Language Models and Human Language Acquisition

Alex Warstadt
ETH Zürich

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)

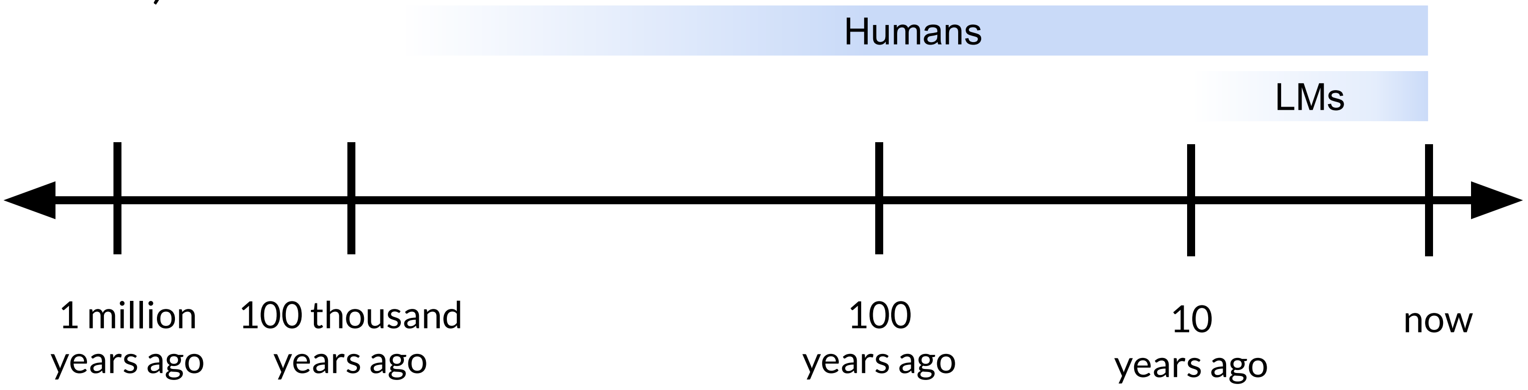


Figure 1: The Transformer - model architecture.

Pharaoh Psamtik
(664 – 610 BCE)

Frederick II
(1194-1250)

James IV
(1473-1513)

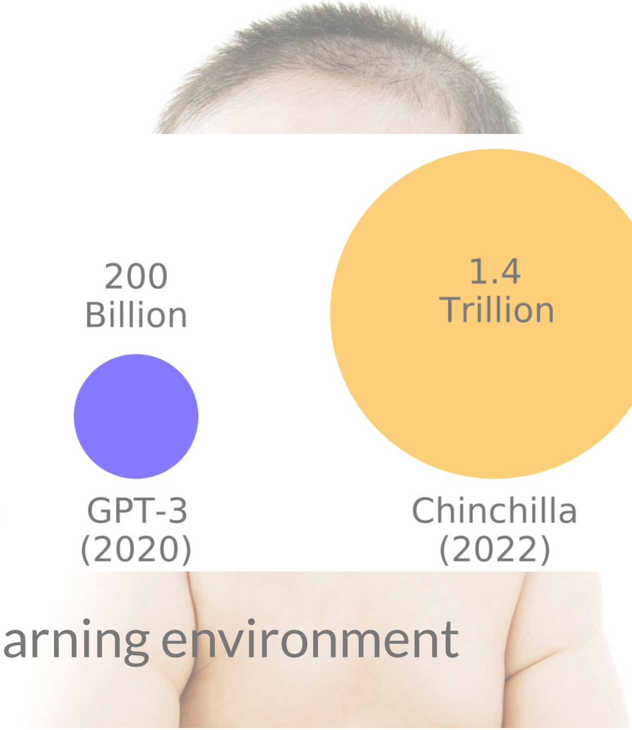
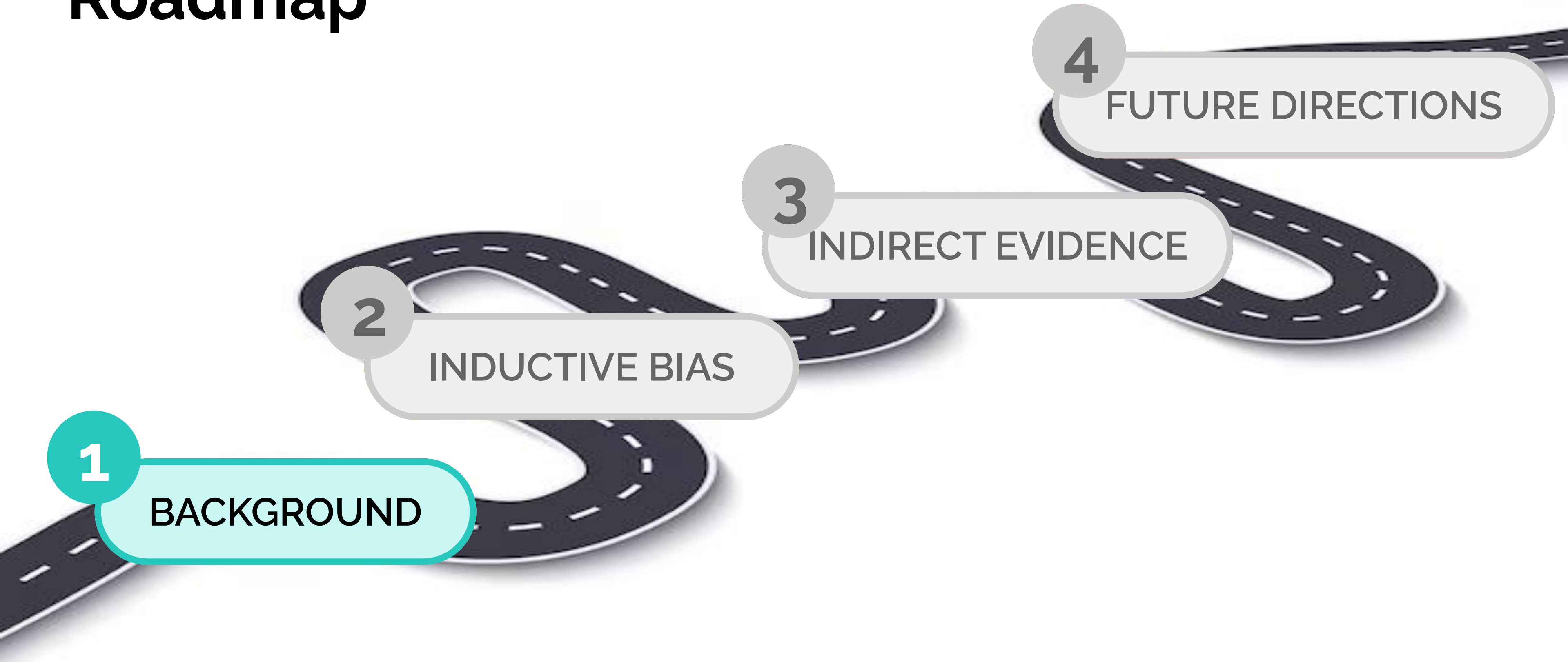


Figure 2: Human baby

Carried out language deprivation experiments

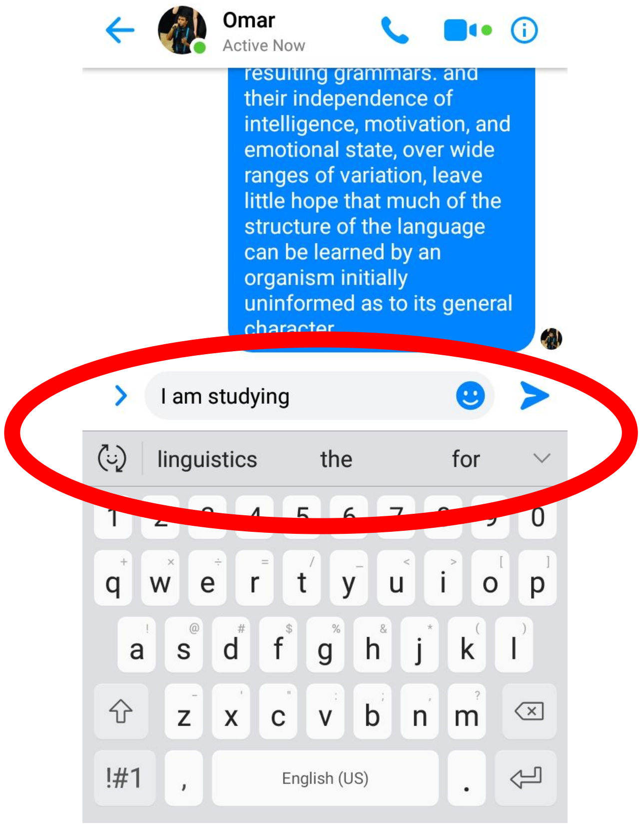
# of words in learning environment

Roadmap

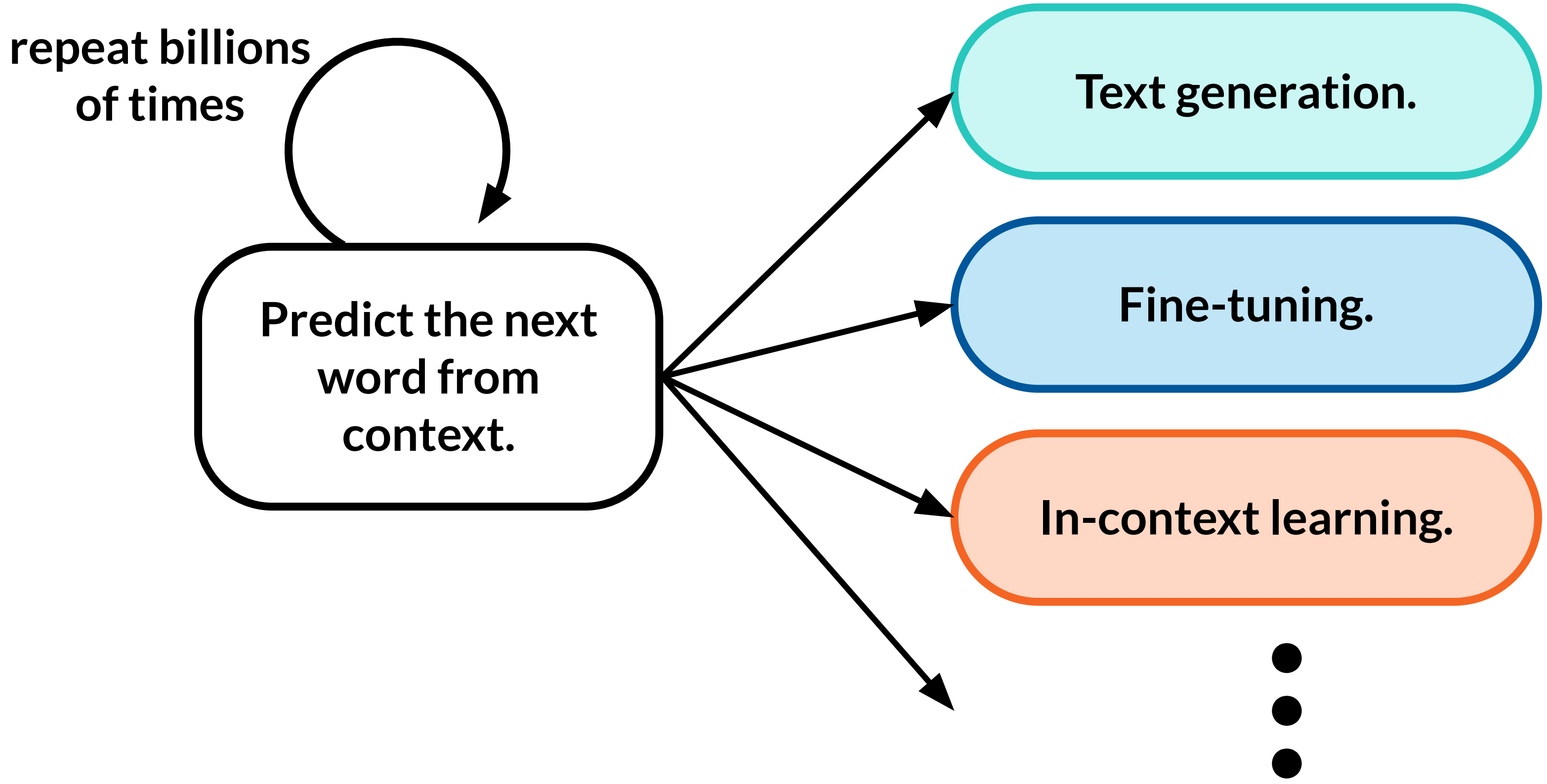

1
BACKGROUND
2
INDUCTIVE BIAS
3
INDIRECT EVIDENCE
4
FUTURE DIRECTIONS

...but first, what is a language model?

$$p(x_1, \ldots, x_T)$$



Language Modeling as Pretraining

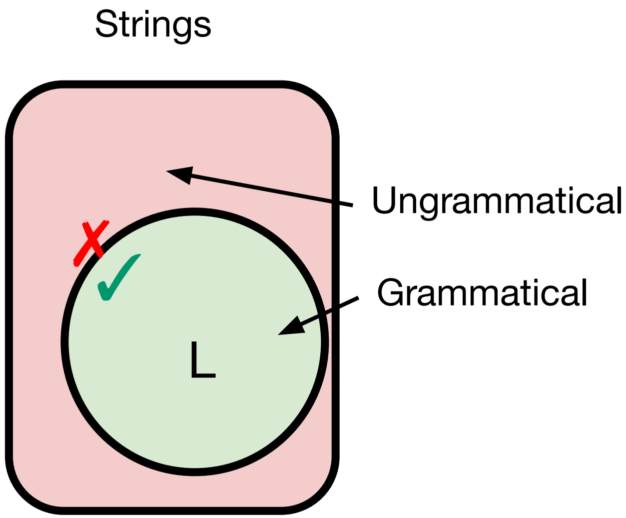


Minimal Pairs

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

✓ Betsy is <u>eager</u> to sleep.

✗ Betsy is <u>easy</u> to sleep.



1. Targeted
2. Reproducible
3. Unsupervised

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

The Benchmark of Linguistic Minimal Pairs (BLiMP)

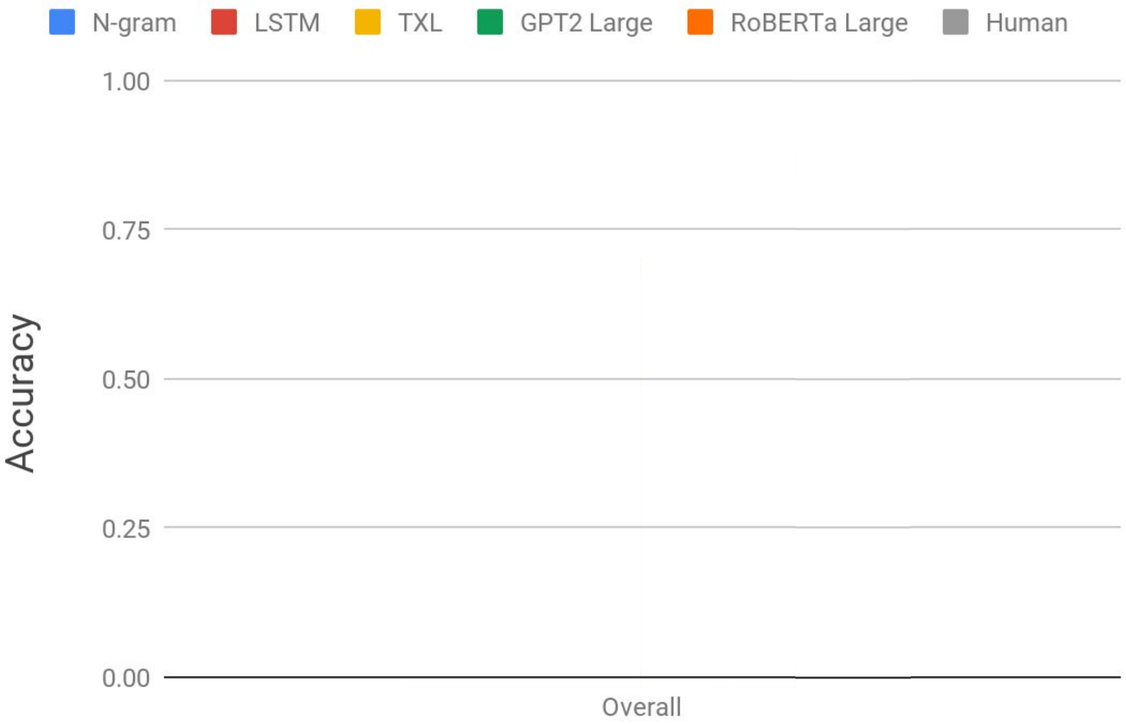
(Warstadt et al., 2020)

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.*

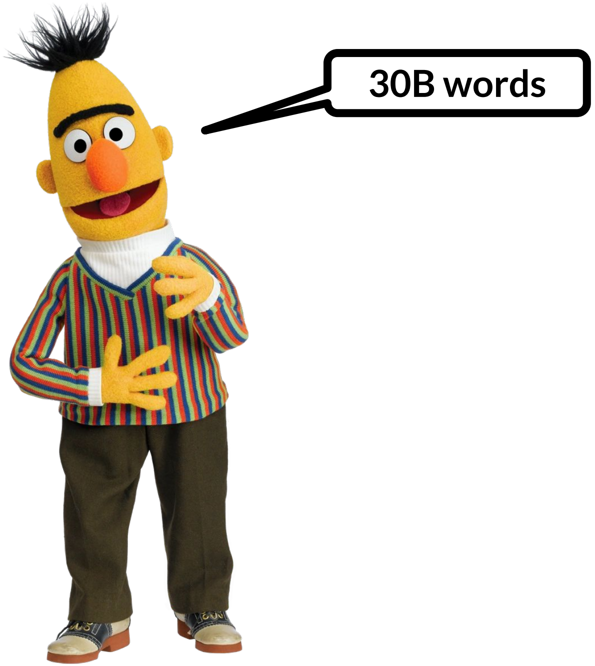
- 67 different minimal pair contrasts
- 1000 sentences each
- 12 broad categories

The Benchmark of Linguistic Minimal Pairs (BLiMP)

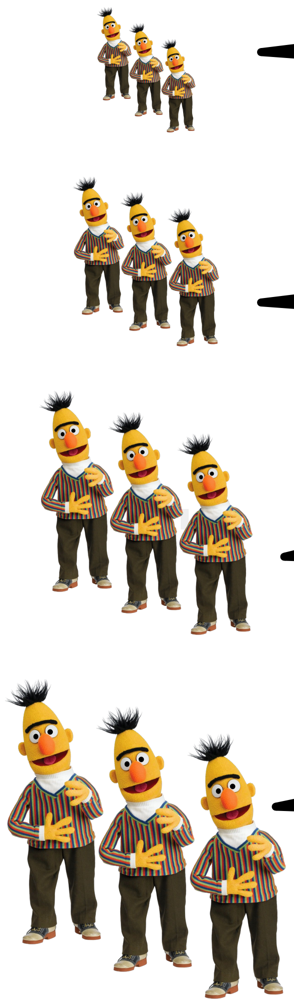
(Warstadt et al., 2020)



