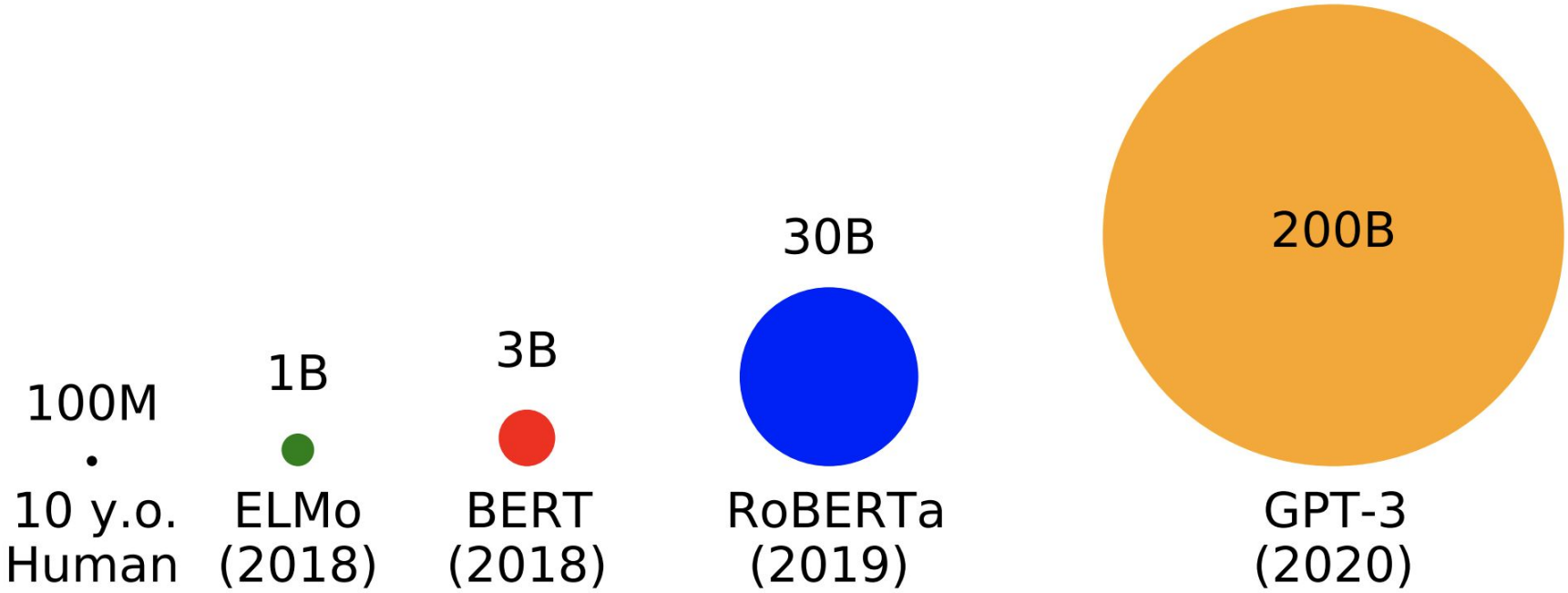
And researchers are racing to evaluate their strengths and limitations.



So what currently holds the state-of-the-art in grammar learning?

So what currently holds the state-of-the-art in grammar learning?

Humans are data-efficient language learners.



Roadmap

1. The Argument

Why study language models as models of human learners? What kinds of questions can they address?

2. The Learner

Training LMs in more plausible learning environments, without major advantages over humans.

3. A Controlled Experiment

A proof-of-concept: Language models unlock new modes of testing learnability hypotheses

Part 1: The Argument

1. The Argument

Why study language models as models of human learners? What kinds of questions can they address?

2. The Learner

Training LMs in more plausible learning environments, without major advantages over humans.

3. A Controlled Experiment

A proof-of-concept: Language models unlock new modes of testing learnability hypotheses

How do we make LMs into plausible cognitive models?

Why study LMs as models of human learners?

What kinds of questions can they address?

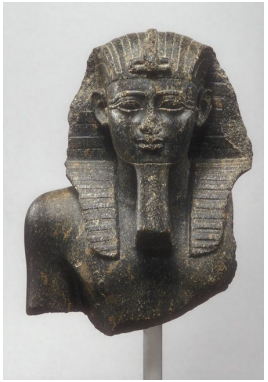
What Artificial Neural Networks Can Tell Us About Human Language Acquisition*

Alex Warstadt, Samuel R. Bowman

August 18, 2022

How does input variable X affect learning outcome Y?

Language Deprivation Experiments



Pharaoh Psamtik
(664 – 610 BCE)

Frederick II
(1194-1250)

James IV
(1473-1513)

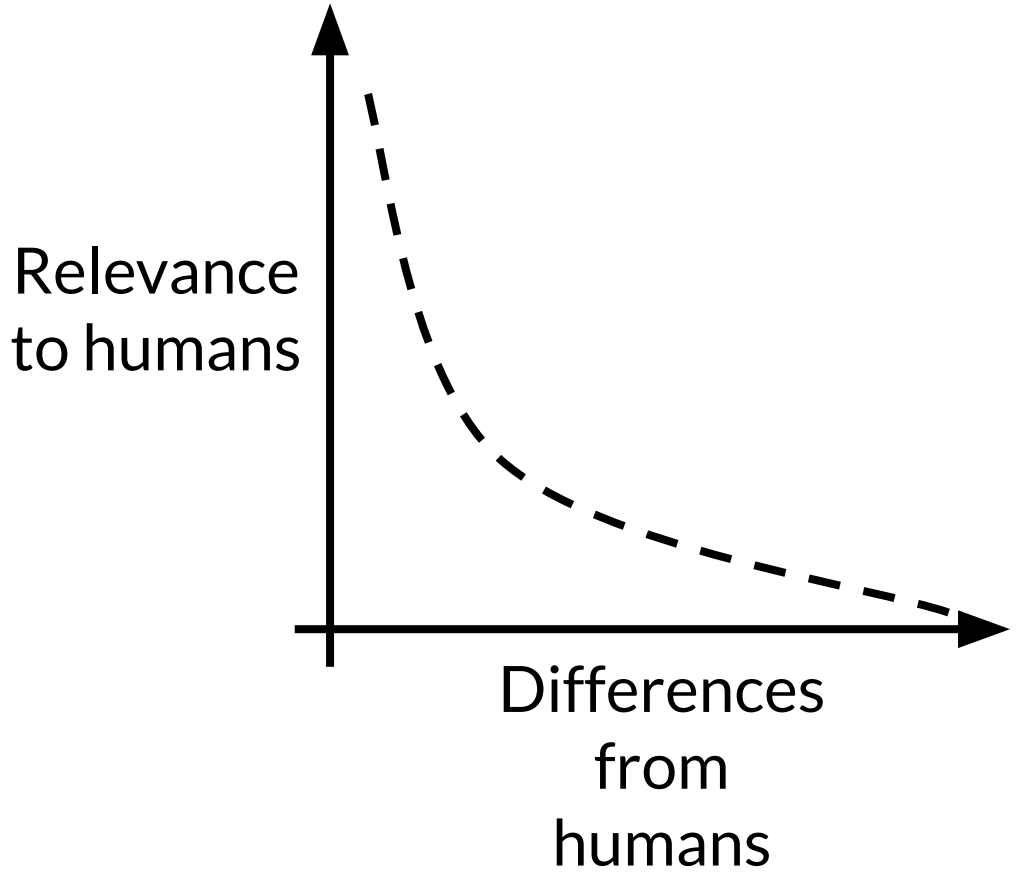
Three reasons to study neural networks instead of humans:

1. Ethics

1. Ethics
2. Expense

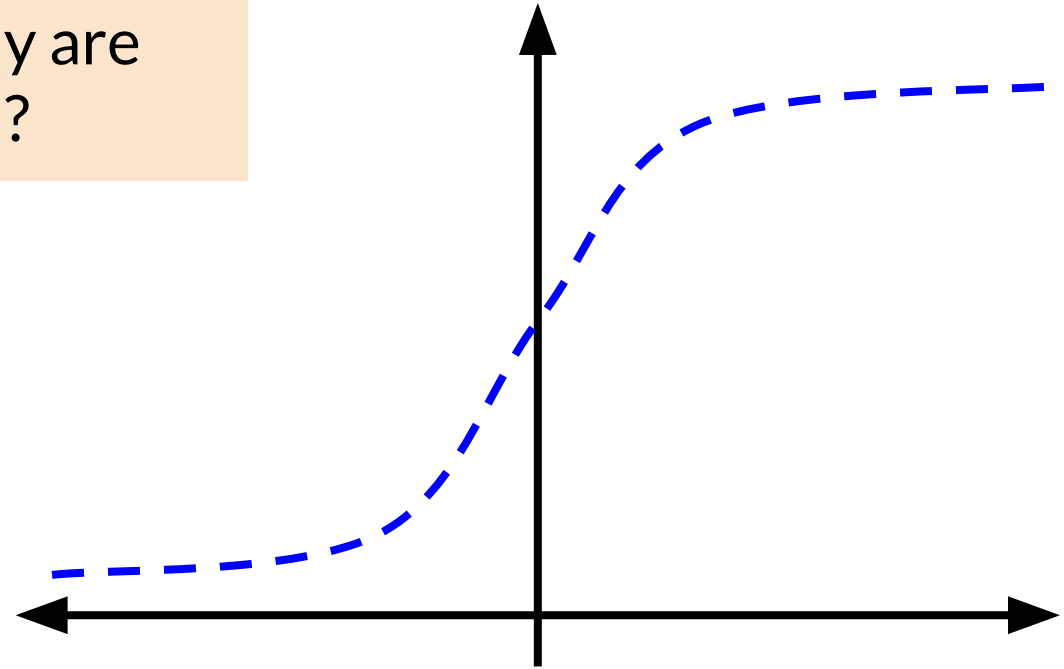
1. Ethics
2. Expense
3. Experimental paradigms

As with any scientific model, there are obvious limitations with LMs.

