Testing the learnability of grammar for humans and machines: Investigations with artificial neural networks

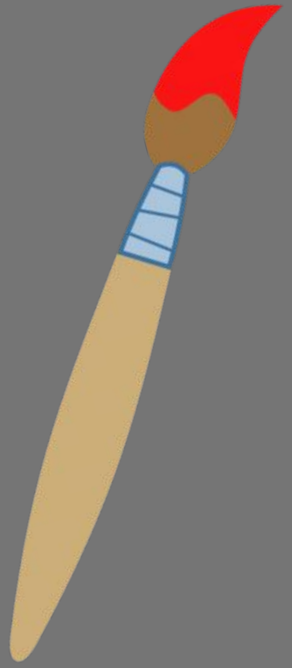
Alex Warstadt
New York University Linguistics
21 October 2021
Text as Data

Center for Data Science
NYU

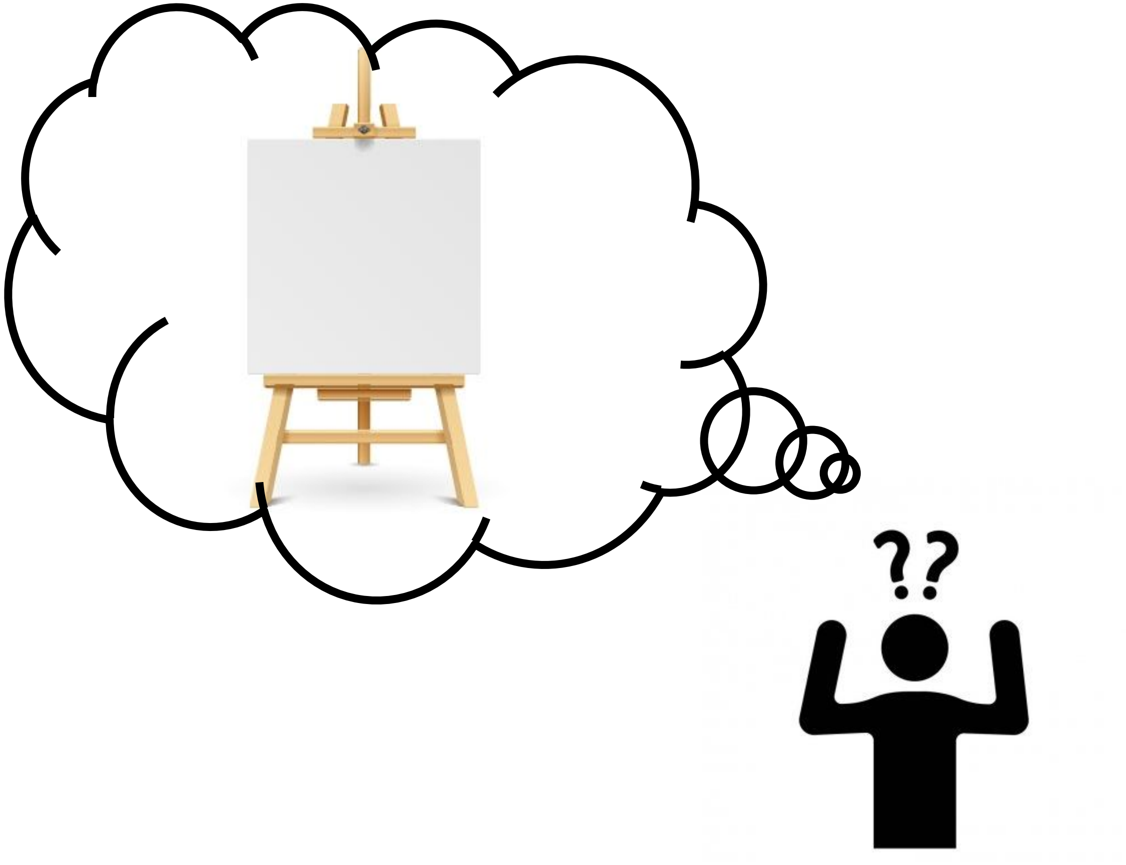
What currently holds the state-of-the-art in language learning?

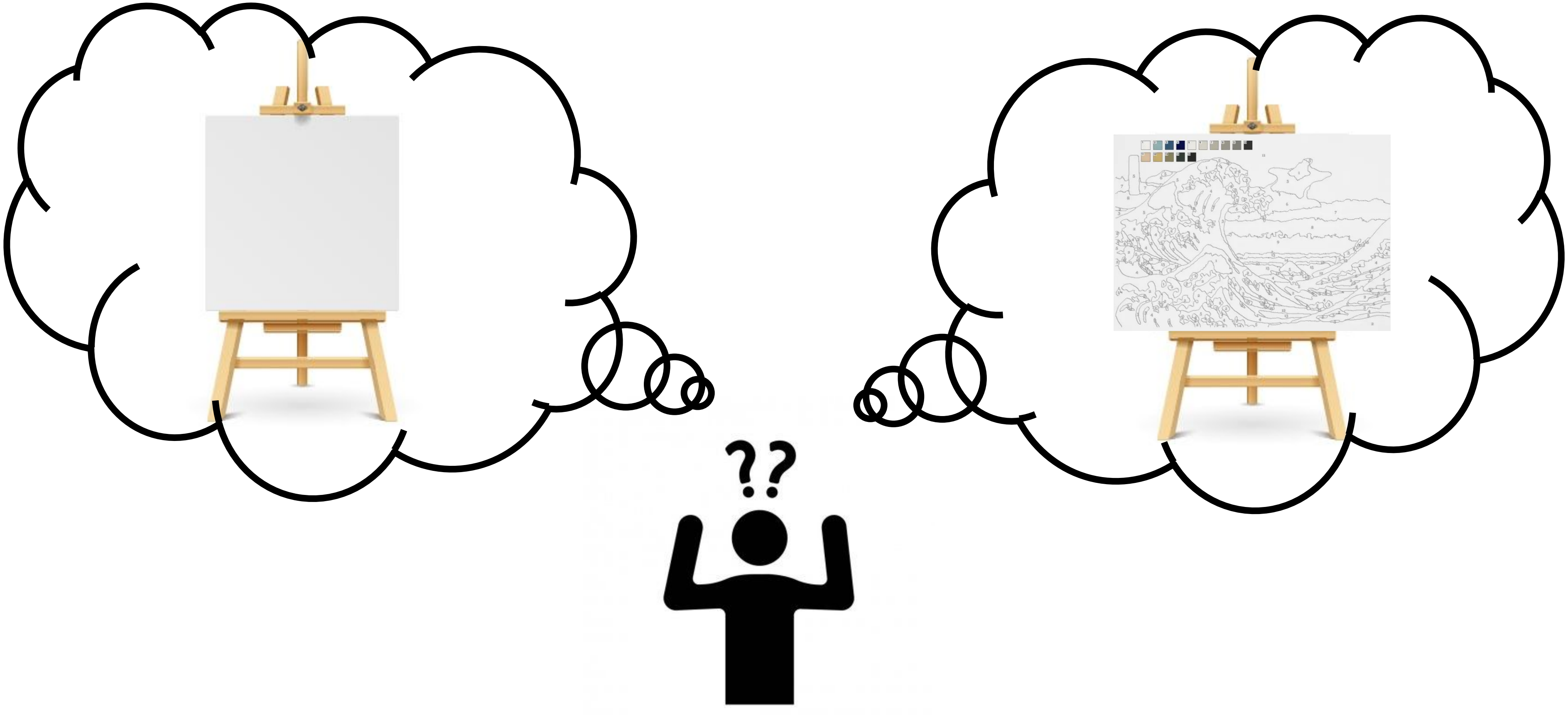
What currently holds the state-of-the-art in language learning?

Our linguistic environments color learning.



What about the canvas?






SLOW
PROCEED
WITH
CAUTION

What currently holds the state-of-the-art in language learning?

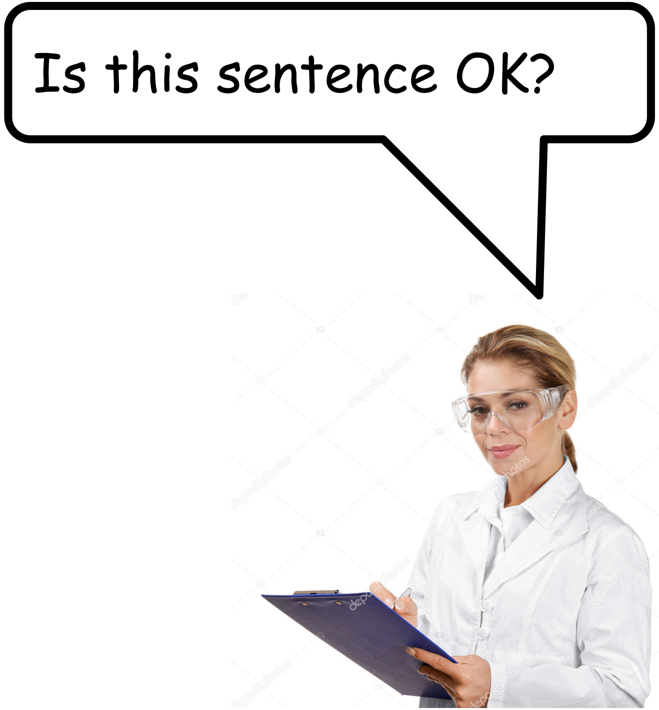
Roadmap

1. Tests for grammatical knowledge
    a. Acceptability judgments
    b. CoLA
    c. BLiMP
2. More human-like training
    a. Probing
    b. Acquiring inductive bias
3. What neural networks can teach us about humans
    a. The idea experiment
    b. Obstacles and opportunities

Part 1:
Tests for grammatical knowledge

Acceptability Judgments

Acceptability Judgments



Acceptability Judgments

Is this sentence OK?

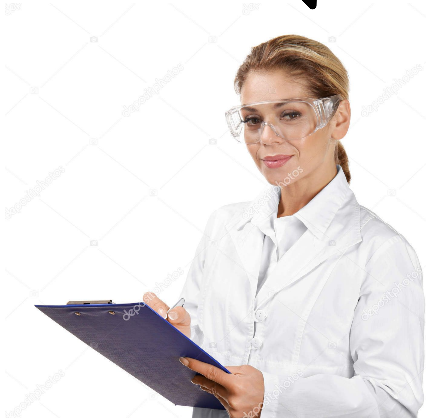
What did Betsy paint a picture of?

Acceptability Judgments

Is this sentence OK?

What did Betsy paint a picture of?

What was a picture of painted by Betsy?



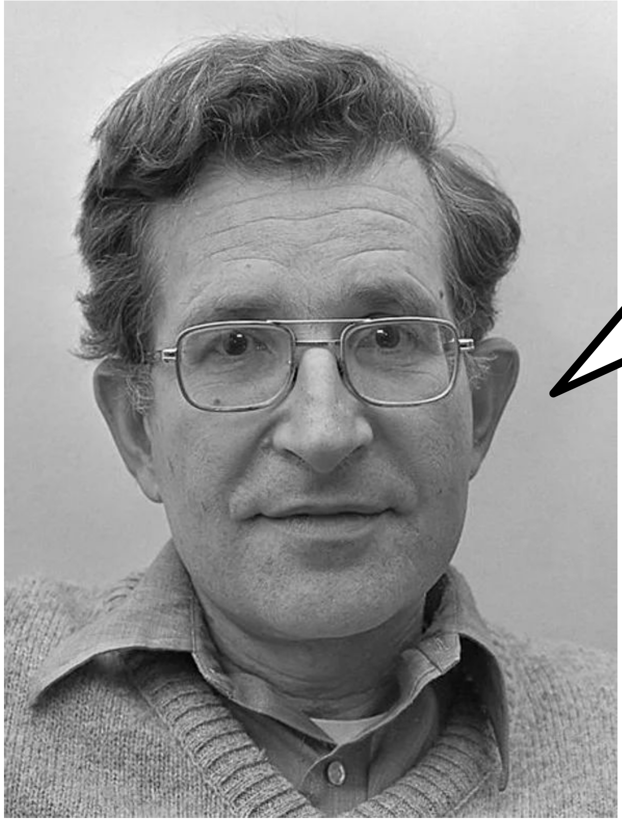
Acceptability Judgments

Is this sentence OK?

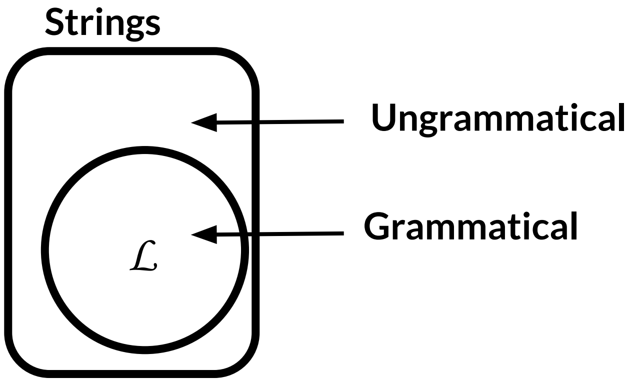
What did Betsy paint a picture of?

What was a picture of painted by Betsy?

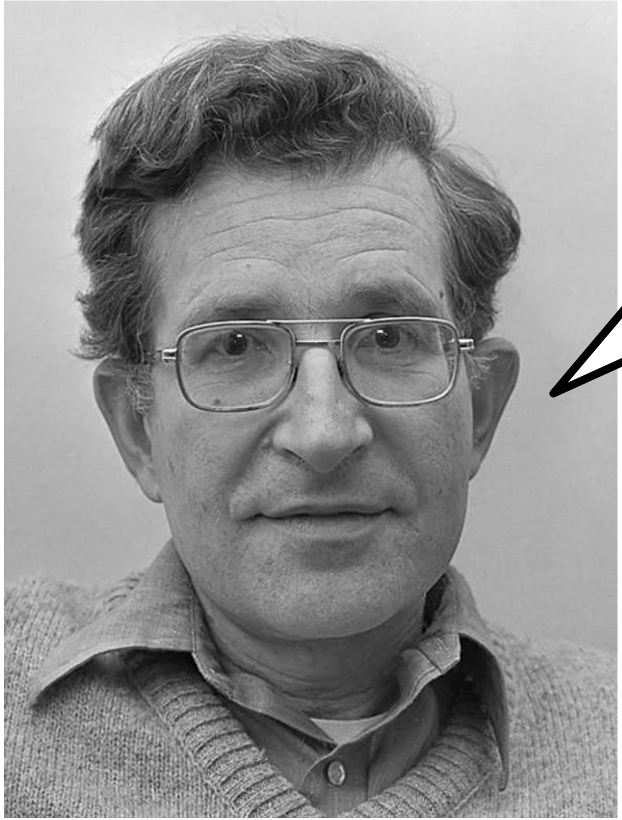
What's the relation between acceptability judgments and grammar?



The fundamental aim in the linguistic analysis of a language L is to separate the grammatical sequences which are the sentences of L from the ungrammatical sequences which are not sentences of L and to study the structure of the grammatical sequences.

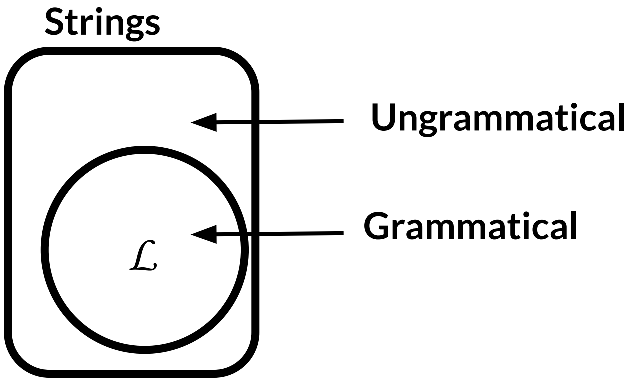


Noam Chomsky, 1957. *Syntactic Structures*.



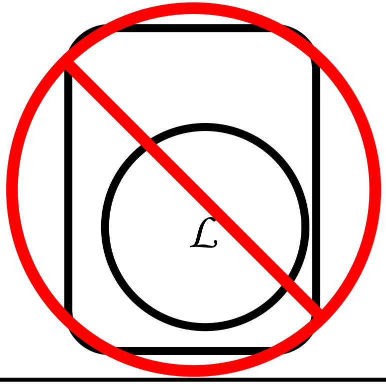
One way to test the adequacy of a grammar proposed for [language] L is to determine whether or not the sequences that it generates are actually grammatical, i.e., acceptable to a native speaker.”

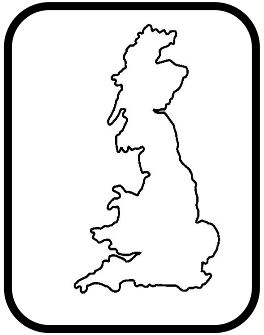
Noam Chomsky, 1957. *Syntactic Structures*.



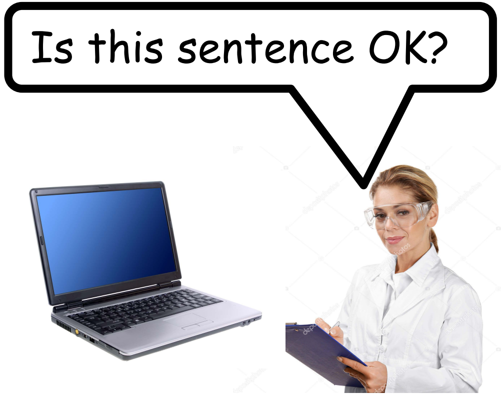
Human grammatical knowledge is:

- Complex
- Strongly held
- Implicit (not taught)
- Widely shared





Linguistic Competence of NNs?



We can compare NNs to humans by recasting acceptability judgments as an NLP task.

An NN with knowledge of grammar should easily learn to make human-like acceptability judgments.

CoLA

The Corpus of Linguistic Acceptability

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

Table 1: Breakdown of CoLA by source.

	n	%label=1	Description
Total	**10657**	**70.5**	
In Domain	**9515**	**71.3**	
Adger (2003)	948	71.9	Syntax textbook
Baltin (1982)	96	66.7	Movement
Baltin and Collins (2001)	880	66.7	Handbook
Bresnan (1973)	259	69.1	Comparatives
Carnie (2013)	870	80.3	Syntax textbook
Culicover and Jackendoff (1999)	233	59.2	Comparatives
Dayal (1998)	179	75.4	Modality
Gazdar (1981)	110	65.5	Coordination
Goldberg and Jackendoff (2004)	106	77.4	Resultative
Kadmon and Landman (1993)	93	81.7	Negative Polarity
Kim and Sells (2008)	1965	71.2	Syntax Textbook
Levin (1993)	1459	69.0	Verb alternations
Miller (2002)	426	84.5	Syntax textbook
Rappaport Hovav and Levin (2008)	151	69.5	Dative alternation
Ross (1967)	1029	61.8	Islands
Sag et al. (1985)	153	68.6	Coordination
Sportiche et al. (2013)	651	70.4	Syntax textbook
Out of Domain	**1049**	**69.2**	
Chung et al. (1995)	148	66.9	Sluicing
Collins (2005)	66	68.2	Passive
Jackendoff (1971)	94	67.0	Gapping
Sag (1997)	112	57.1	Relative clauses
Sag et al. (2003)	460	70.9	Syntax textbook
Williams (1980)	169	76.3	Predication

CoLA

- **>10k sentences from the syntax/semantics literature.**
- **Expert boolean acceptability judgments.**
- **Broad domain of phenomena**
- **>20x larger than similar resources.**

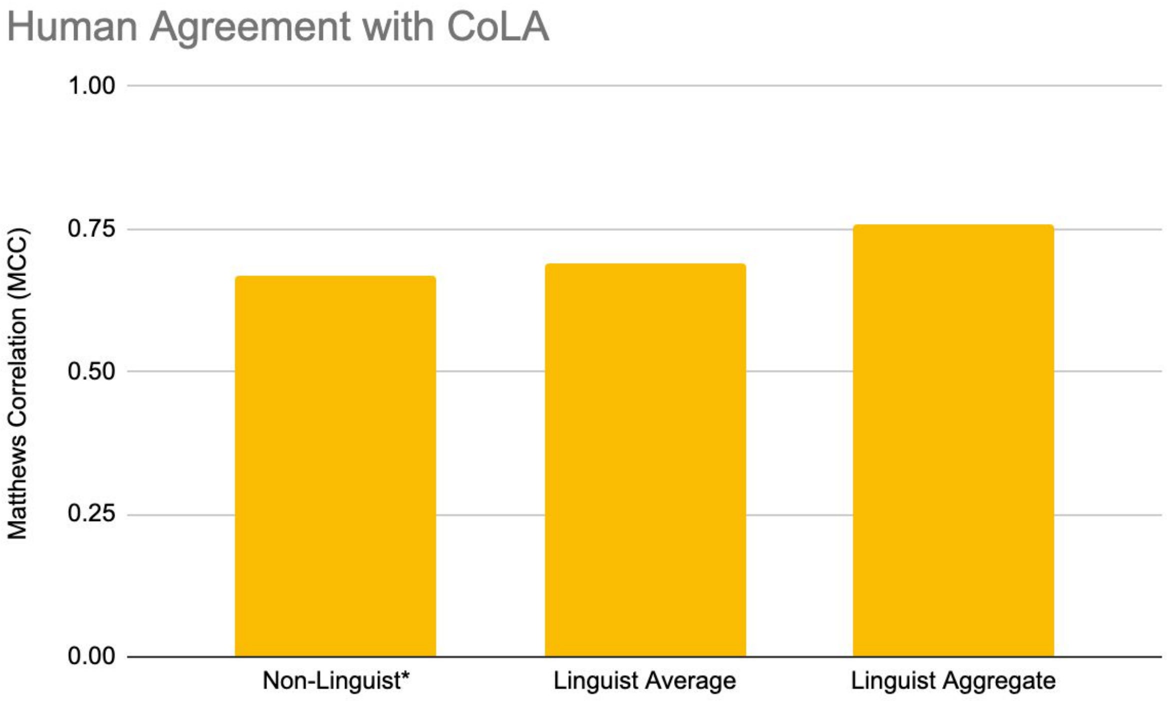
CoLA: Phenomena covered

Included	Morphological Violation	(a)	*Maryann should leaving.
	Syntactic Violation	(b)	*What did Bill buy potatoes and _?
	Semantic Violation	(c)	*Kim persuaded it to rain.
Excluded	Pragmatical Anomalies	(d)	*Bill fell off the ladder in an hour.
	Unavailable Meanings	(e)	*He$_i$ loves John$_i$. (*intended*: John loves himself.)
	Prescriptive Rules	(f)	Prepositions are good to end sentences with.
	Nonce Words	(g)	*This train is arrivable.