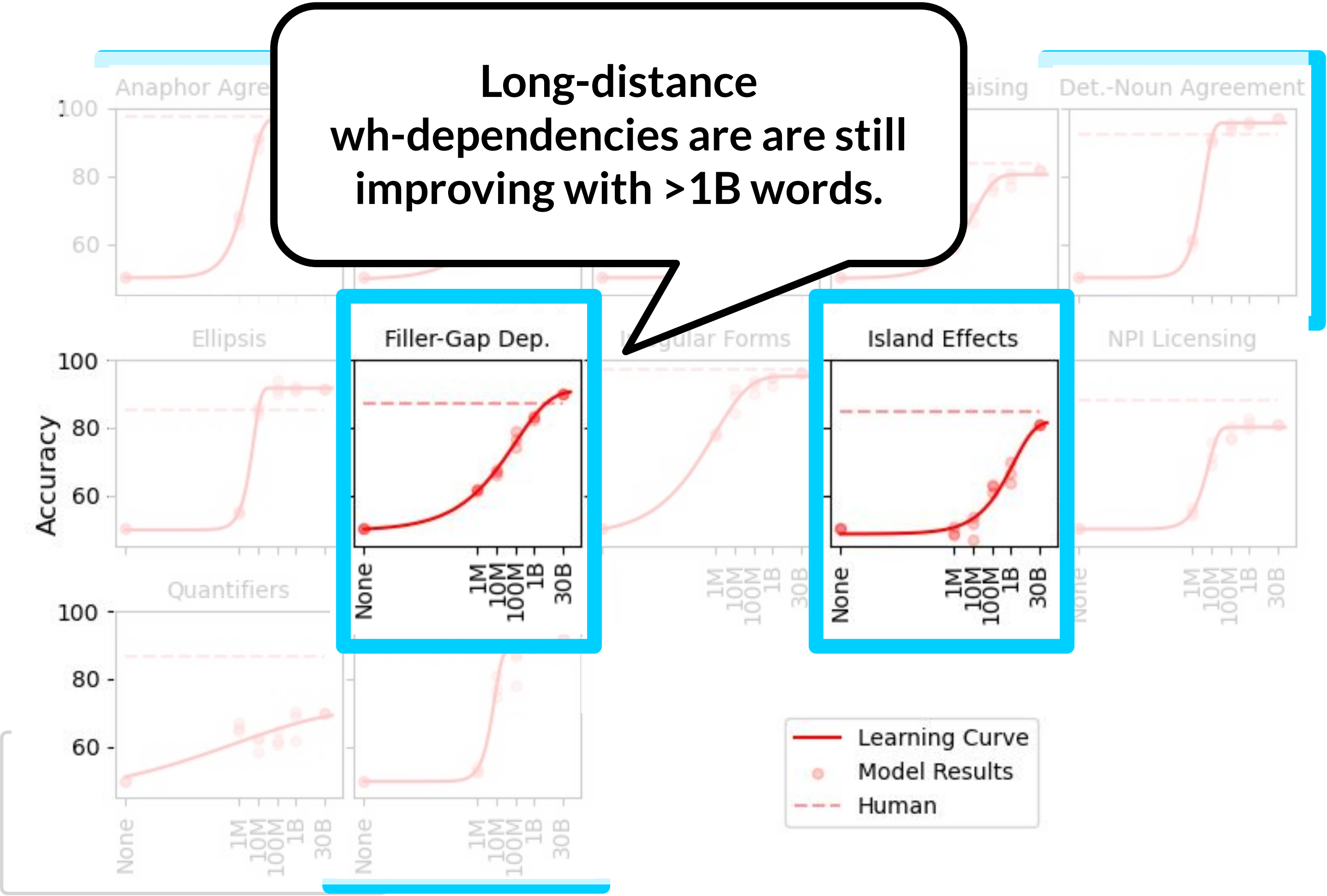
The MiniBERTas



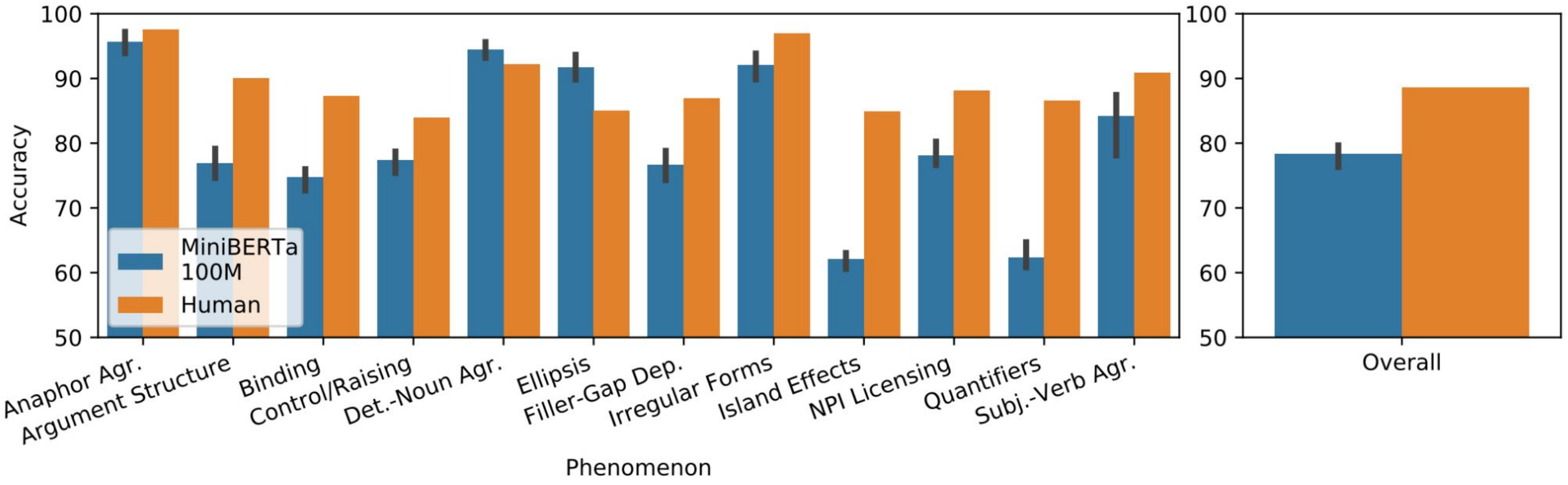
RoBERTa Base



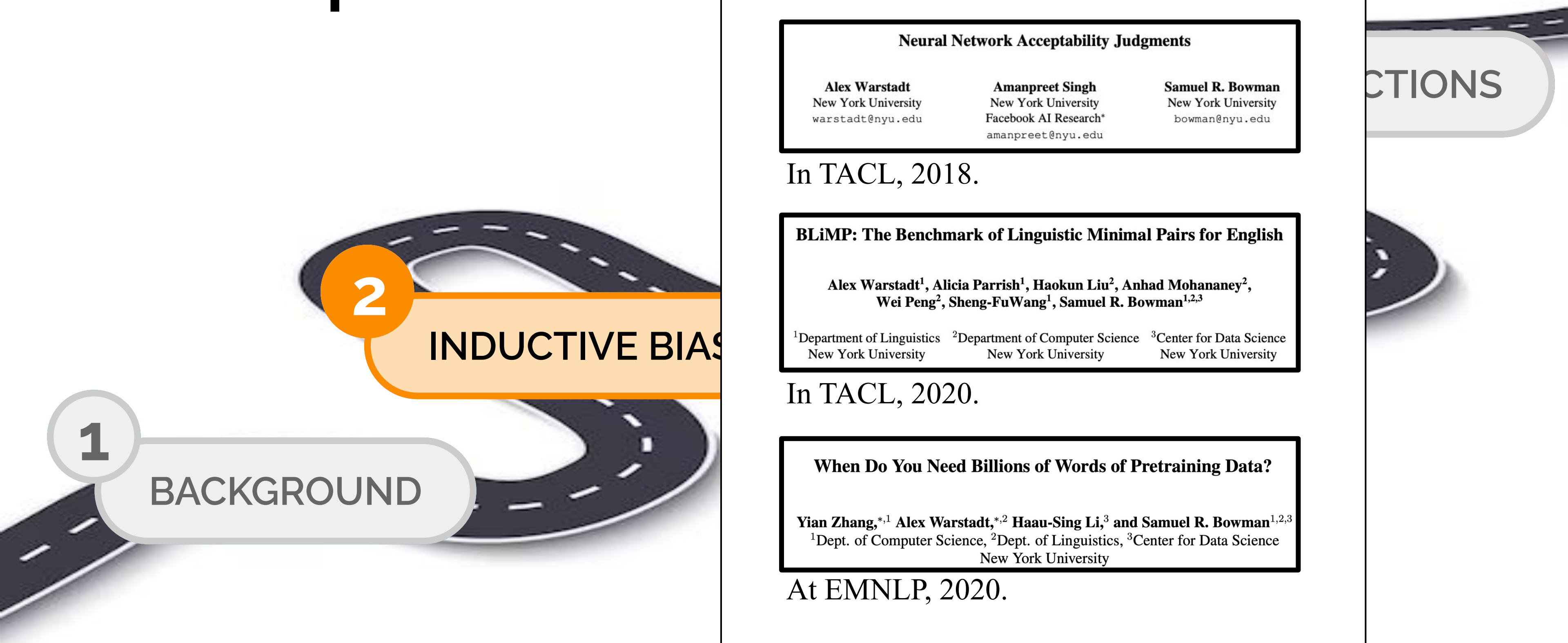
The MiniBERTas on BLiMP



The Data Efficiency Gap



Roadmap



Summary

Neural Network Acceptability Judgments

Alex Warstadt New York University warstadt@nyu.edu
Amanpreet Singh New York University Facebook AI Research* amanpreet@nyu.edu
Samuel R. Bowman New York University bowman@nyu.edu

In TACL, 2018.

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-FuWang[1], Samuel R. Bowman[1,2,3]

[1]Department of Linguistics New York University [2]Department of Computer Science New York University [3]Center for Data Science New York University

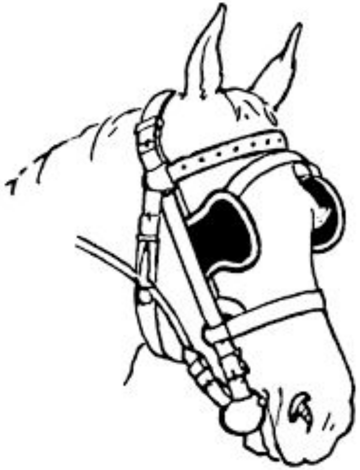
In TACL, 2020.

When Do You Need Billions of Words of Pretraining Data?

Yian Zhang,[*,1] Alex Warstadt,[*,2] Haau-Sing Li,[3] and Samuel R. Bowman[1,2,3]
[1]Dept. of Computer Science, [2]Dept. of Linguistics, [3]Center for Data Science
New York University

At EMNLP, 2020.

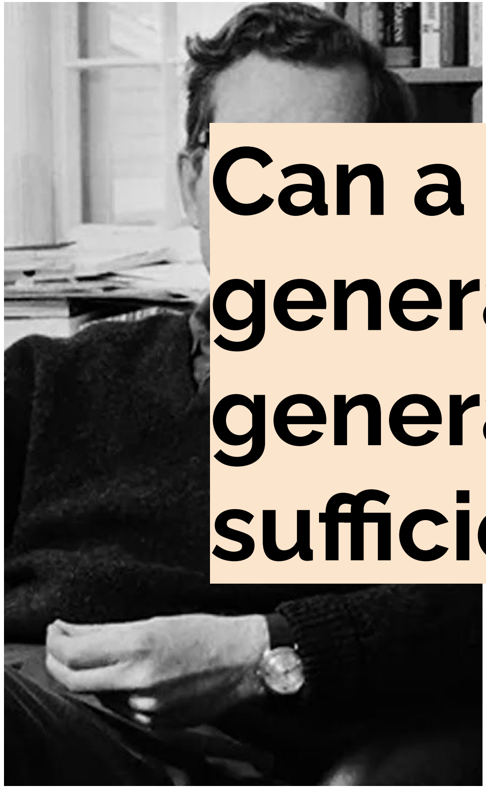
Acquiring Inductive Bias



Inductive biases determine how a learner generalizes given ambiguity in the input.

Language model pretraining is thought to work because it "*induces a hypothesis space H that should be useful for many other NLP tasks*" (Howard & Ruder, 2018)

Linguistic vs. Surface Bias



Can a preference for linguistic generalizations over surface generalizations be acquired with sufficient exposure to language?

des it,

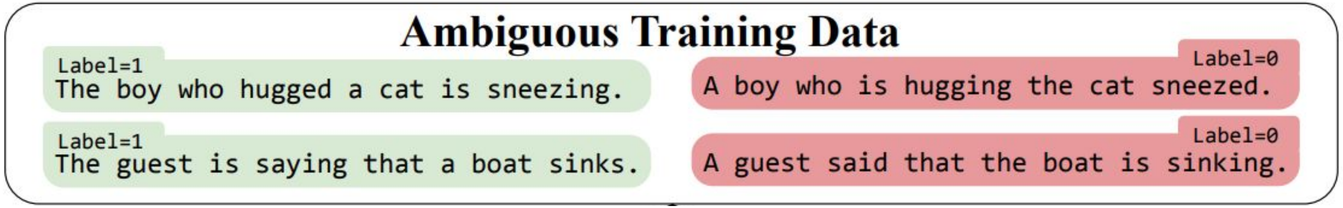
l complexity

on

of a symbol in the middle of a string of even length.

(Chomsky, 1965)

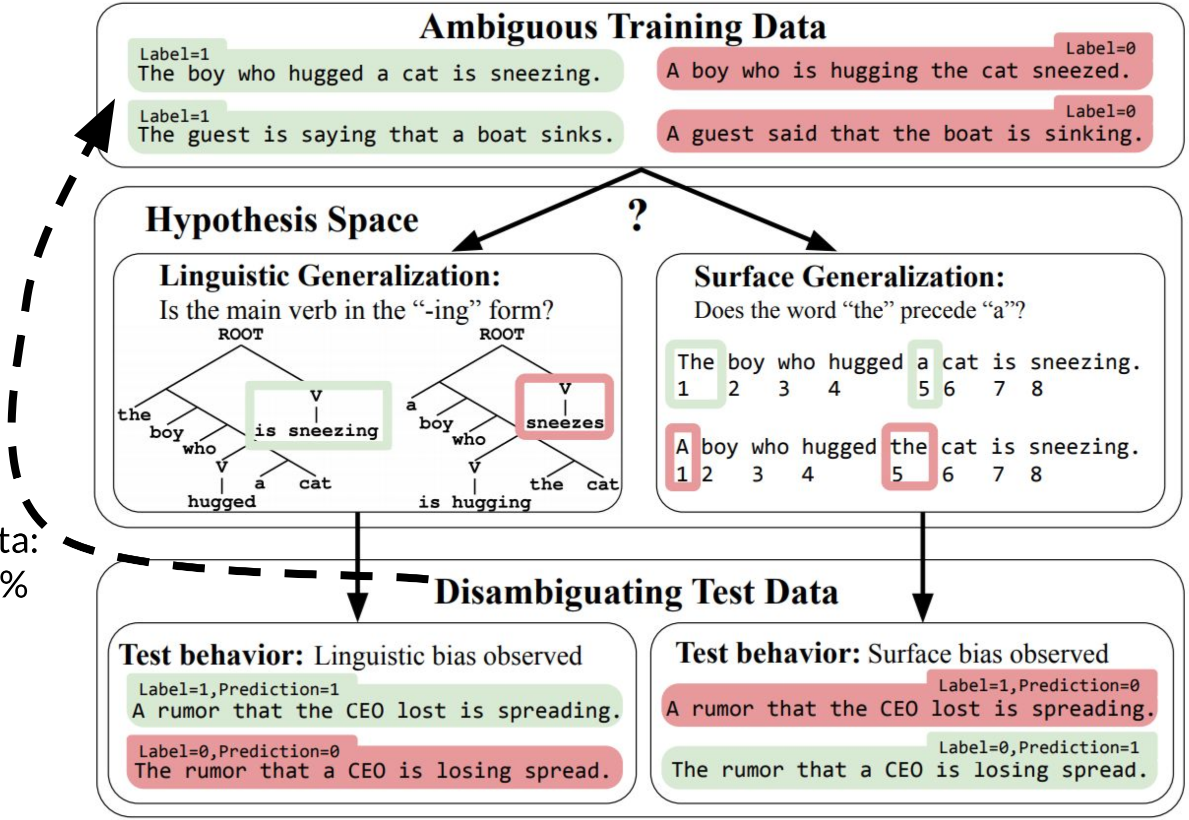
Poverty of the Stimulus Design



Wilson, 2006 (see also McCoy et al. 2018, 2020; Warstadt & Bowman, 2020; Kim & Linzen, 2020; Hupkes et al., 2022; and others)

Poverty of the Stimulus Design +Inoculation

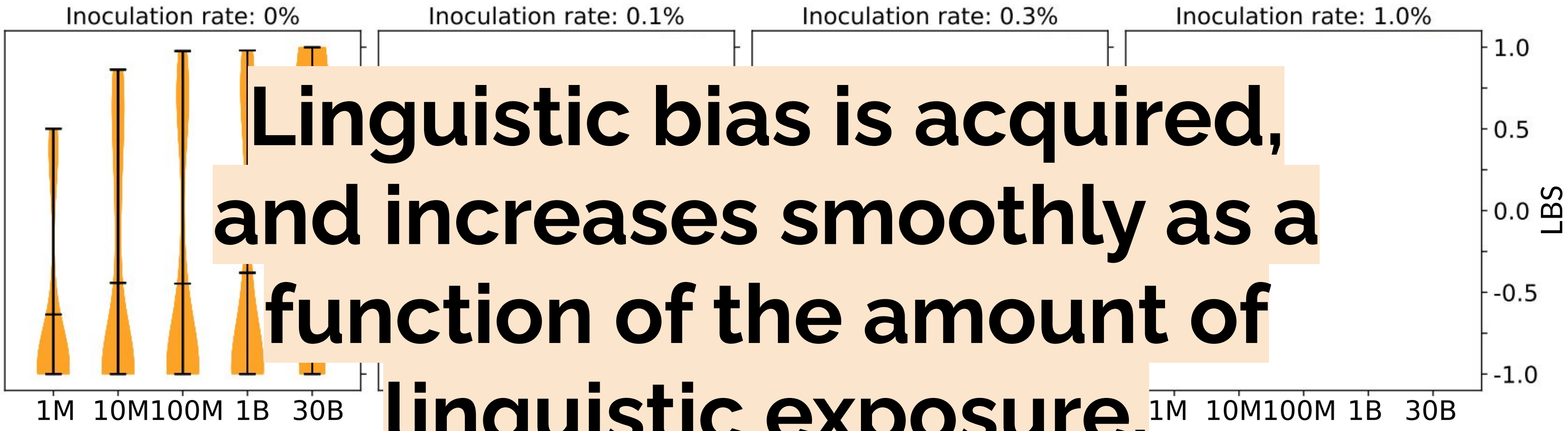
Inoculation data:
0.1% | 0.3% | 1%



Mixed Signals Generalization dataSet (MSGS)

	Feature type	Feature description	Positive example	Negative example
Surface	Absolute position	Is the first token of S "the"?	The cat chased a mouse.	A cat chased a mouse.
	Length	Is S longer than n (e.g., 3) words?	The cat chased a mouse.	The cat meowed.
	Lexical content	Does S contain "the"?	That cat chased the mouse.	That cat chased a mouse.
	Relative position	Does "the" precede "a"?	The cat chased a mouse.	A cat chased the mouse.
	Orthography	Does S appear in title case?	The Cat Chased a Mouse.	The cat chased a mouse.
Linguistic	Morphology	Does S have an irregular past verb?	The cats slept.	The cats meow.
	Syn. category	Does S have an adjective?	Lincoln was tall.	Lincoln was president.
	Syn. construction	Is S the control construction?	Sue is eager to sleep.	Sue is likely to sleep.
	Syn. position	Is the main verb in "ing" form?	Cats who eat mice are purring.	Cats who are eating mice purr.