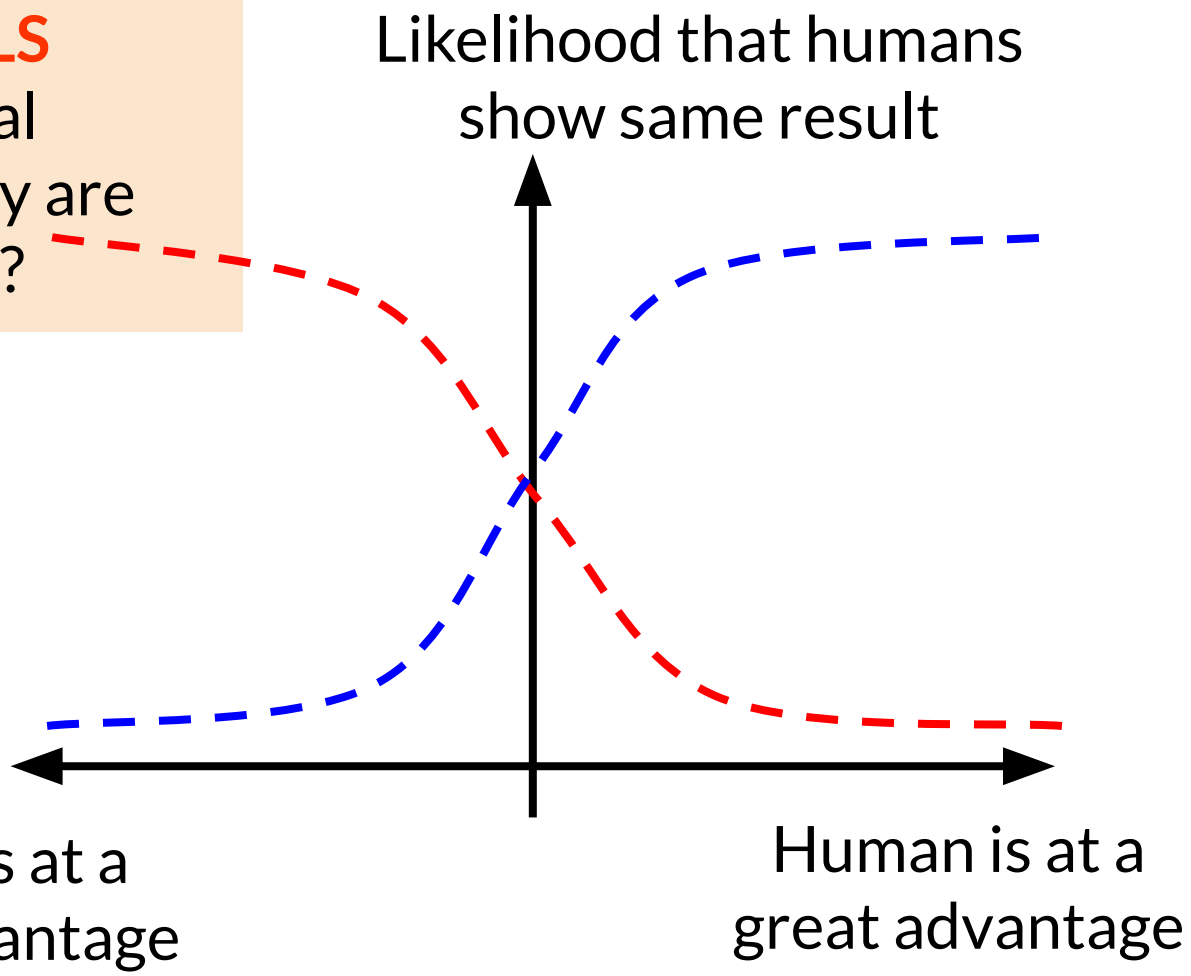

Relevance to humans
Differences from humans

Debates in language acquisition often center around the sufficient conditions for human-learnability.

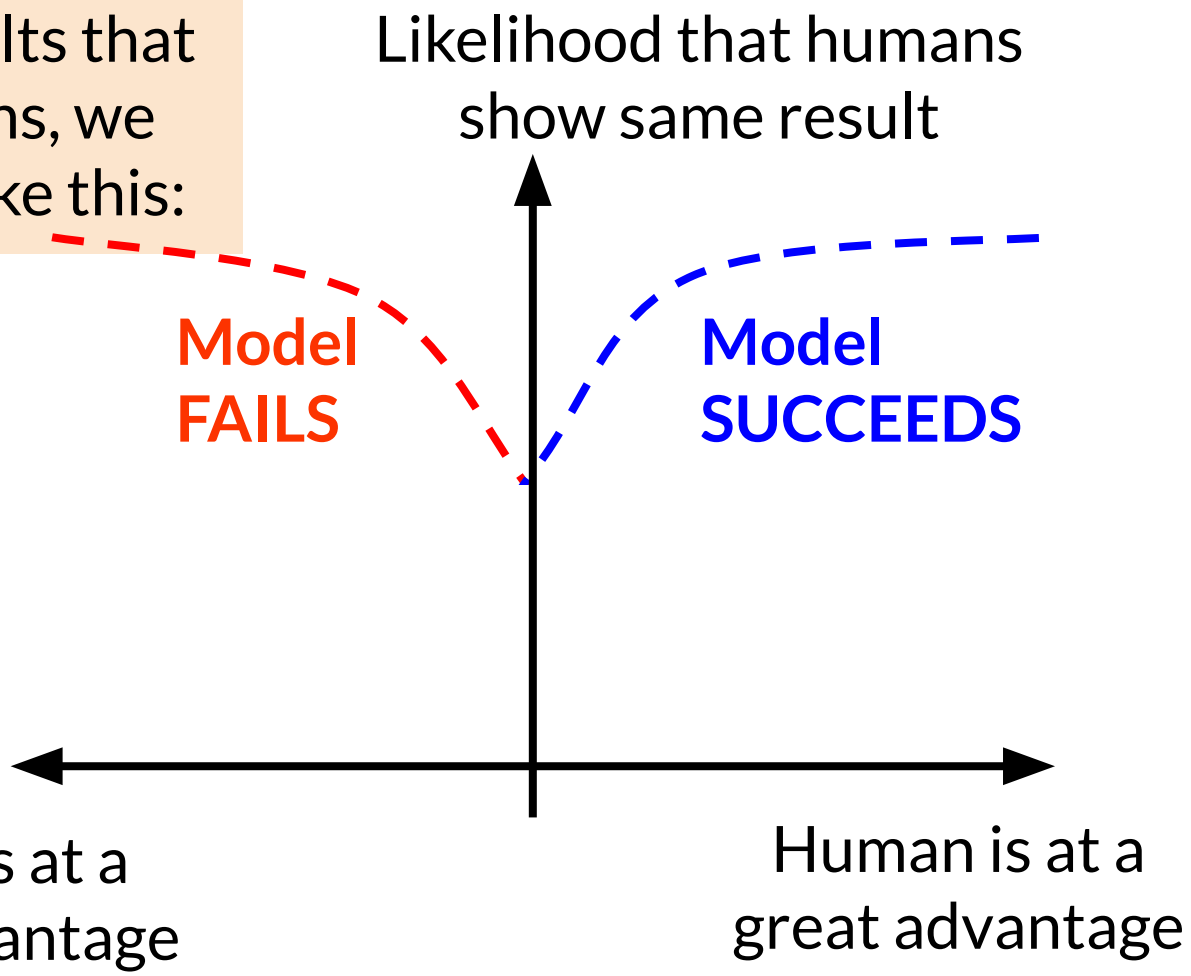
Suppose the model **SUCCEEDS** given some experimental manipulation. How likely are humans also to succeed?



Suppose the model **FAILS** given some experimental manipulation. How likely are humans also to succeed?