CoLA Sample

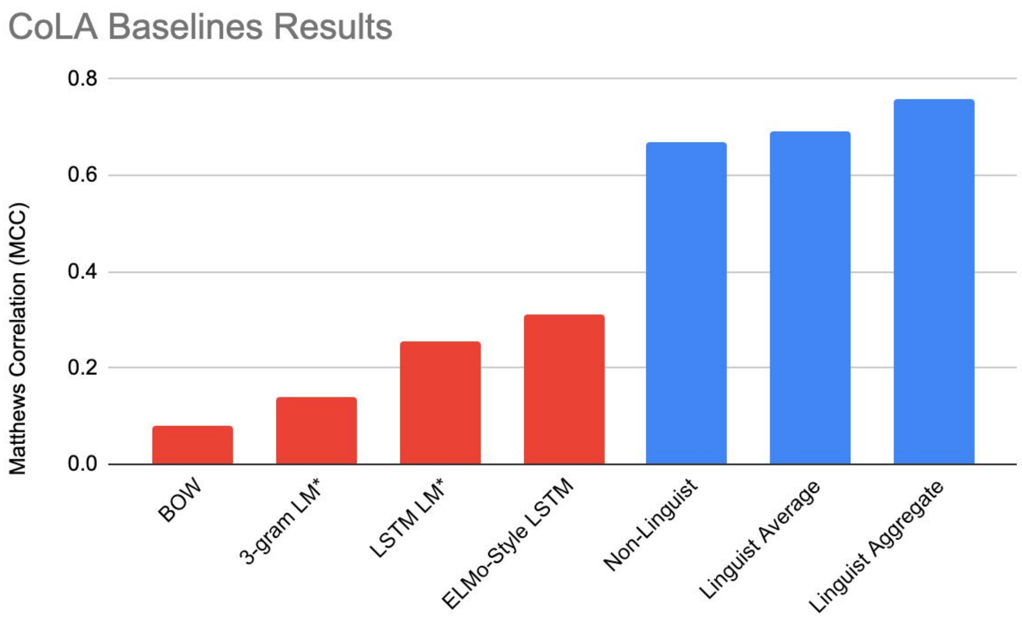
Label	Sentence	Source
0	The ball wiggled itself loose.	gj04
0	The more books I ask to whom he will give, the more he reads.	cj99
1	I said that my father, he was tight as a hoot-owl.	r-67
1	The jeweller inscribed the ring with the name.	l-93
0	We rummaged papers through the desk.	l-93
0	many evidence was provided.	ks08
1	They can sing.	ks08
1	This theorem will take only five minutes to establish that he proved in 1930.	ks08
1	The men would have been all working.	b-82
1	Would John hate that?	b-82
0	Who do you think that will question Seamus first?	c-13
0	Usually, any lion is majestic.	d-98
1	Larry Twentyman hunted all the foxes.	m-02
1	I wrote Blair a letter, but I tore it up before I sent it.	rhl07
1	That's the kindest answer that I ever heard.	b-73

Measuring Human Performance



*Nangia & Bowman (2019)

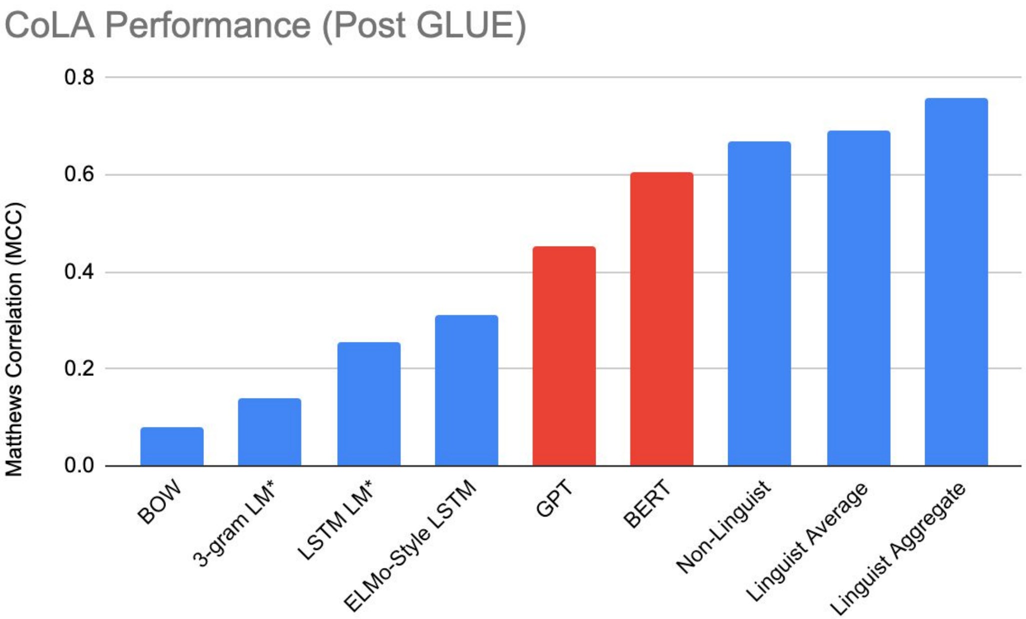
Baselines



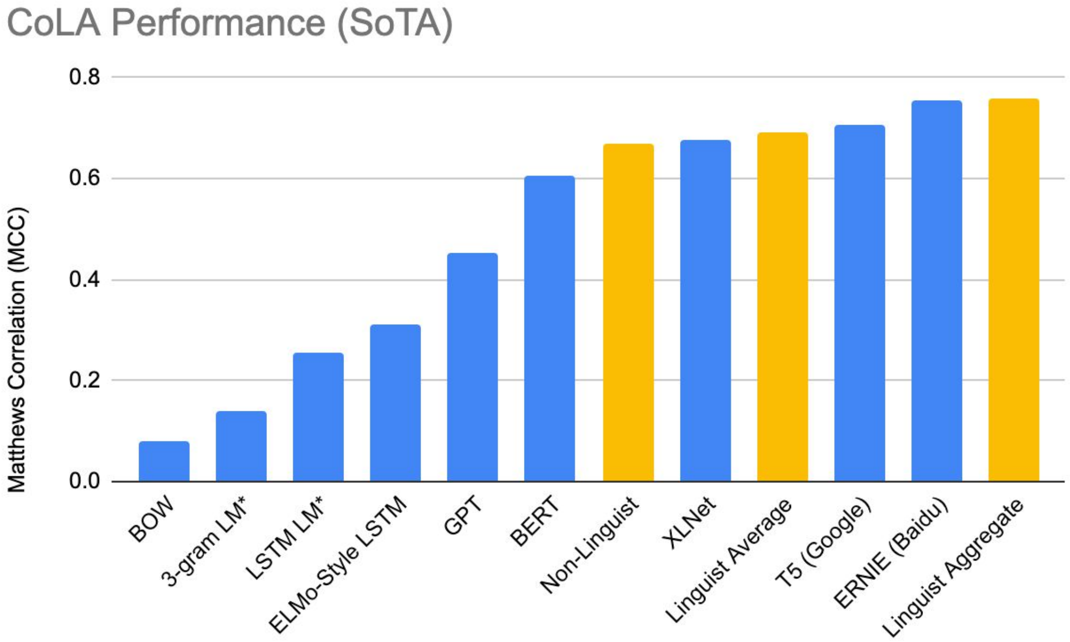
CoLA Baselines Results
Matthews Correlation (MCC)
0.8
0.6
0.4
0.2
0.0
BOW
3-gram LM*
LSTM LM*
ELMo-Style LSTM
Non-Linguist
Linguist Average
Linguist Aggregate

Coca-Cola

Early Transformers



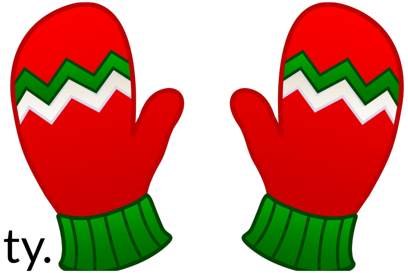
Superhuman Results?



Not so fast...

Evaluating on CoLA requires supervised training, which exposes the model to explicit information about acceptability.

Enter: Minimal Pairs

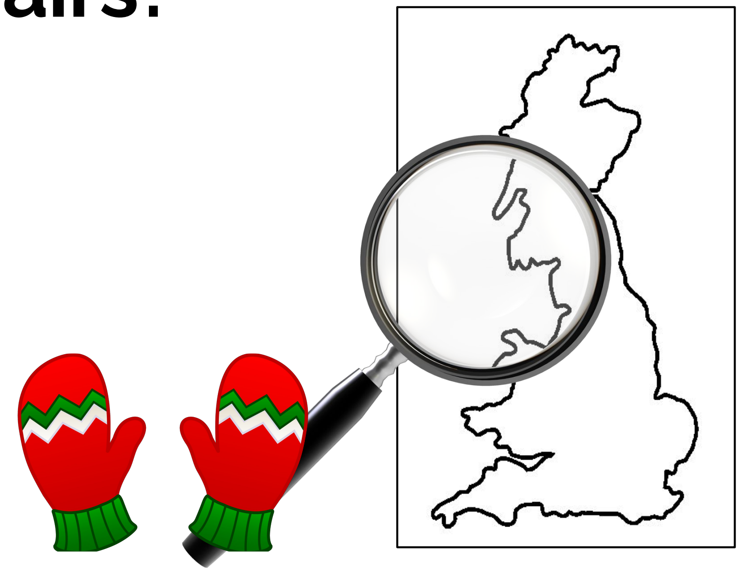


A pair of two nearly identical sentences which differ in acceptability.

Betsy is eager to sleep. ✓

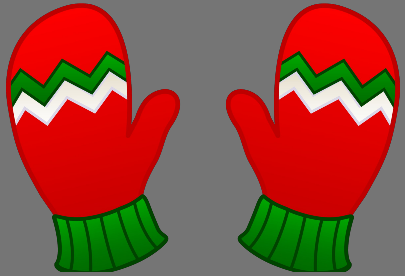
Betsy is easy to sleep. ✗

Why Minimal Pairs?



If $P_{LM}(S_{✓}) > P_{LM}(S_{✗})$, then LM detects a contrast in acceptability.

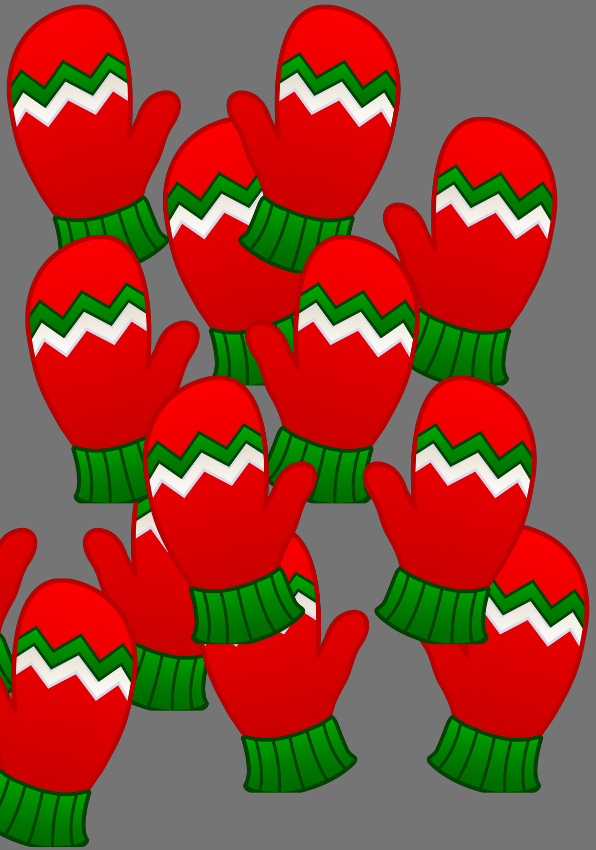
Recently, there's been an abundance of work testing LMs on minimal pairs.



Sample of Work Using Minimal Pairs

Phenomenon	**Relevant work**
Anaphor/binding	Marvin & Linzen (2018); Futrell et al. (2018); Warstadt et al. (2019b)
Subject-verb agreement	Linzen et al. (2016); Futrell et al. (2018); Gulordava et al. (2019); Marvin & Linzen (2018); An et al. (2019); Warstadt et al. (2019b)
Negative polarity items	Marvin & Linzen (2018); Futrell et al. (2018); Jumelet & Hupkes (2018); Wilcox et al. (2019); Warstadt et al. (2019a)
Filler-gap dependencies & islands	Wilcox et al. (2018); Warstadt et al. (2019b); Chowdhury & Zamparelli (2018, 2019); Chaves (to appear); Da Costa & Chaves (to appear)
Argument structure	Kann et al. (2019); Warstadt et al. (2019b); Chowdhury & Zamparelli (2019)

Things are getting a bit complicated...



We need...

1 dataset to rule them all.

BLiMP: The Benchmark of Linguistic Minimal Pairs






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

Enter: BLiMP

A wide-coverage dataset of targeted minimal pairs.

67 unique paradigms with 1000 minimal pairs each, organized into 12 categories.

Evaluation is simple: just compare LM probabilities on the good and bad sentences.

All minimal pairs in BLiMP:

(a) Are equal in length.
(b) Differ in at most 1 vocabulary item.

Data -- Coverage

Phenomenon	N	Acceptable example	Unacceptable example
Anaphor agreement	2	Many girls insulted themselves.	Many girls insulted herself.
Argument structure	9	Rose wasn’t disturbing Mark.	Rose wasn’t boasting Mark.
Binding	7	It’s himself who Robert attacked.	It’s himself who attacked Robert.
Control/Raising	5	Kevin isn’t irritating to work with.	Kevin isn’t bound to work with.
Determiner-N 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	Which bikes is John fixing?	Which is John fixing bikes?
NPI licensing	7	The truck has clearly tipped over.	The truck has ever tipped over.
Quantifiers	4	There was a cat annoying Alice.	There was each cat annoying Alice.
Subject-Verb agr.	6	These casseroles disgust Kayla.	These casseroles disgusts Kayla.

Data Generation

Data generation allows for large, syntactically controlled datasets.

We use a hand-crafted vocabulary of >3K items.

- More comprehensive than similar resources.
- >70 morphological, syntactic, and semantic features.

expression	category	category_2	verb	noun	non_v_pred	frequent	sg	pl	mass	animate	properNoun	finite	bare	pres	past	ing	en	3sg	arg_1	arg_2	arg_3
skateboard	N			1		1	1	0	0	0	0										
skateboards	N			1		1	0	1	0	0	0										
wheelbarrow	N			1		1	1	0	0	0	0										
wheelbarrows	N			1		1	0	1	0	0	0										
computer	N			1		1	1	0	0	0	0										
computers	N			1		1	0	1	0	0	0										
screen	N			1		1	1	0	0	0	0										
screens	N			1		1	0	1	0	0	0										
heal	(S\NP)/NP		1			1						0	1	0	0	0	0	0	animate=1	animate=1;animal=1	
heal	(S\NP)/NP		1			1						1	0	1	0	0	0	0	sg=0^animate=1	animate=1;animal=1	
heals	(S\NP)/NP		1			1						1	0	1	0	0	0	1	sg=1^animate=1	animate=1;animal=1	
healed	(S\NP)/NP		1			1						1	0	0	1	0	0	0	animate=1	animate=1;animal=1	
healed	(S\NP)/NP		1			1						0	0	0	0	0	1	0	animate=1	animate=1;animal=1	
healing	(S\NP)/NP		1			1						0	0	0	0	1	0	0	animate=1	animate=1;animal=1	
sick	N/N	adjective			1	1													animate=1;animal=1		
ill	N/N	adjective			1	1													animate=1;animal=1		
cure	(S\NP)/NP		1			1						0	1	0	0	0	0	0	animate=1	animate=1;animal=1	
cure	(S\NP)/NP		1			1						1	0	1	0	0	0	0	sg=0^animate=1	animate=1;animal=1	
cures	(S\NP)/NP		1			1						1	0	1	0	0	0	1	sg=1^animate=1	animate=1;animal=1	
cured	(S\NP)/NP		1			1						1	0	0	1	0	0	0	animate=1	animate=1;animal=1	
cured	(S\NP)/NP		1			1						0	0	0	0	0	1	0	animate=1	animate=1;animal=1	
curing	(S\NP)/NP		1			1						0	0	0	0	1	0	0	animate=1	animate=1;animal=1	

Data -- Generation procedure

Sentences are generated according to simple templates

```python
def sample(self):
    # What did      John read  before filing the book?
    # Wh   Aux_mat  Subj V_mat ADV    V_emb  Obj
    # What did      John read  the book before filing?
    # Wh   Aux_mat  Subj V_mat Obj      ADV    V_emb

    V_mat = choice(all_non_finite_transitive_verbs)
    Subj = N_to_DP_mutate(choice(get_matches_of(V_mat, "arg_1", all_nouns)))
    Aux_mat = return_aux(V_mat, Subj, allow_negated=False)
    Obj = N_to_DP_mutate(choice(get_matches_of(V_mat, "arg_2", all_nouns)))
    V_emb = choice(get_matched_by(Obj, "arg_2", get_matched_by(Subj, "arg_1", self.all_ing_transitives)))
    Wh = choice(get_matched_by(Obj, "arg_1", all_wh_words))
    Adv = choice(self.adverbs)

    data = {
        "sentence_good": "%s %s %s %s %s %s %s?" % (Wh[0], Aux_mat[0], Subj[0], V_mat[0], Adv, V_emb[0], Obj[0]),
        "sentence_bad": "%s %s %s %s %s %s %s?" % (Wh[0], Aux_mat[0], Subj[0], V_mat[0], Obj[0], Adv, V_emb[0])
    }
    return data, data["sentence_good"]
```

Data -- Human validation

Via Amazon Mechanical Turk, 20 English speaking annotators evaluate 5 pairs from each paradigm (6700 total judgments).

Forced choice task: annotators select the more acceptable sentence from a pair.

Inclusion criteria: Majority vote agreement with 4/5 pairs in the paradigm.

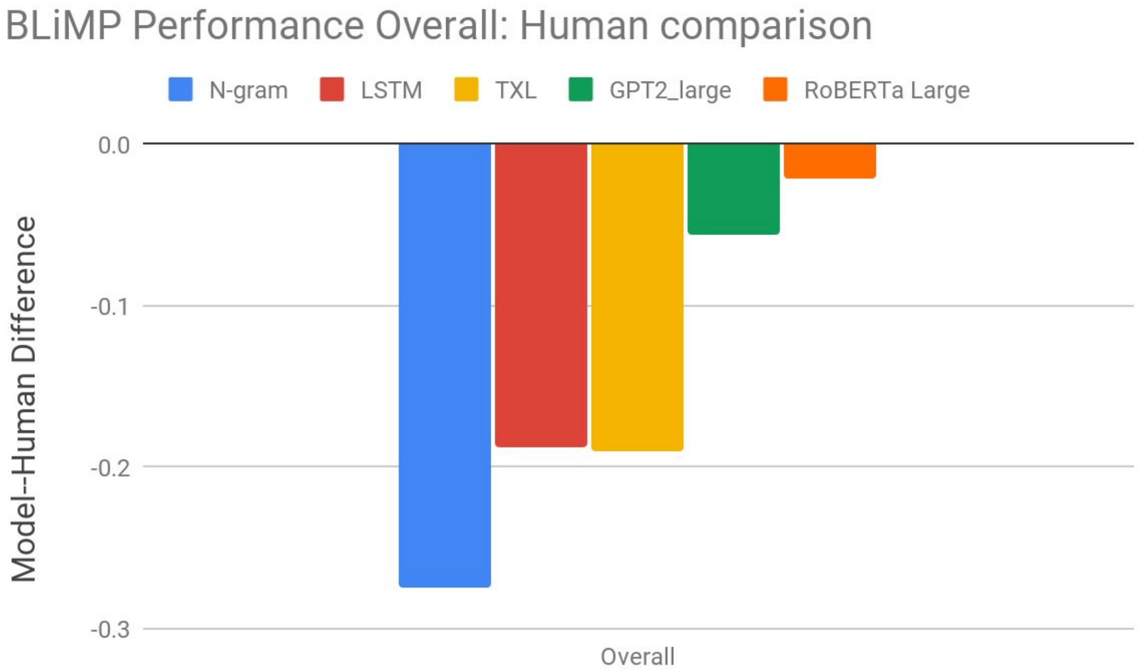
Majority vote human agreement with our annotations is 96.4% overall; individual human agreement is 88.6%.

Models

1. *N*-gram (5-gram)
   - English Gigaword (3.07B tokens)
2. LSTM
   - English Wikipedia (83M tokens), trained by Gulordava et al. (2018)
3. Transformer
   - Transformer-XL: Trained on WikiText-103 (103M tokens) by Dai et al. (2019)
   - GPT-2: Trained on WebText (~8B tokens) by Radford et al. (2019)
   - RoBERTa: Trained on Wikipedia, web data, and books (30B tokens) by Liu et al. (2020)*

* results from Salazar et al (2020)

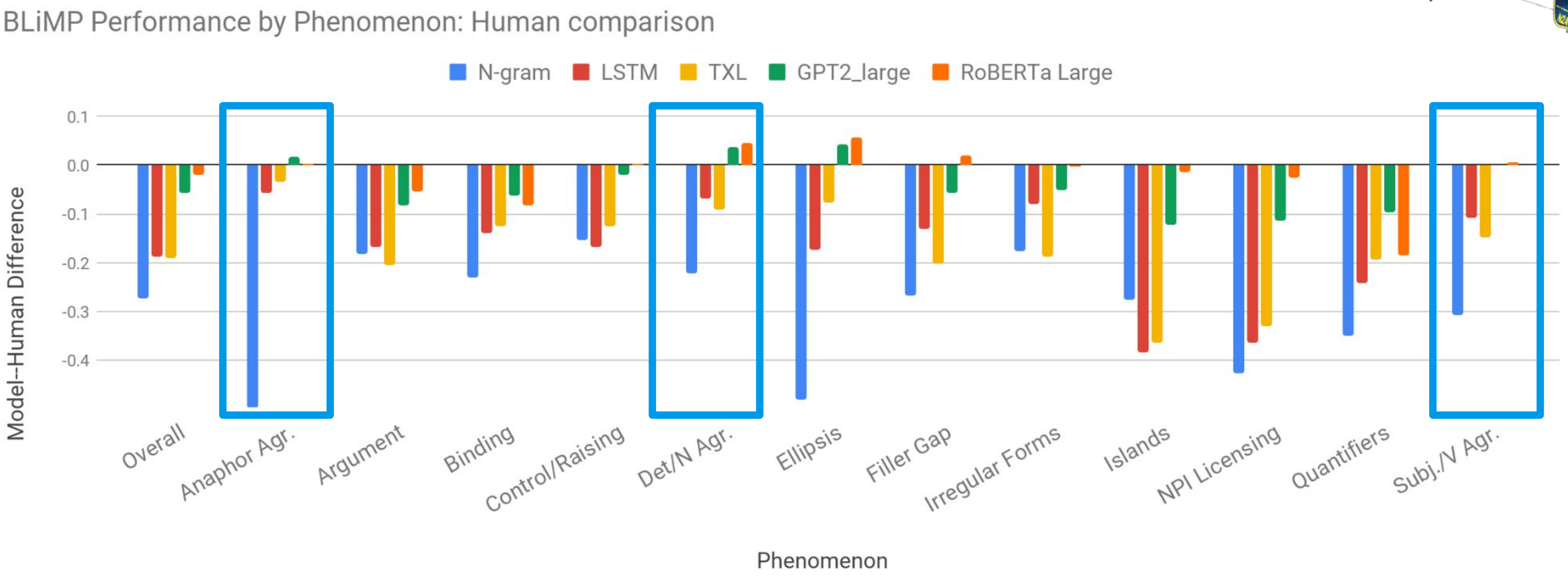
Overall Results



Agreement Results

Phenomenon	N	Acceptable example	Unacceptable example
Anaphor agreement	2	Many girls insulted themselves.	Many girls insulted herself.
Argument structure	9	Rose wasn't disturbing Mark.	Rose wasn't boasting Mark.
Binding	7	It's himself who Robert attacked.	It's himself who attacked Robert.
Control/Raising	5	Kevin isn't irritating to work with.	Kevin isn't bound to work with.
Determiner-N 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	Which bikes is John fixing?	Which is John fixing bikes?
NPI licensing	7	The truck has clearly tipped over.	The truck has ever tipped over.
Quantifiers	4	There was a cat annoying Alice.	There was each cat annoying Alice.
Subject-Verb agr.	6	These casseroles disgust Kayla.	These casseroles disgusts Kayla.