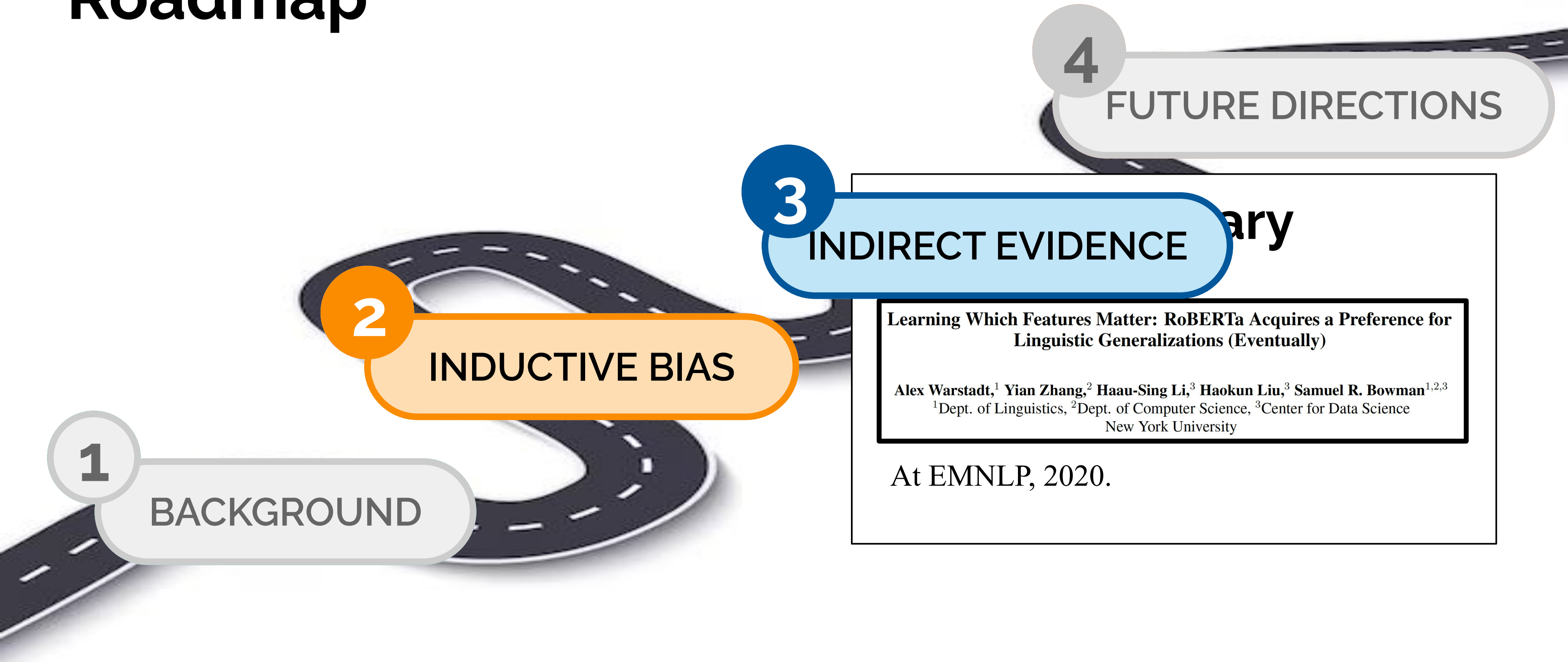
Results on MSGS



Linguistic bias is acquired, and increases smoothly as a function of the amount of linguistic exposure.

Linguistic bias score (LBS) = { 1, if fully linguistic; -1, if fully surface }

Roadmap



Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations (Eventually)

Alex Warstadt,[1] Yian Zhang,[2] Haau-Sing Li,[3] Haokun Liu,[3] Samuel R. Bowman[1,2,3]
[1]Dept. of Linguistics, [2]Dept. of Computer Science, [3]Center for Data Science
New York University

At EMNLP, 2020.

Why should we conduct language acquisition exp

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

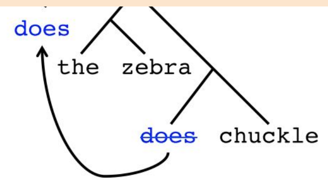
al design

Subject Auxiliary Inversion

The zebra does chuckle → Does the zebra chuckle?

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

(Crain and Nakayama, 1987) ~~does~~ chuckle



Example: McCoy et al. (2020)

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.”

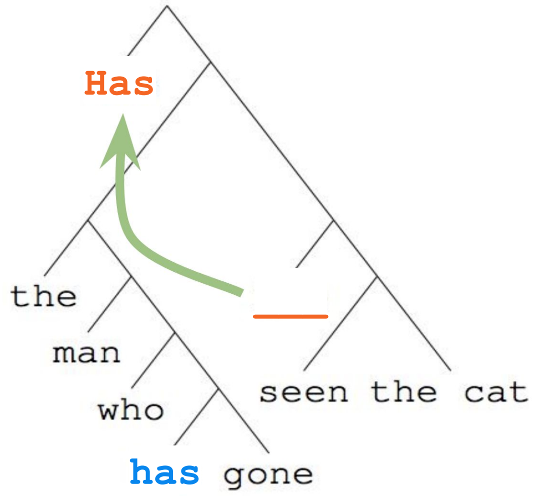
(Legate & Yang, 2001)

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

Surface Generalization:
Move the <u>first</u> auxiliary to the front.

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

Linguistic Generalization:
Move the <u>structurally highest</u> 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)

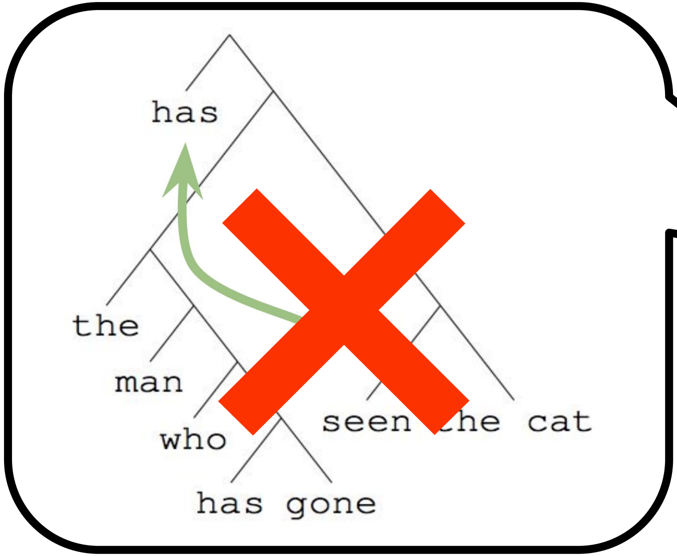
LMs and Subject Auxiliary Inversion

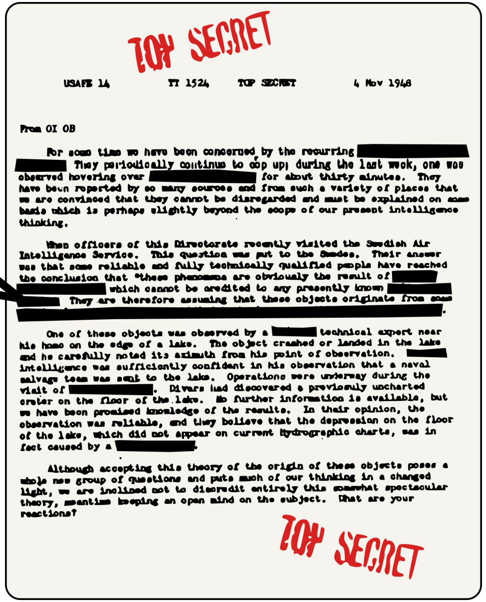
Earlier findings:

- 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?

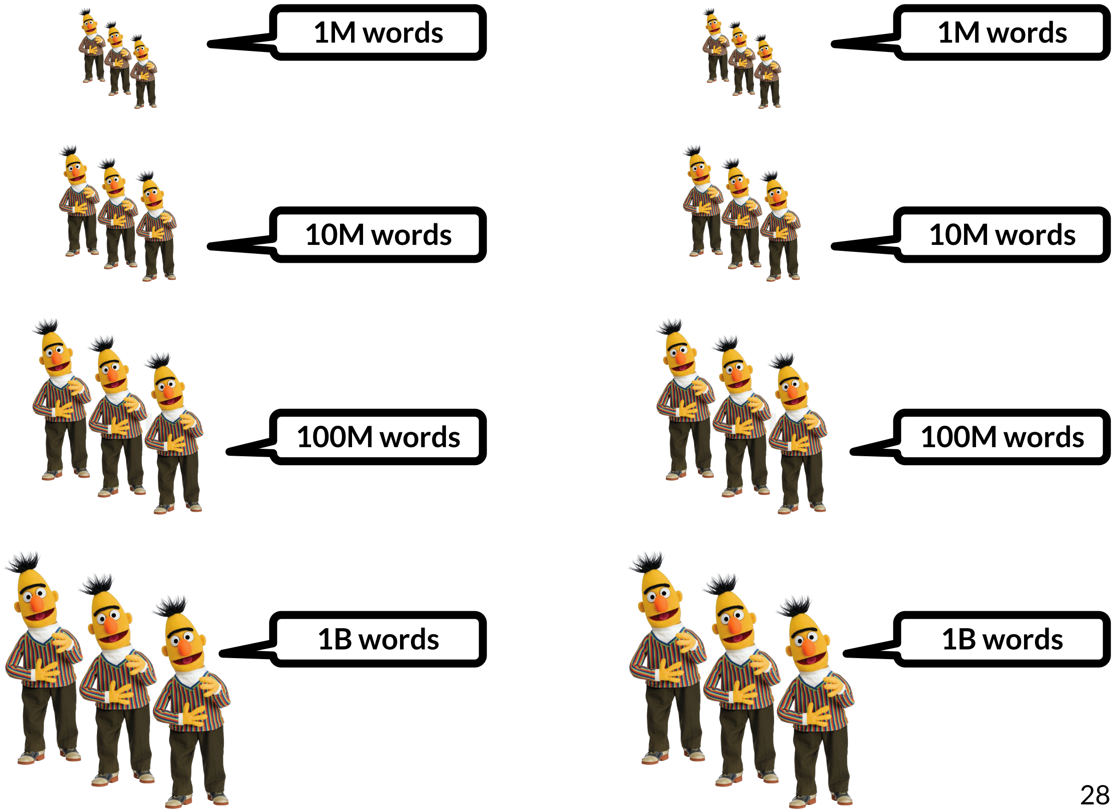
Models

Filtered Condition

Unfiltered Condition (control)

48 RoBERTa models pretrained from scratch

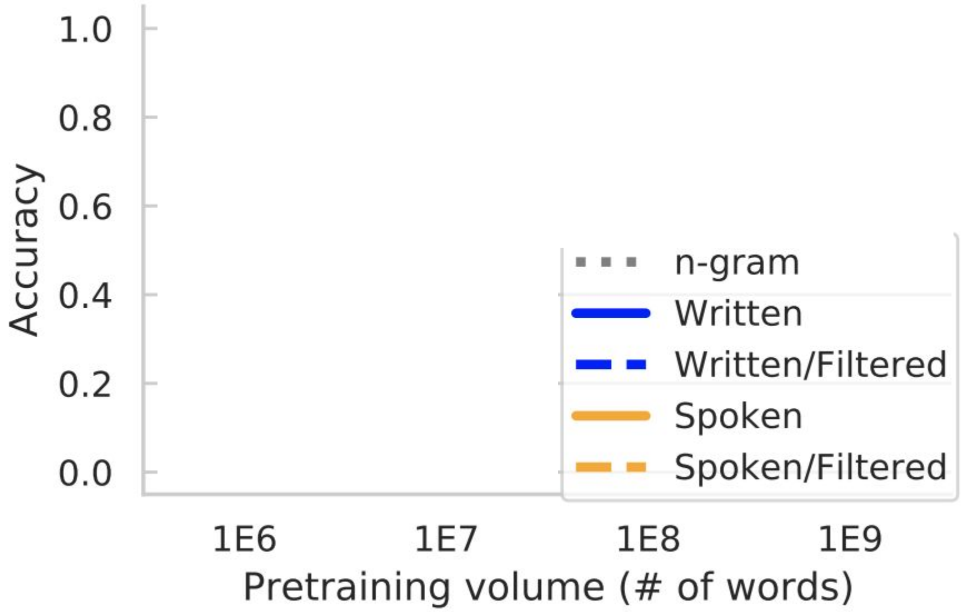
- 2 main conditions
- 4 sizes
- 3 runs (failed runs discarded)
- 2 domains (written, spoken)



Results: General acceptability judgments on BLiMP

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

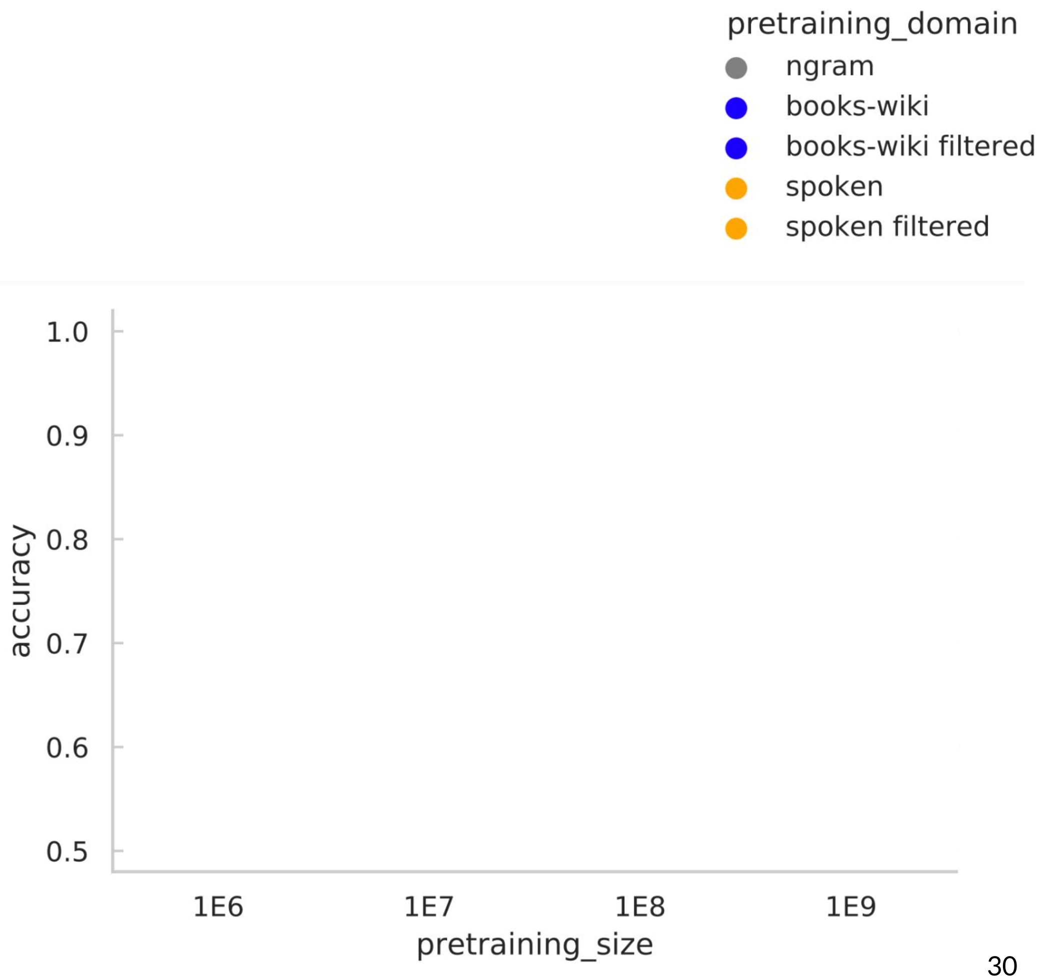
Answer: No



Results: Subject Aux Inversion

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

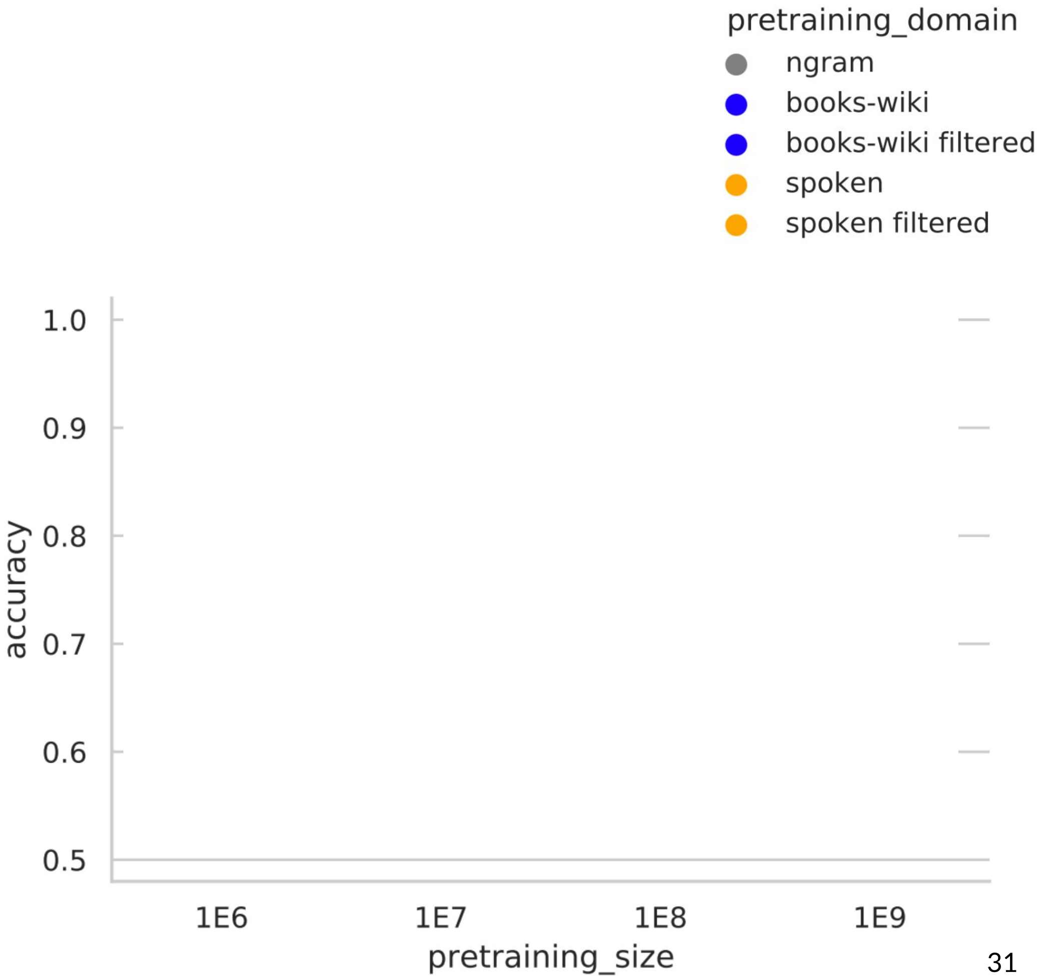
Answer: Slightly, only in the written domain.



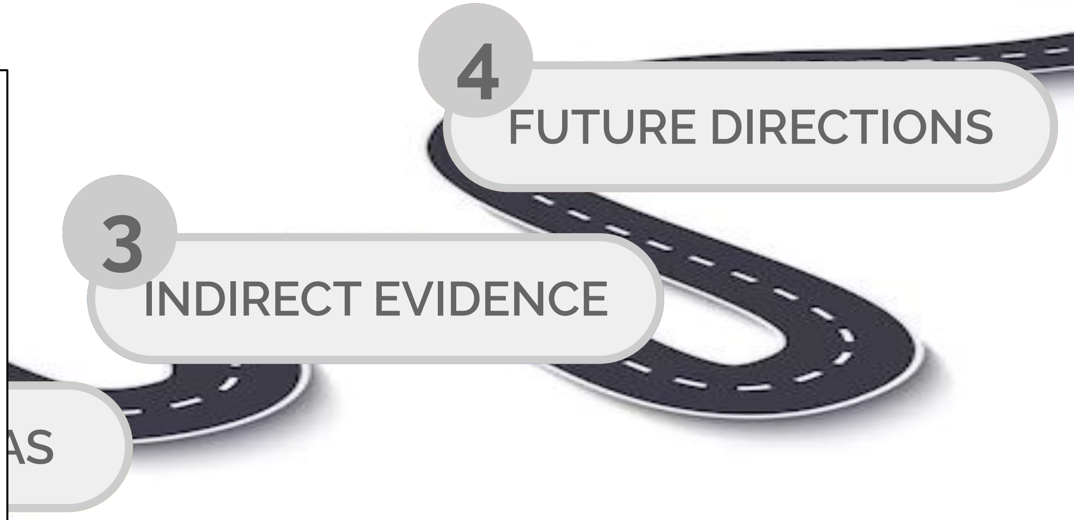
Results: Subject Aux Inversion

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

Answer: Yes, but only in the best case.



Roadmap



Summary

CHAPTER 6

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

Dissertation, NYU, 2022.