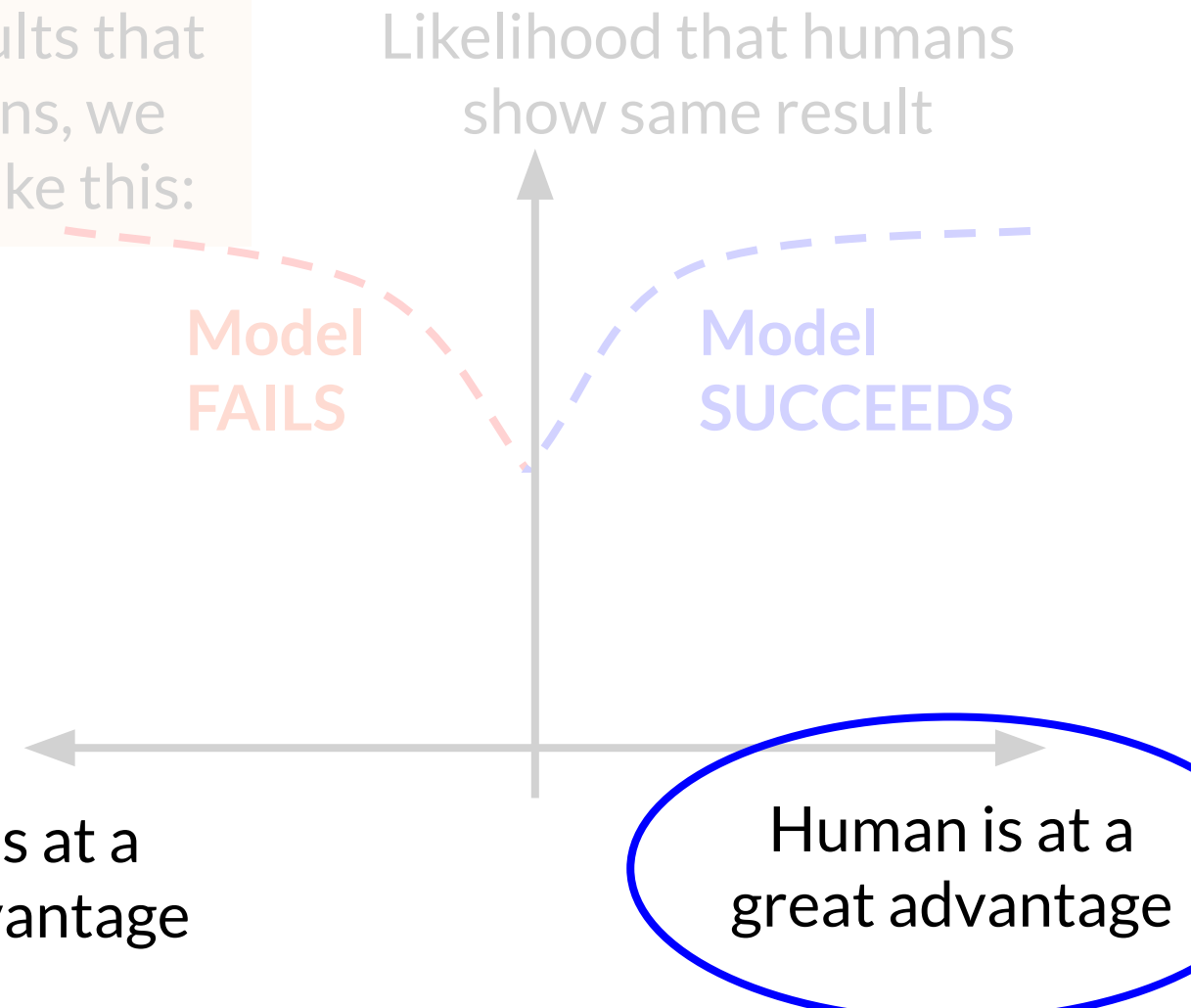


Assuming we want results that will generalize to humans, we should aim for results like this:



Assuming we want results that will generalize to humans, we should aim for results like this:

Assuming we want results that will generalize to humans, we should aim for results like this:



A recipe for relevant model learners:

A recipe for relevant model learners:

- Increase the relevance of positive by results by reducing the advantages that models have over humans.

A recipe for relevant model learners:

- Increase the relevance of positive by results by reducing the advantages that models have over humans.
- Increase the chance of positive results by providing models with advantages that humans have.

A recipe for relevant model learners:

- Increase the relevance of positive by results by reducing the advantages that models have over humans.
- Increase the chance of positive results by providing models with advantages that humans have.

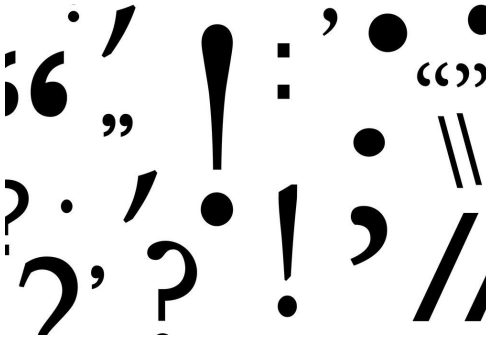
In other words, the best model learners will be ones whose environments and innate abilities are as rich as possible without being richer than those of humans.

Advantages ANNs Have

Data quantity

Data domain

Orthography



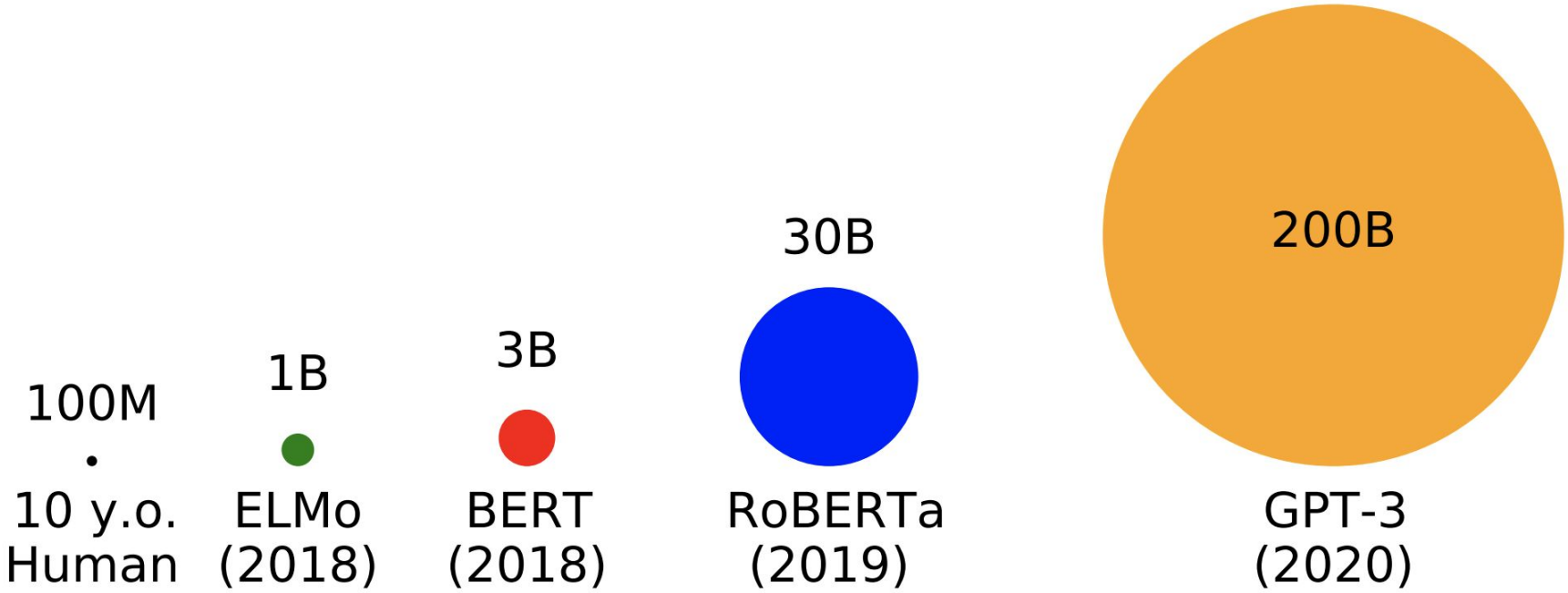
Advantages ANNs Have

Data quantity

Data domain

Orthography

Human vs. model linguistic input (# of word tokens)



Part 2: The Learner

1. The Argument

Why study language models as models of human learners? What kinds of questions can they address?

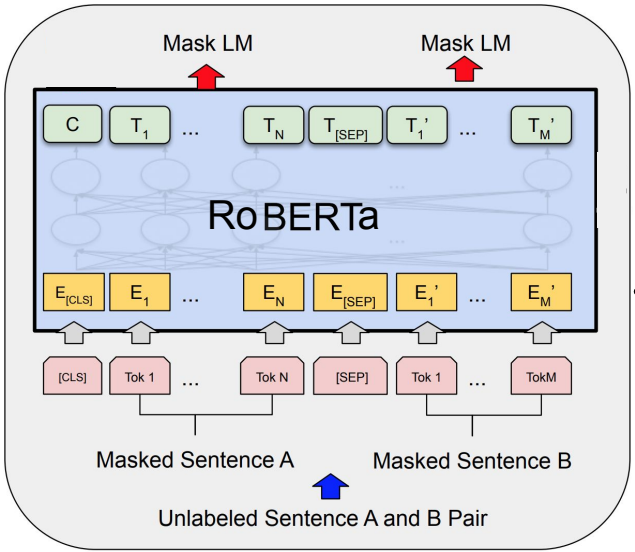
2. The Learner

Training LMs in more plausible learning environments, without major advantages over humans.