Agreement Results

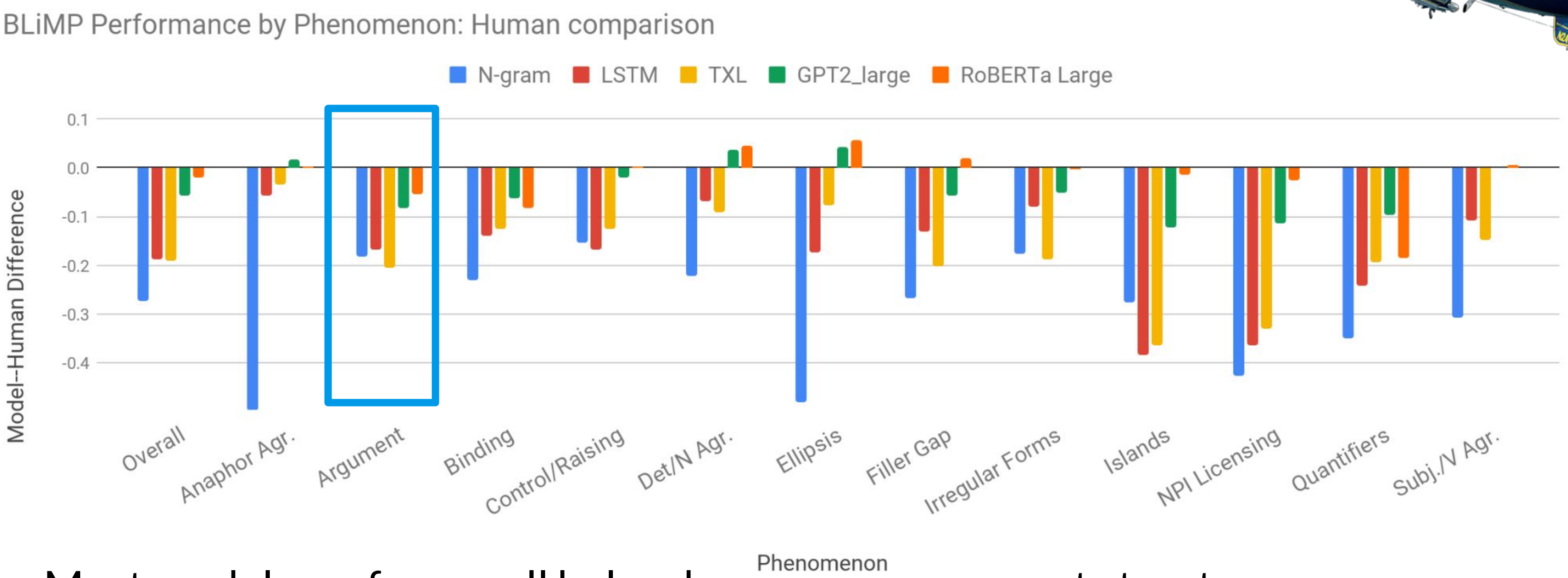


Agreement phenomena tend to show the highest performance across models.

Argument Structure Results

Phenomenon	N	Acceptable example	Unacceptable example
Anaphor agreement	2	Many girls insulted themselves.	Many girls insulted herself.
Argument structure	9	Rose wasn't disturbing Mark.	Rose wasn't boasting Mark.
Binding	7	It's himself who Robert attacked.	It's himself who attacked Robert.
Control/Raising	5	Kevin isn't irritating to work with.	Kevin isn't bound to work with.
Determiner-N 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	Which bikes is John fixing?	Which is John fixing bikes?
NPI licensing	7	The truck has clearly tipped over.	The truck has ever tipped over.
Quantifiers	4	There was a cat annoying Alice.	There was each cat annoying Alice.
Subject-Verb agr.	6	These casseroles disgust Kayla.	These casseroles disgusts Kayla.

Argument Structure Results



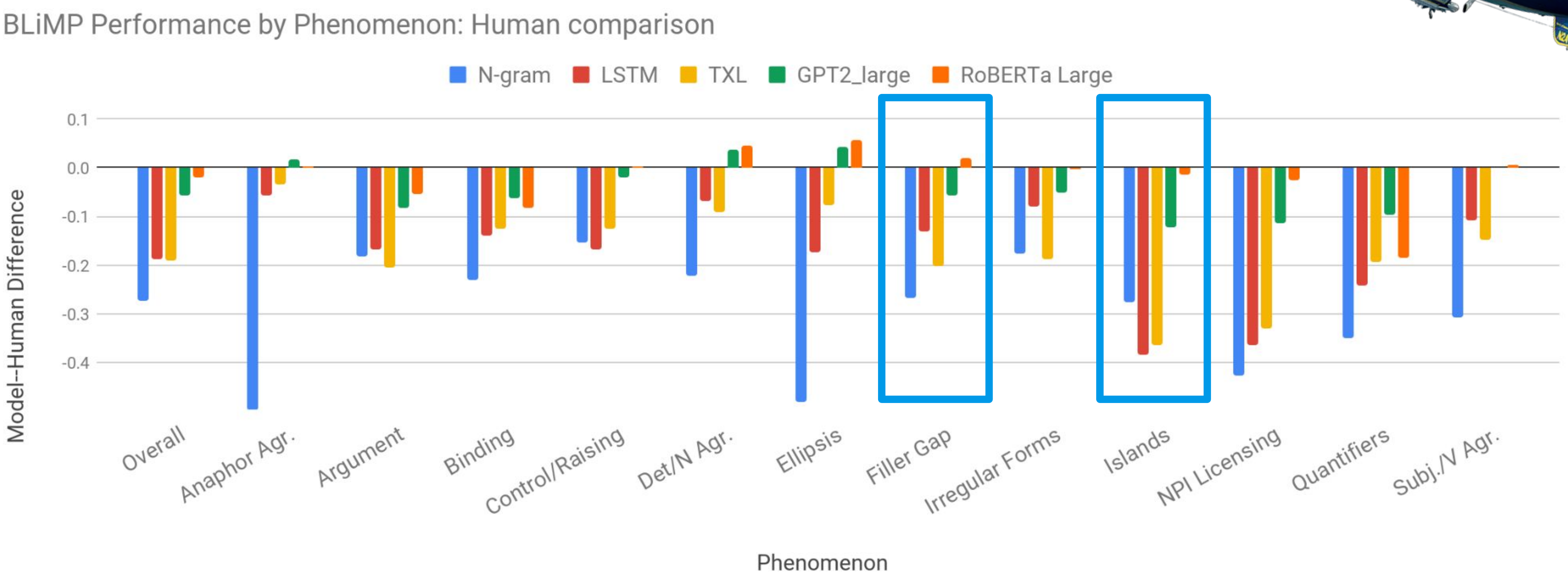
Most models perform well below humans on argument structure.

Even GPT-2 is **not much better than the *n*-gram LM.**

Filler-Gap Dependency Results

Phenomenon	N	Acceptable example	Unacceptable example
Anaphor agreement	2	Many girls insulted themselves.	Many girls insulted herself.
Argument structure	9	Rose wasn't disturbing Mark.	Rose wasn't boasting Mark.
Binding	7	It's himself who Robert attacked.	It's himself who attacked Robert.
Control/Raising	5	Kevin isn't irritating to work with.	Kevin isn't bound to work with.
Determiner-N 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	Which bikes is John fixing?	Which is John fixing bikes?
NPI licensing	7	The truck has clearly tipped over.	The truck has ever tipped over.
Quantifiers	4	There was a cat annoying Alice.	There was each cat annoying Alice.
Subject-Verb agr.	6	These casseroles disgust Kayla.	These casseroles disgusts Kayla.

Filler-Gap Dependency Results

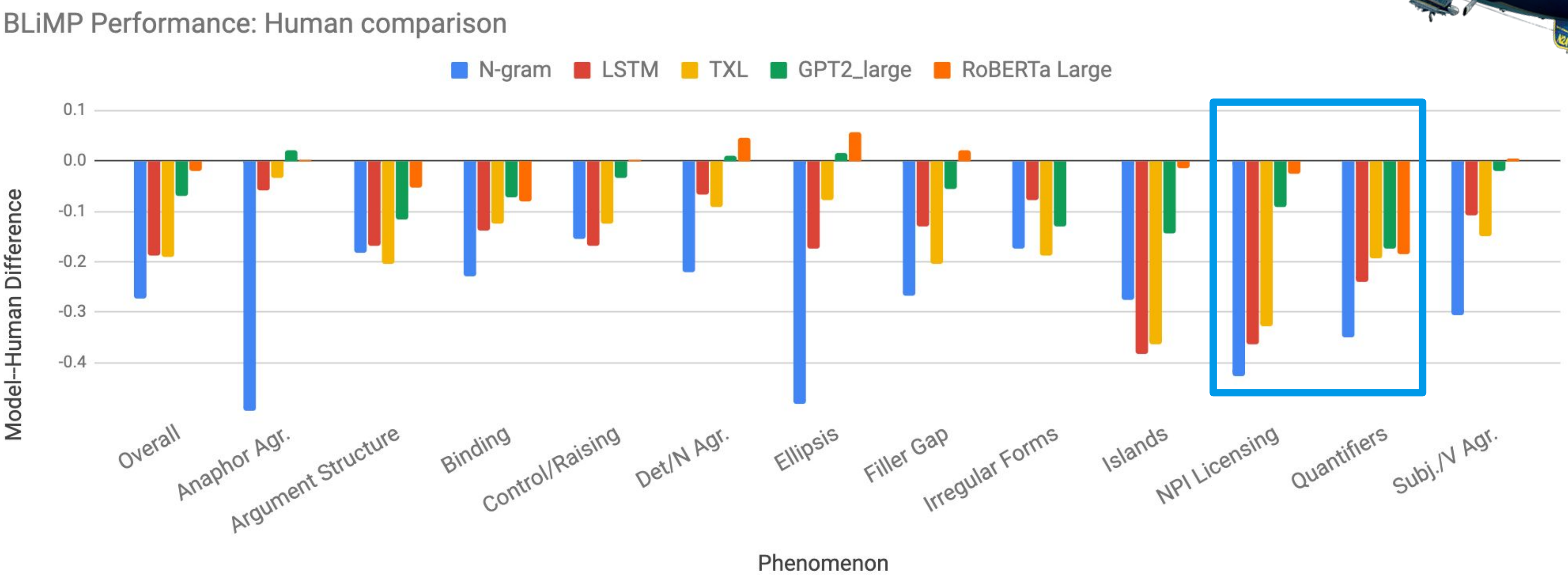


Wh-phenomena are not hard in general, but island effects are hard for most neural models.

Quantifiers and NPIs results

Phenomenon	N	Acceptable example	Unacceptable example
Anaphor agreement	2	Many girls insulted themselves.	Many girls insulted herself.
Argument structure	9	Rose wasn't disturbing Mark.	Rose wasn't boasting Mark.
Binding	7	It's himself who Robert attacked.	It's himself who attacked Robert.
Control/Raising	5	Kevin isn't irritating to work with.	Kevin isn't bound to work with.
Determiner-N 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	Which bikes is John fixing?	Which is John fixing bikes?
NPI licensing	7	The truck has clearly tipped over.	The truck has ever tipped over.
Quantifiers	4	There was a cat annoying Alice.	There was each cat annoying Alice.
Subject-Verb agr.	6	These casseroles disgust Kayla.	These casseroles disgusts Kayla.

Quantifiers and NPIs results



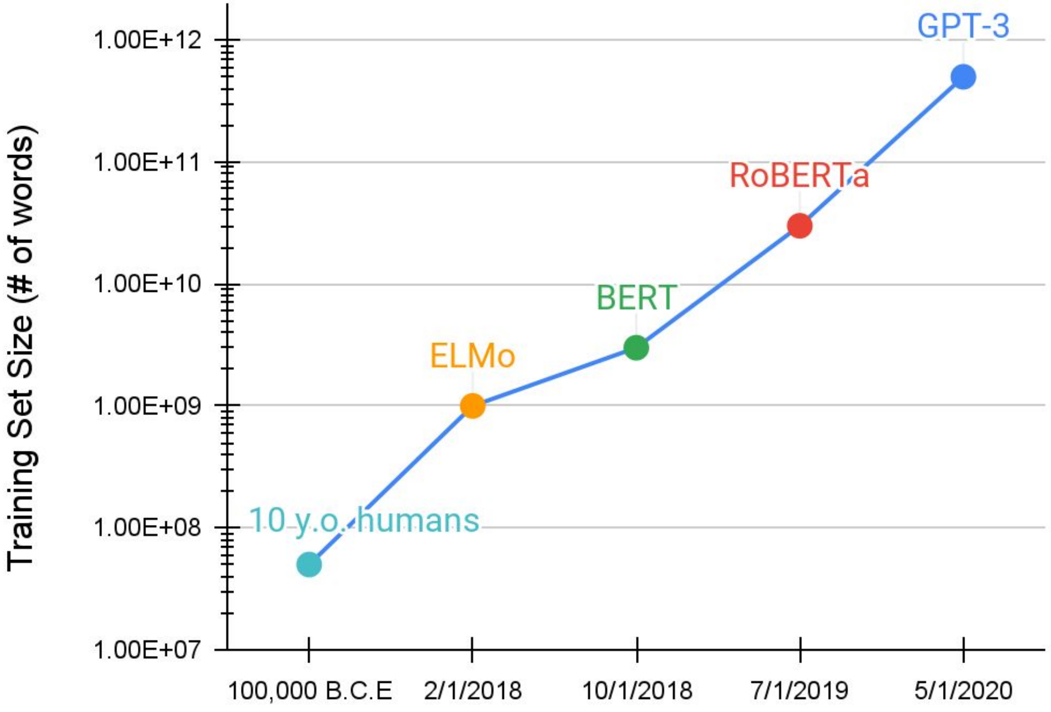
Semantic restrictions on quantifiers and NPIs are challenging for most models.
Quantifier distributions are the hardest phenomenon for RoBERTa

Part 2
More human like learning environments

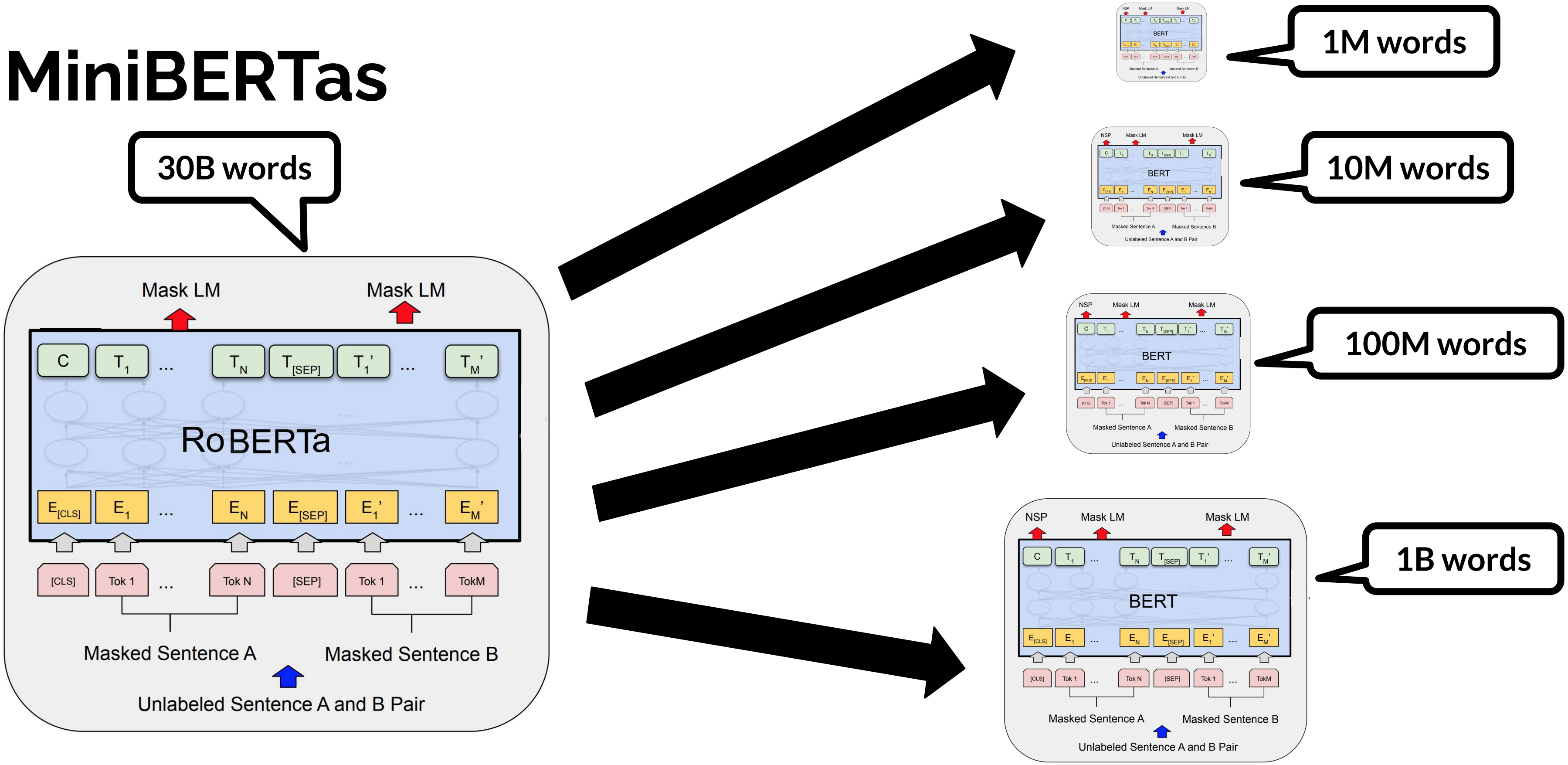
Near-human results on BLiMP from RoBERTa are impressive.

But how does RoBERTa's learning environment compare to humans'?

Growth in LM Training Sets (2018-2020)



MiniBERTas



30B words
Mask LM
Mask LM
C
T_1
...
T_N
$T_{[SEP]}$
T_1'
...
T_M'
RoBERTa
$E_{[CLS]}$
E_1
...
E_N
$E_{[SEP]}$
E_1'
...
E_M'
[CLS]
Tok 1
...
Tok N
[SEP]
Tok 1
...
TokM
Masked Sentence A
Masked Sentence B
Unlabeled Sentence A and B Pair
1M words
10M words
100M words
1B words
NSP
Mask LM
Mask LM
BERT
Masked Sentence A
Masked Sentence B
Unlabeled Sentence A and B Pair

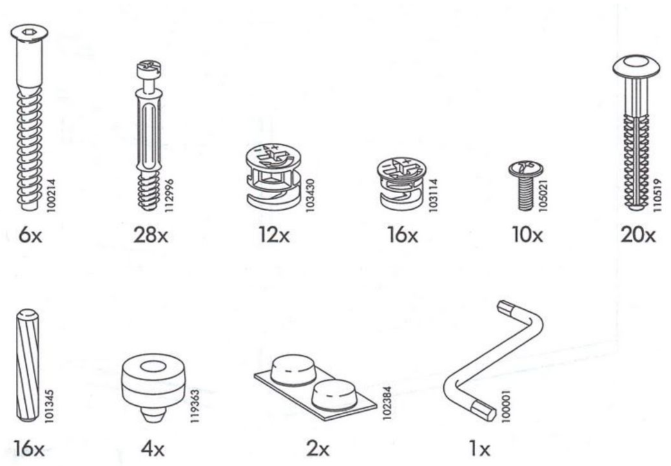
Training

- 1M, 10M, 100M, 1B words of training data
- 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 size, we train >= 10 models & select 3 with best PPL.

The 12 MiniBERTas on Transformers

https://huggingface.co/nyu-mll

Probing for features





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
{yian.zhang, warstadt, xl3119, bowman}@nyu.edu

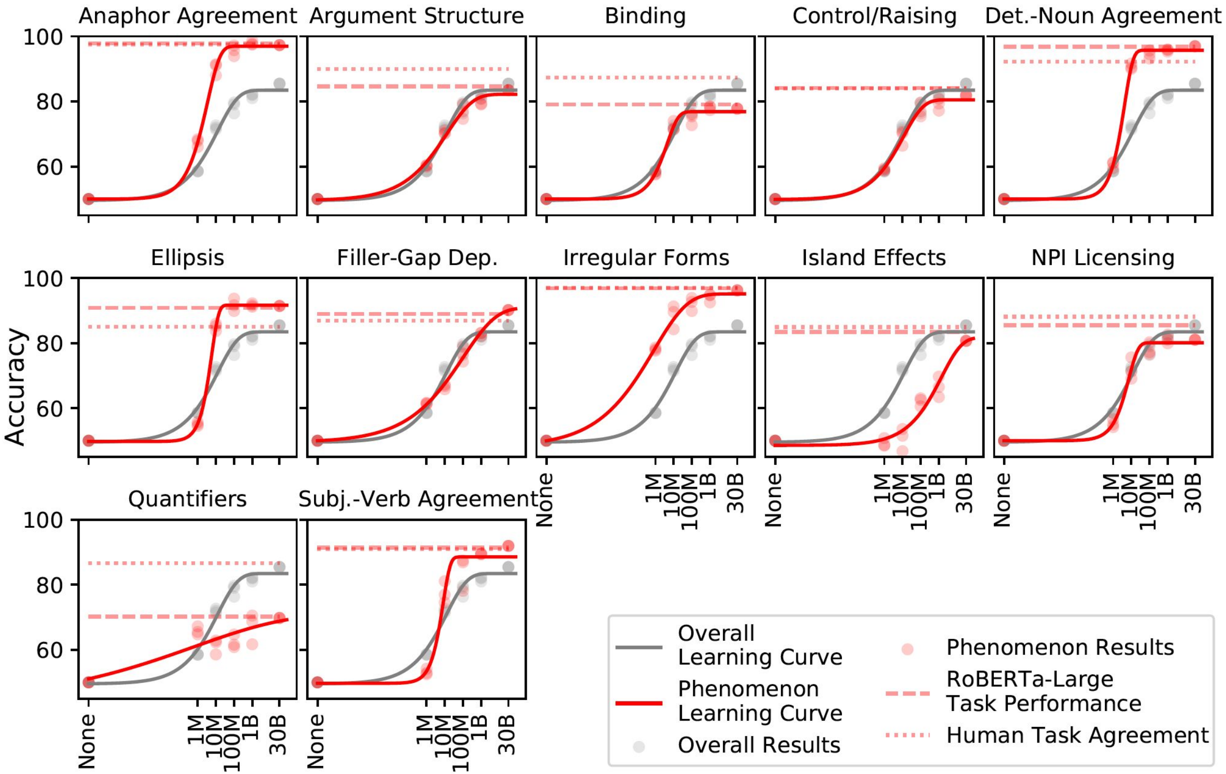
Five Sets of Probing Methods

1. "Standard" classifier probing
2. "Information theoretic" probing
3. Unsupervised acceptability judgments
4. Unsupervised commonsense knowledge test
5. Fine-tuning on downstream NLU tasks