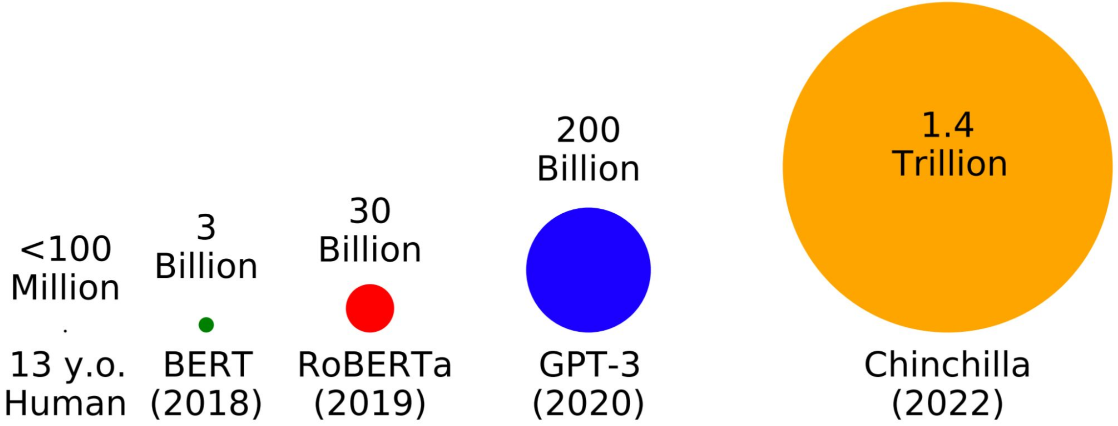
Can neural networks acquire a structural bias from raw linguistic data?

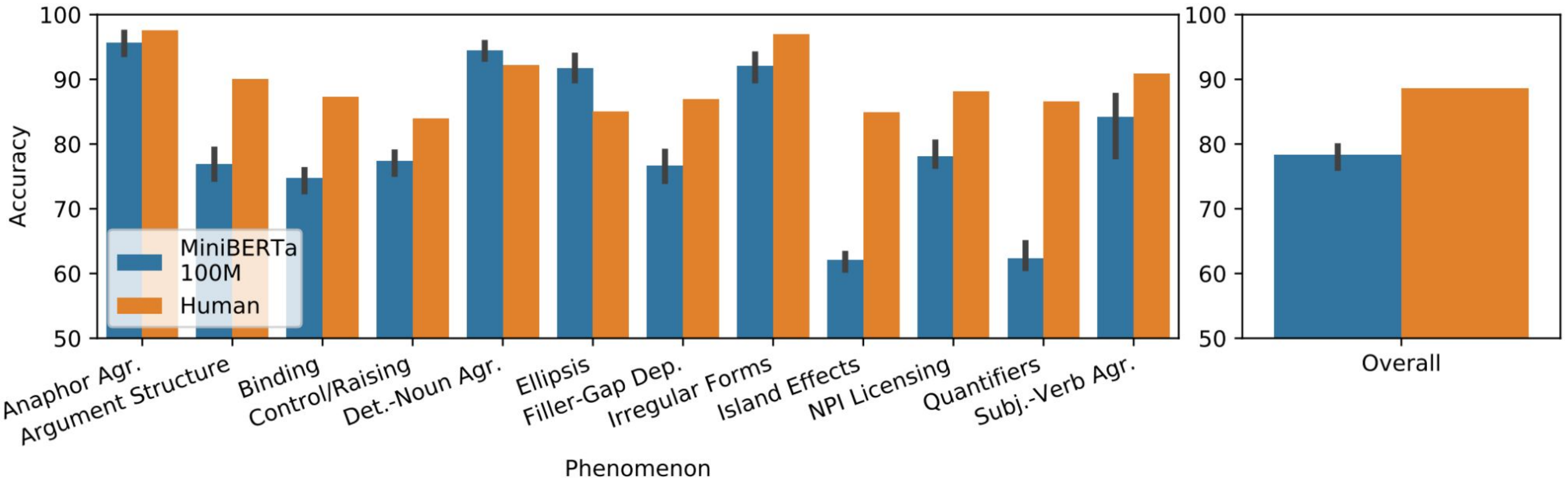
Alex Warstadt (warstadt@nyu.edu)
Department of Linguistics, New York University
New York, NY 10003 USA

Samuel R. Bowman (bowman@nyu.edu)
Department of Linguistics & Center for Data Science & Department of Computer Science, New York University
New York, NY 10003 USA

CogSci, 2020.

Advantages & Disadvantages: The data efficiency gap





UPCOMING Shared task @ CMCL/CoNLL 2023

Objectives: Challenge

plausible corpus

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

- 100 milli…
- Mostly tr… speech
- Test on accepta… and downstream tasks

Track 1: Strict

… and downstream tasks

Track 2: Strict-small

- <100 million words
- Unlimited non-linguistic data
- Unlimited model-generated data

Track 3: Loose

Is a Picture Really Worth a Thousand Words

(with Theodor Amariucai & Ryan Cotterell)

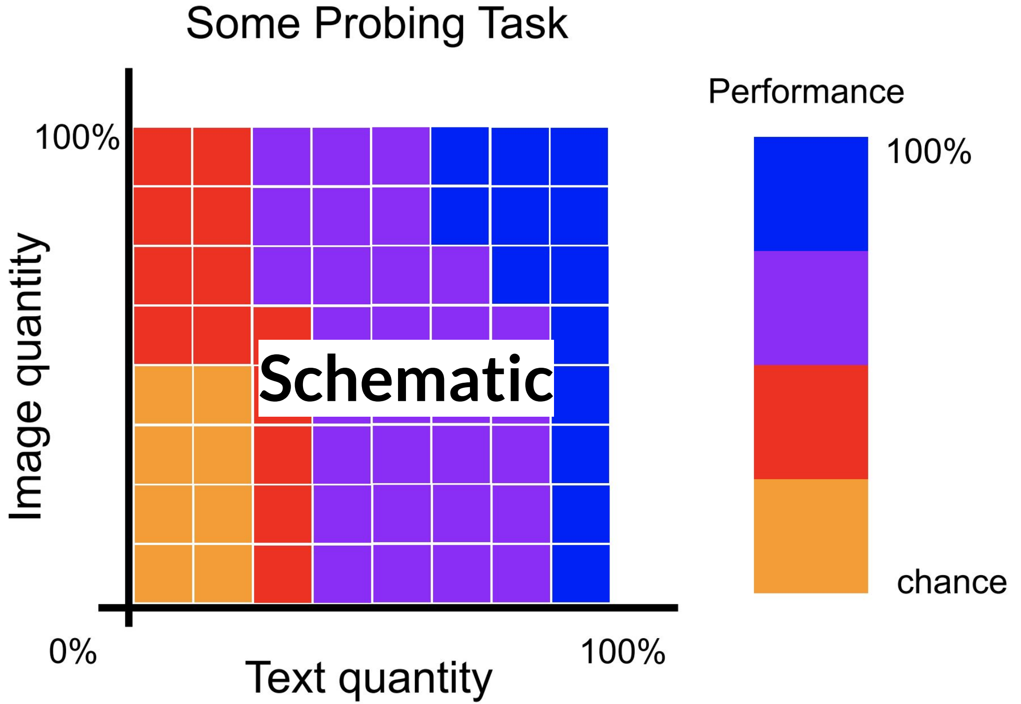
Big Question: How much can we close data-efficiency gap using multimodal input?

Prior work:

- Multimodal vision + text models are becoming ubiquitous.
- Models are typically pretrained LMs, fine-tuned on captions data.
- Models are rarely tested in a language-only setting.

Our approach: Multitask multimodal learning on complex and abstract texts.

Is a Picture Really Worth a Thousand Words
(with Theodor Amariucai & Ryan Cotterell)

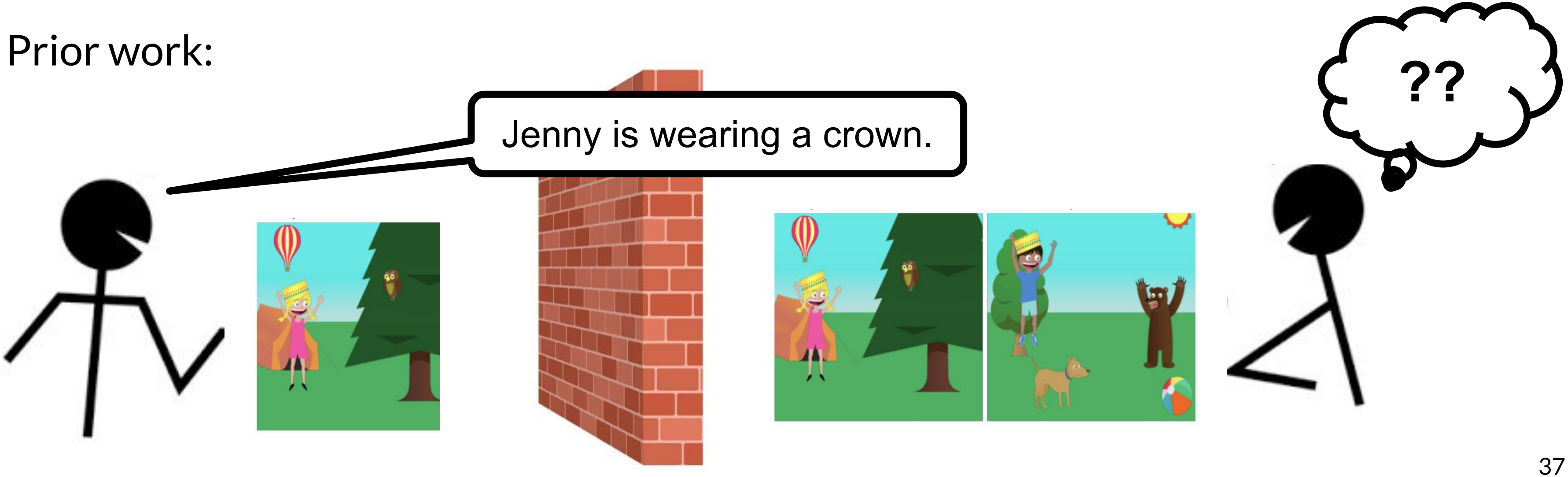


Interactive Language Mode

(With Lennart Stoepler, Mitja Nikolaus, and Ryan Cotterell)

Big Question: How much can we close data-efficiency
using inter-agent interaction

Prior work:

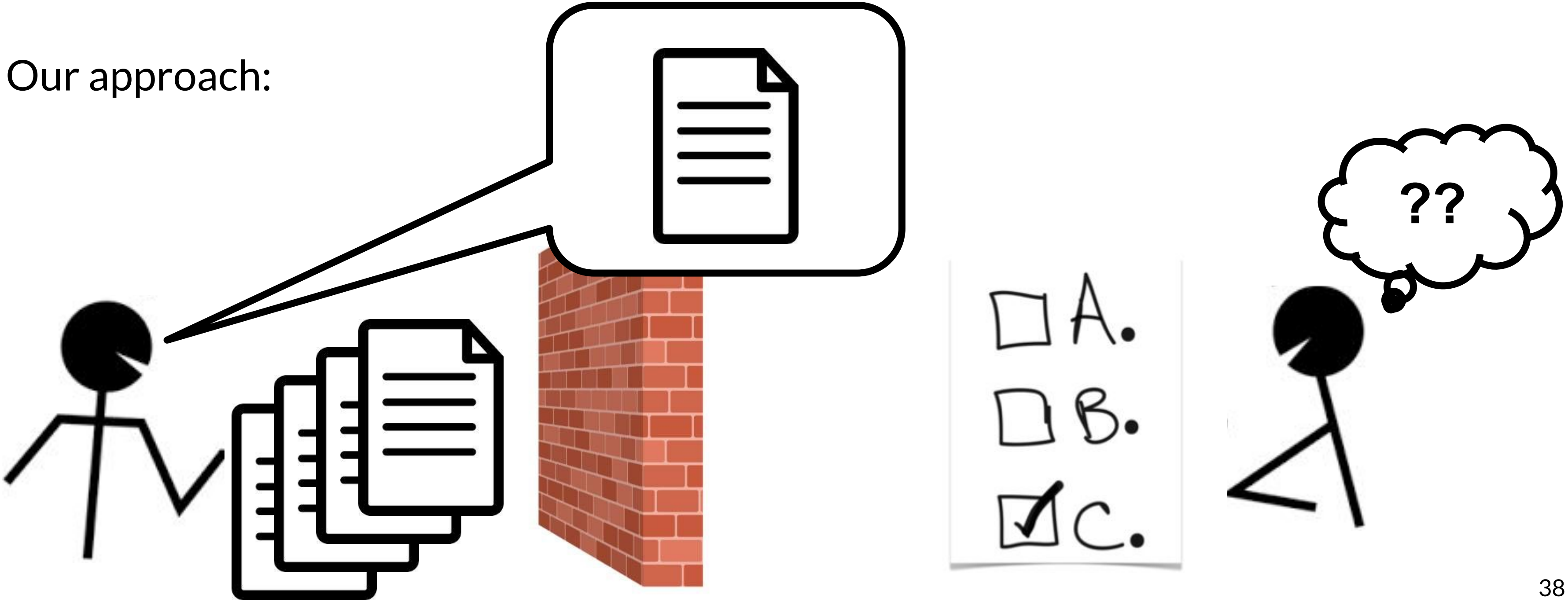


Lazaridou et al., 2020; Nikolaus & Fourtassi, 2021

Interactive Language Mode

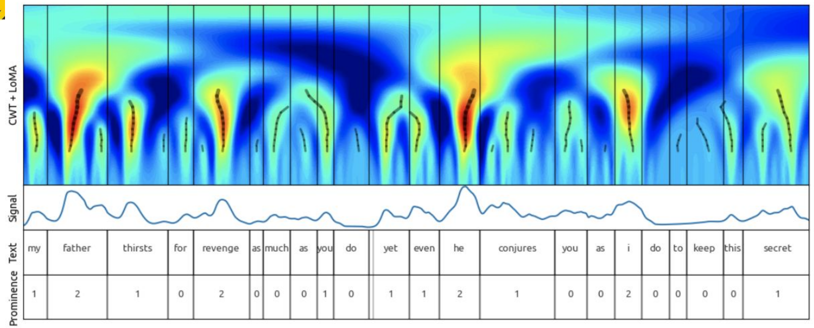
(With Lennart Stoepler, Mitja Nikolaus, and Ryan Cotterell)

Our approach:



Prosody and LMs

(With Lukas Wolf, Tamar Regev, Eghbal Hosseini
Ethan Wilcox, & Ev Fedorenko)



Question 1: How much information does prosody encode that isn't in the text?

An utterance can be decomposed into two variables:

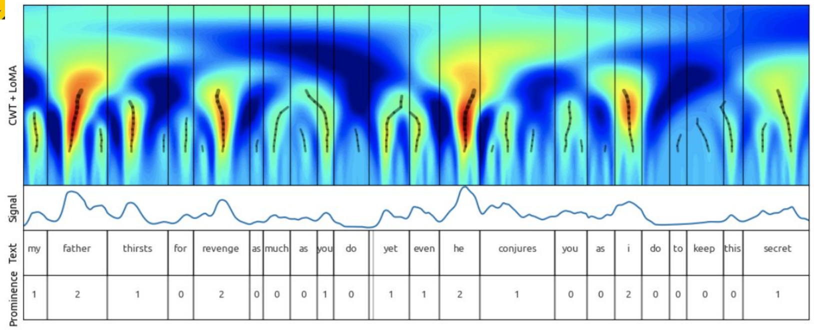
- T = the text (i.e., a string of words)
- P = the prosody (i.e., pitch + loudness + duration)

What is MI(T; P)?

Method: Train the most powerful possible probe to predict prosodic features from text (Pimentel et al., 2020)

Prosody and LMs

(With Lukas Wolf, Tamar Regev, Eghbal Hosseini Ethan Wilcox, & Ev Fedorenko)



Question 2: How much can we close the data-efficiency gap by adding prosodic information to LM training data.

Methods:

1. Extract text & prosody from audio corpus.
2. Predict prosody from our probe for a text-only corpus, and give those representations to the LM during training.

Roadmap

4 FUTURE DIRECTIONS

ECT EVIDENCE

1 BACKG

Summary

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

Alex Warstadt, Samuel R. Bowman

In *Algebraic Structures in Natural Language*, 2022.

Call for Papers - The BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus

https://babylm.github.io/

Alex Warstadt ETH Zürich

Leshem Choshen IBM Research

Aaron Mueller Johns Hopkins University

Ethan Wilcox ETH Zürich

Adina Williams Meta AI

Chengxu Zhuang MIT

At CoNLL and CMCL, *forthcoming in 2023*.

Conclusions

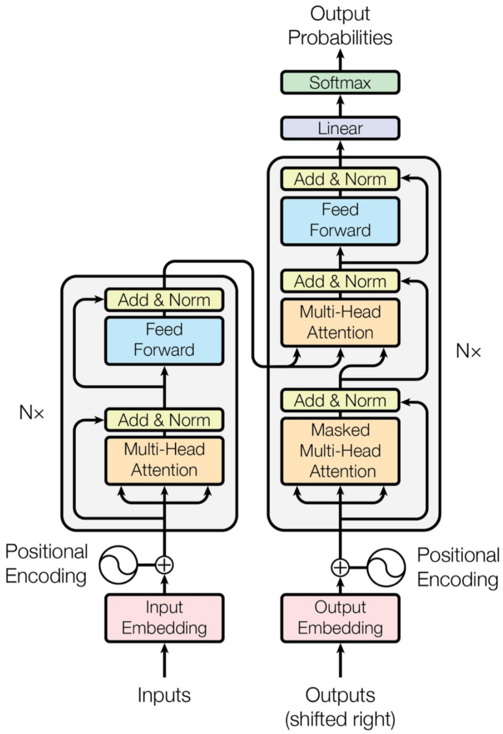


Figure 1: The Transformer - model architecture.

Figure 2: Human baby

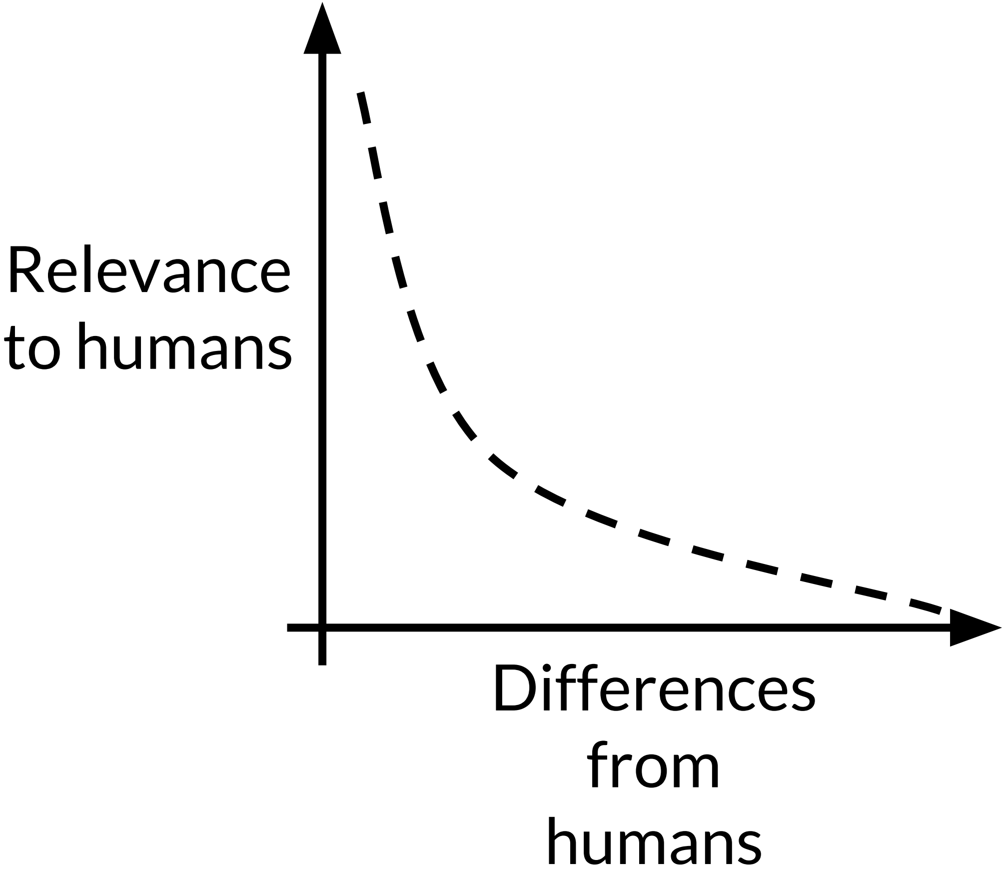
Thank you!