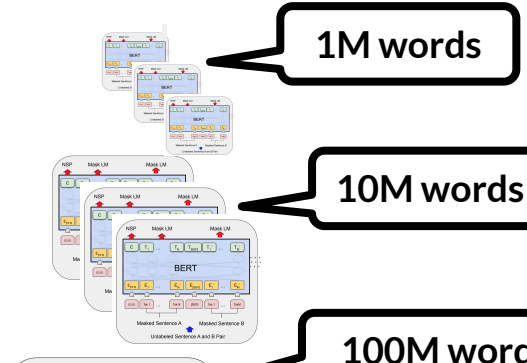
3. A Controlled Experiment

A proof-of-concept: Language models unlock new modes of testing learnability hypotheses

MiniBERTas



30B words



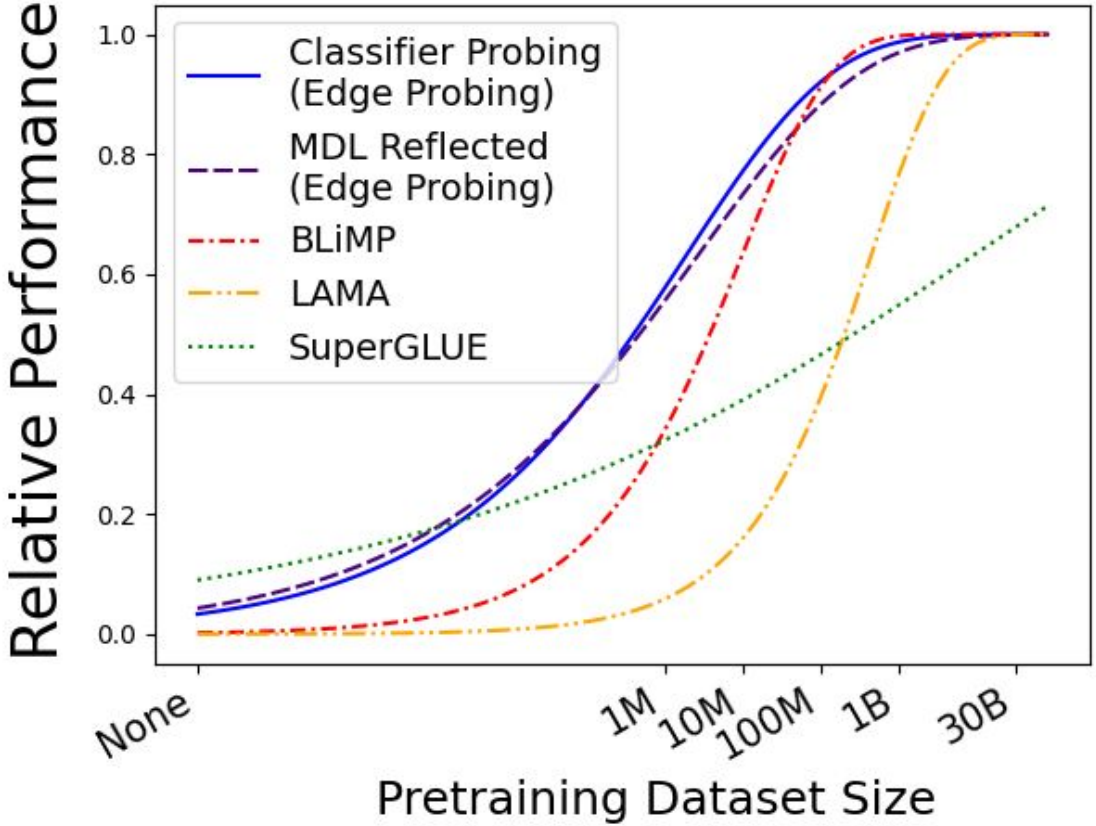
1M words

10M words

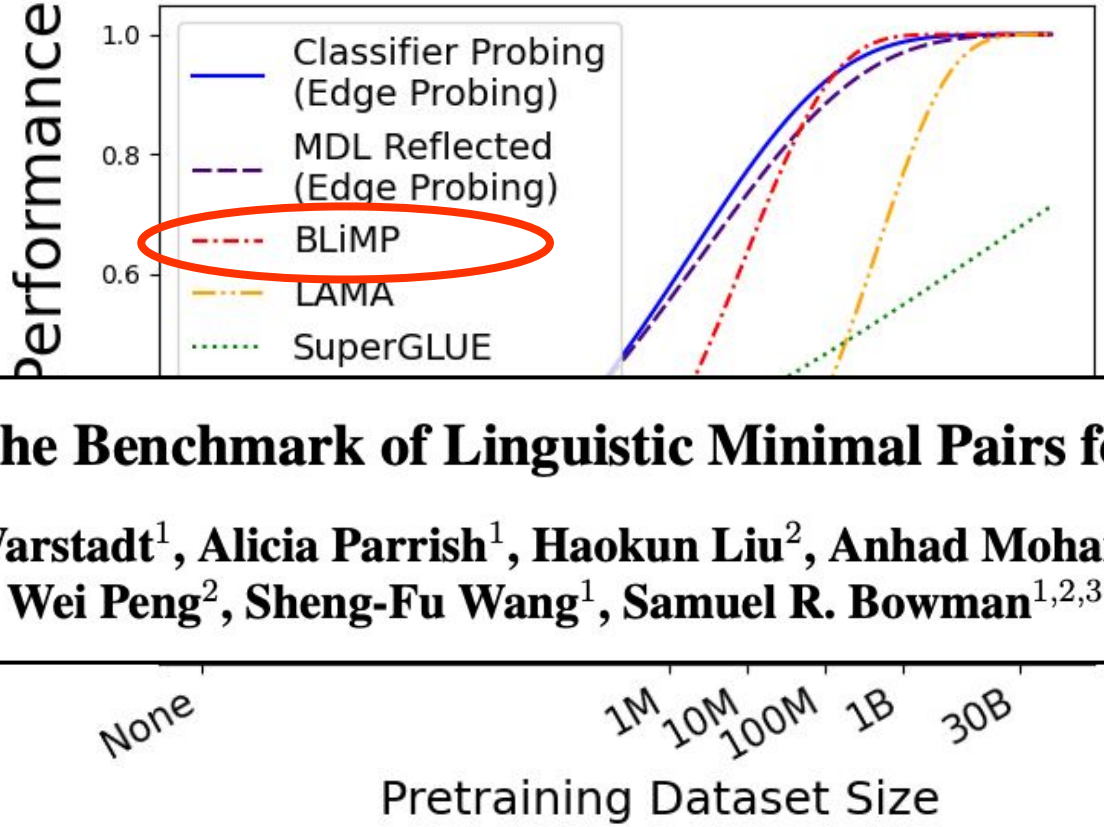
100M words

1B words

Five Sets of Probing Methods



Five Sets of Probing Methods



BLiMP: The Benchmark of Linguistic Minimal Pairs for English

Alex Warstadt[1], Alicia Parrish[1], Haokun Liu[2], Anhad Mohananey[2], Wei Peng[2], Sheng-Fu Wang[1], Samuel R. Bowman[1,2,3]

Minimal Pairs

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

Betsy is eager to sleep. ✓

Betsy is easy to sleep. ✗

BLiMP Categories

Morphology

- Anaphor agr.
- Determiner-noun agr.
- Irregular forms
- Subj-verb agr

Syntax

- Argument structure
- Binding
- Control/raising
- Ellipsis
- Filler-gap (wh) deps.
- Island constraints

Semantics

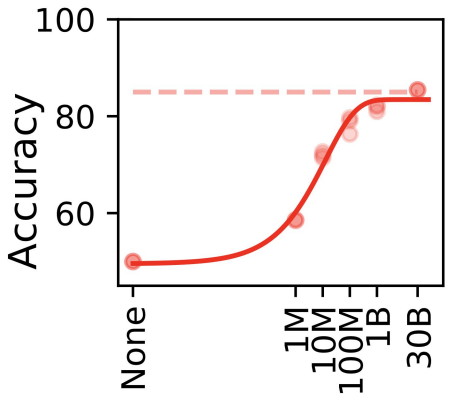
- NPIs
- Quantifiers

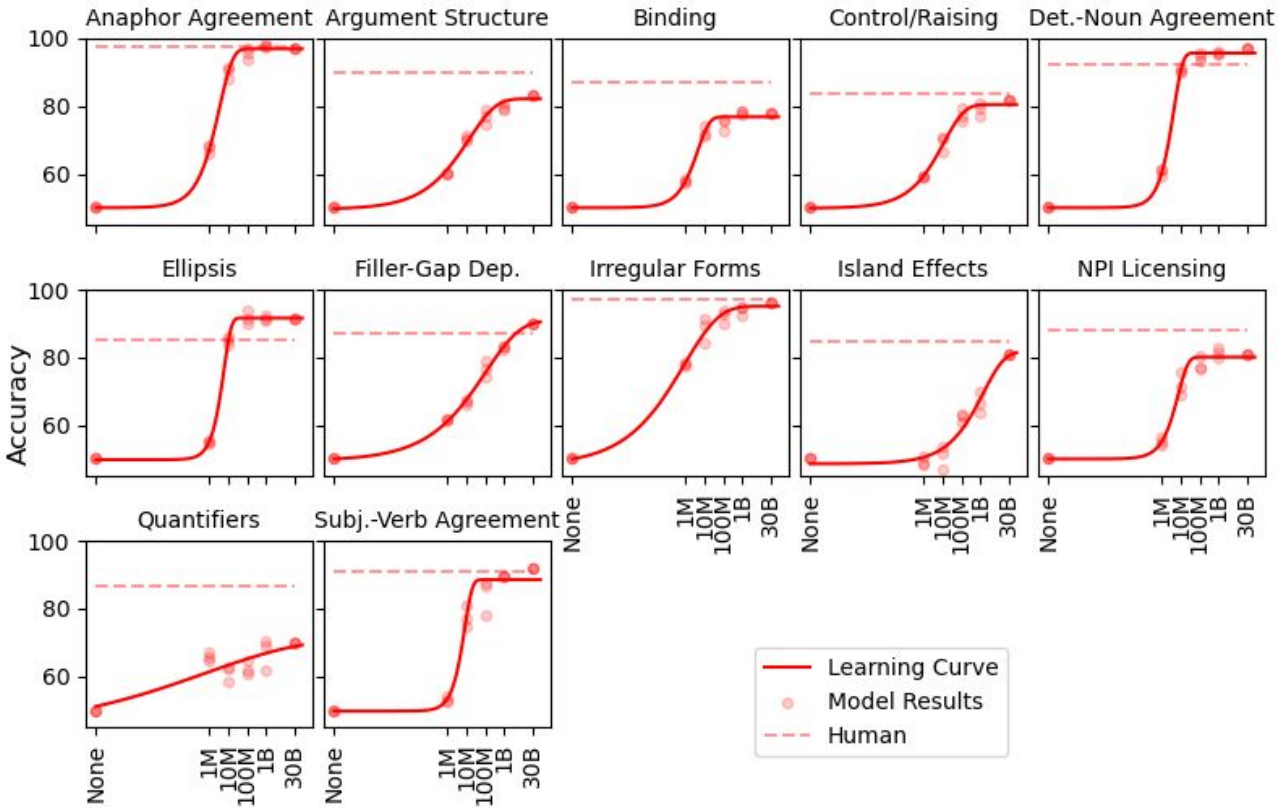
Data Sample

Phenomenon	Acceptable example	Unacceptable example
Anaphor agr.	Many girls insulted themselves.	Many girls insulted herself.
Det.-noun agr.	Rachelle had bought that chair.	Rachelle had bought that chairs.
Subject-verb agr.	These casseroles disgust Kayla.	These casseroles disgusts Kayla.
Filler-gap	Brett knew what many waiters find.	Brett knew that many waiters find.
Island effects	Which bikes is John fixing?	Which is John fixing bikes?