Five Sets of Probing Methods

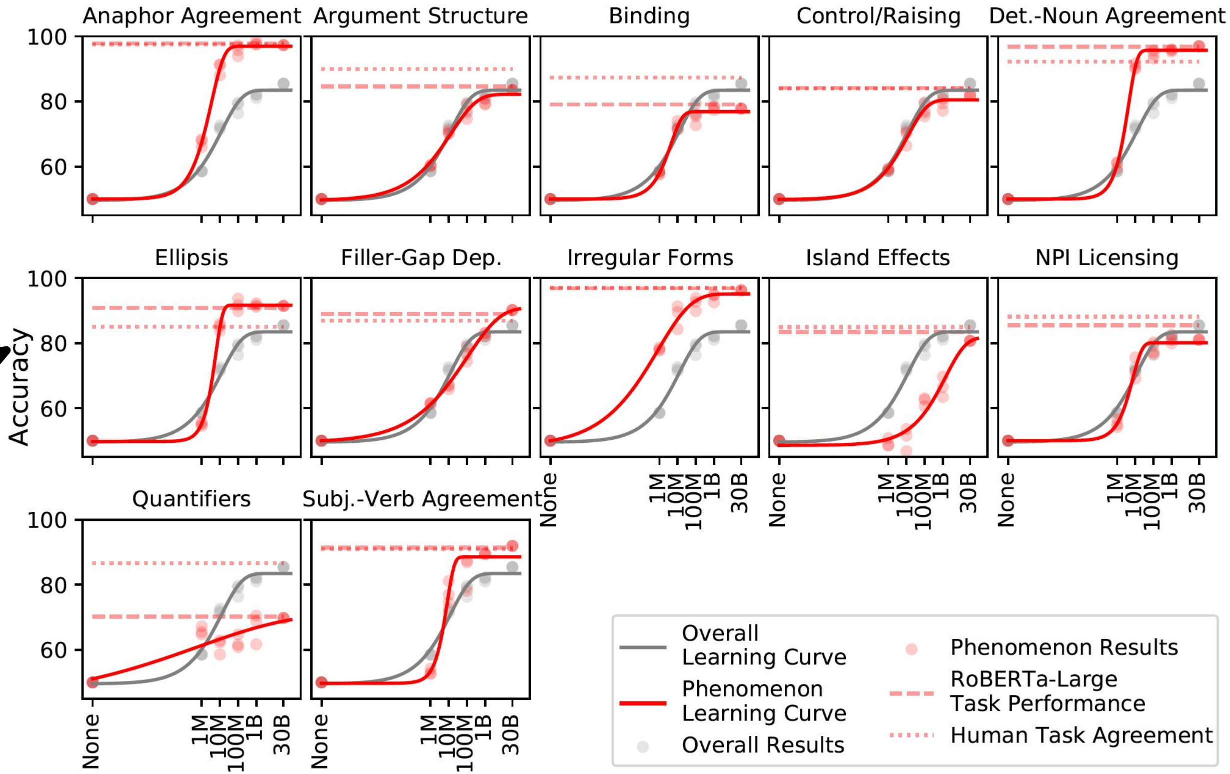
1. “Standard” classifier probing
2. “Information theoretic” probing
3. Unsupervised acceptability judgments
4. Unsupervised commonsense knowledge test
5. Fine-tuning on downstream NLU tasks

3. BLiMP: Unsupervised Acceptability Judgments



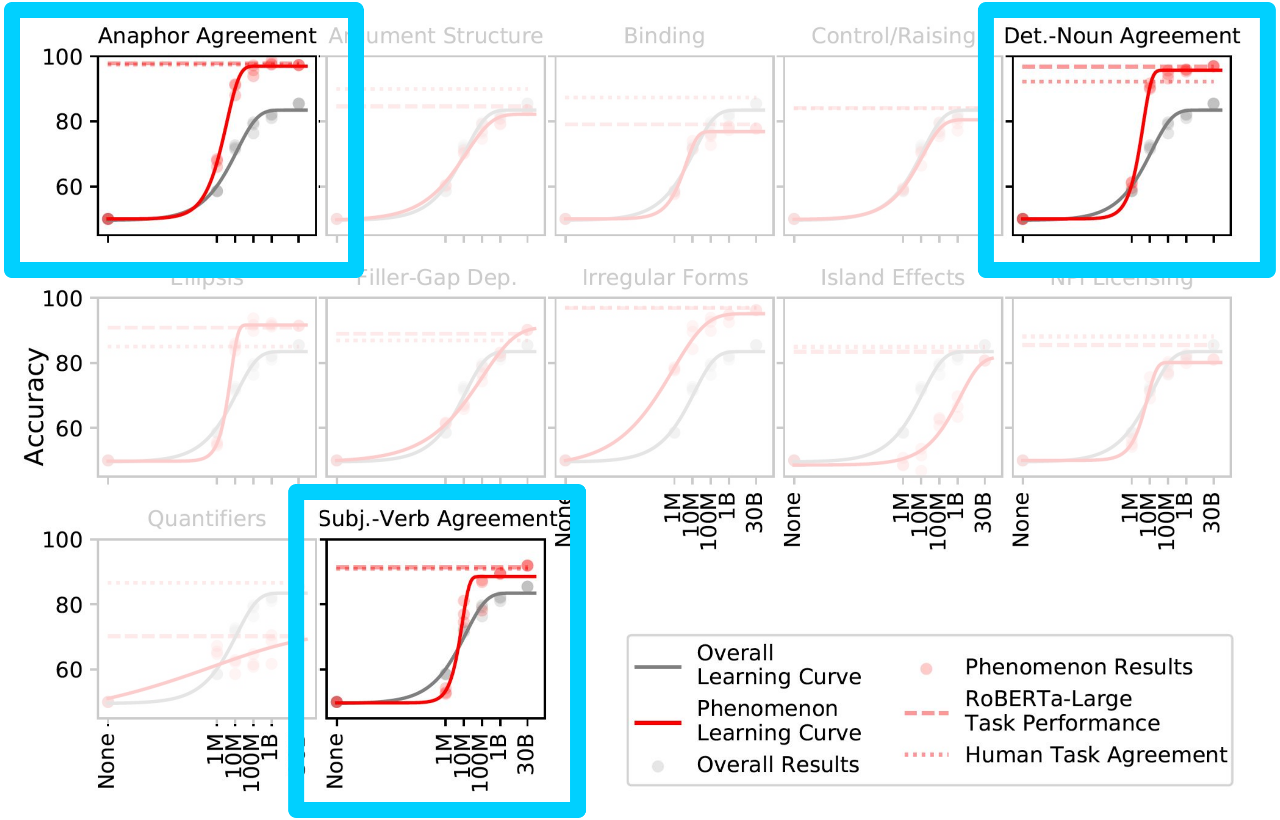
3. BLiMP: Unsupervised Acceptability Judgments

Overall grammatical knowledge increases mainly between 1M and 100M words.



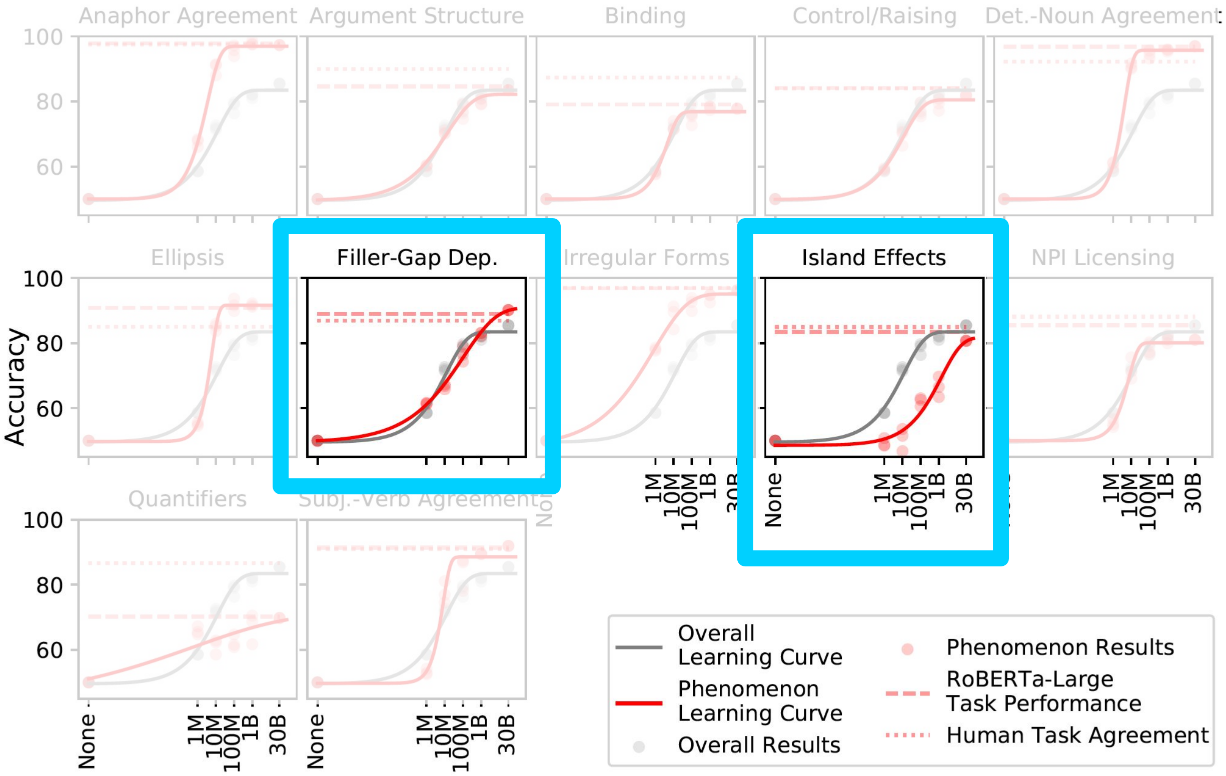
3. BLiMP: Unsupervised Acceptability Judgments

Agreement phenomena are learned with only ~10M words (and often with very high accuracy)

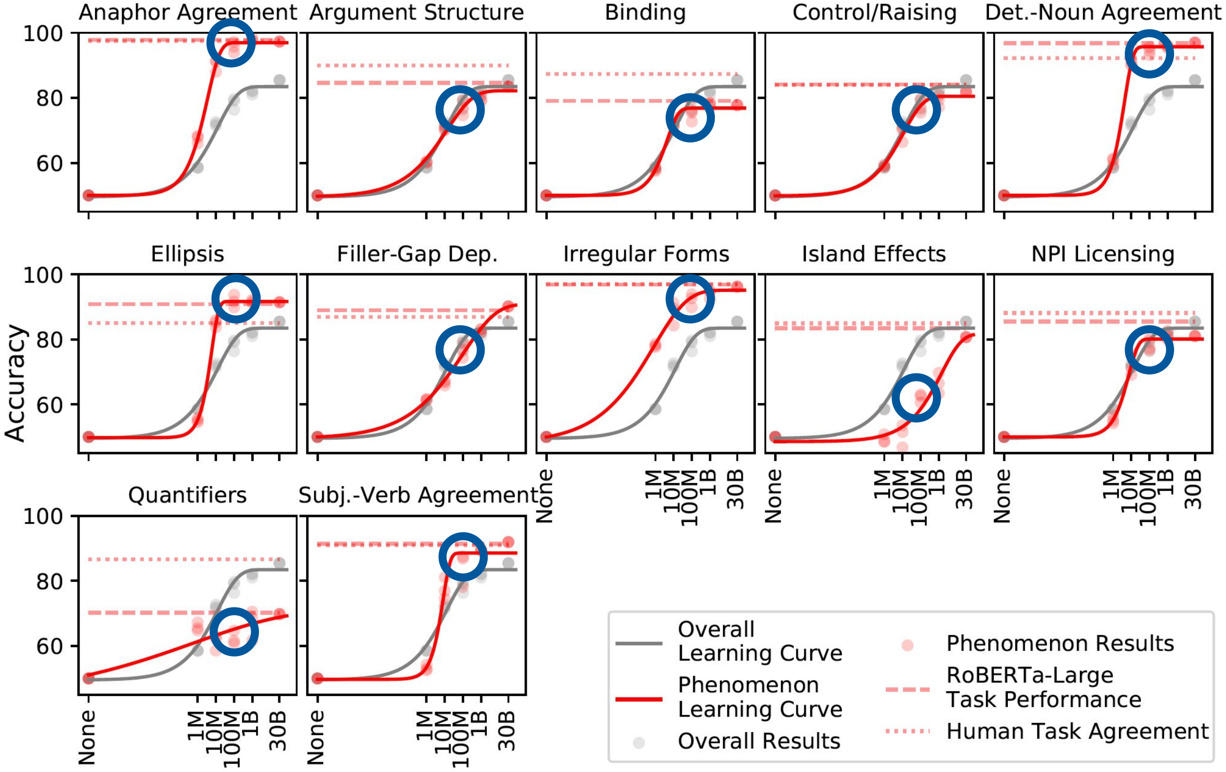


3. BLiMP: Unsupervised Acceptability Judgments

Long-distance wh-dependencies are are still improving with >1B words.

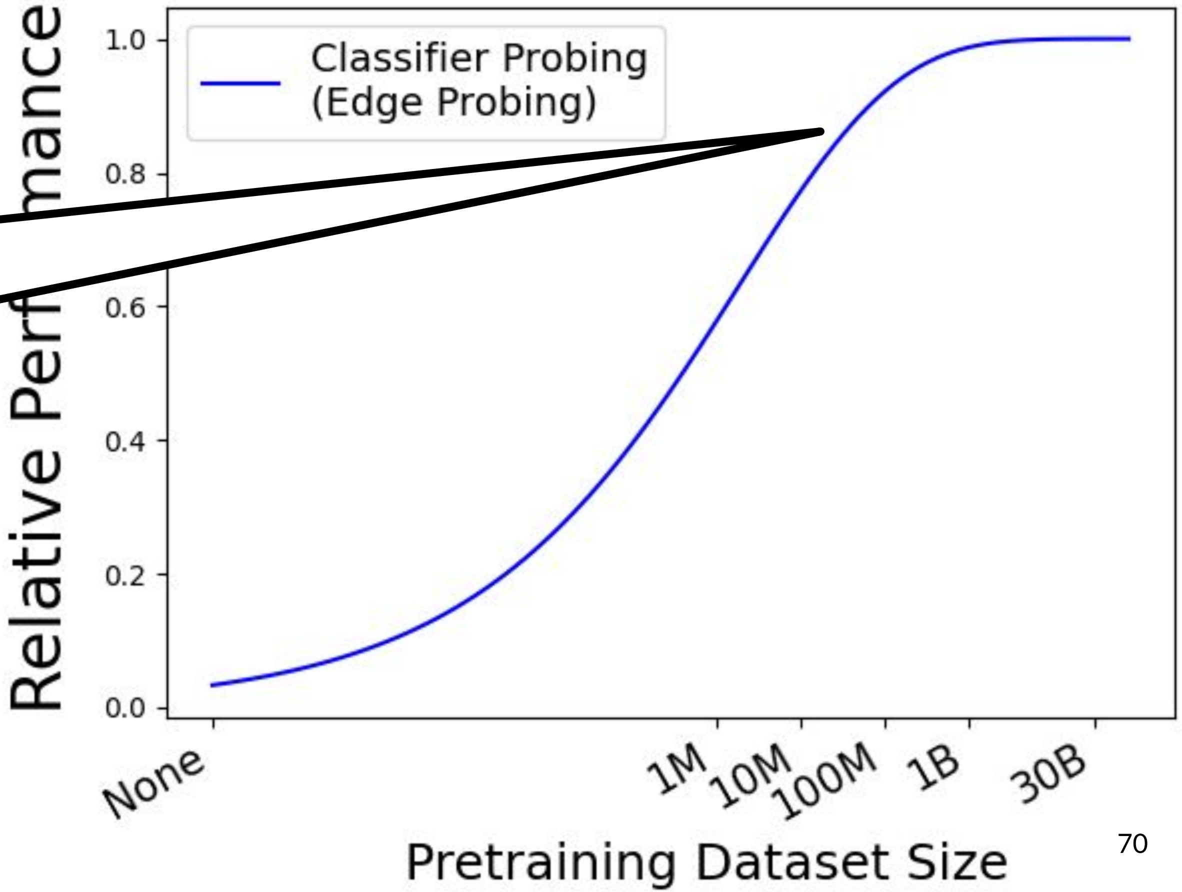


3. BLiMP: Unsupervised Acceptability Judgments



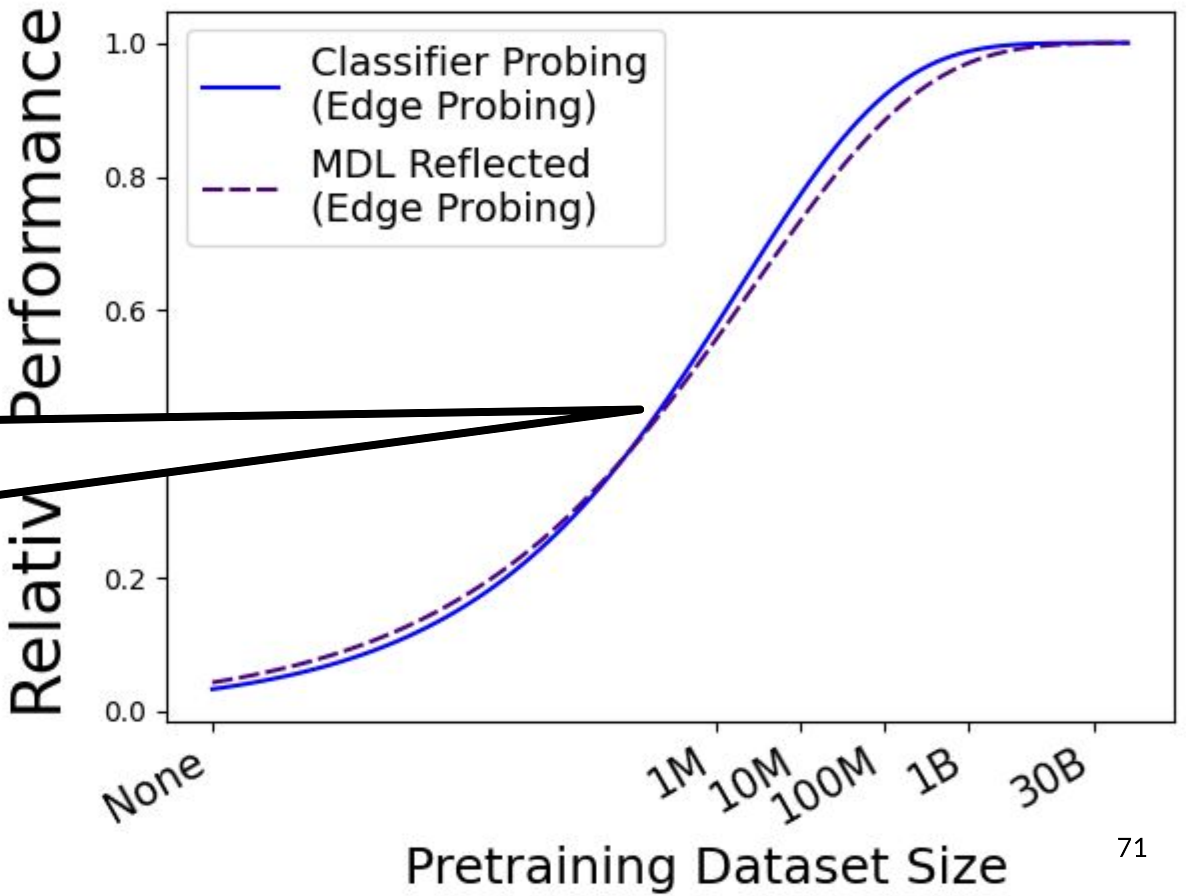
Overall Comparison

Core NLP features are learned with 10M-100M words.



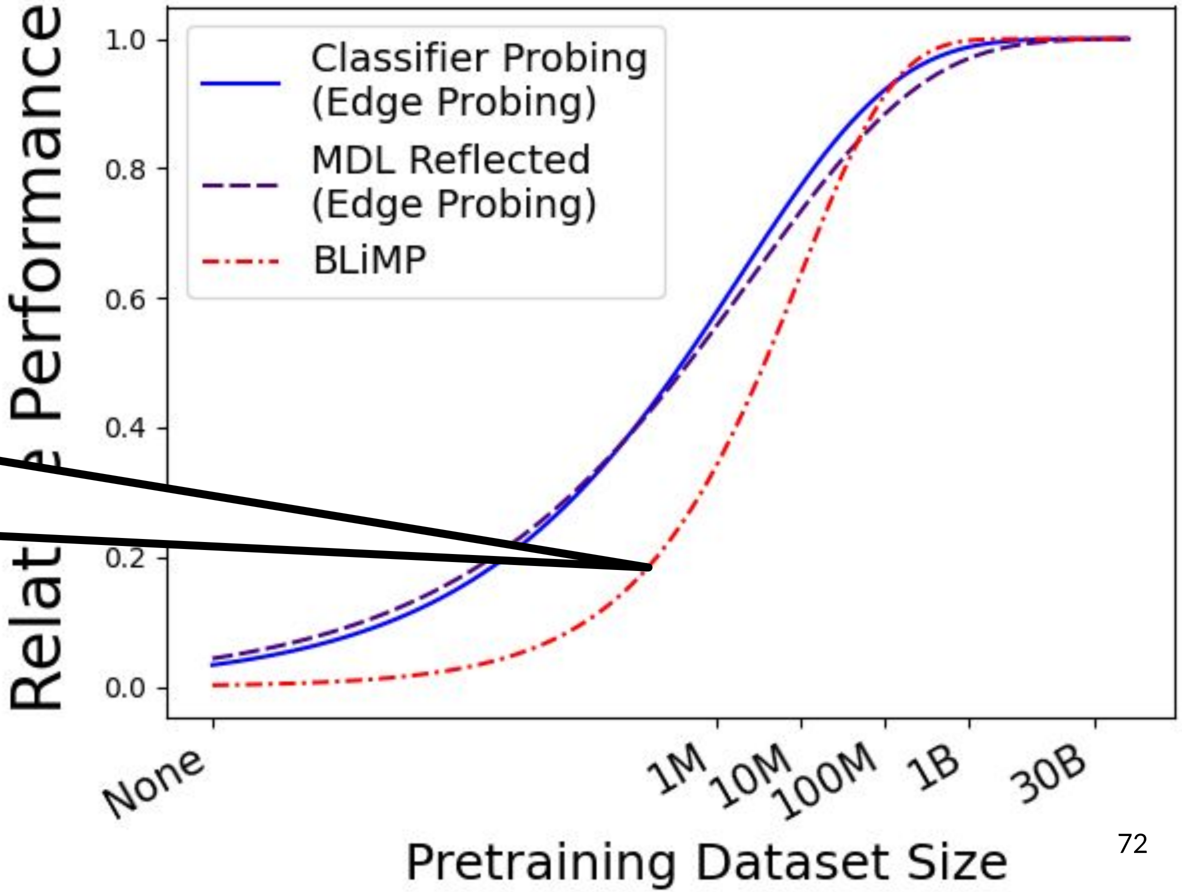
Overall Comparison

Information theoretic probing looks the same as “standard” classifier probing.