Collaborators: Sam Bowman, Amanpreet Singh, Alicia Parrish, Yian Zhang, Haokun Liu, Haau-Sing Li, Sheng-Fu Wang, Anhad Mohananey, Wei Peng, Theodor Amariucai, Lennart Stoepler, Ryan Cotterell

Bonus Slides

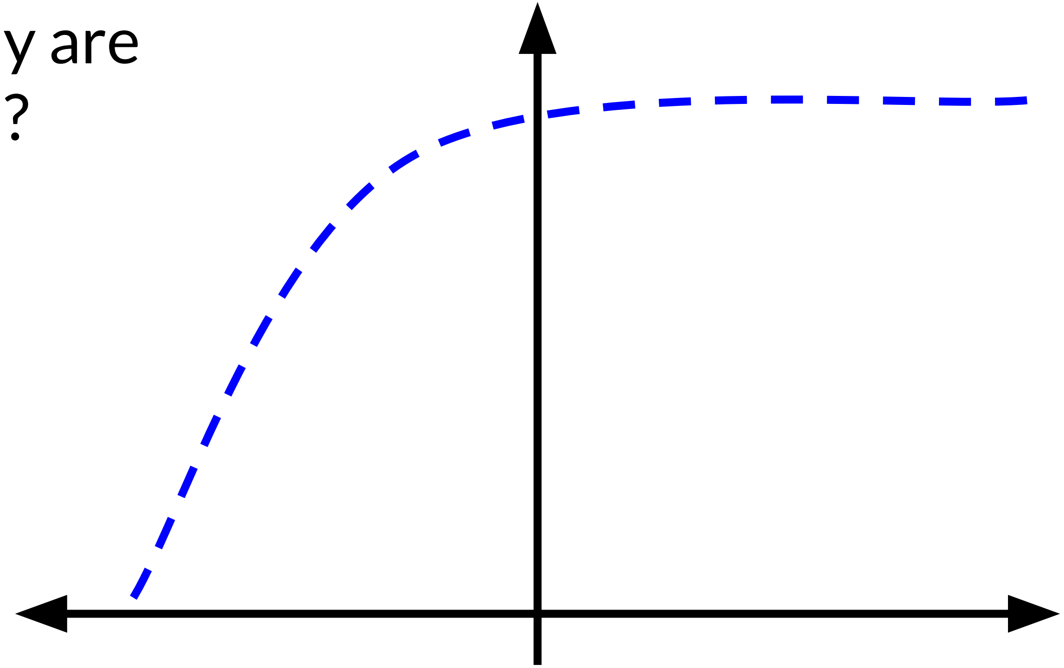
The Recipe for Model Learners

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


Relevance to humans
Differences from humans

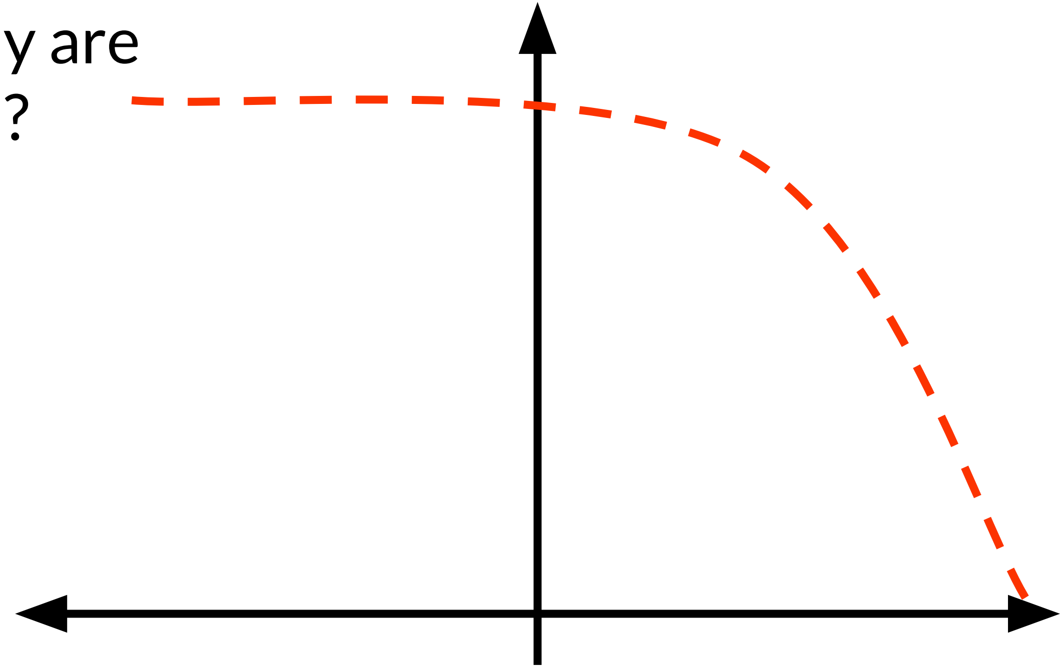
Debates in language acquisition often center around the necessary and sufficient conditions for human-learnability.

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?

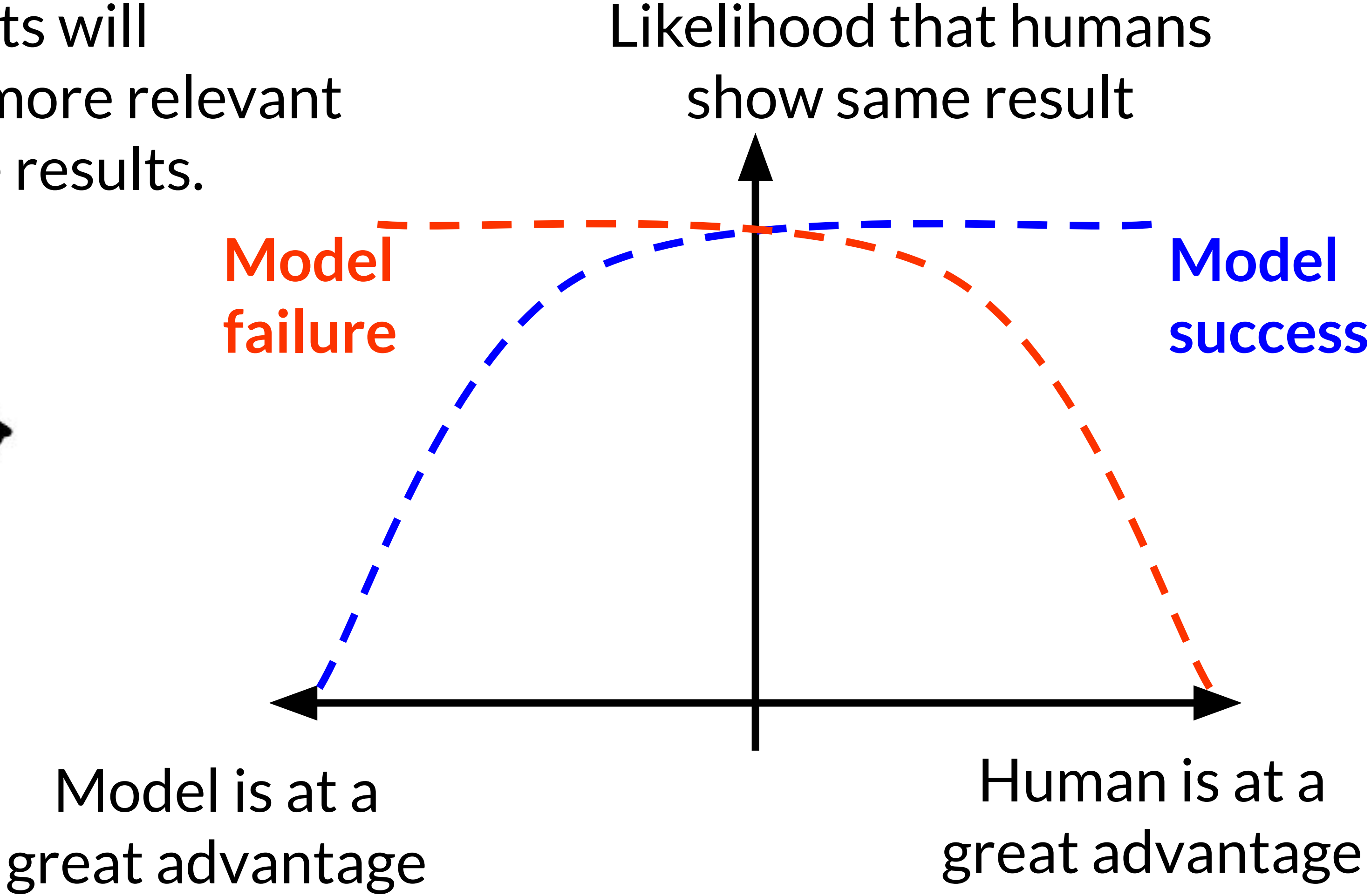
Likelihood that humans show same result



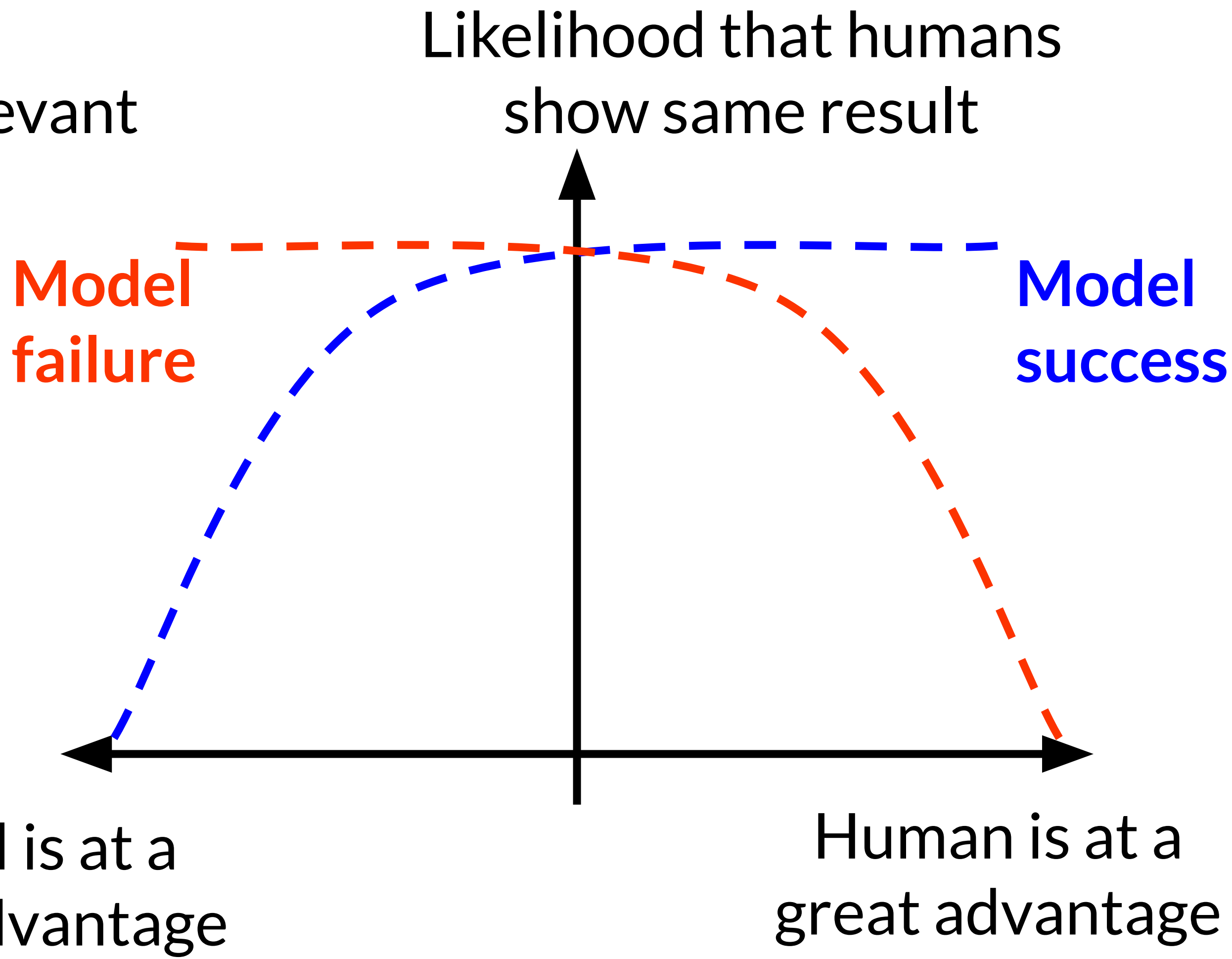
Model is at a great advantage

Human is at a great advantage

Positive results will generally be more relevant than negative results.



Positive results will generally be more relevant than negative results.



A recipe for relevant model learners:

- Maximize relevance of positive results by minimize advantages that models have over humans.
- Maximize chances of positive results by minimizing advantages that humans have over models.

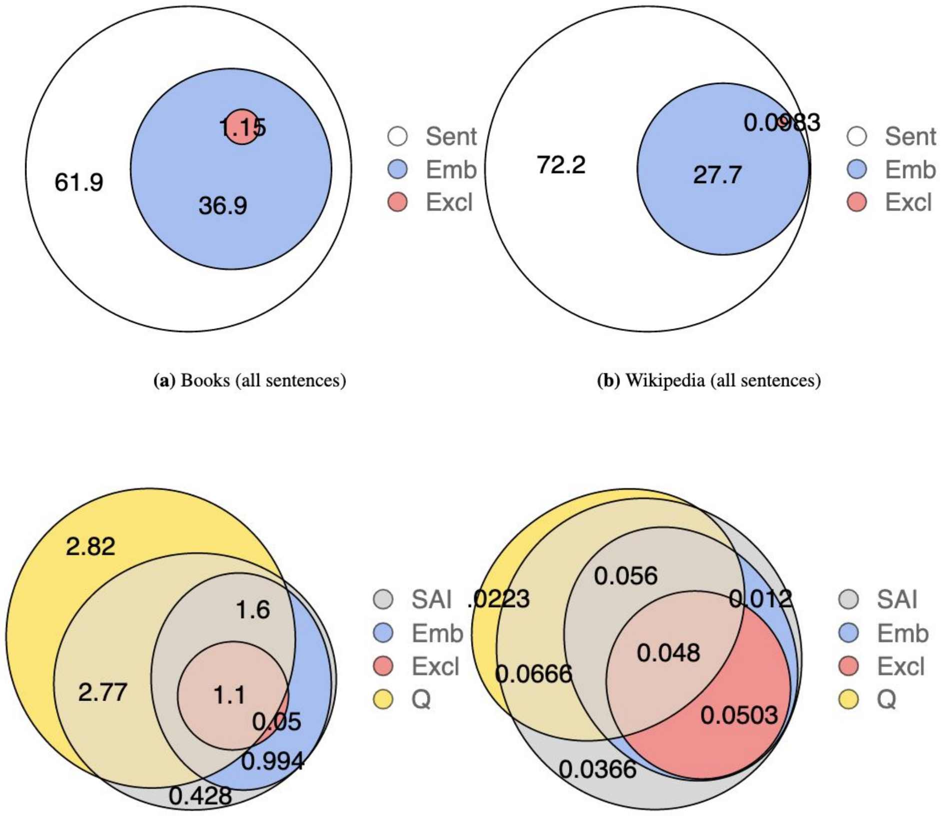
Environmental vs. Innate advantages

- It’s relatively obvious how to apply this recipe to environmental advantages.
- But how do we apply this recipe to innate properties of the learner?

Typical ANNs appear to have weak language-specific advantages. But measuring and manipulating inductive bias is a serious problem where we don’t have great solutions.

Indirect evidence

Distribution of direct evidence (by domain)



(a) Books (all sentences)

(b) Wikipedia (all sentences)

(c) Books (SAI ∪ Q)

(d) Wikipedia (SAI ∪ Q)

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

1. Classic constituency tests

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

Coordination
<u>John</u> and [the man who was here this afternoon] are friends.

Pronominalization
[The man who was here this afternoon] left. <u>He</u> 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] <u>is</u> here.

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

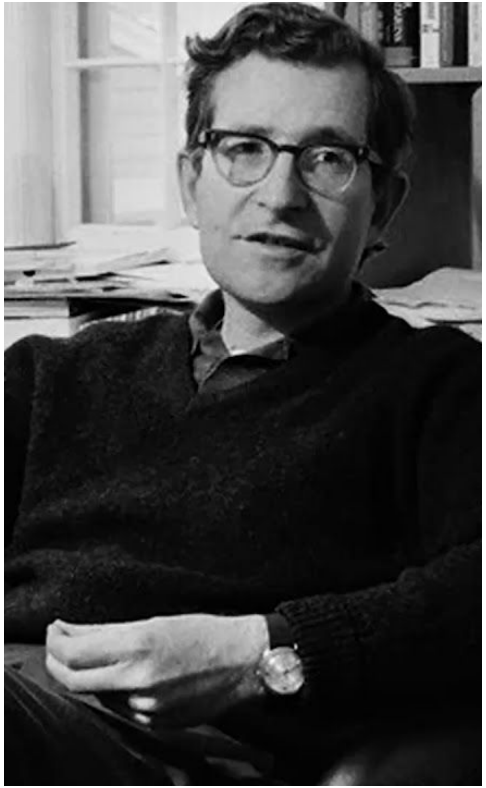
Intro stuff

The Mystery of Human Language Acquisition

Thousands of linguists have spent decades trying to describe the grammar of human language (and only partly succeeding).

How does a single child acquire the grammar of their native language in a matter of years?

Richness of Grammar vs. Poverty of Stimulus



[L]anguage acquisition is based on the child's discovery of what from a formal point of view is a deep and abstract theory a generative grammar of his language — many of the concepts and principles of which are only remotely related to experience by long and intricate chains of unconscious quasi-inferential steps.

A consideration of the character of the grammar that is acquired, the degenerate quality and narrowly limited extent of the available data, the striking uniformity of the resulting grammars, and their independence of intelligence, motivation, and emotional state, over wide ranges of variation, leave little hope that much of the structure of the language can be learned by an organism initially uninformed as to its general character.

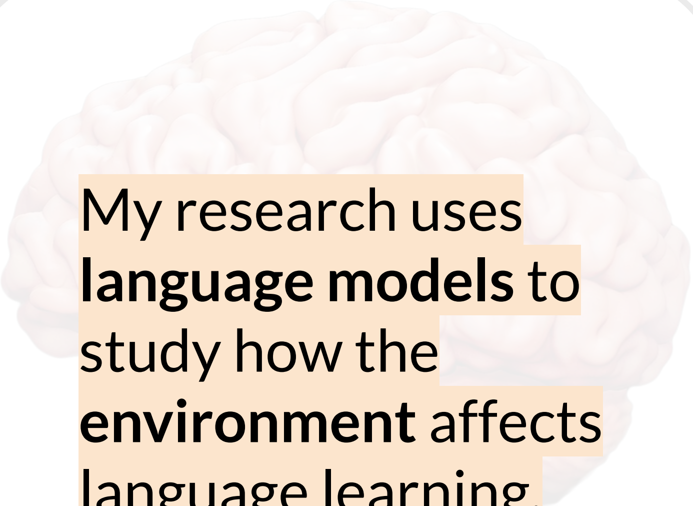
(Chomsky, 1965)

Two Sources of Grammatical Knowledge

Innate Bias

The Environment

My Research



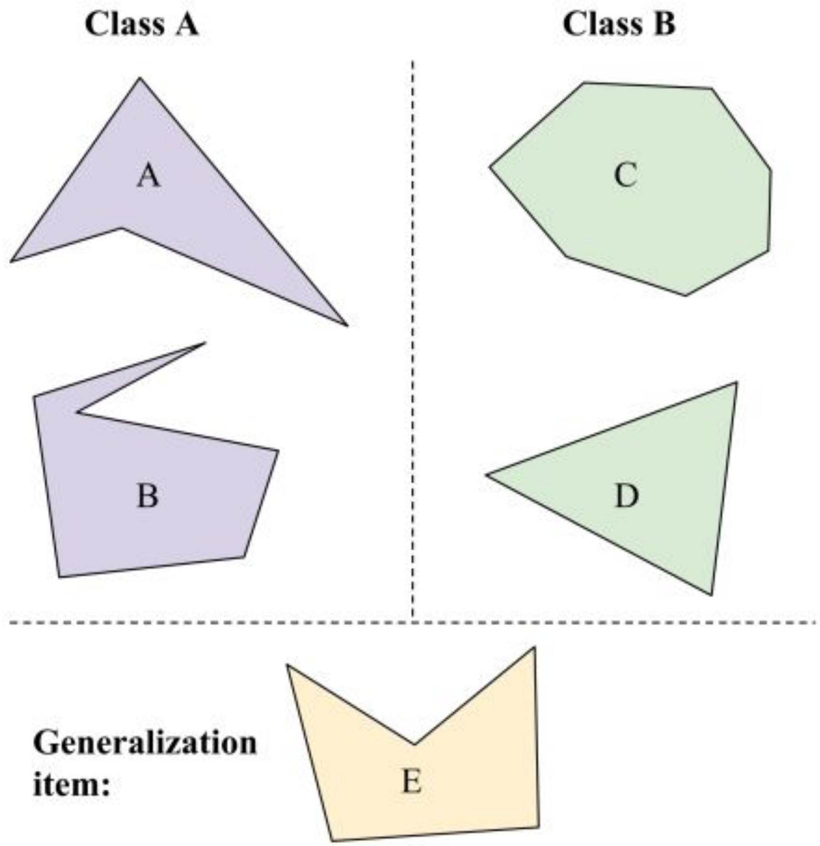
My research uses **language models** to study how the **environment** affects language learning.