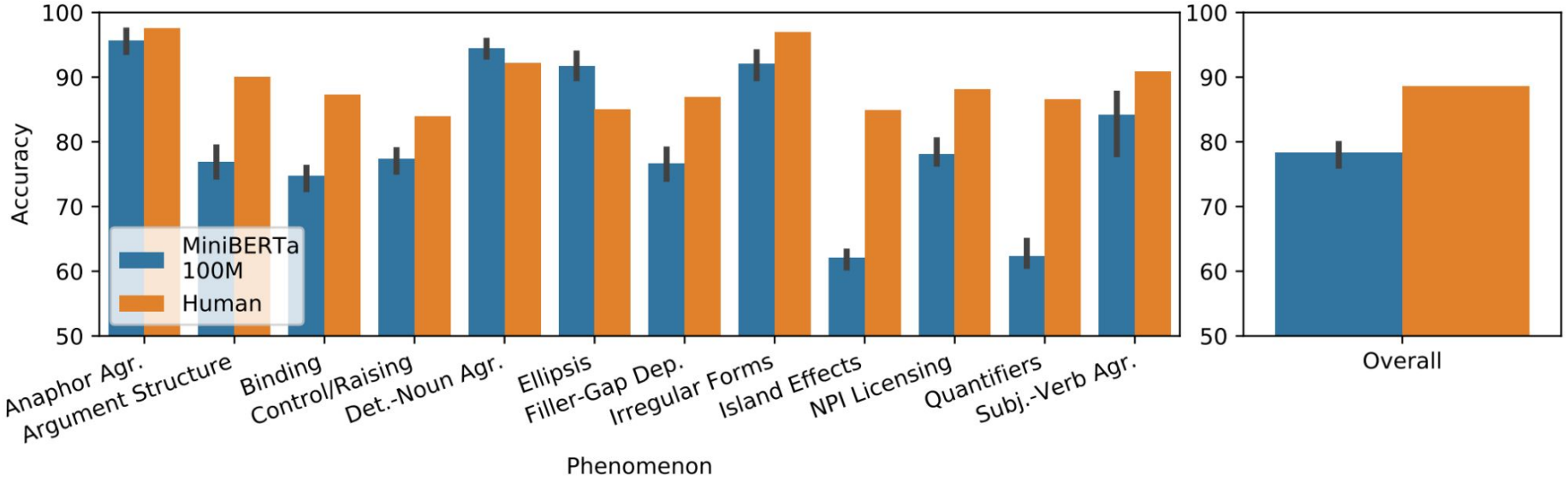
BLiMP Learning Curves

BLiMP Learning Curves





BLiMP:
Human vs. Human-scale model



UPCOMING Shared task @ CMCL/CoNLL 2023

BabyLM Challenge

Sample-efficient pretraining on a developmentally plausible corpus

Approximate Timeline:

December 2022: Training data released
February 2023: Shared evaluation pipeline published
June 2023: Submissions for presentation at CMCL due
September 2023: CMCL, Submissions for presentation at CoNLL
November 2023: CoNLL

Part 3: A Controlled Experiment

1. The Argument

Why study language models as models of human learners? What kinds of questions can they address?

2. The Learner

Training LMs in more plausible learning environments, without major advantages over humans.

3. A Controlled Experiment

A proof-of-concept: Language models unlock new modes of testing learnability hypotheses

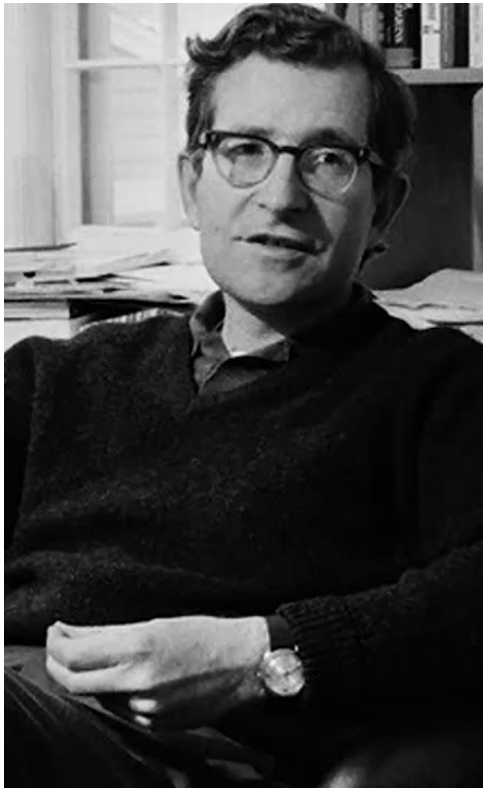
Testing the Poverty of the Stimulus:

Controlled experiments at the scale of human language learning

with Yian Zhang, Haau-Sing Li, and Samuel R. Bowman

There's a long history of debate about the role of innate bias in the acquisition of hierarchical syntactic rules in favor of alternative linear rules.

Chomsky (1965, 1971); Crain & Nakayama (1987); Lewis & Elman (2001); Pullum & Scholz (2002); Legate & Yang (2002); Reali & Christiansen (2005); Perfors, Tenenbaum, & Regier (2011); Berwick et al. (2011); Hsu, Chater, & Vitanyi (2013); McCoy, Frank, & Linzen (2018, 2020); Warstadt & Bowman (2020)



A theory that attributes possession of certain linguistic universals to a language-acquisition system [...] implies that only certain kinds of symbolic systems can be acquired [....] Specifically, grammatical transformations are necessarily "structure-dependent" [i.e., hierarchical] [....] It is impossible, however, [for the language-acquisition system to learn] a transformation such a simple operation as reflection of an arbitrary string.

(Chomsky, 1965)

Subject Auxiliary Inversion

The zebra **does** chuckle.

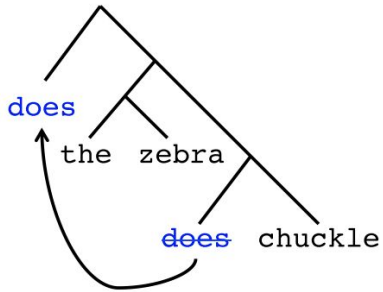
Does the zebra chuckle?

Example: McCoy et al. (2020)

MOVE-FIRST: Move the linearly first auxiliary to the front of the sentence.

MOVE-FIRST: Move the linearly first auxiliary to the front of the sentence.

MOVE-MAIN: Move the main verb's auxiliary to the front of the sentence.



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

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

MOVE-FIRST: Move the linearly first auxiliary to the front of the sentence.

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

MOVE-FIRST: Move the linearly first auxiliary to the front of the sentence.

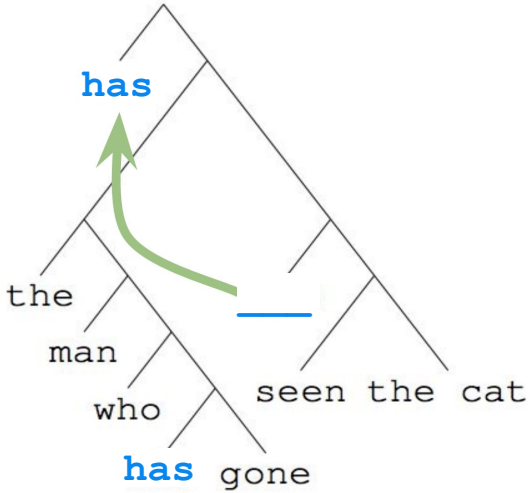
Has the man who ___ gone **has** seen the cat?

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

MOVE-MAIN: Move the main verb's auxiliary to the front of the sentence.

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

MOVE-MAIN: Move the main verb's auxiliary to the front of the sentence.



The Poverty of the Stimulus Argument relies on quantifying “sufficient evidence”:

“Surely, if children hear enough sentences like those [below], then they could reject the [move-first] hypothesis. But if such evidence is virtually absent from the linguistic data, one cannot but conclude that children do not entertain the [move-first] hypothesis, because the knowledge of structure dependency is innate.”