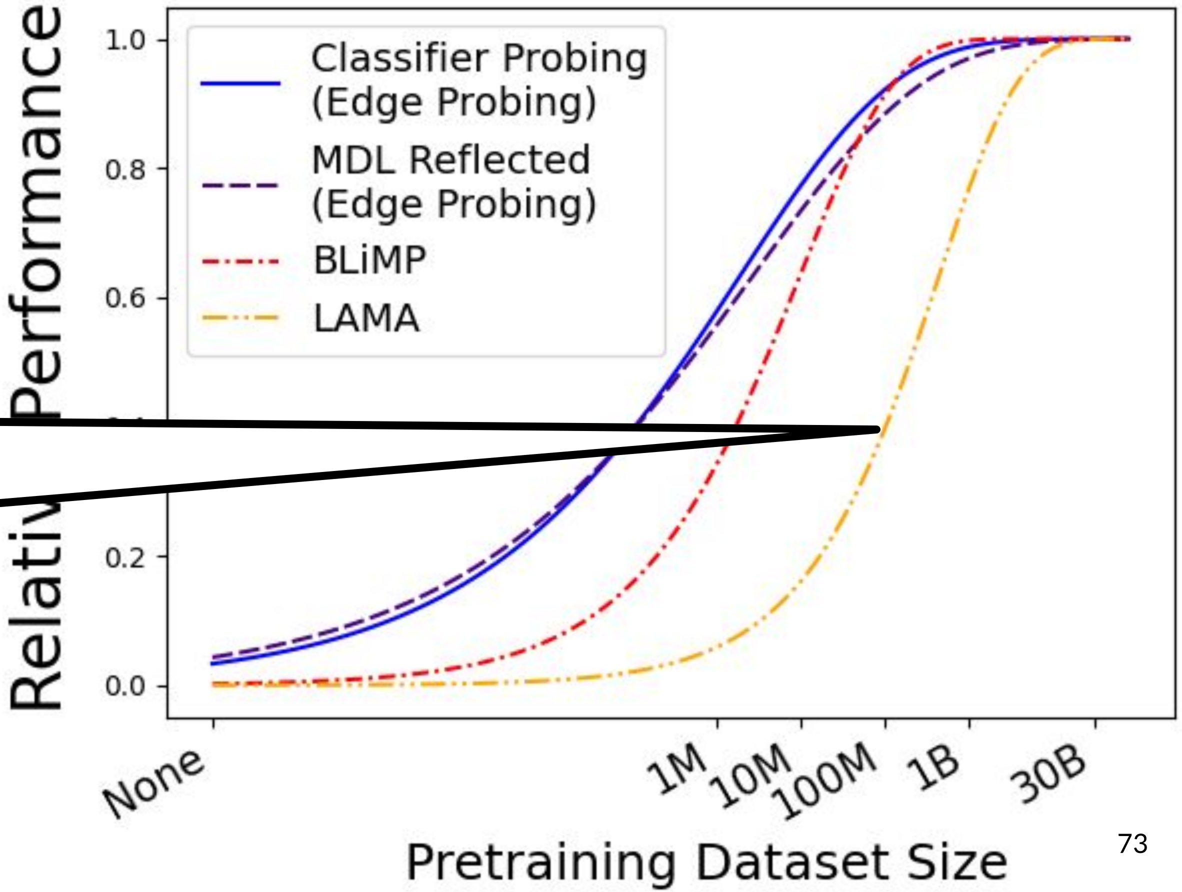
Overall Comparison

Grammaticality knowledge requires more data to start acquiring.



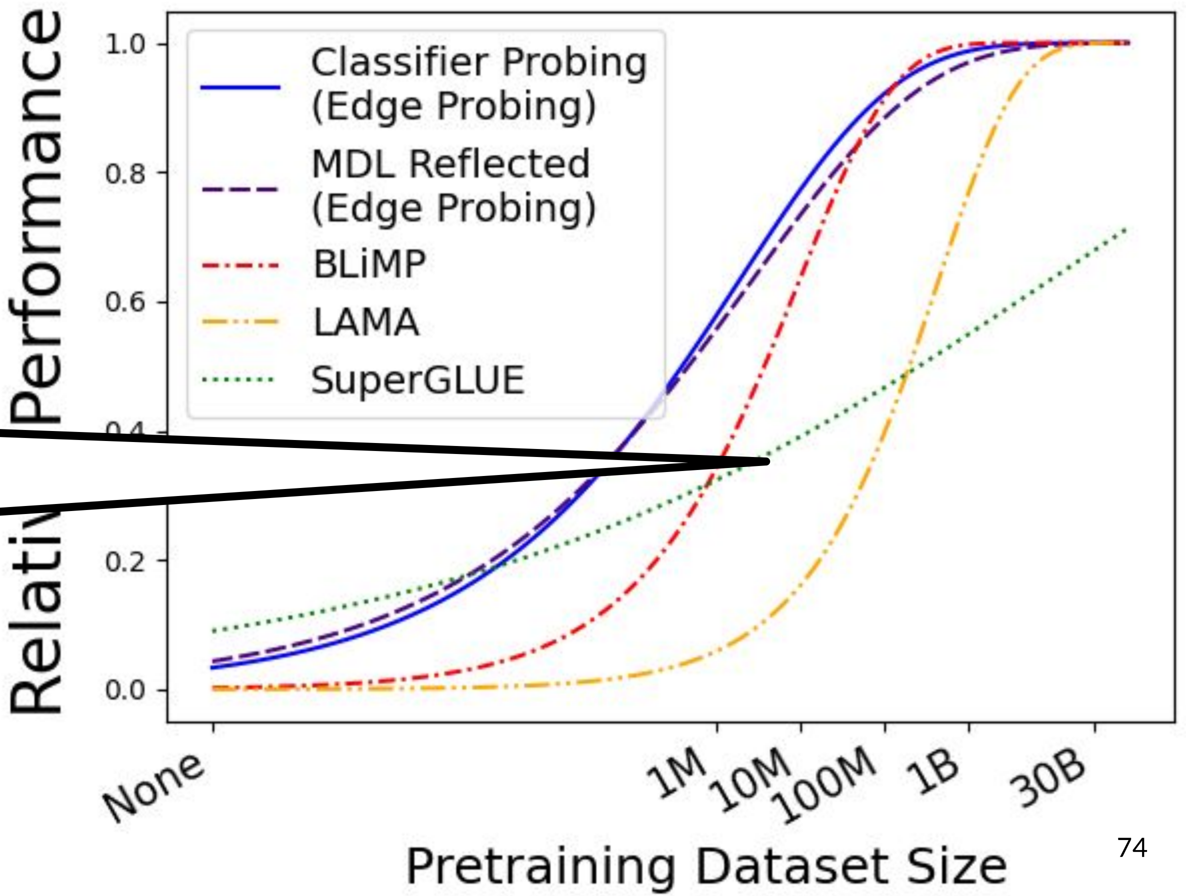
Overall Comparison

World knowledge/ commonsense reasoning requires ~1B words.

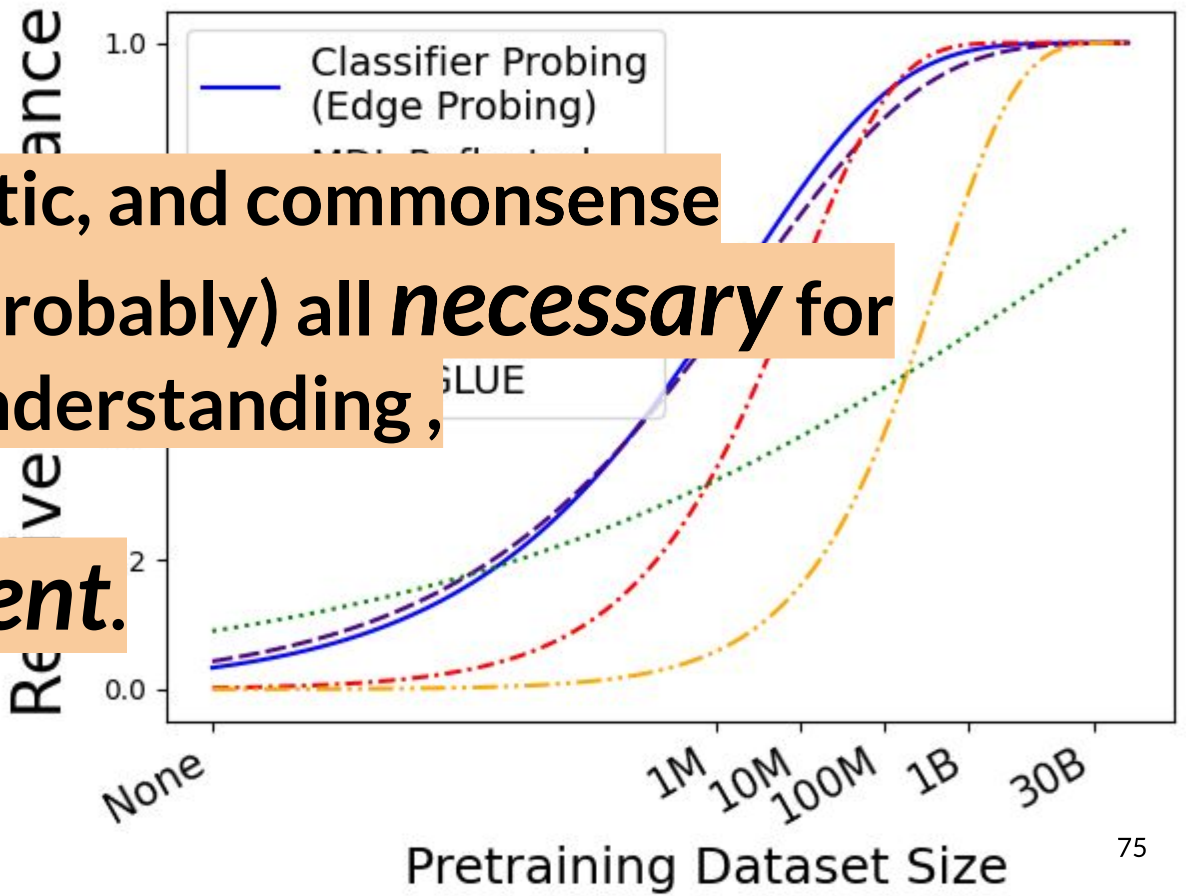


Overall Comparison

Strong performance on downstream tasks requires billions of words.



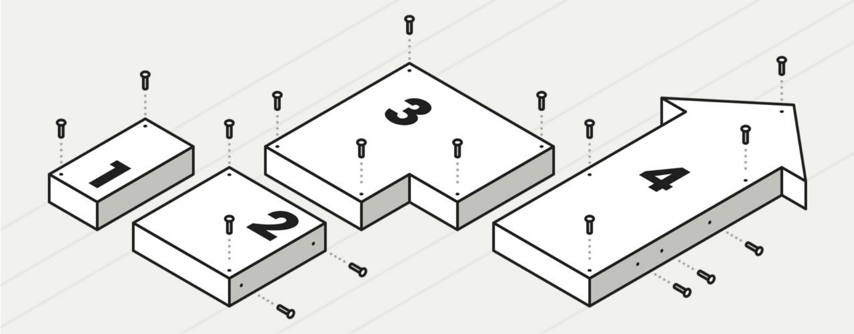
Overall Comparison



Syntactic, semantic, and commonsense knowledge are (probably) all ***necessary*** for good language understanding ,

...but not ***sufficient***.

Acquiring Inductive Bias






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

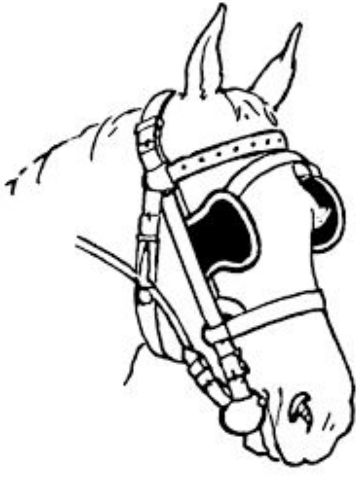
Alex Warstadt,[1] **Yian Zhang,**[2] **Haau-Sing Li,**[3] **Haokun Liu,**[3] **Samuel R. Bowman**[1,2,3]
[1]Dept. of Linguistics, [2]Dept. of Computer Science, [3]Center for Data Science
New York University
Correspondence: warstadt@nyu.edu

Feature learning isn’t everything.

Feature learning isn't everything.

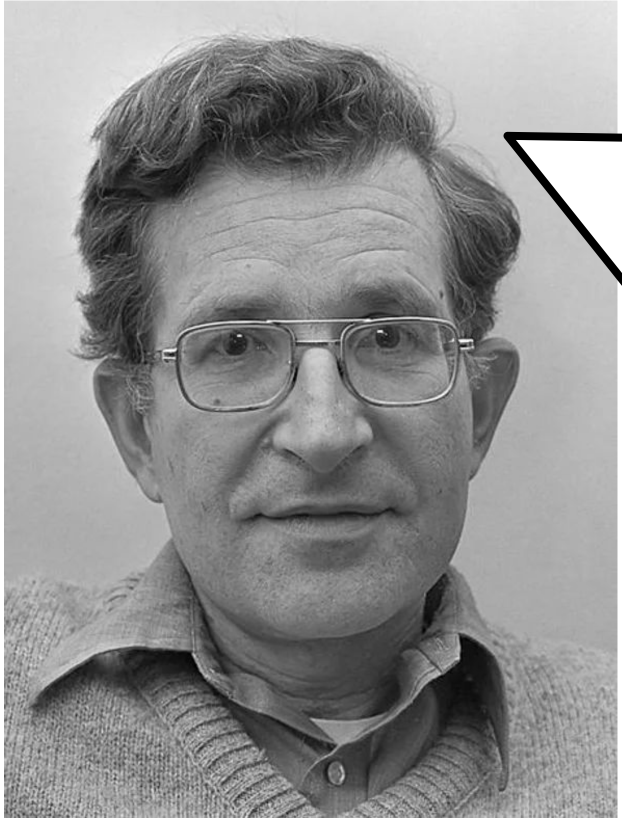
...You have to know how/when to use 'em.

Learning Inductive Biases



Inductive biases limit the learner's hypothesis space.

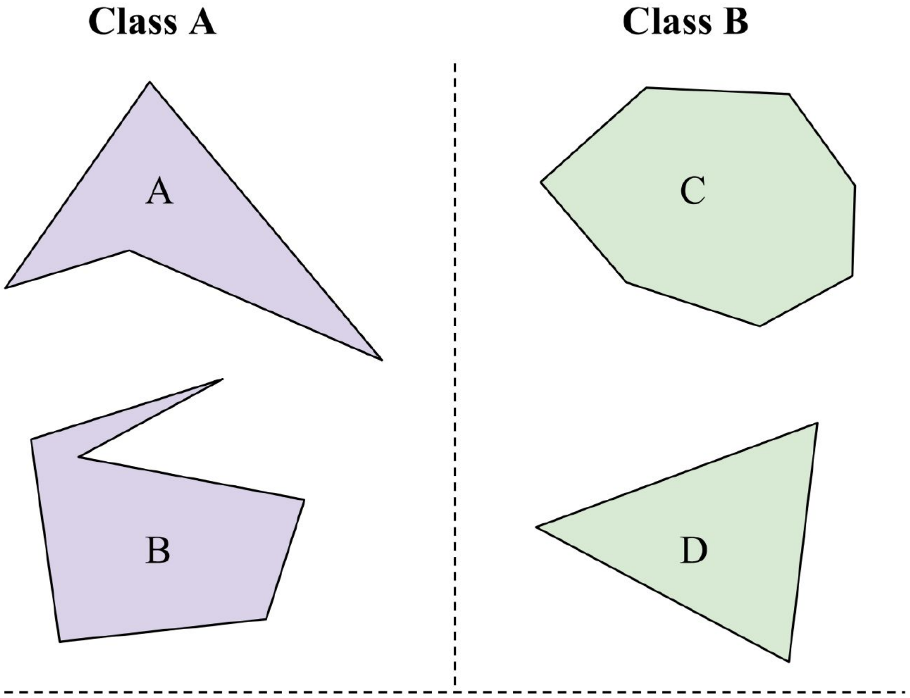
Language model pretraining "*induces a hypothesis space H that should be useful for many other NLP tasks*" (Howard & Ruder, 2018)



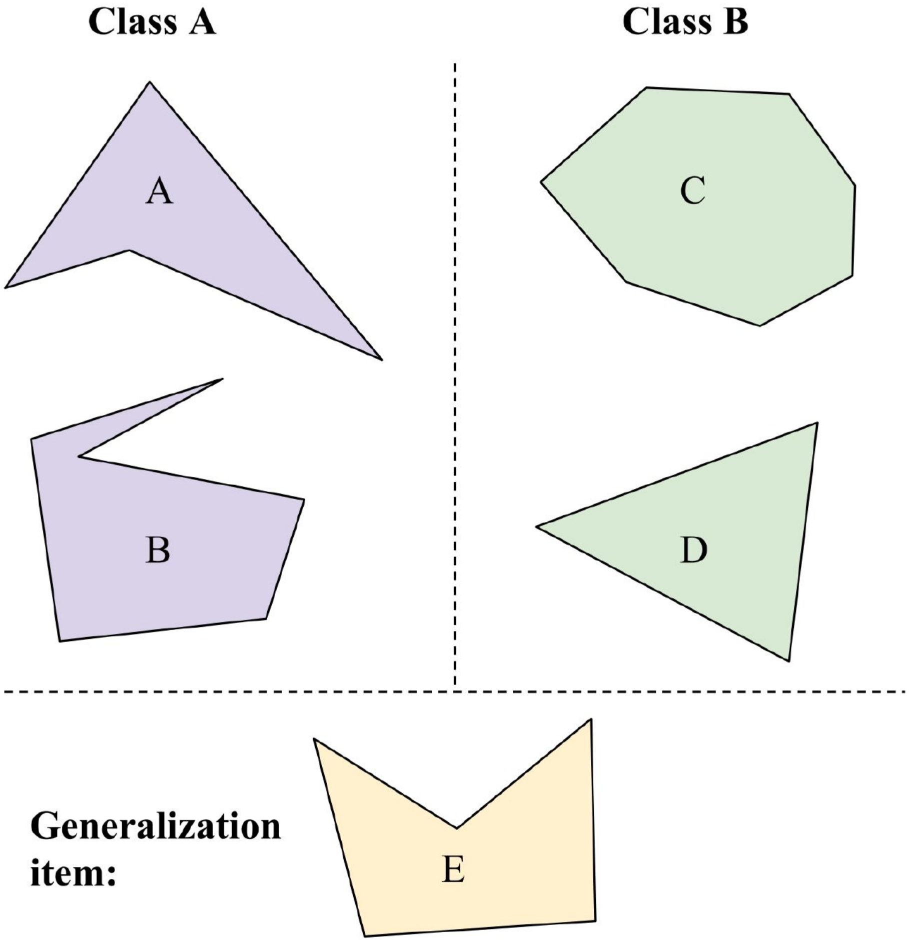
[I]t is possible [in human language] to formulate a transformation [...] independently of what the length or internal complexity of the strings belonging to these categories may be. It is impossible, however, to formulate as a transformation such a simple operation as reflection of an arbitrary string [...], or interchange of the $(2n - i)^{th}$ word with the $2n^{th}$ word throughout a string of arbitrary length [...].

Noam Chomsky, 1957. *Syntactic Structures*.

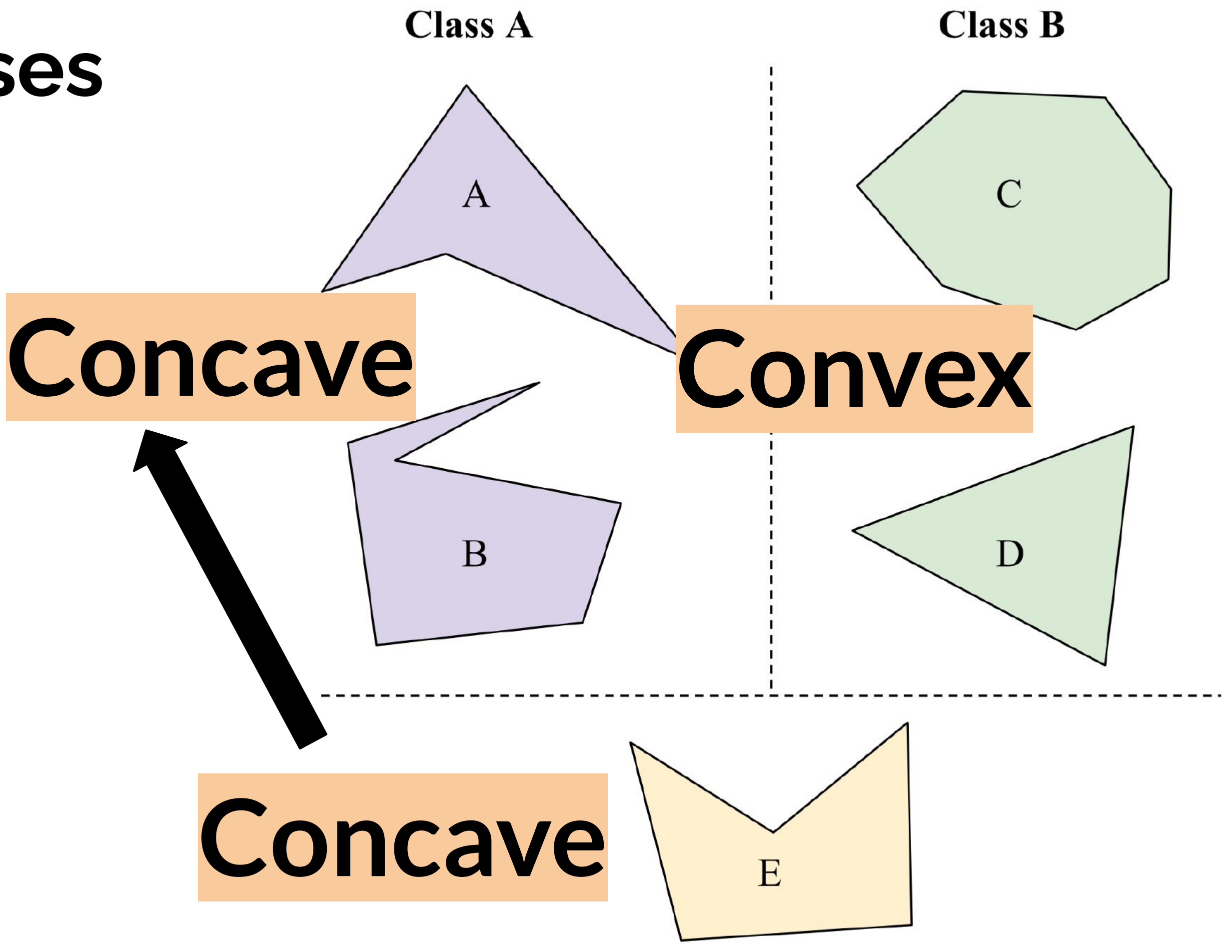
Inductive biases



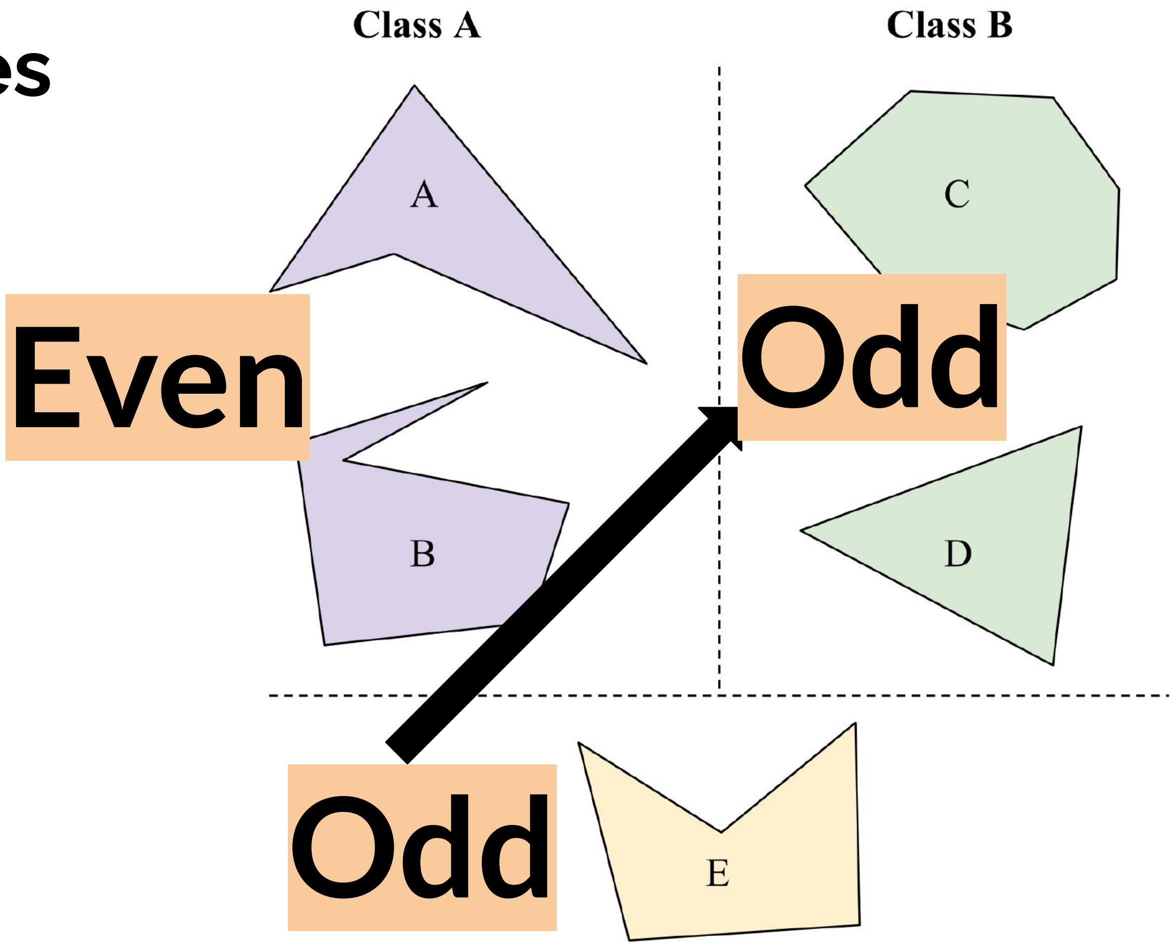
Inductive biases



Inductive biases



Inductive biases



Representing *F* ≠ Using *F*

Our questions

1. Can a preference for linguistic features over surface features be acquired with sufficient data?

Our questions

1. Can a preference for linguistic features over surface features be acquired with sufficient data?
2. How do feature preferences change as the volume of pretraining data increases?