Innate Bias

The Environment

Learning which features matter

An example



Pretraining → Feature Learning

- Dependency structures can be extracted from BERT (Hewitt & Manning, 2019)
- Contextual embeddings contain POS, semantic roles, coreference, etc. (Tenney et al., 2019a/b)

...*and many more* (see Rogers et al., 2020)

But feature learning isn't everything.

Representing *F* ≠ Using *F*

Models that represent linguistic features can still fail to use them during fine-tuning (McCoy et al., 2019).

Representing *F* ≠ Using *F*

Models that represent linguistic features can still fail to use them during fine-tuning (McCoy et al., 2019).

Inductive bias is also crucial to good generalization.

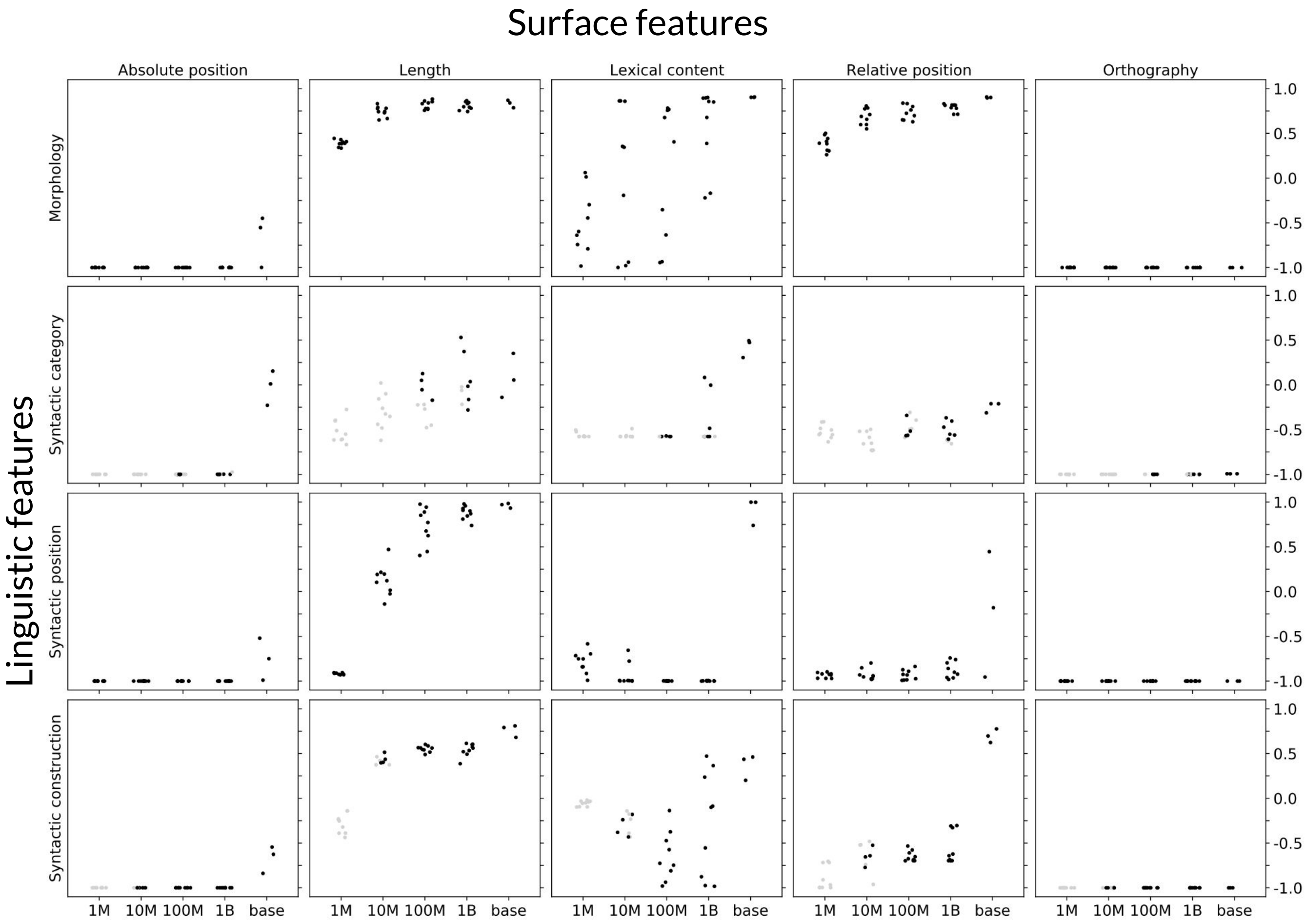
Learning which feature matter

New work in probing emphasizes feature *accessibility*:

- Minimum description length probing (Voita & Titov, 2020)
- Amnesic probing (Elazar et al., 2020)
- The classic probing paradigm is trivial when taken to the extreme (Pimentel et al., 2020)

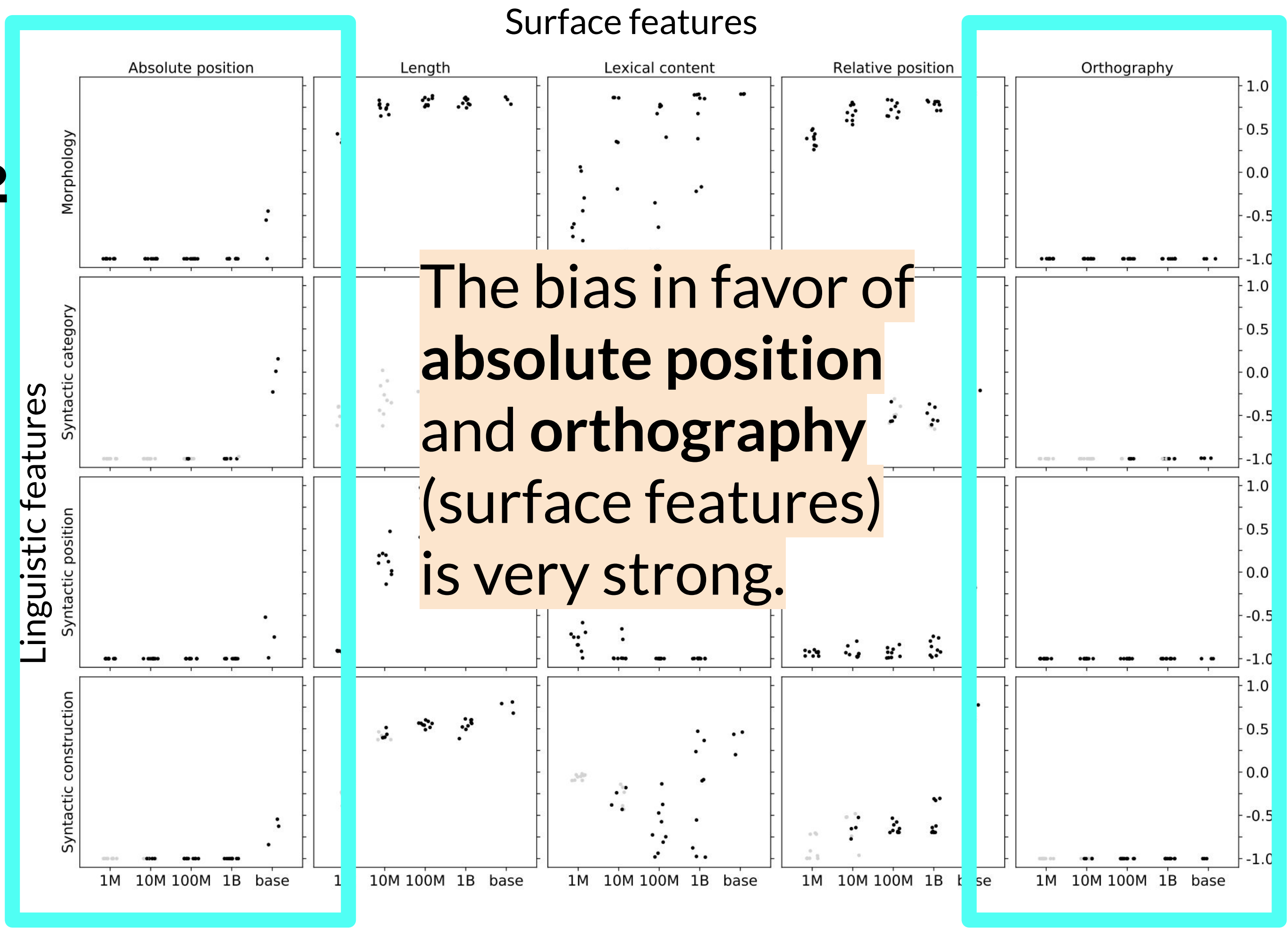
We probe feature preference explicitly.

Results: Experiment 2 (Ambiguous)
(Fine-grained)



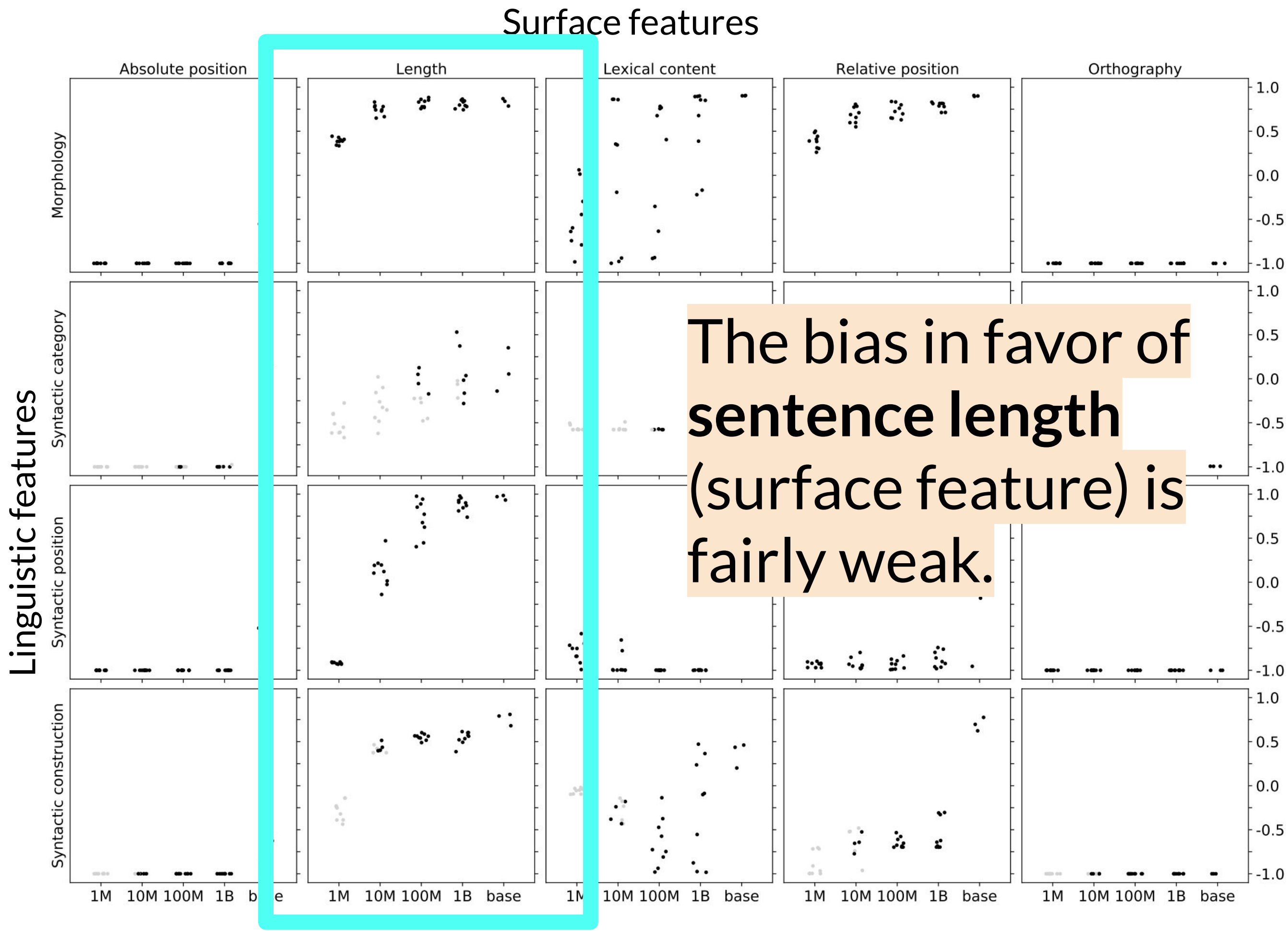
Results: Experiment 2 (Ambiguous)

(Fine-grained)



The bias in favor of **absolute position** and **orthography** (surface features) is very strong.

Results: Experiment 2 (Ambiguous)
(Fine-grained)



The bias in favor of **sentence length** (surface feature) is fairly weak.

Data Generation

- The MSGS data is generated from templates.
- We always test classifiers' ability to generalize out-of-domain.

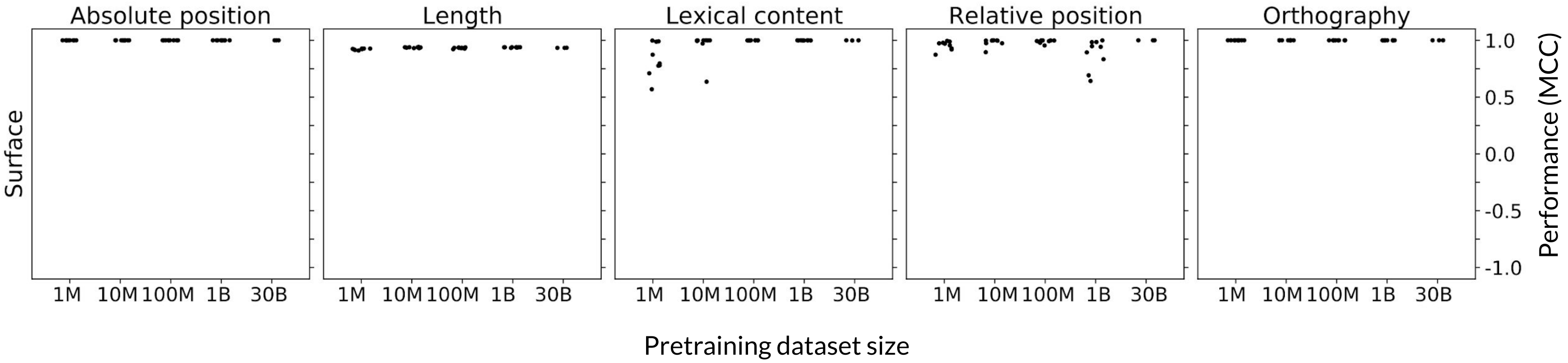
In domain: *The big dog is yawning.*

Out of domain: *The dog in the dark forest yawned.*

Fine-tuning

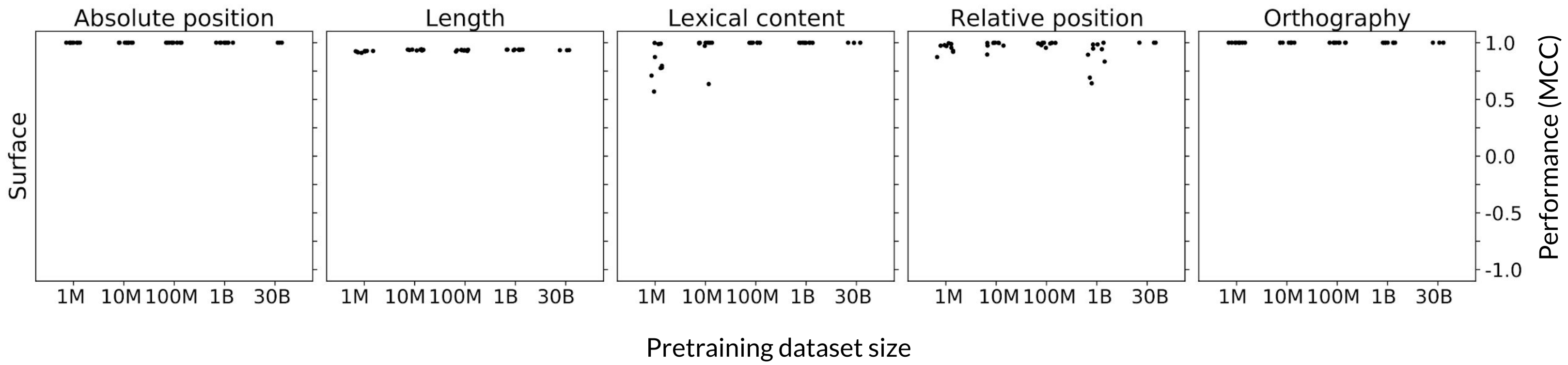
- 9 tasks (4 linguistic + 5 surface)
- 12 miniBERTas + original RoBERTa$_{BASE}$ (~30B words)
- The training sets are 10k sentences each

Results: Experiment 1 (Feature Learning)

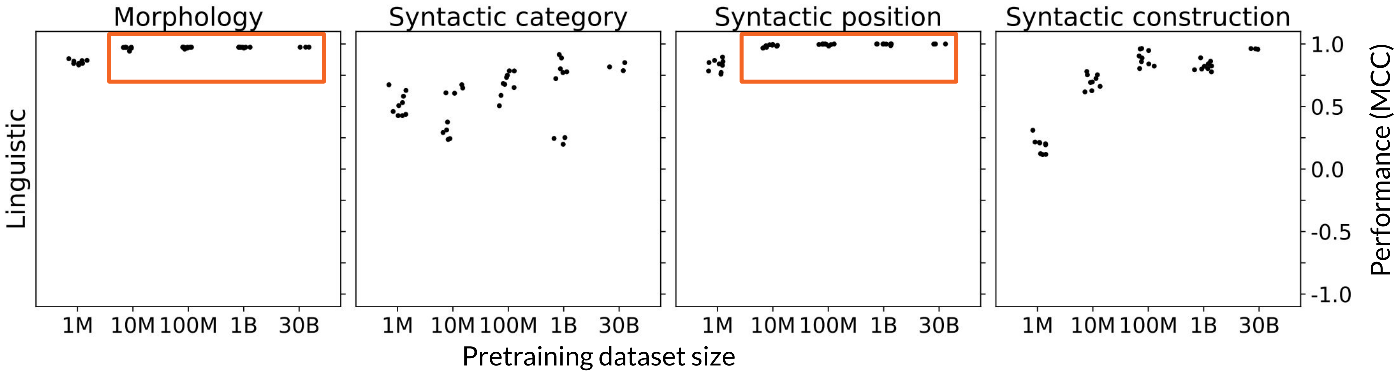


Surface features: Performance is at ceiling.