(Legate & Yang, 2001)

e.g., *Has the man who has gone seen the cat?*

The INDIRECT Evidence Hypothesis:

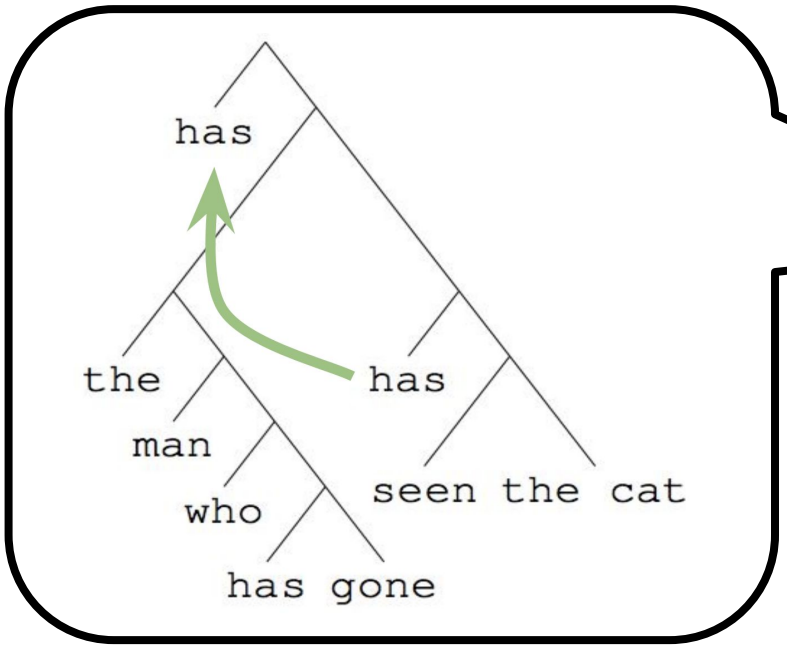
Indirect evidence may be sufficient for a learner without hierarchical bias to eliminate move-first.

Proponents of Indirect Evidence

“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 et al., 2011)

Syntactic filtering



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].

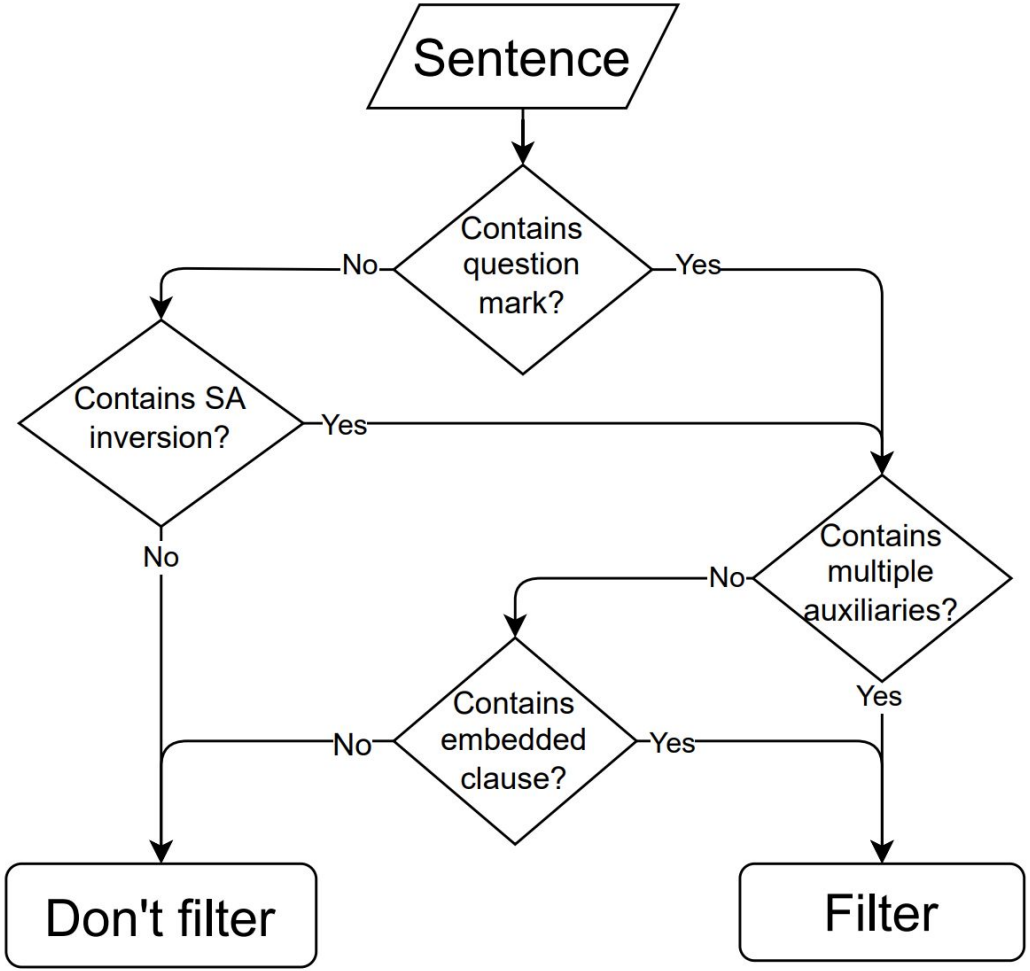
Although accepting this theory of the origin of these objects poses a whole new group of questions and puts much of our thinking in a changed light, we are inclined not to discredit entirely this somewhat spectacular theory, meantime keeping an open mind on the subject. What are your reactions?

TOP SECRET

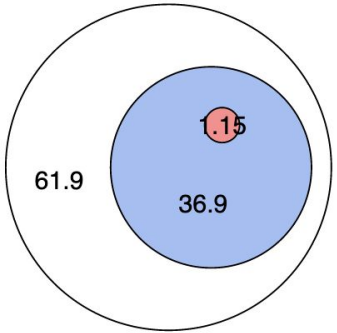
Syntactic filtering

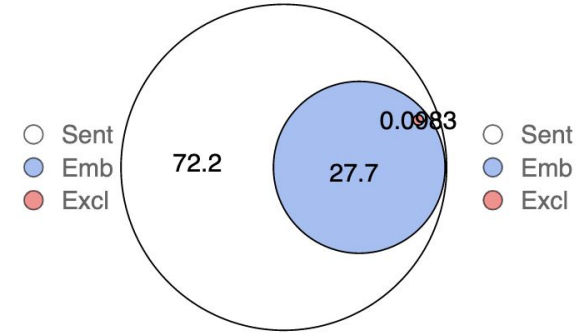
Training data: 1B words from books & Wikipedia

- Percent filtered: 1.7%
- Accuracy: 98%
- Recall (% of direct evidence removed): 99%



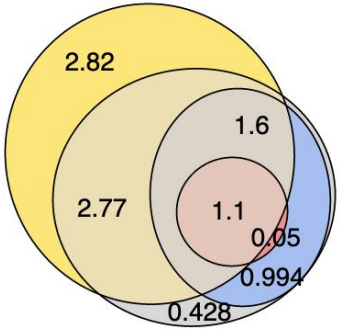
Distribution of direct evidence (by domain)

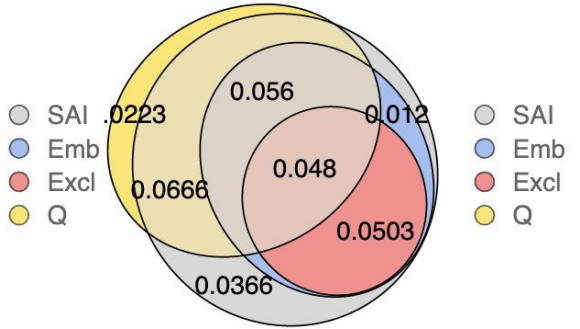




(a) Books (all sentences)

(b) Wikipedia (all sentences)





(c) Books (SAI ∪ Q)

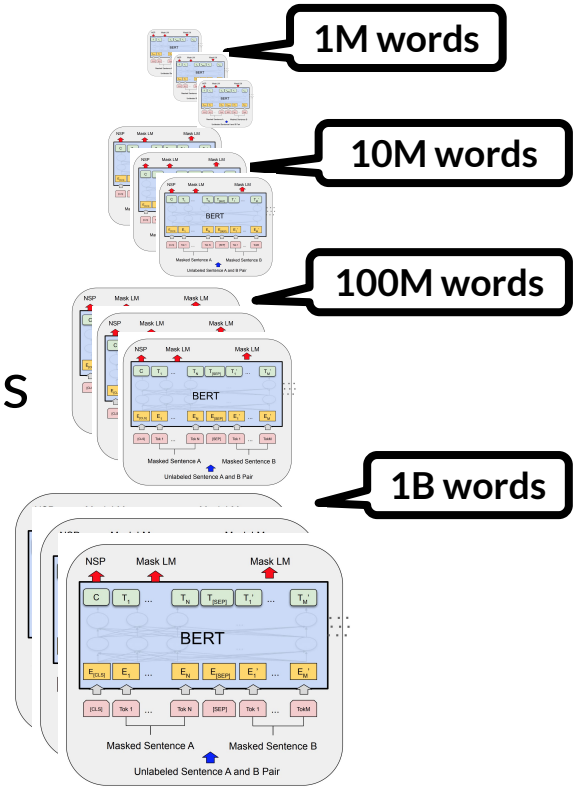
(d) Wikipedia (SAI ∪ Q)

Models

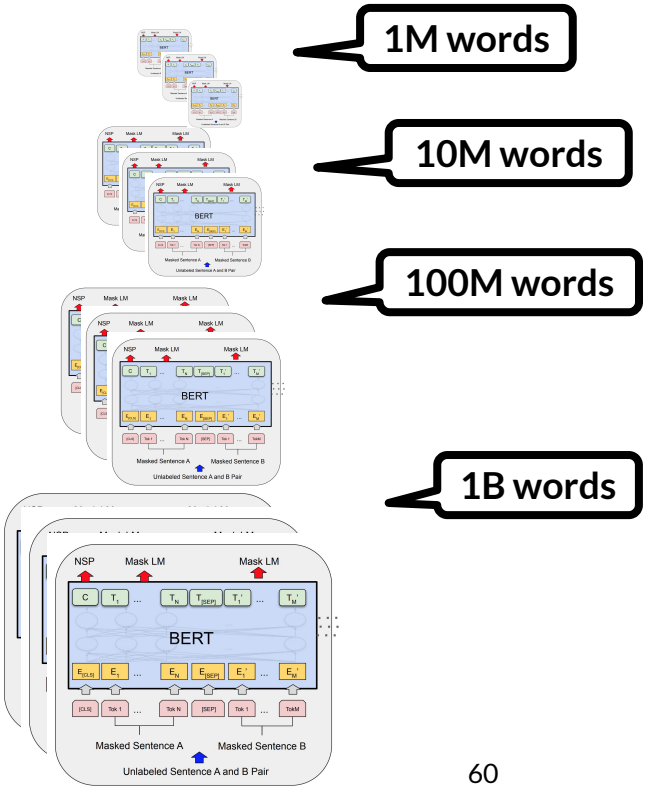
24 RoBERTa models pretrained from scratch.