Our questions

1. Can a preference for linguistic features over surface features be acquired with sufficient data?
2. How do feature preferences change as the volume of pretraining data increases?
3. How does the acquisition of feature preferences differ from the acquisition of (mere) feature representations.

Ambiguous Experiments

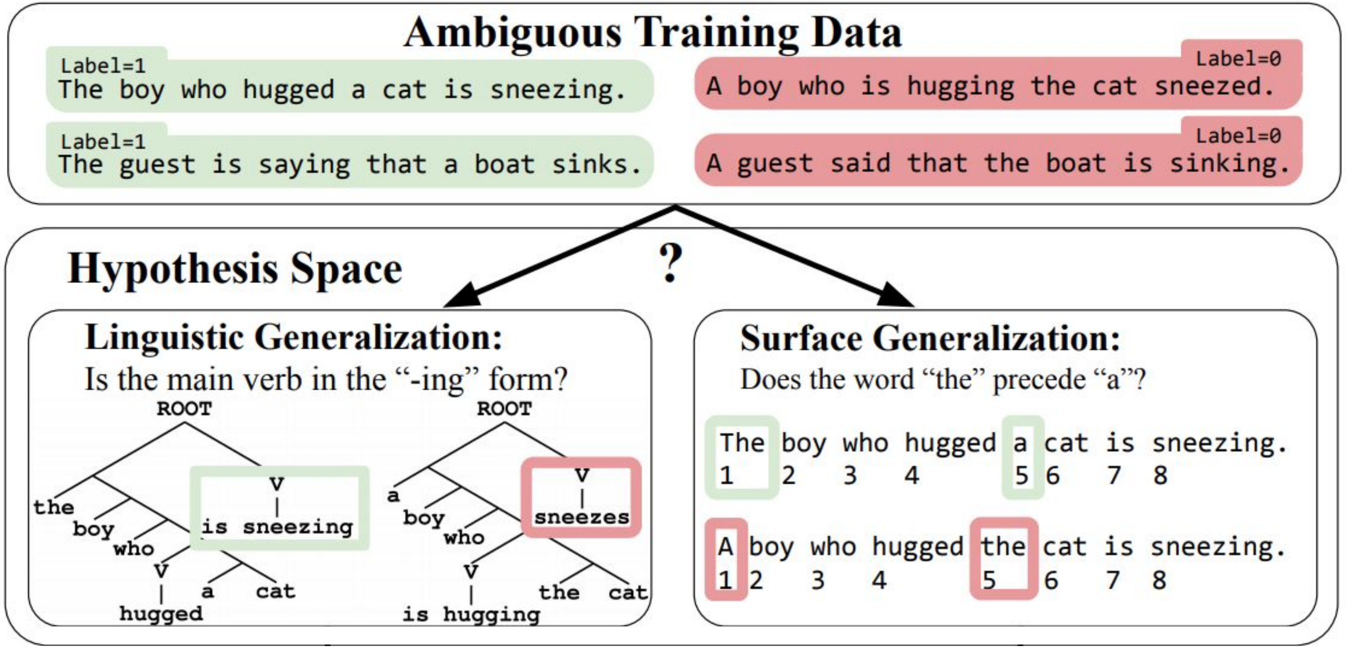
*Does model X **ever** prefer linguistic feature A or surface feature B?*

Ambiguous Experiments

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

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

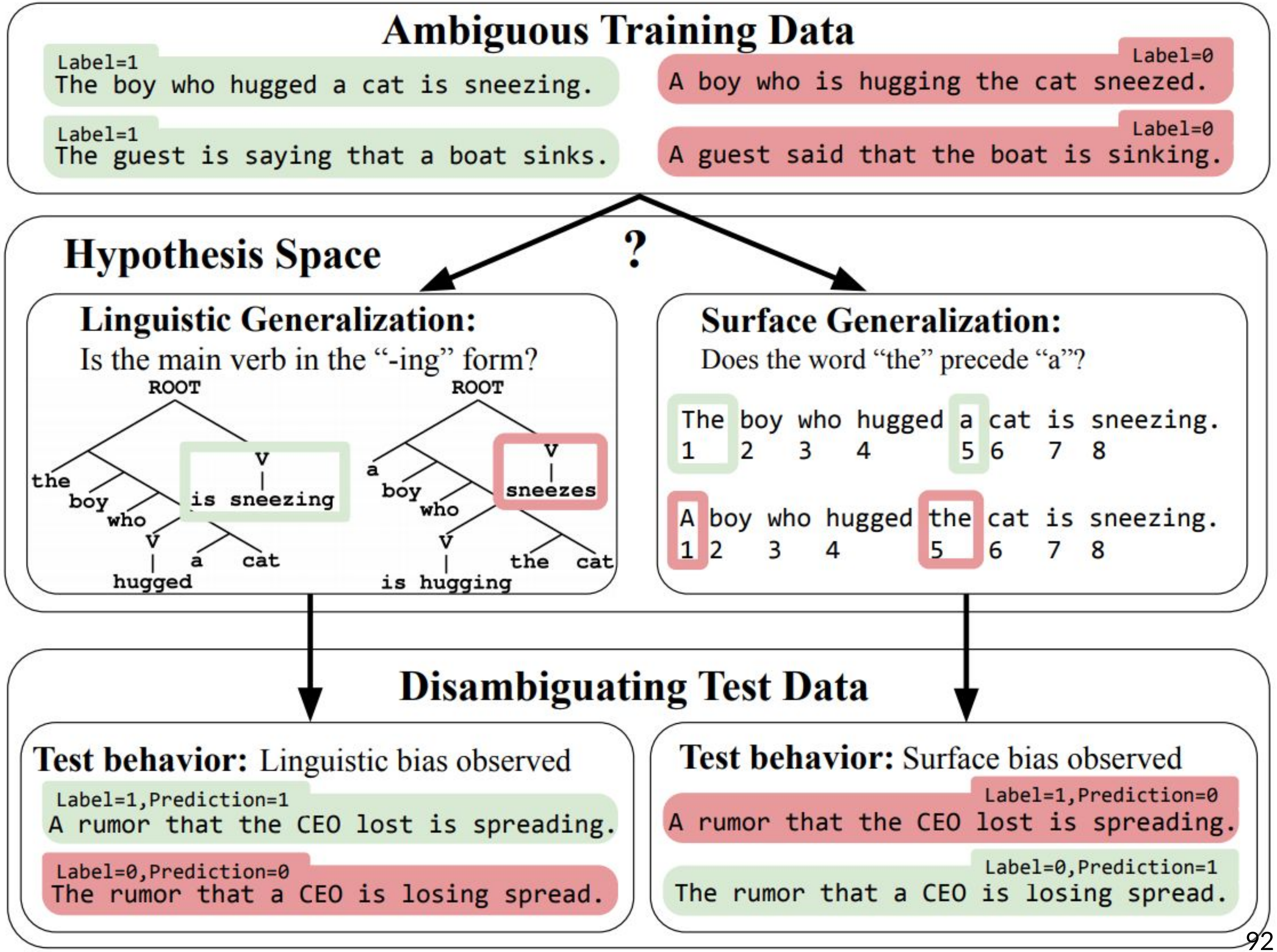
Poverty of the Stimulus Design



Example from the **SYNTACTIC POSITION × RELATIVE (LINEAR) POSITION** task

Poverty of the Stimulus Design

Example from the SYNTACTIC POSITION × RELATIVE (LINEAR) POSITION task



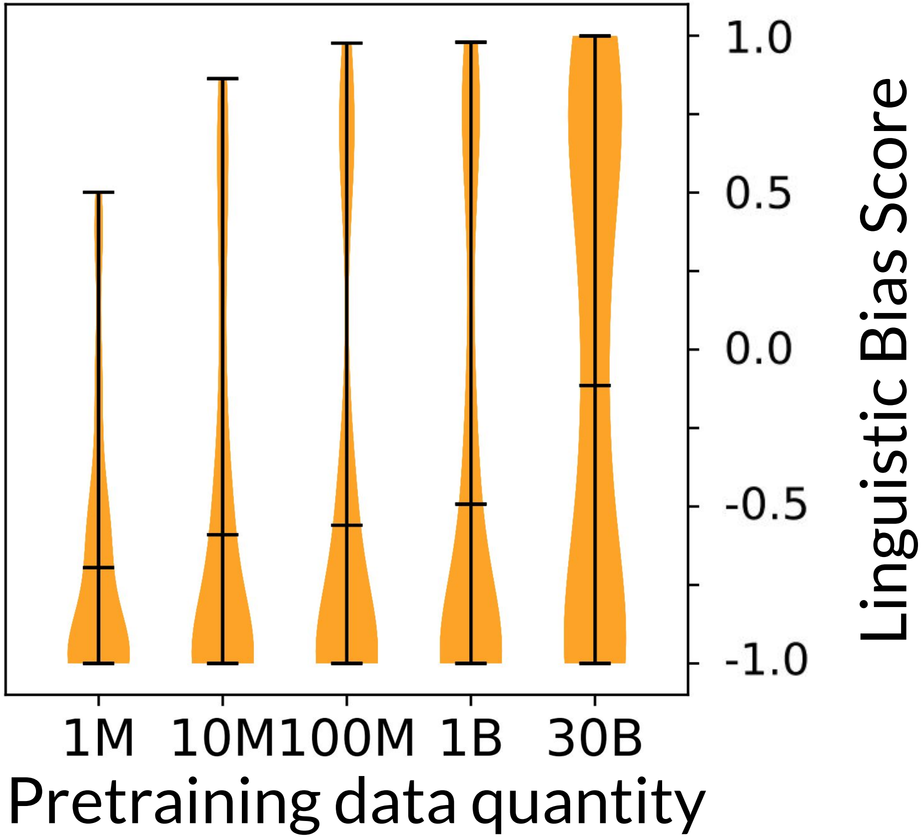
Surface vs. Linguistic Features

	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.

5 surface × 4 linguistic features = 20 ambiguous tasks

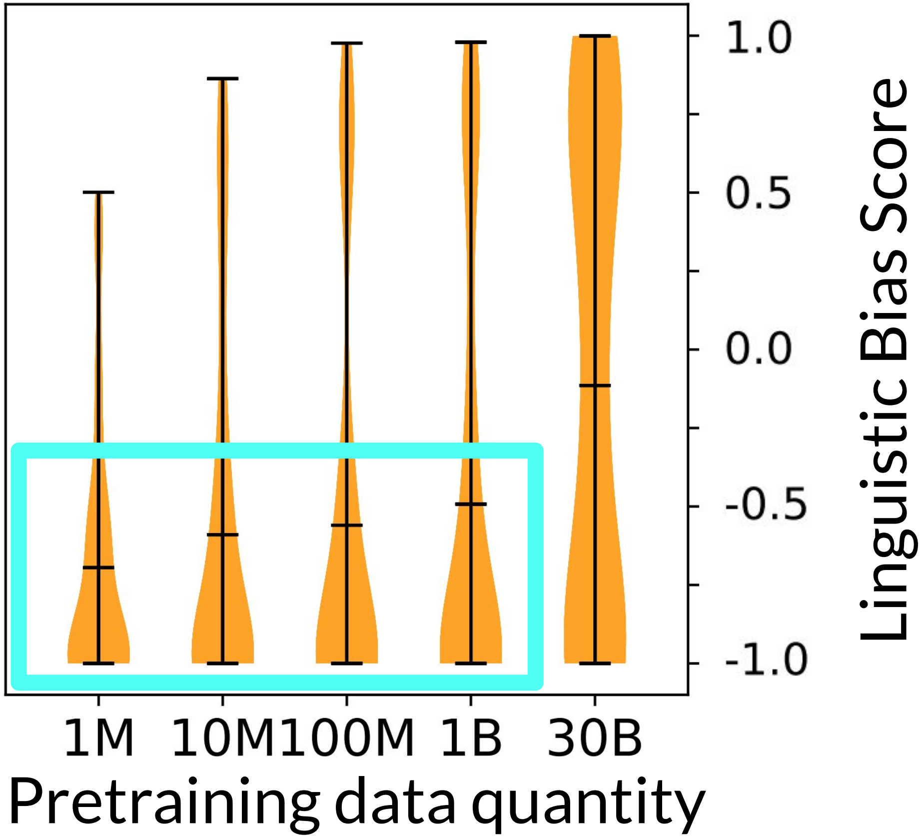
Results: Ambiguous Experiment

Aggregate results over all tasks, separated by pretraining dataset size.



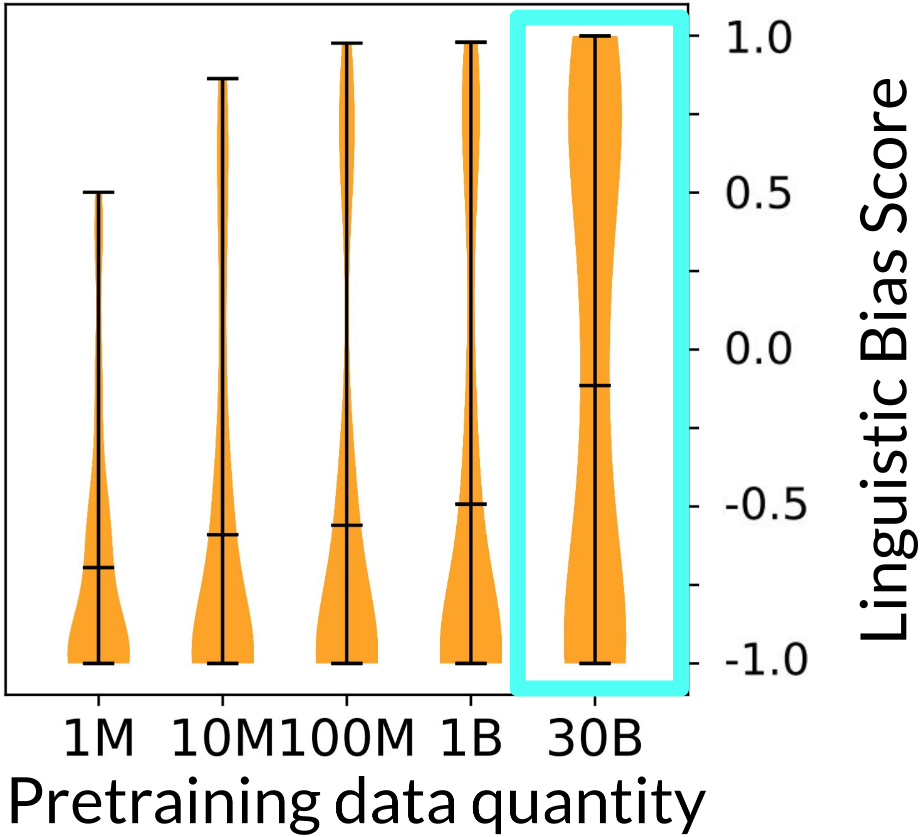
Results: Ambiguous Experiment

Models trained on 1B words or less almost always choose the surface generalization.



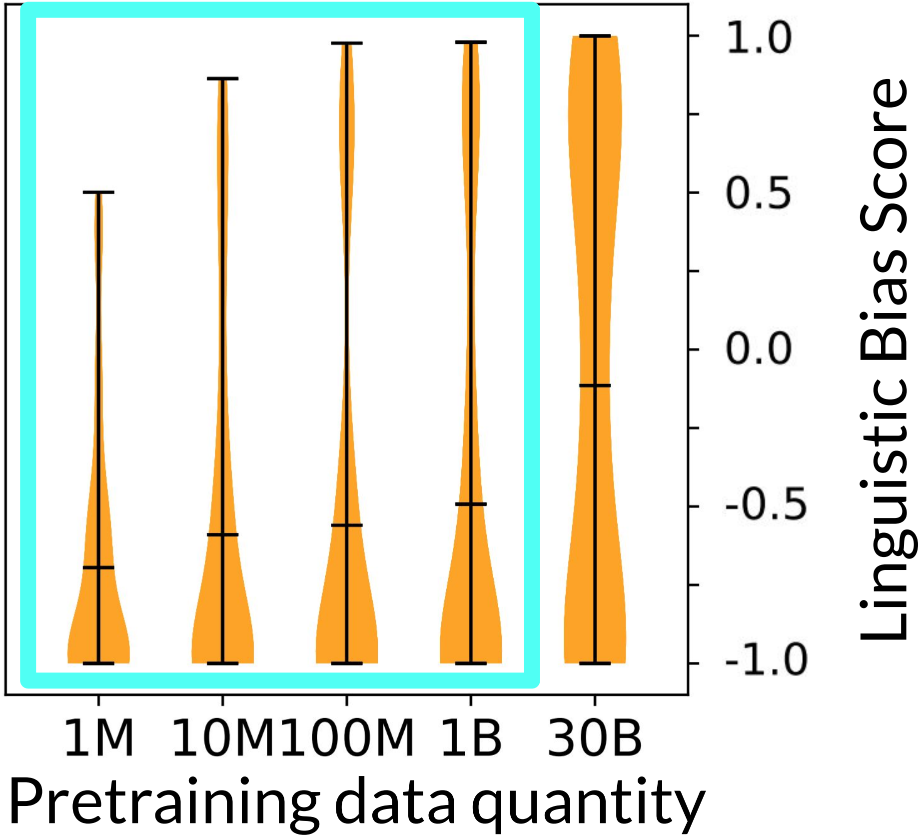
Results: Ambiguous Experiment

RoBERTa-base (trained on ~30B words) chooses the linguistic generalization about half the time.

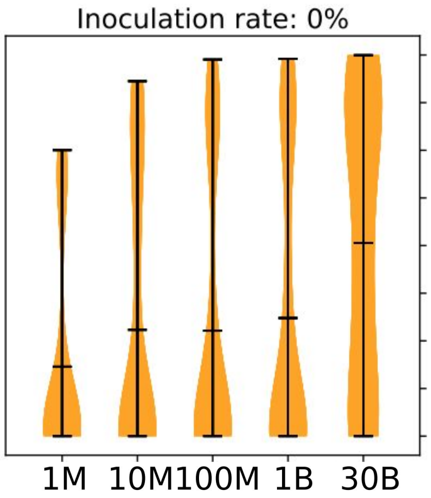


Results: Ambiguous Experiment

The remaining models show similar results. *Does this mean they have similar inductive biases?*

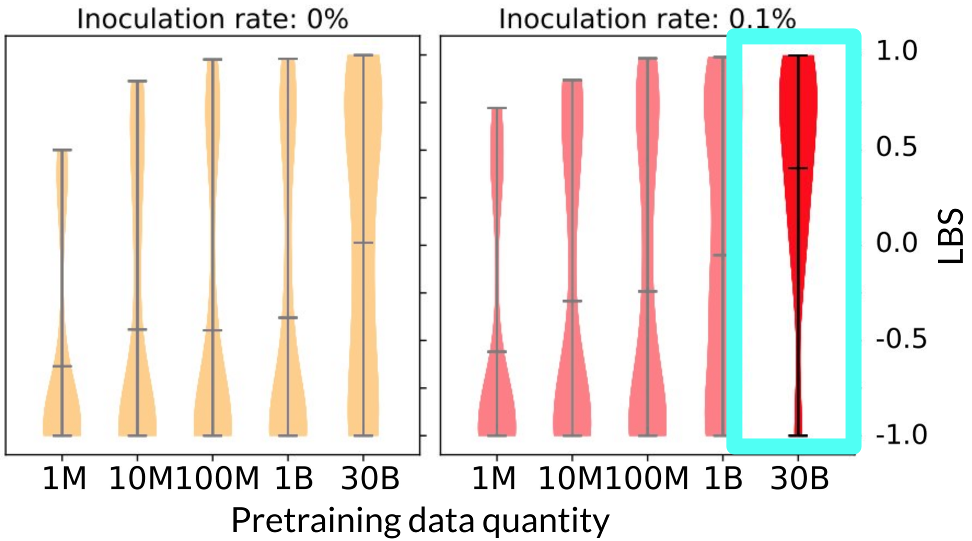


Inoculation Experiments



- We replace 0.1%, 0.3%, or 1% of the training data with **inoculation data**.
- We can quantify how strong a bias is by how much counter-evidence is needed to override it.

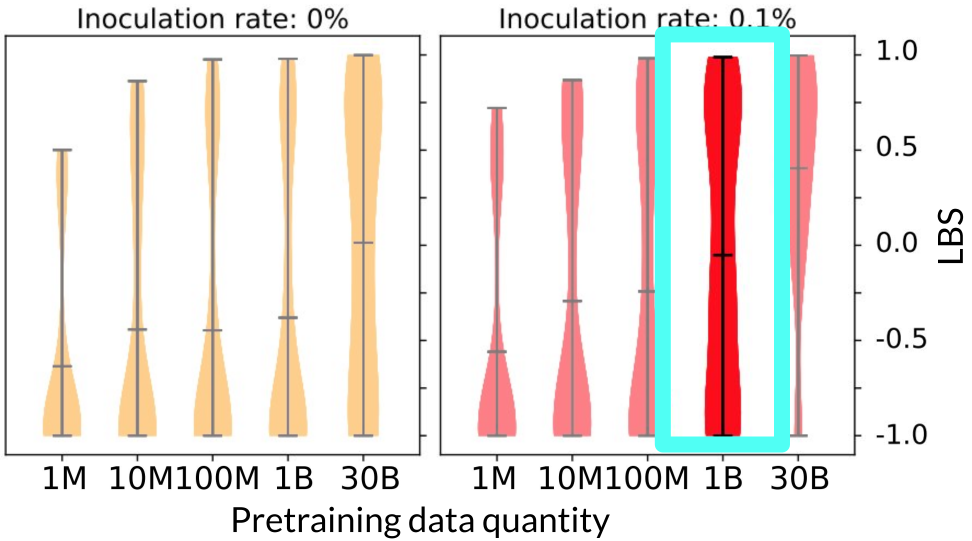
Results: Inoculation Experiments



Add 0.1% inoculation
(10 examples/10k)

RoBERTa base shows a more systematic linguistic bias.