Results: Experiment 1 (Feature Learning)

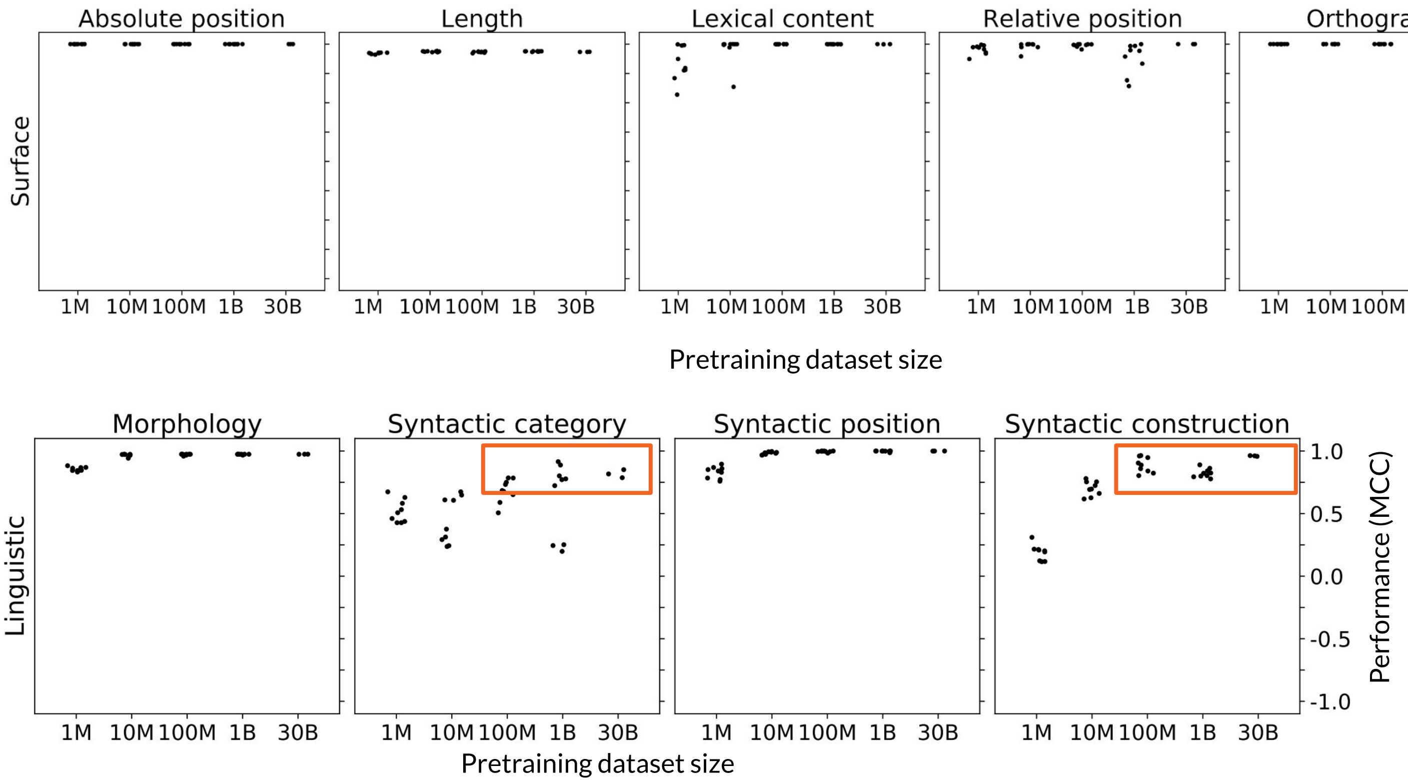


Surface features: Performance is at ceiling.



Linguistic features: Performance is near ceiling for morphology & syntactic position >1M words.

Results: Experiment 1 (Feature Learning)

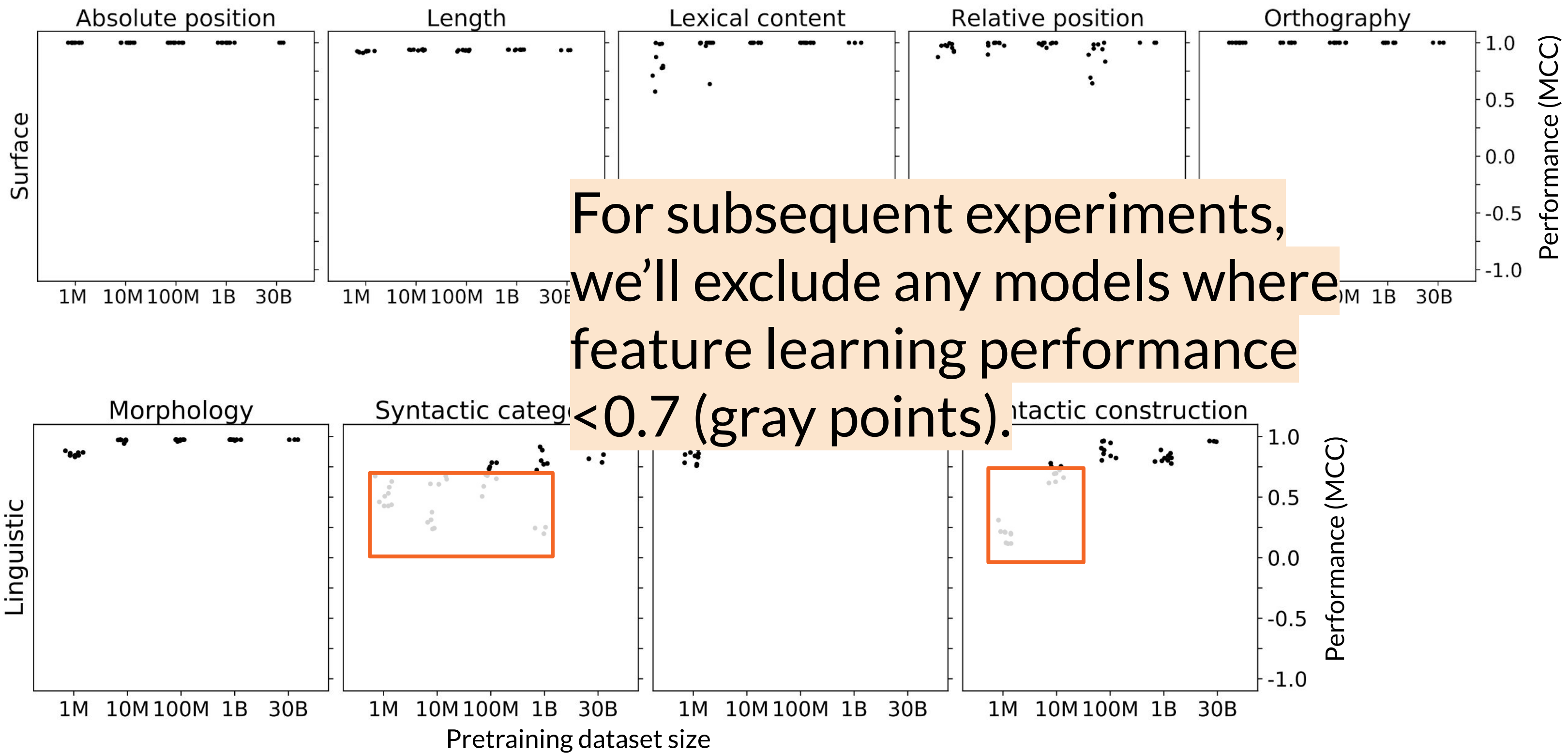


Surface features: Performance is at ceiling.

Linguistic features: Performance is near ceiling for morphology & syntactic position >1M words.

Performance for syntactic category & construction is high for >100M words.

Results: Experiment 1 (Feature Learning)



For subsequent experiments, we'll exclude any models where feature learning performance <0.7 (gray points).

Experiment 2: Ambiguous Data

Does model X prefer linguistic feature A or surface feature B?

Experiment 2: Ambiguous Data

Does model X prefer linguistic feature A or surface feature B?

We fine-tune X on an ambiguous binary classification task.

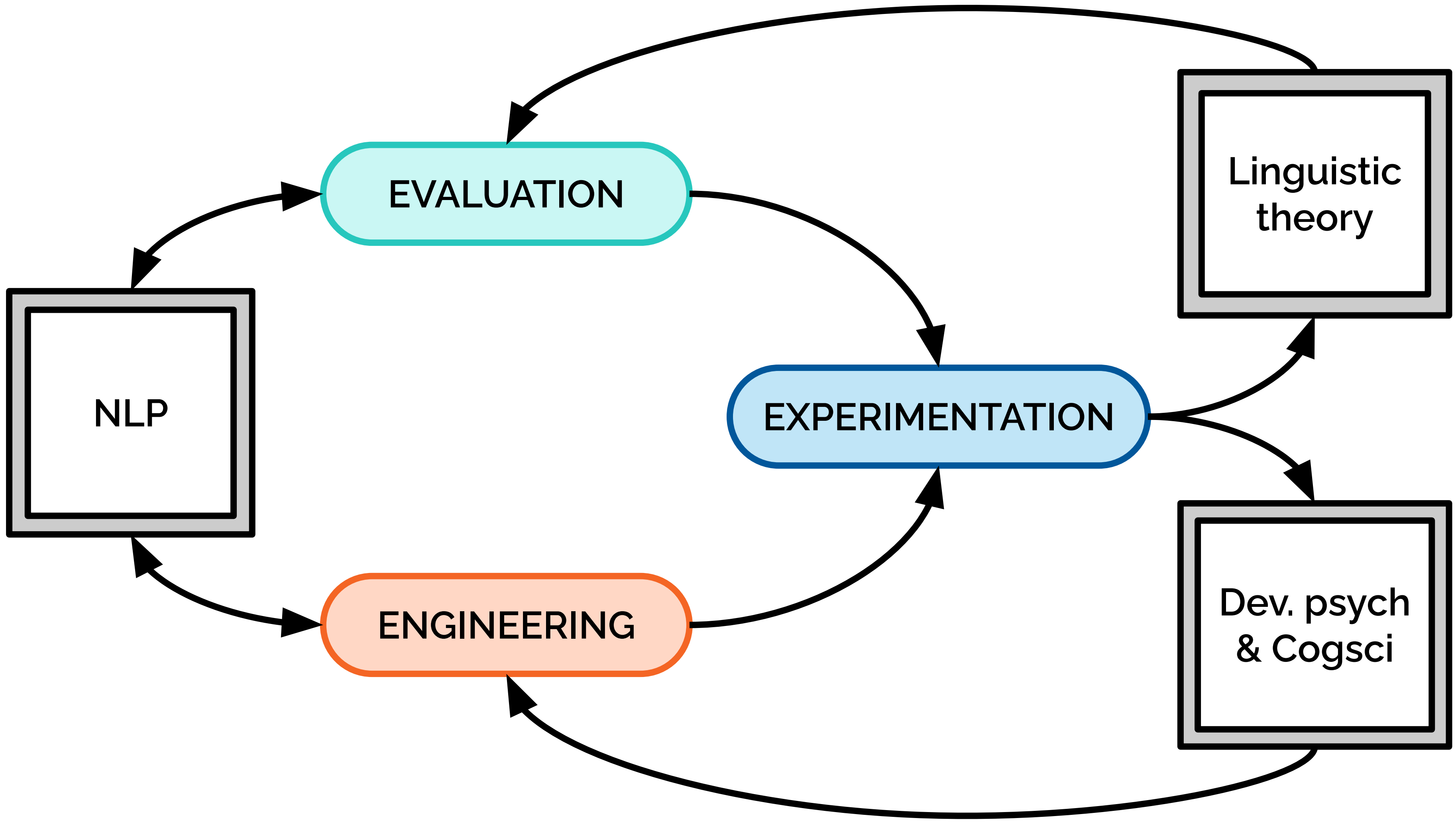
Experiment 2: Ambiguous Data

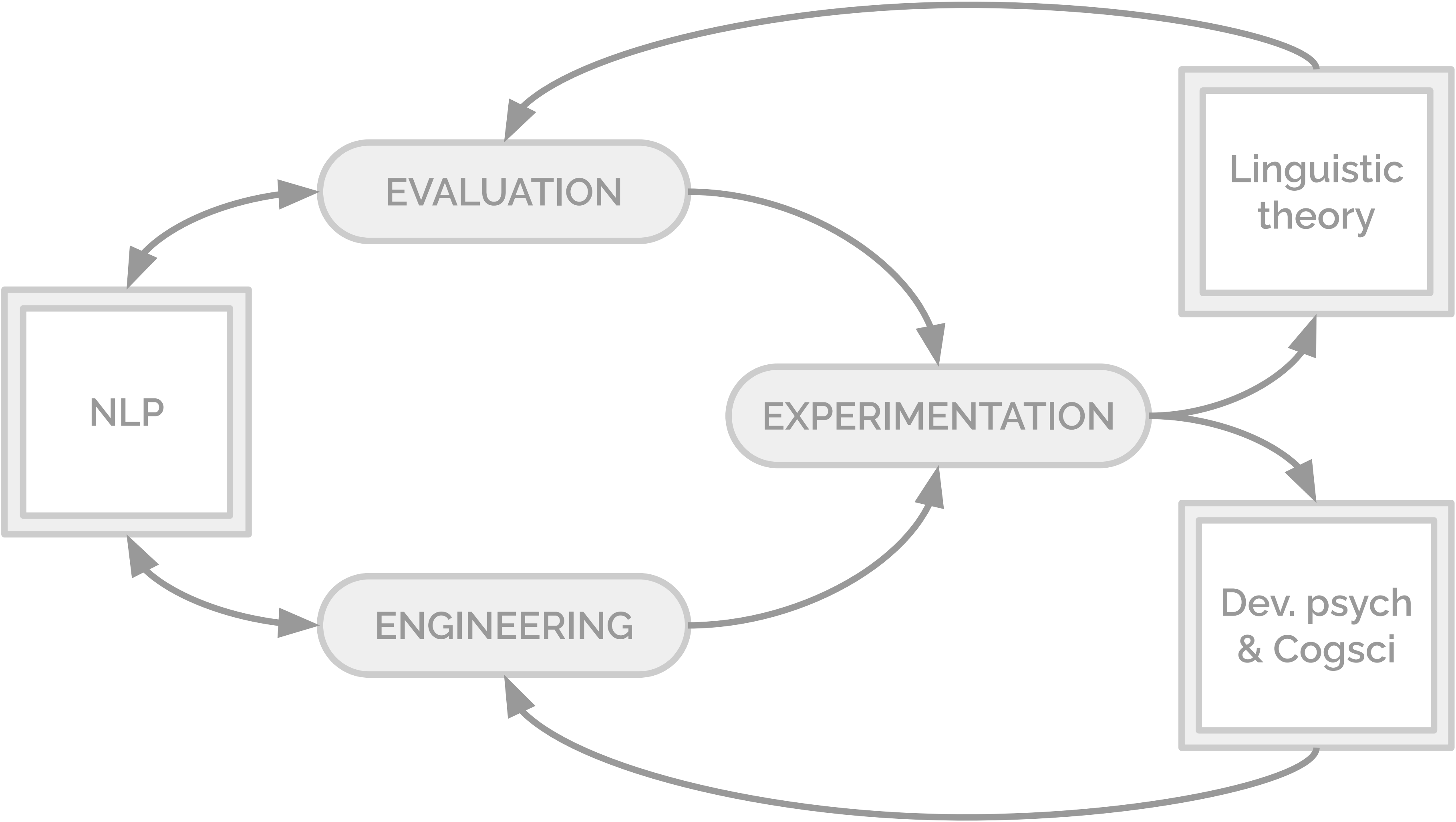
Does model X prefer linguistic feature A or surface feature B?

We fine-tune X on an ambiguous binary classification task.

Poverty of the Stimulus design (Wilson, 2006)

- Also used by McCoy et al. (2018, 2020), Warstadt & Bowman (2020), and others.

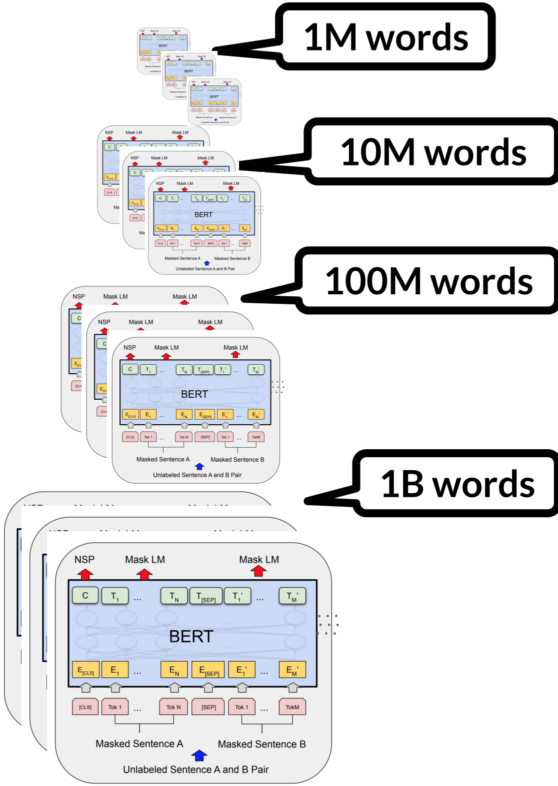

EVALUATION
Linguistic theory
NLP
EXPERIMENTATION
Dev. psych & Cogsci
ENGINEERING


EVALUATION
Linguistic theory
NLP
EXPERIMENTATION
ENGINEERING
Dev. psych & Cogsci

The MiniBERTas

https://huggingface.co/nyu-mll

- 4 incremental datasets: 1M, 10M, 100M, 1B words
- We simulate the original BERT training set:
  - ~¾ English Wikipedia
  - ~¼ self-published books from Smashwords
- We mostly follow the original RoBERTa training procedure.
- For each data size, we train at least 10 models with varying hyperparameters (e.g., # of parameters) & select the best 3.