- 2 main conditions
- 4 sizes
- 3 runs (failed runs discarded)

Filtered Condition



Unfiltered Condition (control)



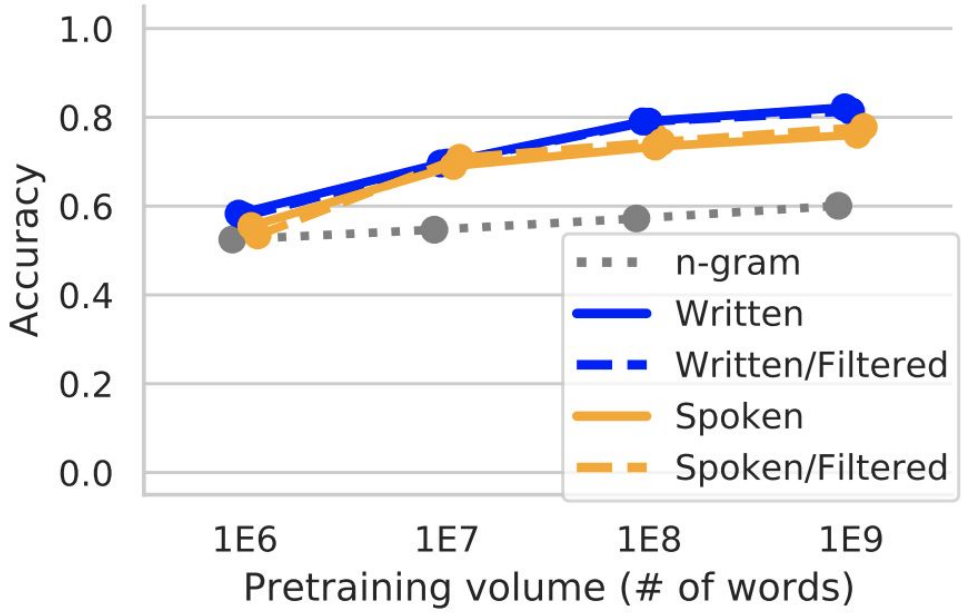
Results: General acceptability judgments on BLiMP

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

Results: General acceptability judgments on BLiMP

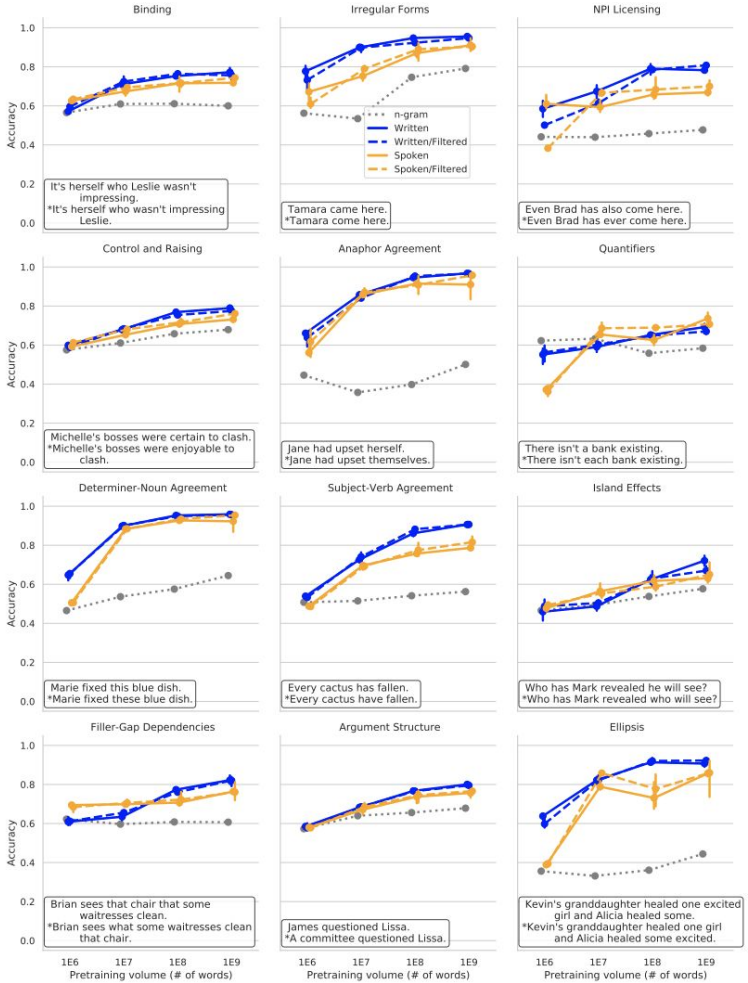
Question: Did the removal of direct evidence have effects on unrelated phenomena?

Answer: No



Results: General acceptability judgments on BLiMP

This result holds across all phenomena in BLiMP.



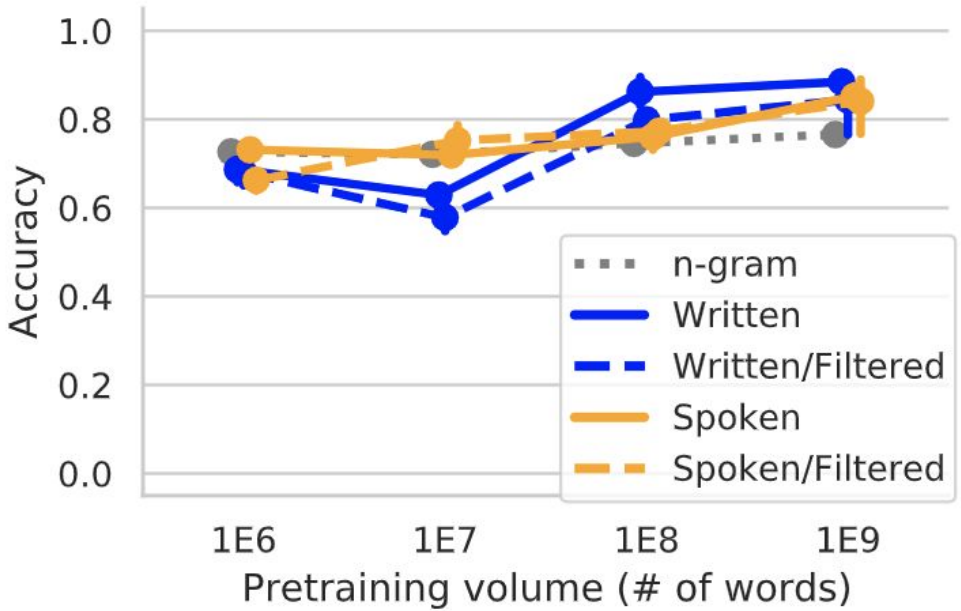
Results: Subject Aux Inversion

Question: Did the removal of direct evidence affect learning of the target phenomenon?

Results: Subject Aux Inversion

Question: Did the removal of direct evidence affect learning of the target phenomenon?

Answer: Yes, in the written domain.



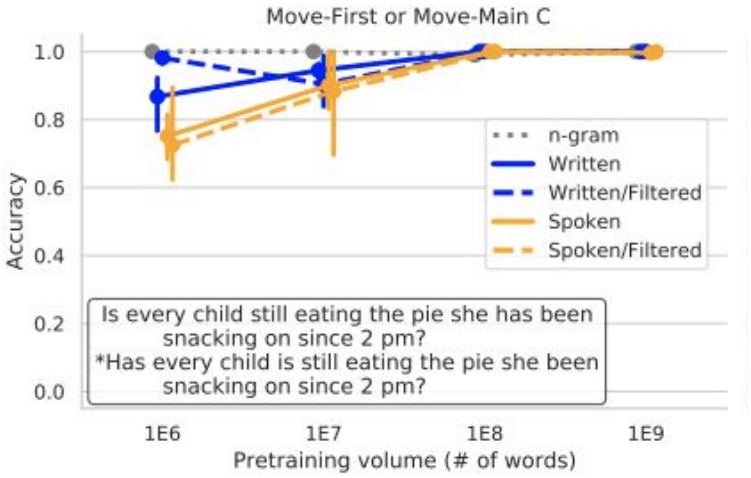
Results: Subject Aux Inversion (fine-grained)

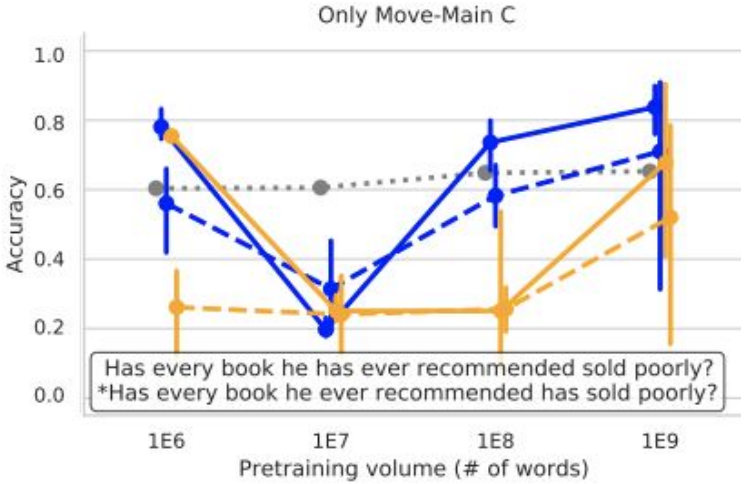
Question: Did the removal of direct evidence differentially affect Only Move-Main examples?