Results: Inoculation Experiments

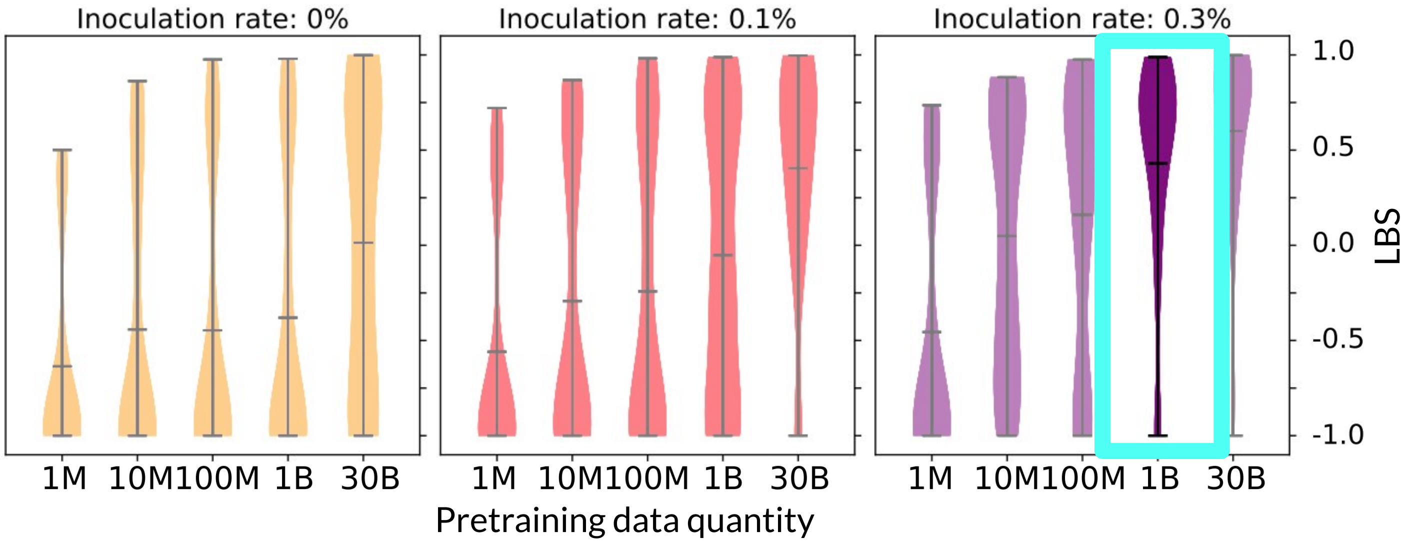


Add 0.1% inoculation (10 examples/10k)

RoBERTa base shows a more systematic linguistic bias.

The 1B models start to adopt the linguistic generalization fairly often.

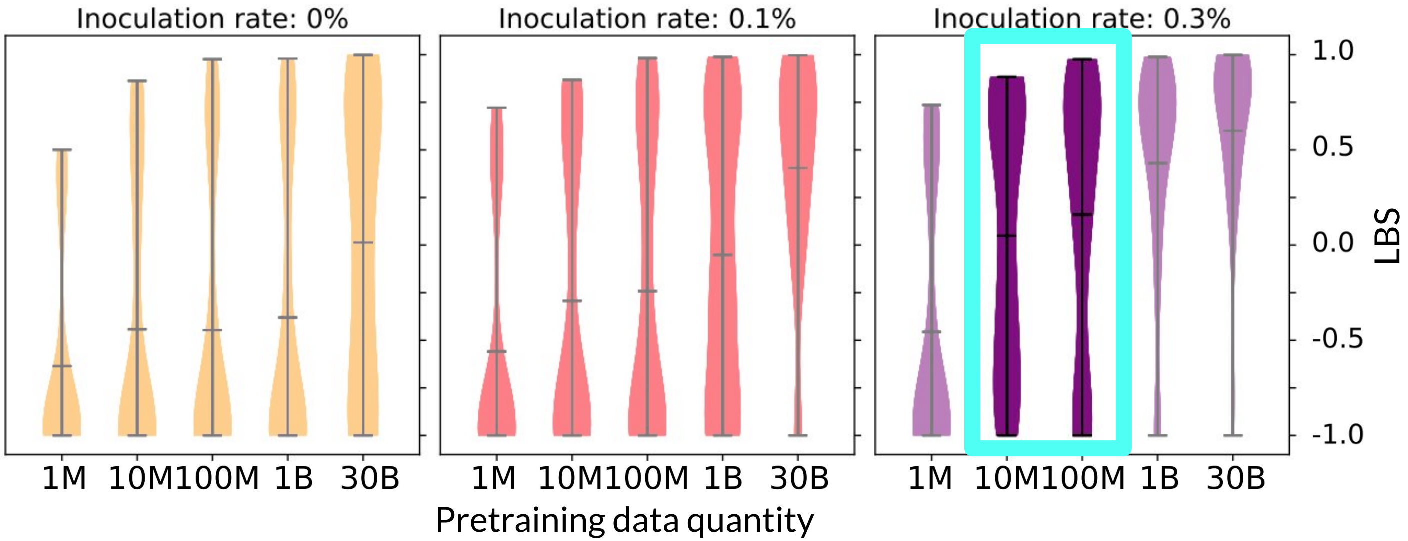
Results: Inoculation Experiments



Add 0.3% inoculation (30 examples/10k)

1B model shows a systematic linguistic bias.

Results: Inoculation Experiments

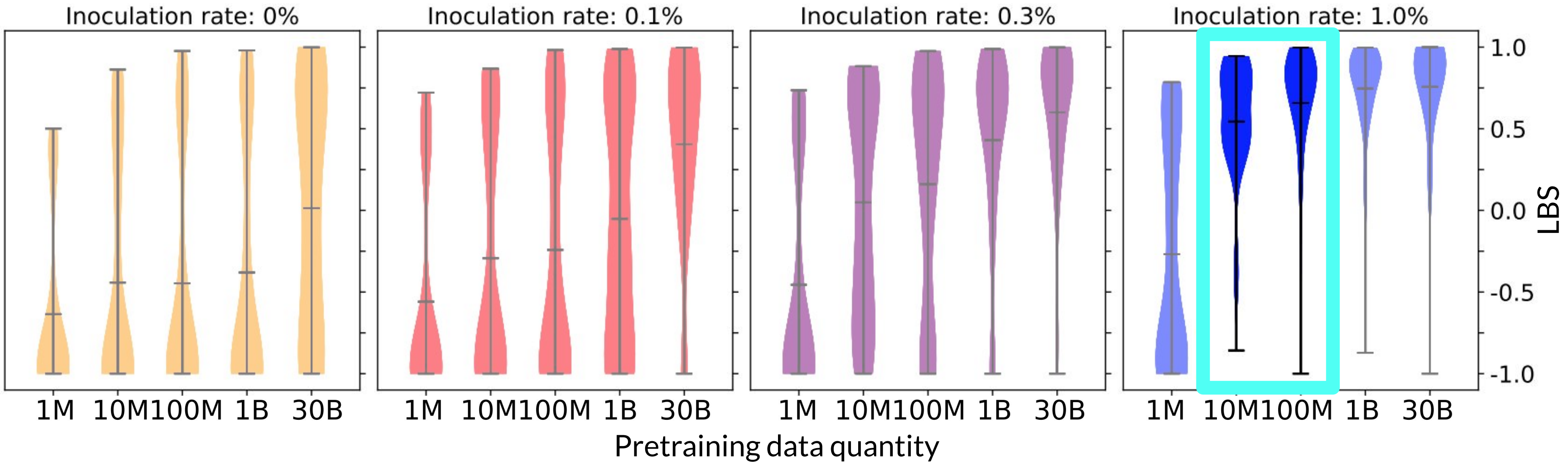


Add 0.3% inoculation (30 examples/10k)

1B model shows a systematic linguistic bias.

The 10M and 100M models start to consistently make the linguistic generalization.

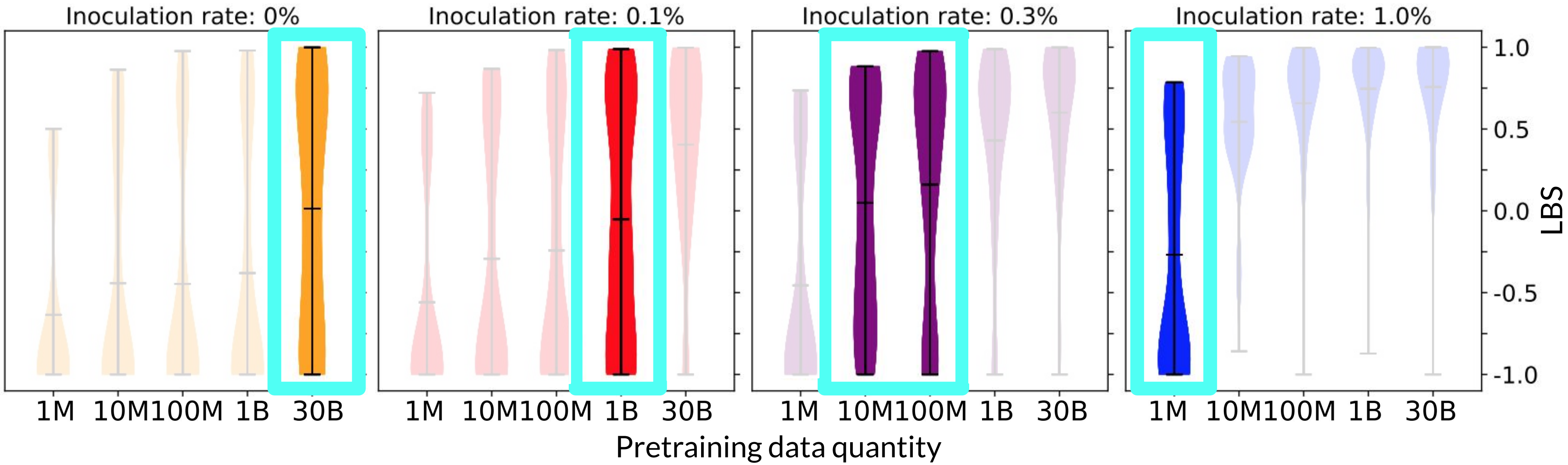
Results: Inoculation Experiments



Add 1% inoculation (100 examples/10k)

The 10M and 100M models systematically make the linguistic generalization.

Results: Inoculation Experiments



A "phase shift" where inoculation starts to change the model behavior happens more easily for models with more pretrainind data.

Part 3
What can neural networks teach us about humans?

The ideal experiment

The ideal experiment

What are the necessary conditions for human language acquisition?

Deprivation experiments

What are the necessary conditions for human language acquisition?

Pharaoh Psamtik
(664 – 610 BCE)

Frederick II
(1194-1250)

James IV
(1473-1513)

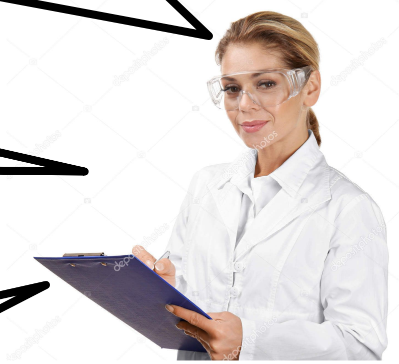
Deprivation experiments

What are the necessary conditions for human language acquisition?

Is hypothesized advantage *A* necessary for acquiring linguistic fact *F*.

•••

Is hypothesized advantage *B* necessary for acquiring linguistic fact *G*.



Is hypothesized advantage *B* necessary for acquiring linguistic fact *G*?

1. **Train artificial learner *L* without advantage *A*.**
2. **Check if *L* can acquire fact *F*.**
3. **If *L* succeeds, *and doesn't have any additional advantage over humans*, then *A* is not necessary to explain human acquisition of *F*.**

Is hypothesized advantage *B* necessary for acquiring linguistic fact *G*?

1. Train BERT without advantage *A*.
2. Check if BERT can acquire fact *F*.
3. If BERT succeeds, ***and doesn't have any additional advantage over humans***, then *A* is not necessary to explain human acquisition of *F*.

Is hypothesized advantage *B* necessary for acquiring linguistic fact *G*?

1. Train BERT without innate structural bias.
2. Check if BERT can acquire fact *F*.
3. If BERT succeeds, *and doesn't have any additional advantage over humans*, then innate structural bias is not necessary to explain human acquisition of *F*.

Is hypothesized advantage *B* necessary for acquiring linguistic fact *G*?

1. Train BERT without innate structural bias.
2. Check if BERT can acquire subject aux inversion.
3. If BERT succeeds, *and doesn't have any additional advantage over humans*, then innate structural bias is not necessary to explain human acquisition of subject aux inversion.

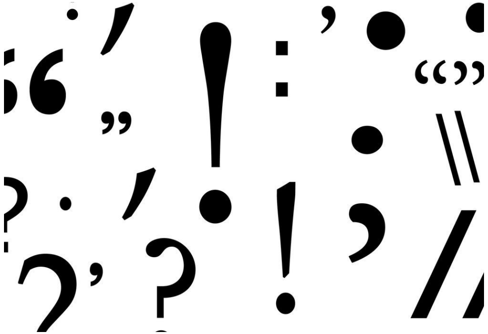
… if the learner doesn't have any additional advantage over humans

Advantages ANNs Have

Data quantity

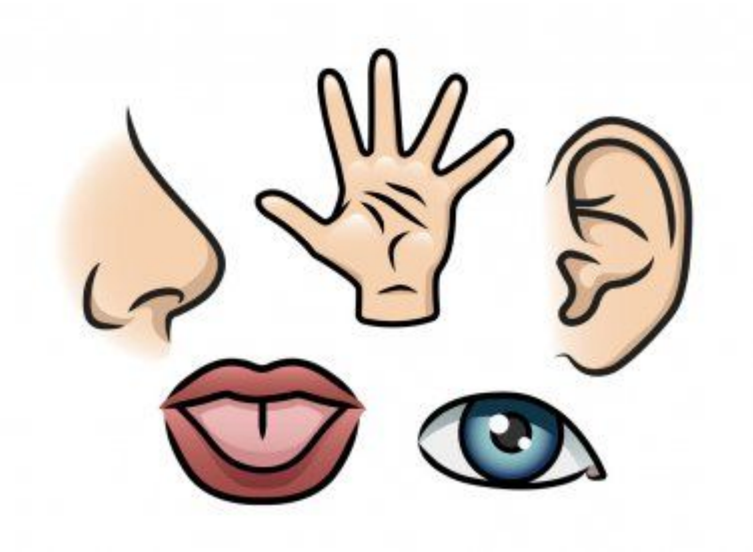
Data domain

Orthography



Advantages Humans Have

Multimodal input



Interactive learning

Resources

1. miniBERTas [link]
2. MSGS data/code [link]
3. Probing code [link]

Questions?

Bonus slides

Conclusions

Main Findings

Support for two different stages of learning as data quantity grows:

Main Findings

Support for two different stages of learning as data quantity grows:

1. Linguistic feature learning needs 1M-100M words of data.

Main Findings

Support for two different stages of learning as data quantity grows:

1. Linguistic feature learning needs 1M-100M words of data.
2. Linguistic bias and strong generalization on NLU tasks requires >1B words.

Lessons for Pretraining

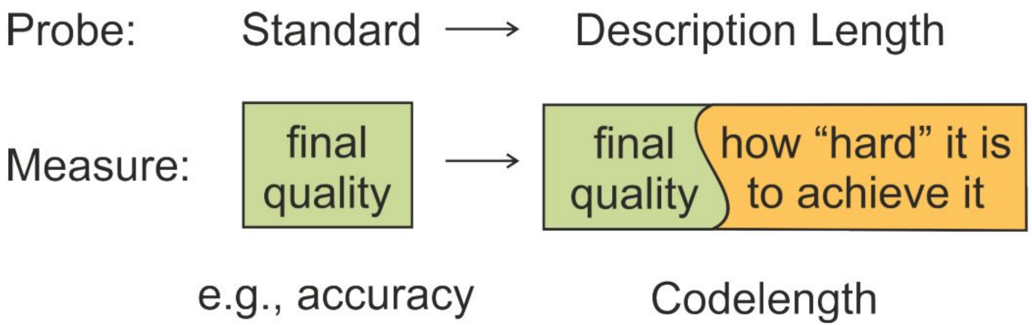
...So an LM trained on trillions of words will be better at linguistic generalization?!

Lessons for Pretraining

...So an LM trained on trillions of words will be better at linguistic generalization?!

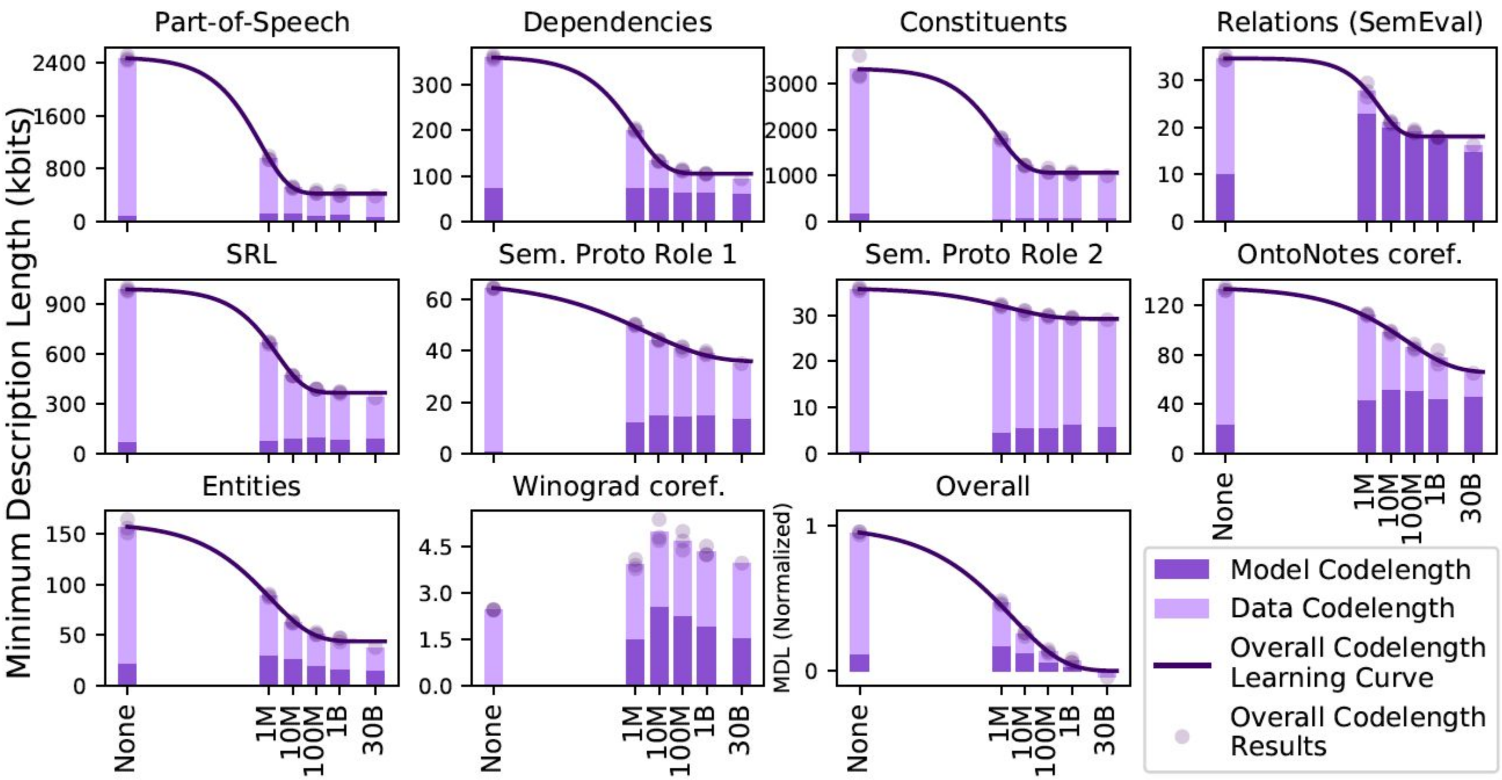
More important: If we want to improve pretraining, we should make feature preference learning more efficient.

2. Information theoretic MDL probing

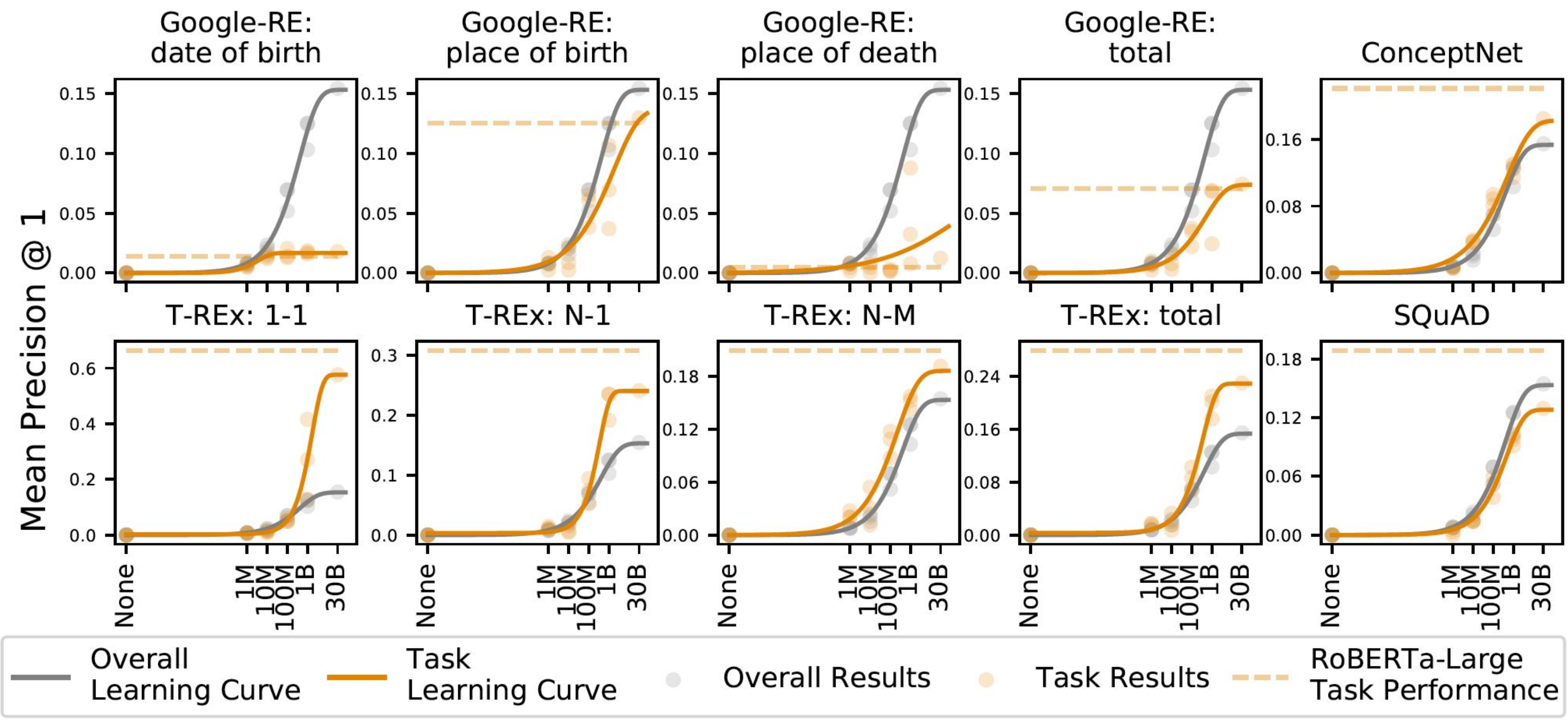


Source: Voita & Titov (2020)

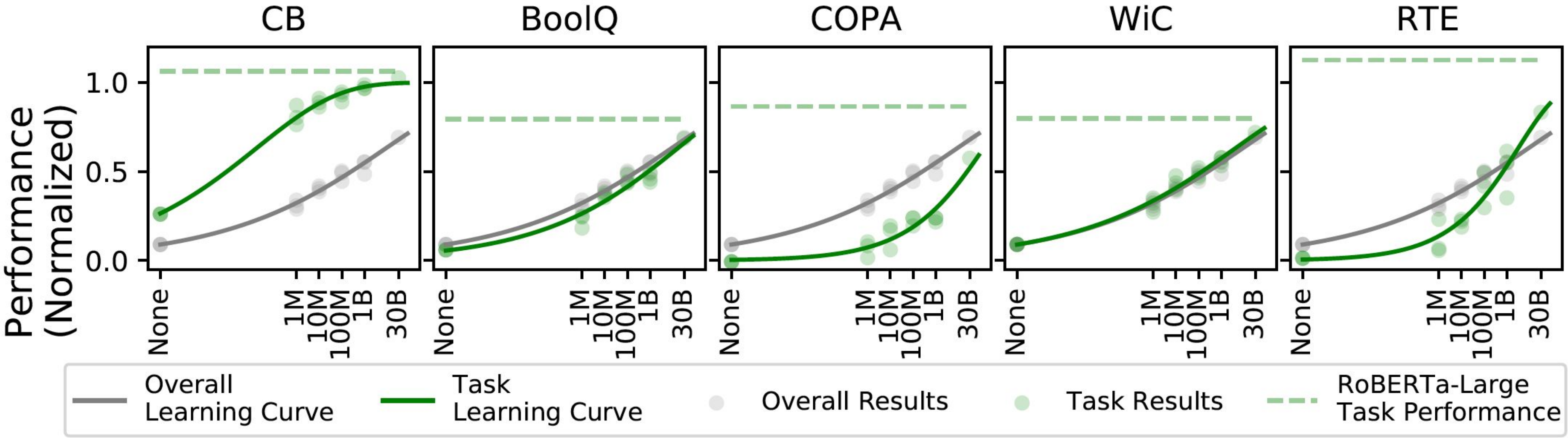
2. Information theoretic MDL probing



4. Unsupervised Commonsense Knowledge



5. SuperGLUE: Downstream NLU Tasks



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.