Hypothetical Human Inductive Biases

Linguistic features

- Inflectional form
- Syntactic category
- Syntactic position
- Semantic roles

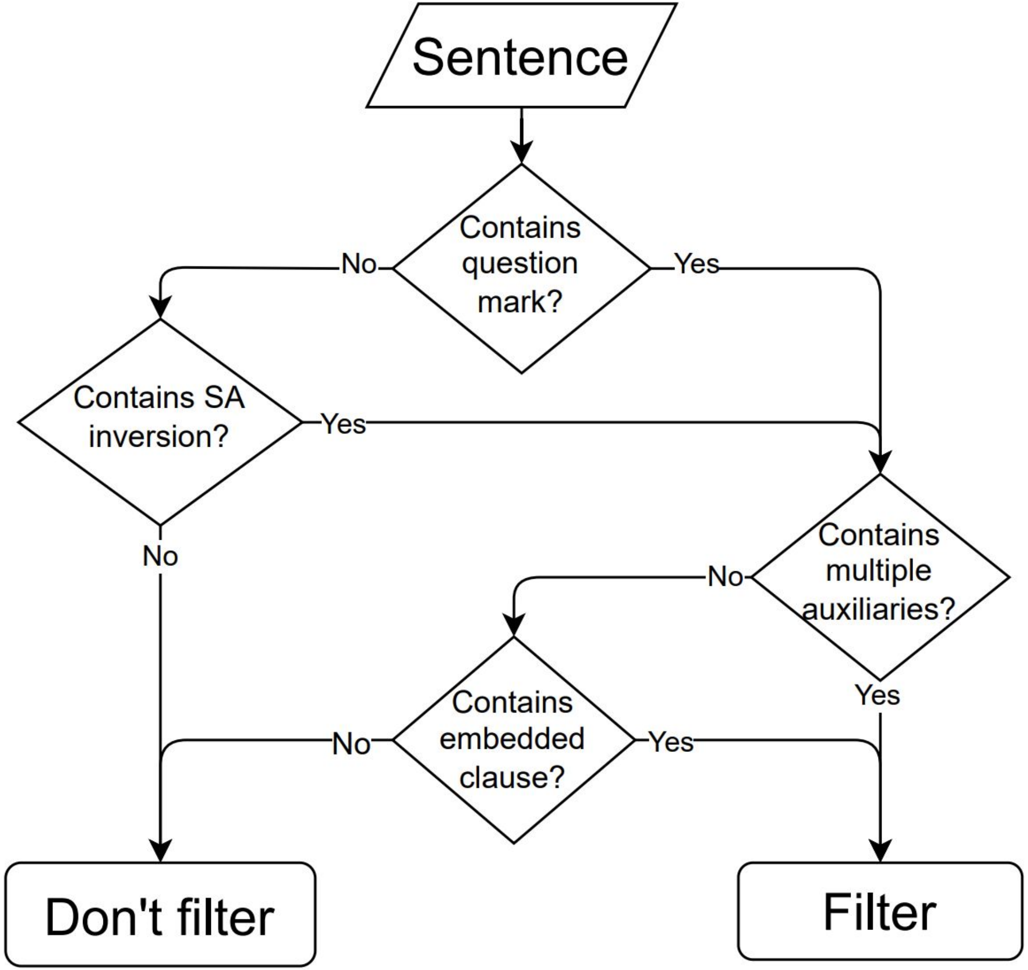
Surface features

- Linear position
- Length
- Lexical content
- Orthography
- Linear precedence

Syntactic filtering

Training data: 1B words from books & Wikipedia

- Percent filtered: 1.7%
- Recall (% of direct evidence removed): 99%
- Precision (% of removed data that is direct evidence): 51%

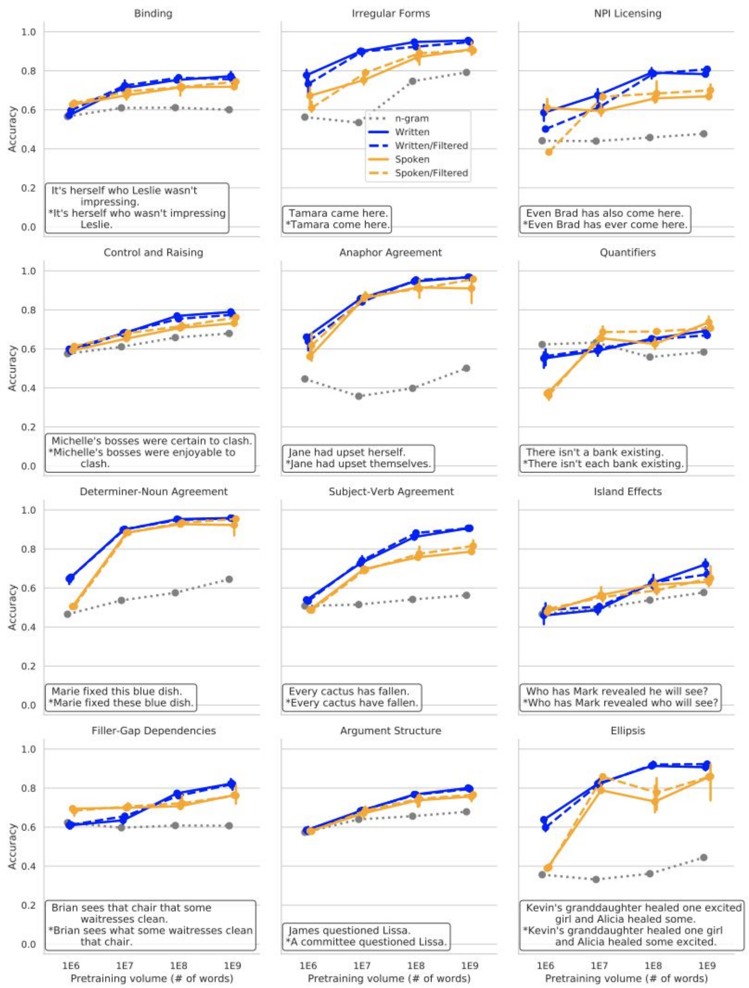


Evaluation

We do BLiMP-style evaluation on a hand-crafted test suite of subject-auxiliary inversion minimal pairs.

We designed minimal pairs following 8 different templates to probe generalization to different syntactic structures, and compared LM scores for the good and bad sentences.

Results: General acceptability judgments on BLiMP



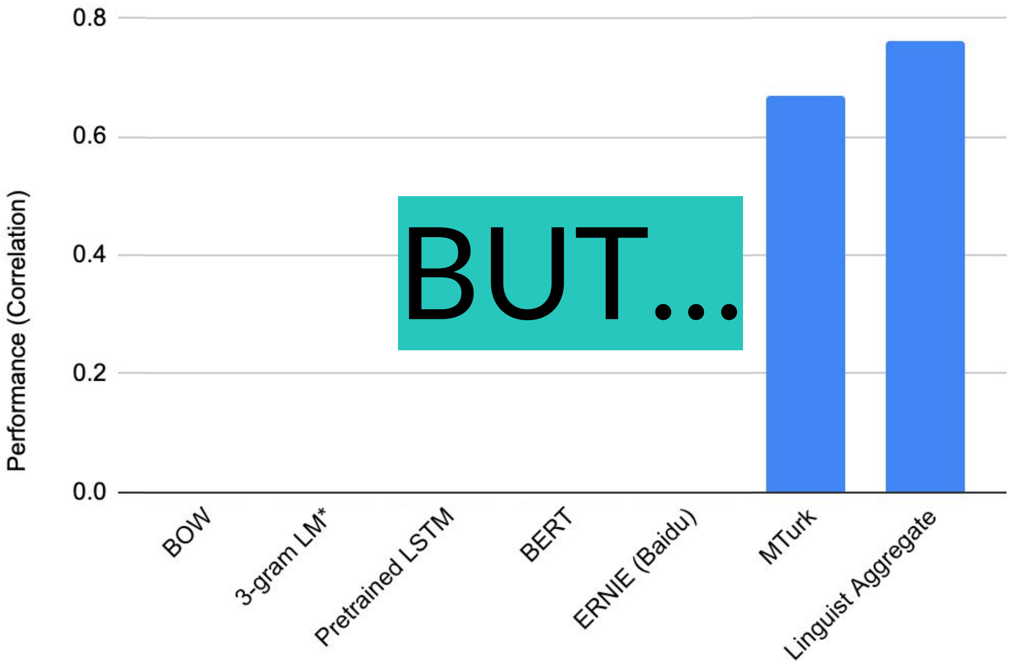
This result holds across all phenomena in BLiMP.

Takeaways

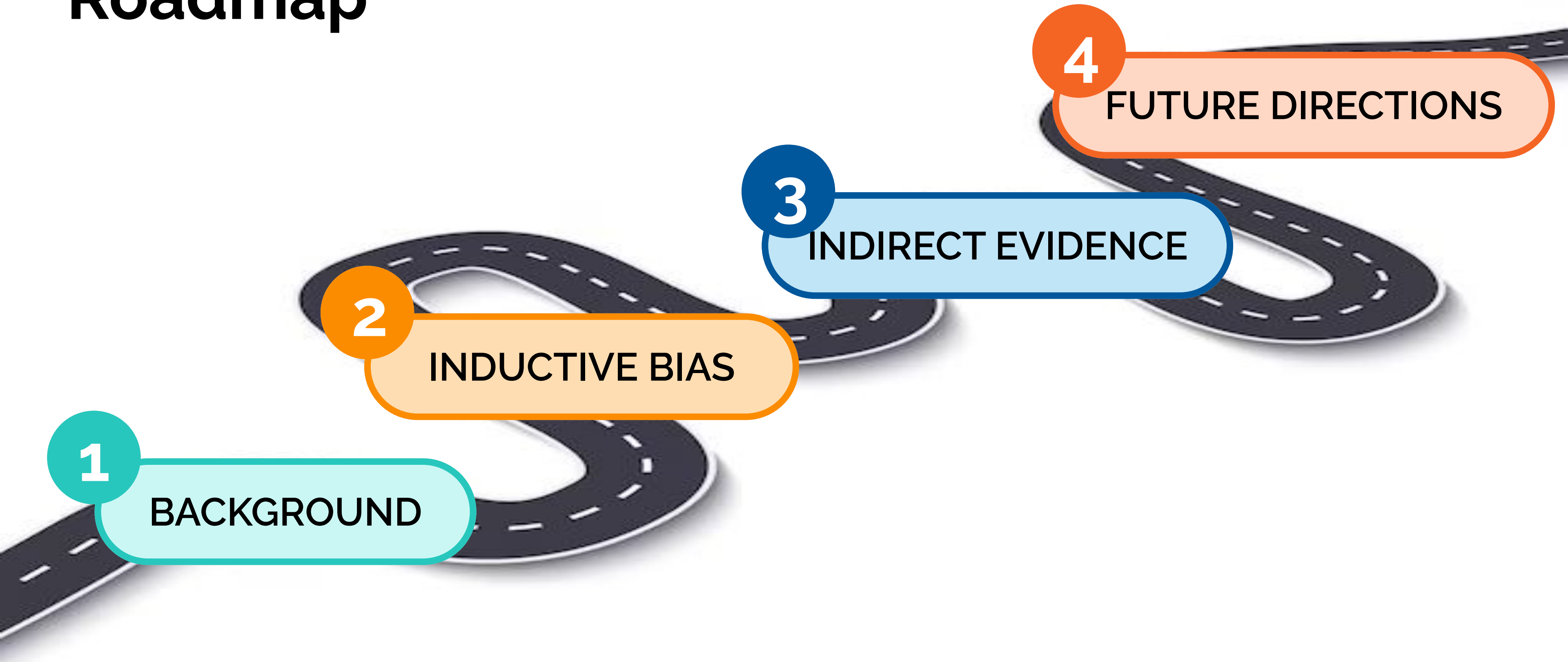
The results support the indirect evidence hypothesis, but with important caveats.

- 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?
- Can we identify and quantify indirect evidence?

The Corpus of Linguistic Acceptability (CoLA)



Roadmap


1
BACKGROUND
2
INDUCTIVE BIAS
3
INDIRECT EVIDENCE
4
FUTURE DIRECTIONS

Developments in text generation (2015-now)

how it started

===Widely accepted grammars===

There are twelve dialects which concern under the language of which which in sufficient, areas will be surprising before the racial controversy, probably those who in history, and no consensual is sincere.

Karpathy (2015)
http://karpathy.github.io/2015/05/21/rnn-effectiveness/
(h/t Will Merrill)

how it's going

Generate a wikipedia article titled:
===Widely accepted grammars===

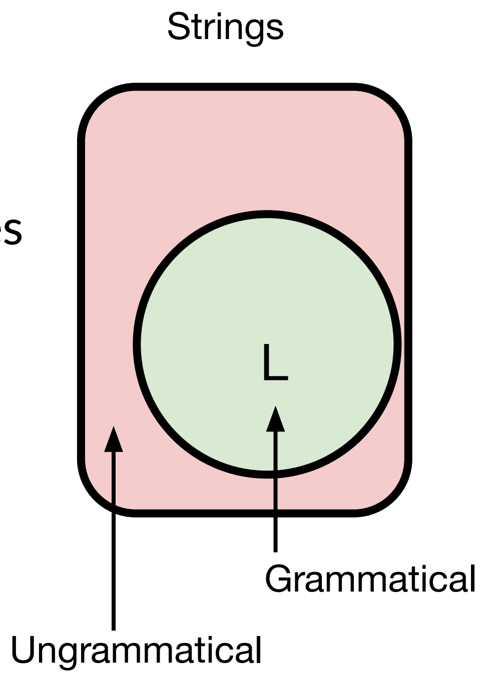
In linguistics, grammar refers to the set of rules that govern the structure of a language…. One of the most well-known grammars is the generative grammar proposed by Noam Chomsky in the 1950s.

GPT-4 (OpenAI, 2023)

Acceptability Judgments

An empirically adequate grammar of a language L generates all and only the grammatical strings of L.

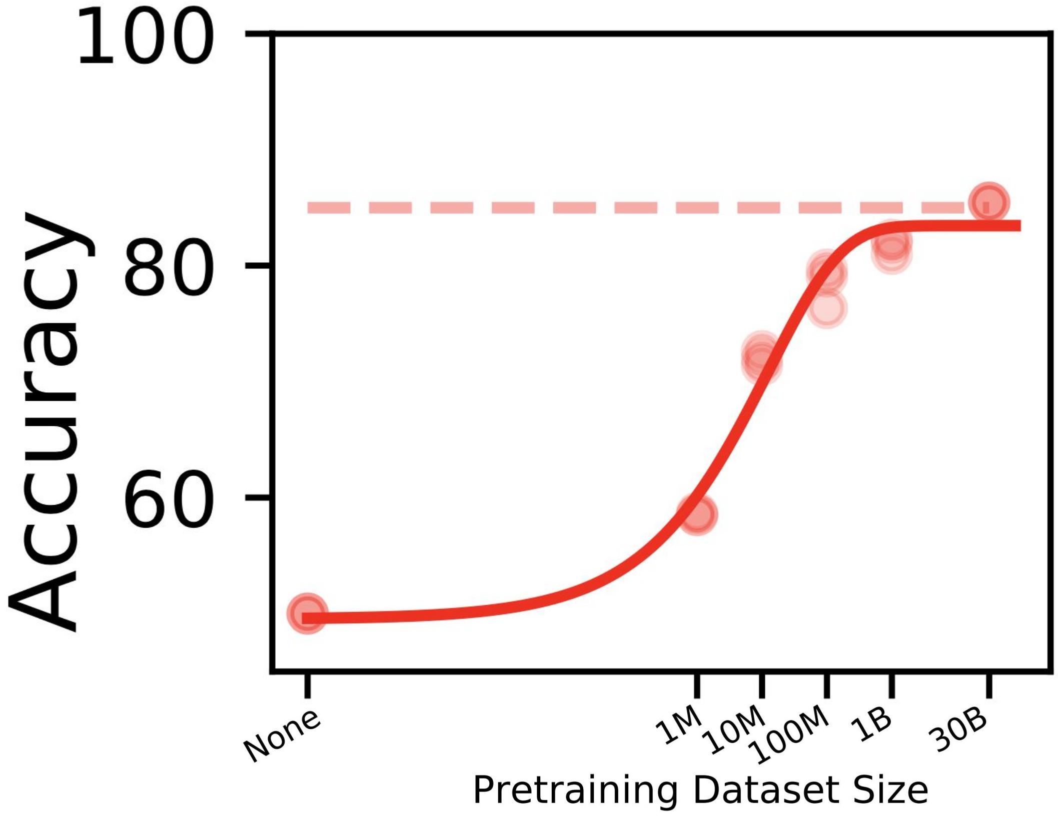
Acceptability judgments are the primary behavioral test of grammatical theories in linguistics.



Examples from linguistics publications

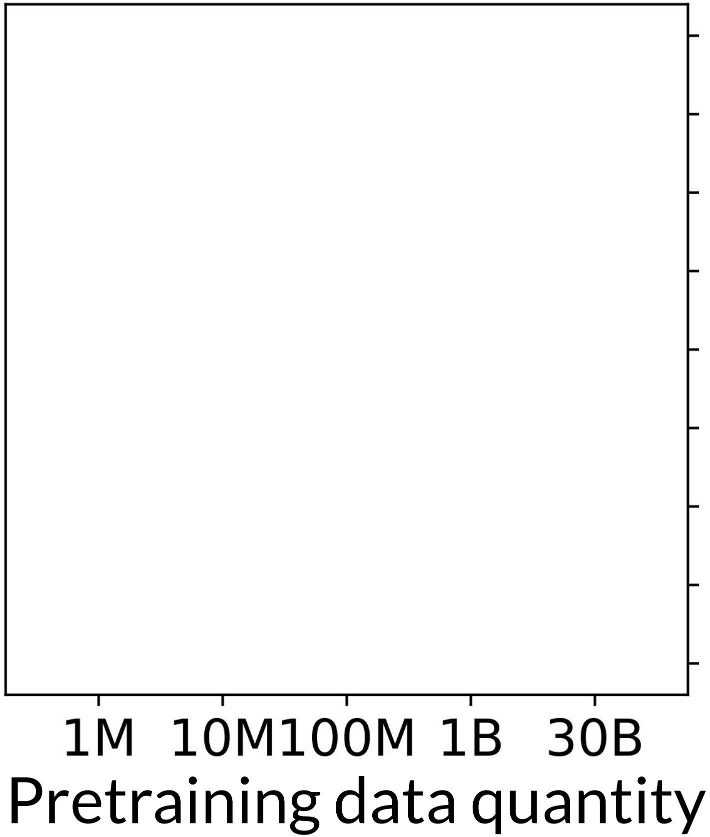
✓ Mary should know that you must go to the station.
✓ I promised that around midnight he would be there.
✓ Susan whispered the news to Rachel.
✗ When time will you be there?
✗ Patrick is likely that left.
✗ Harry coughed us into a fit.

The MiniBERTas on BLiMP

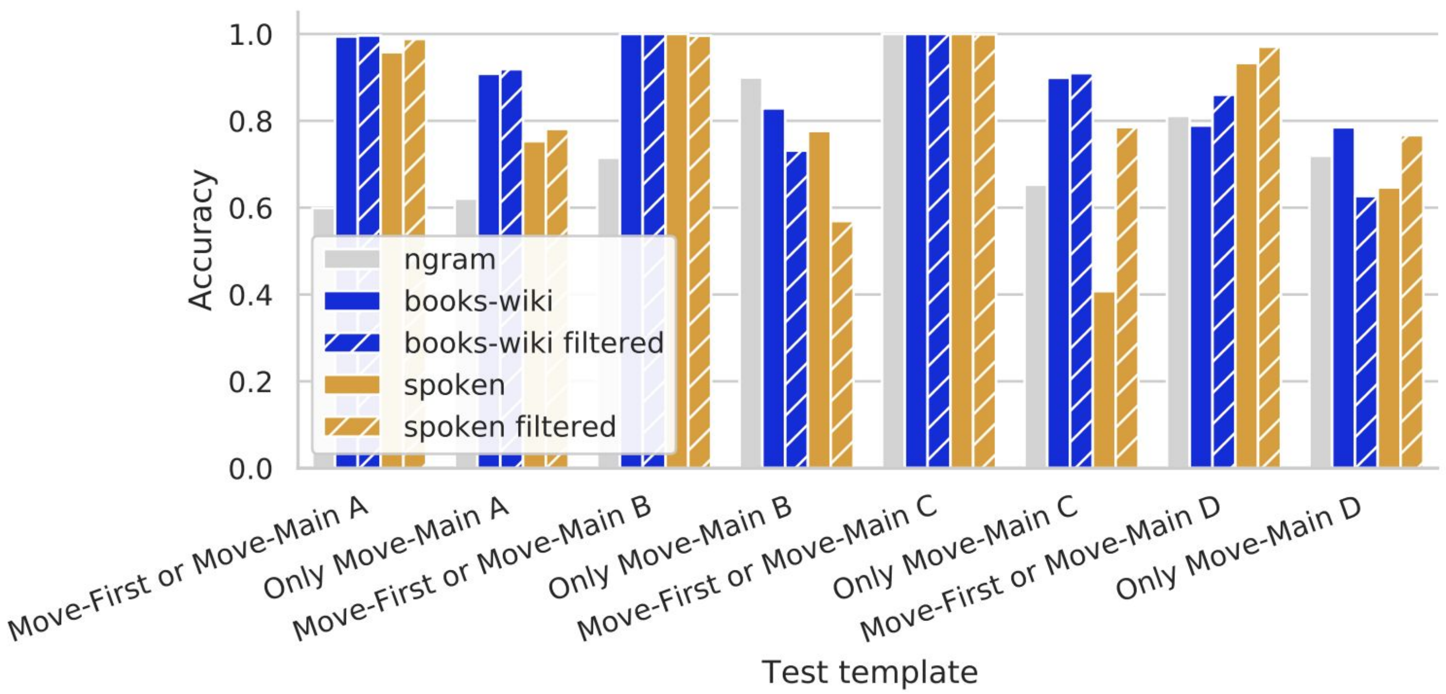


Results: Experiment 1 (Fully Ambiguous)

- 20 tasks * (12 miniBERTas + RoBERTa base)
- Linguistic bias score = 1 if linguistic, -1 if surface.
- <1B words: surface bias
- RoBERTa base: 50/50



Results: Subject Aux Inversion (BEST CASE)



The Recipe for Model Learners

1. Minimize any advantages that language models have over humans learners.
2. Provide language models with more of the advantages that we know humans have.
3. Gather training data from developmentally plausible sources.