Results: Subject Aux Inversion (fine-grained)

Question: Did the removal of direct evidence differentially affect Only Move-Main examples?

Answer: Yes





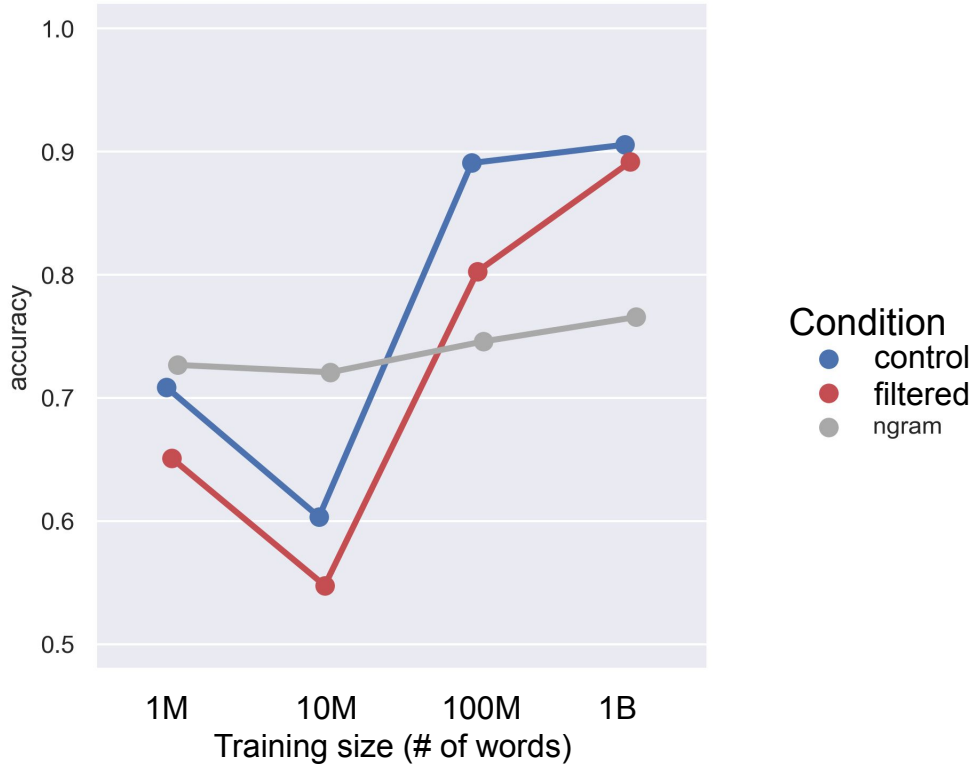
Results: Subject Aux Inversion (BEST CASE)

Question: Is indirect evidence sufficient to acquire the hierarchical rule?

Results: Subject Aux Inversion (BEST CASE)

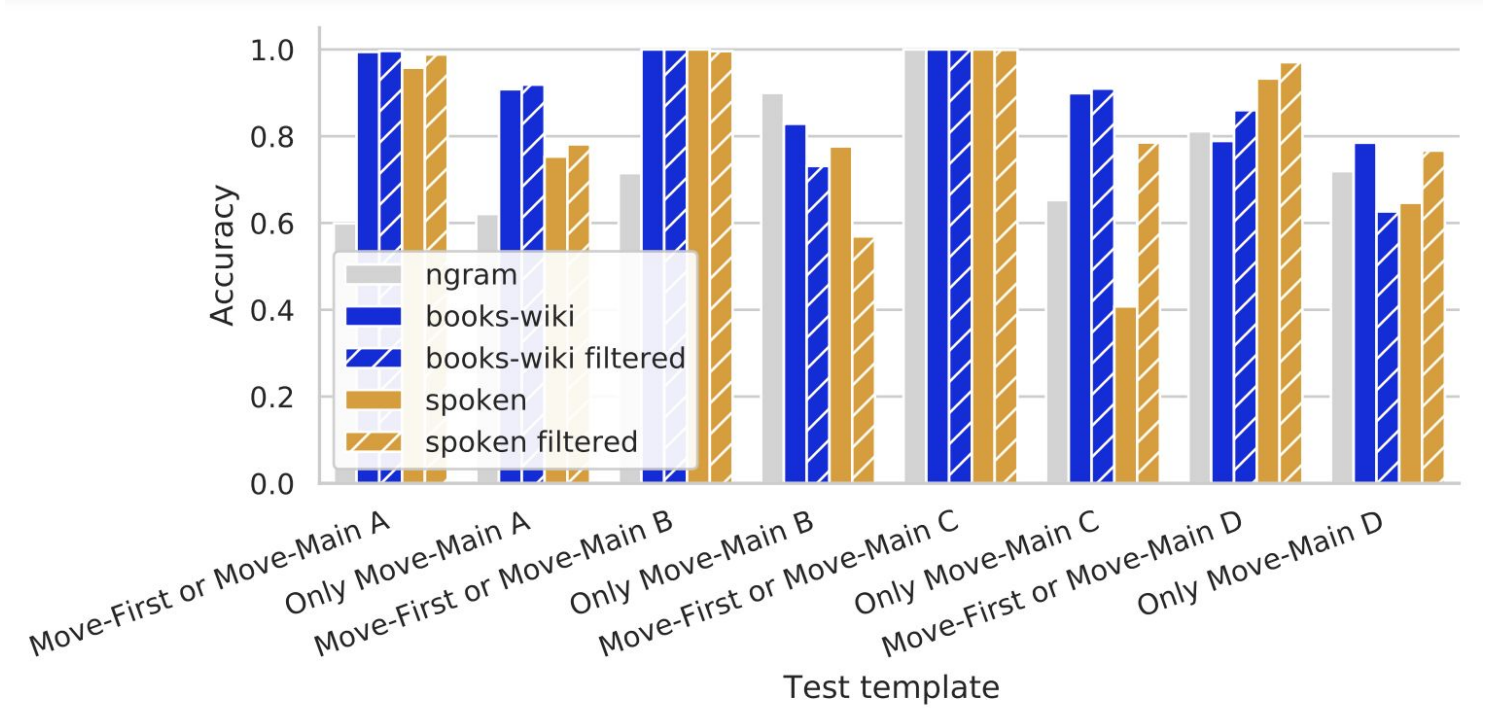
Question: Is indirect evidence sufficient to acquire the hierarchical rule?

Answer: Yes



Results: Subject Aux Inversion (BEST CASE)

This result holds across all test cases in the written domain.



Takeaways

The results support the indirect evidence hypothesis, but with important caveats.

- How reproducible is the best model's success?
- How important are small amounts of direct evidence that passed through the filter?
- Can models succeed with the same data-volume limitations as humans?

Discussion: What does indirect evidence for hierarchical structure look like?

1. Classic constituency tests

Fragment answers
Who has seen the cat? [The man who was here this afternoon]

Coordination
John and [the man who was here this afternoon] are friends.

Pronominalization
[The man who was here this afternoon] left. He saw the cat.

Discussion: What does indirect evidence for hierarchical structure look like?

2. Other hierarchical rules

Subject Verb Agreement
[The man who saw the cats] is here.

Passivization
I greeted [the man who saw the cat.] → [The man who saw the cat] was greeted by me.

Conclusion

Computational model learners allow us to:

- **Make causal inferences about the effects of the environment on language learning.**

- **Conduct controlled experiments on the scale of human language learning.**

BUT we are still far from having developmentally plausible learners.

Why are humans more data efficient than LMs?

BUT we are still far from having developmentally plausible learners.

Why are humans more data efficient than LMs?

Multimodal input

Interactive learning

Open questions

- How do we quantify indirect evidence, and locate relevant indirect evidence in the input?
- Does indirect evidence drive typological tendencies across the world's languages?
- How do we compare human and model developmental trajectories?
- How do we optimize LM training at small scales?
- How do we incorporate interactive learning signals into LM training?

Thank you!

TALEP Marseille for hosting me, especially Mitja Nikolaus, Abdellah Fourtassi, Carlos Ramisch, and Sylvie Ros!

Collaborators: Sam Bowman, Amanpreet Singh, Alicia Parrish, Yian Zhang, Haokun Liu, Haau-Sing Li, Sheng-Fu Wang, Anhad Mohananey, Wei Peng.

Audiences who helped improve previous versions of this talk: NYU Text-as-Data, Stanford Language & Cognition Lab, CoLaLa (UC Berkeley), UT Austin, ETH Zurich, ENS Paris, IST Unbabel Lisbon,

This work was funded in part by the NSF.