Data Generation

- The MSGS data is generated from templates.

Data Generation

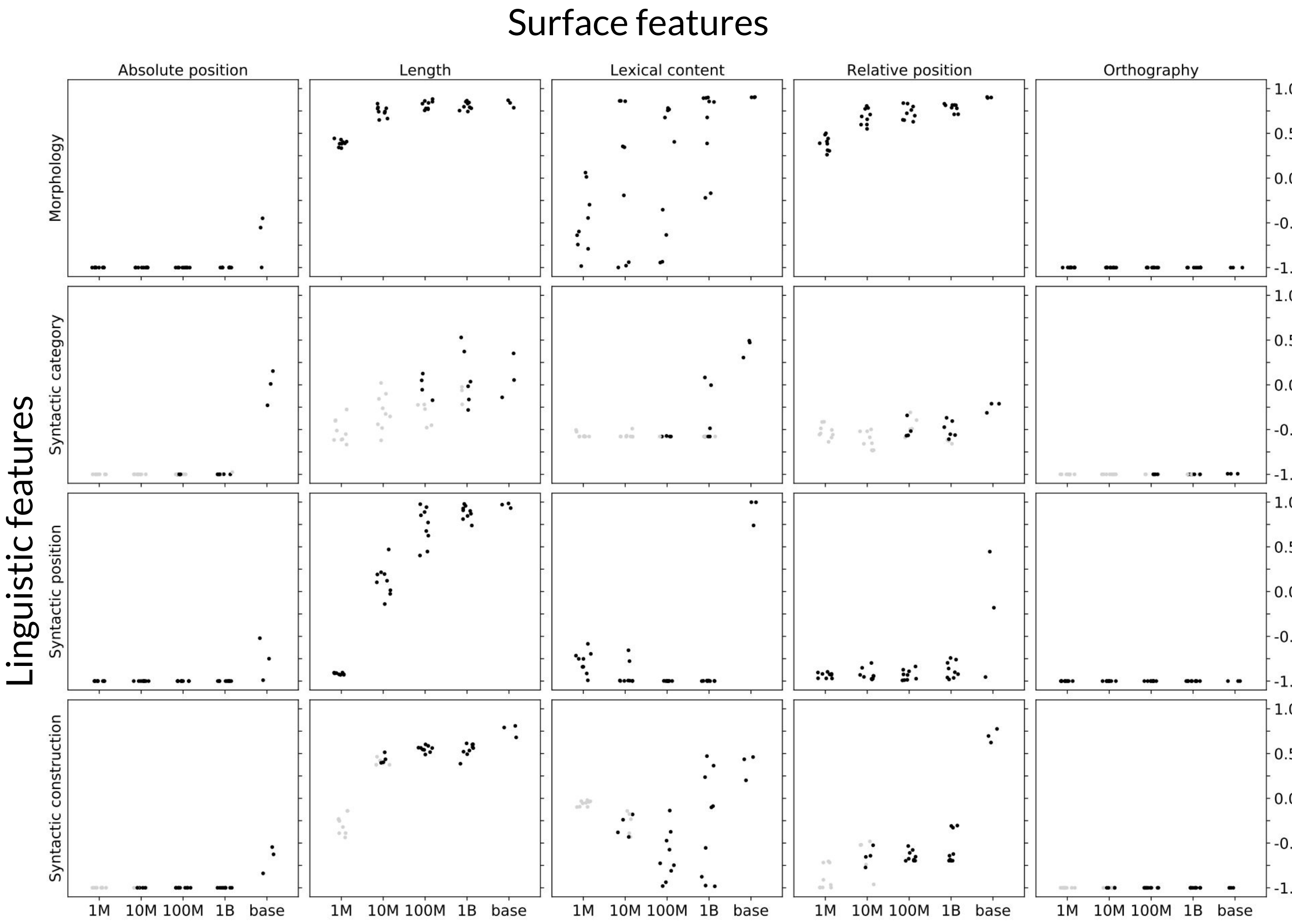
- The MSGS data is generated from templates.
- We always test classifiers' ability to generalize out-of-domain.

Example: In-domain vs. Out-of-Domain

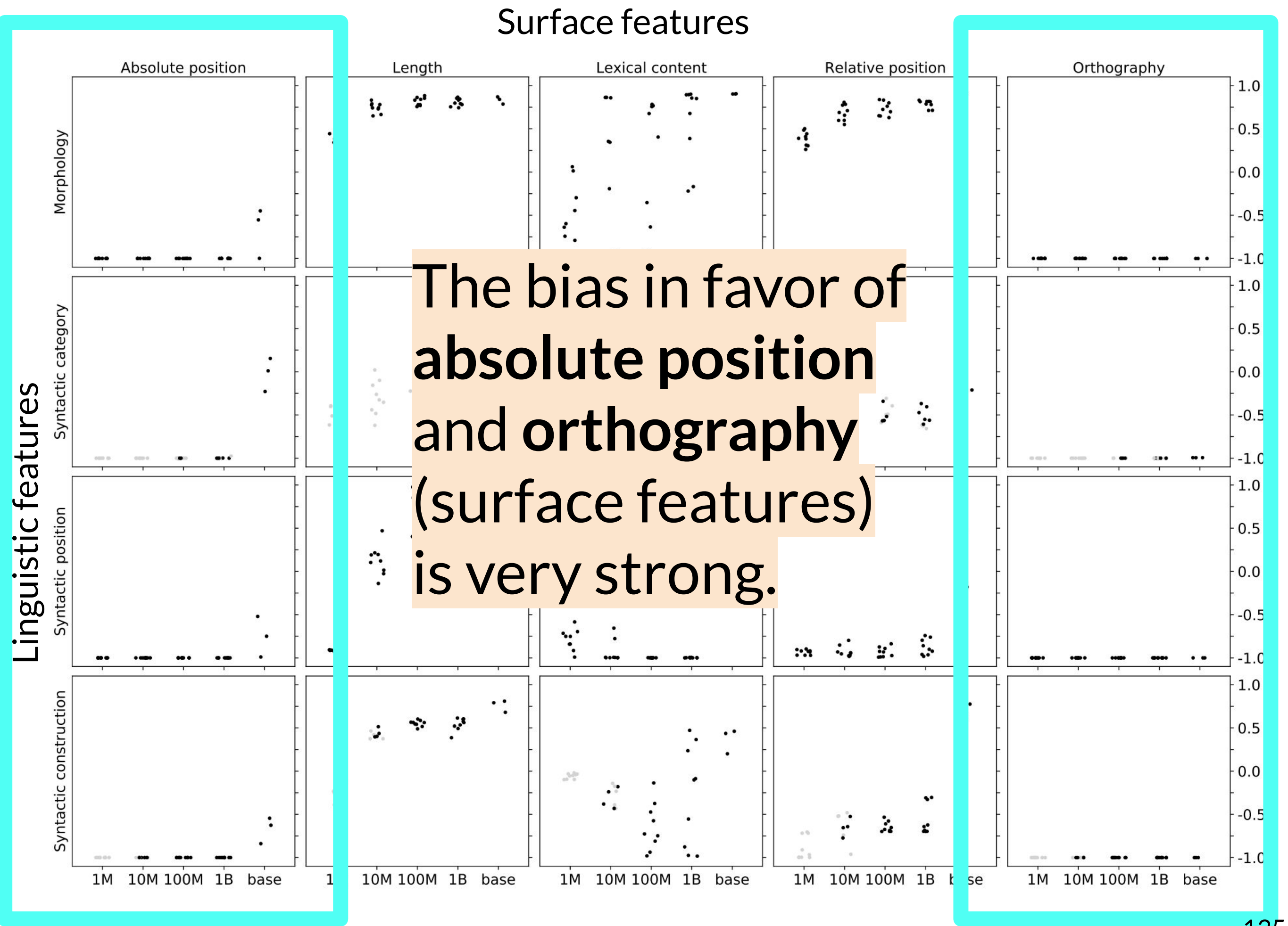
In domain: *The big dog is yawning.*

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

Results: Ambiguous Experiment
(Fine-grained)

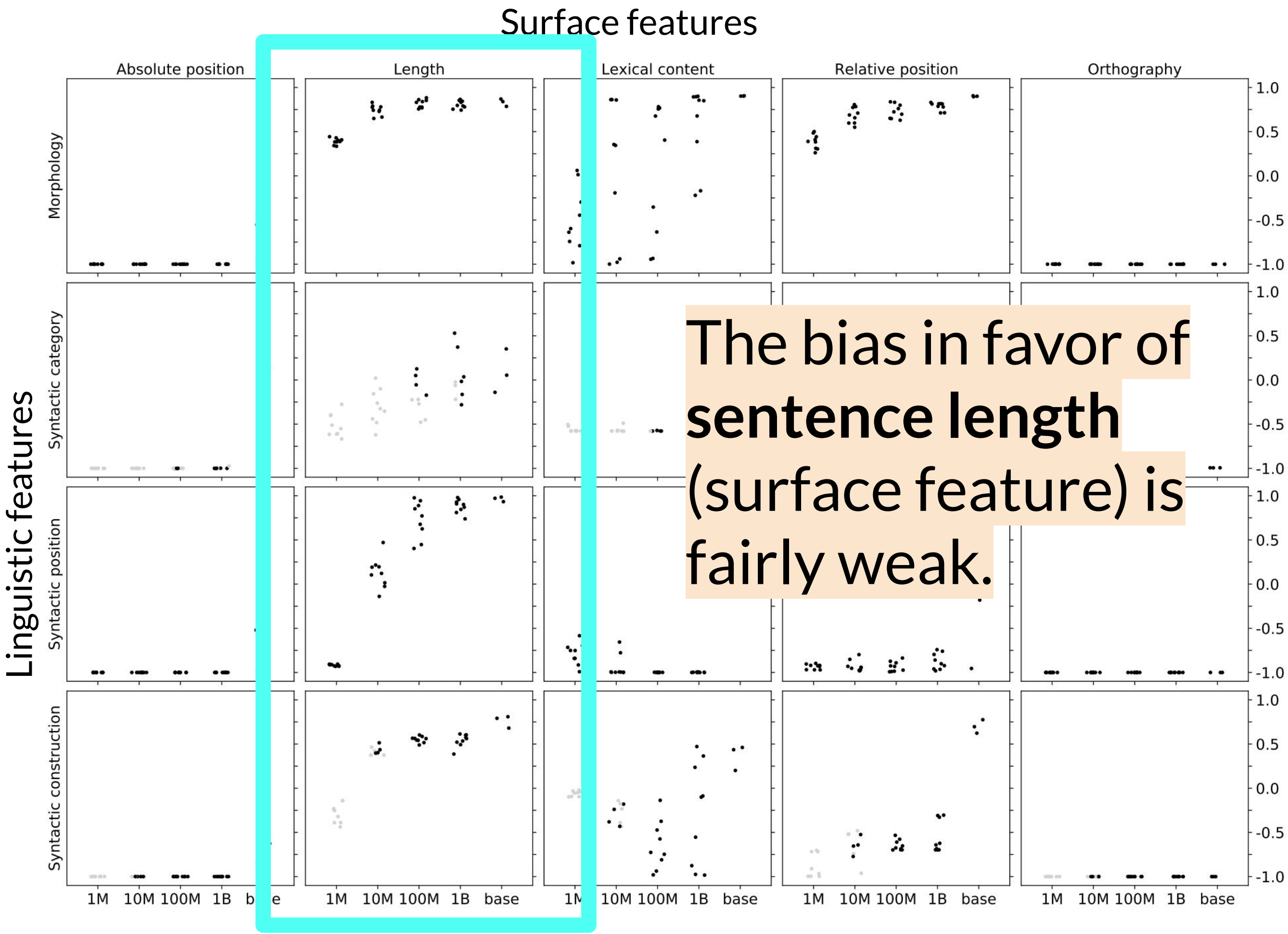


Results: Ambiguous Experiment (Fine-grained)



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

Results: Ambiguous Experiment
(Fine-grained)



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

Part I: Features/Data/Methods

Feature Learning Experiments

Does model X represent linguistic/surface feature Y?

Feature Learning Experiments

Does model X represent linguistic/surface feature Y?

Two motivations:

1. Feature preferences only make sense for features that are represented.

Feature Learning Experiments

Does model X represent linguistic/surface feature Y?

Two motivations:

1. Feature preferences only make sense for features that are represented.
2. We can compare the difficulty of feature learning and preference learning.

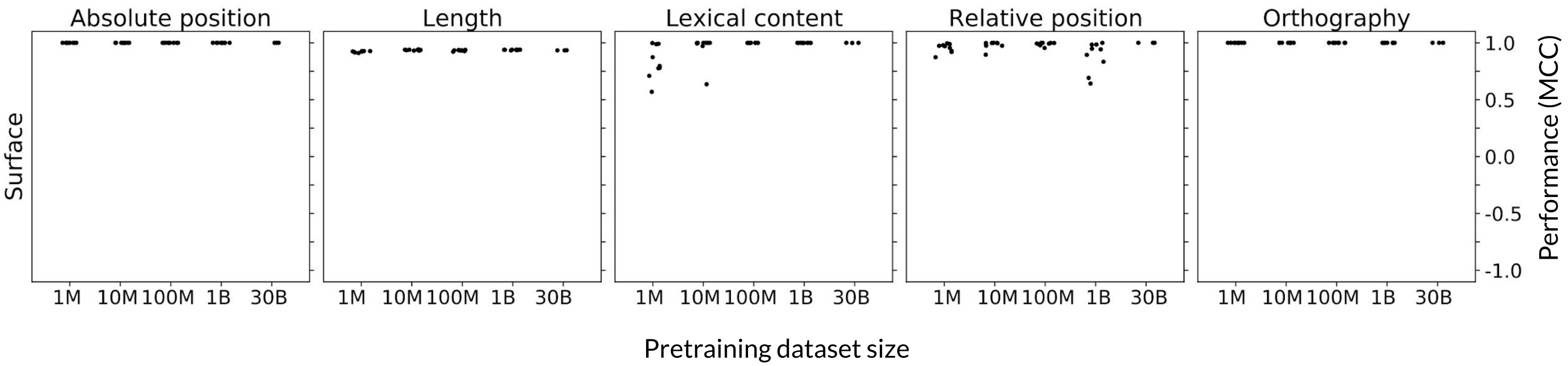
Surface vs. Linguistic Features

	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.

Fine-tuning

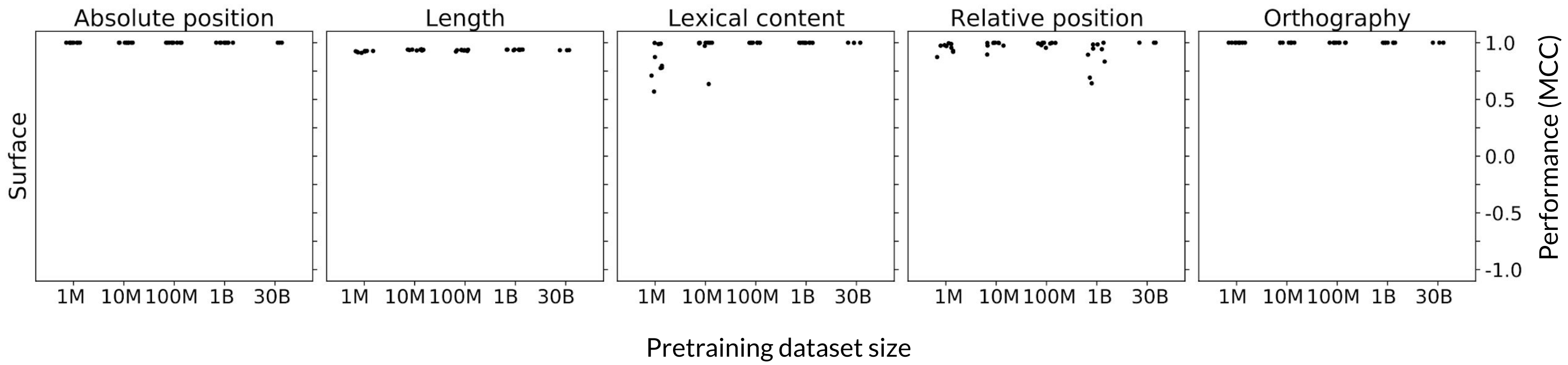
- 9 tasks (4 linguistic + 5 surface)
- 12 miniBERTas + original $\text{RoBERTa}_{\text{BASE}}$ (~30B words)
- The training sets are 10k sentences each

Results: Feature Learning Experiments

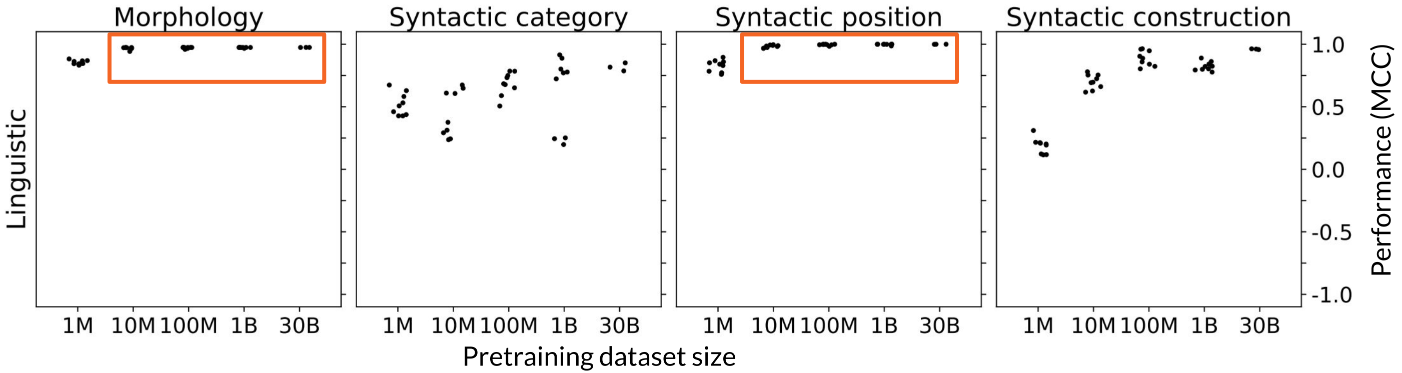


Surface features: Performance is at ceiling.

Results: Feature Learning Experiments

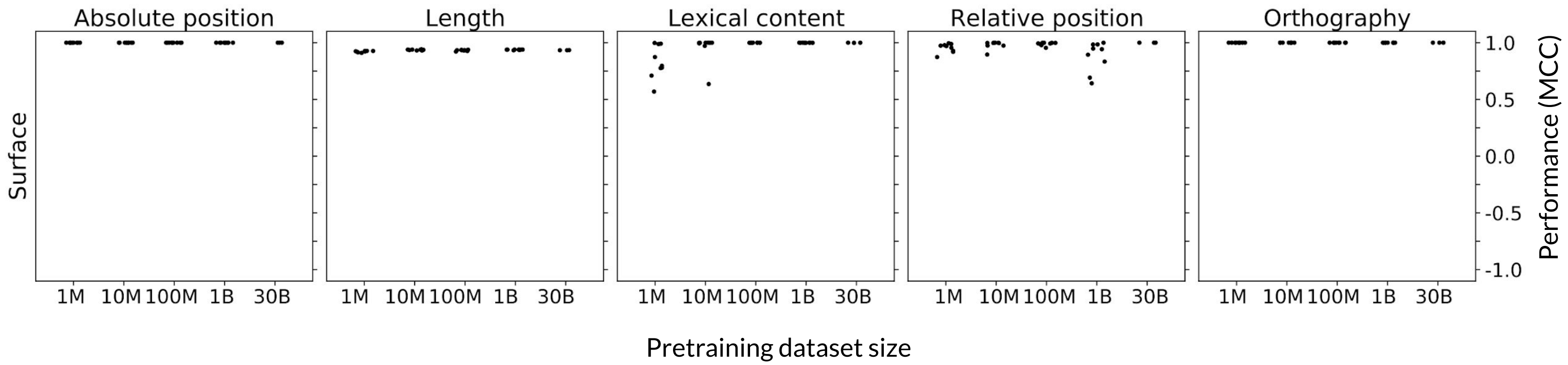


Surface features: Performance is at ceiling.

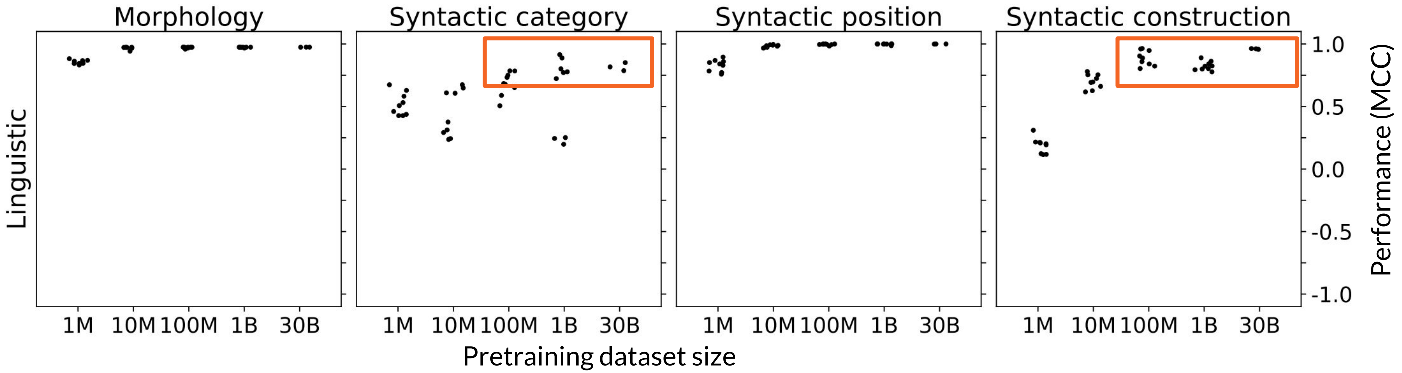


Linguistic features: Performance is near ceiling for morphology & syntactic position >1M words.

Results: Feature Learning Experiments



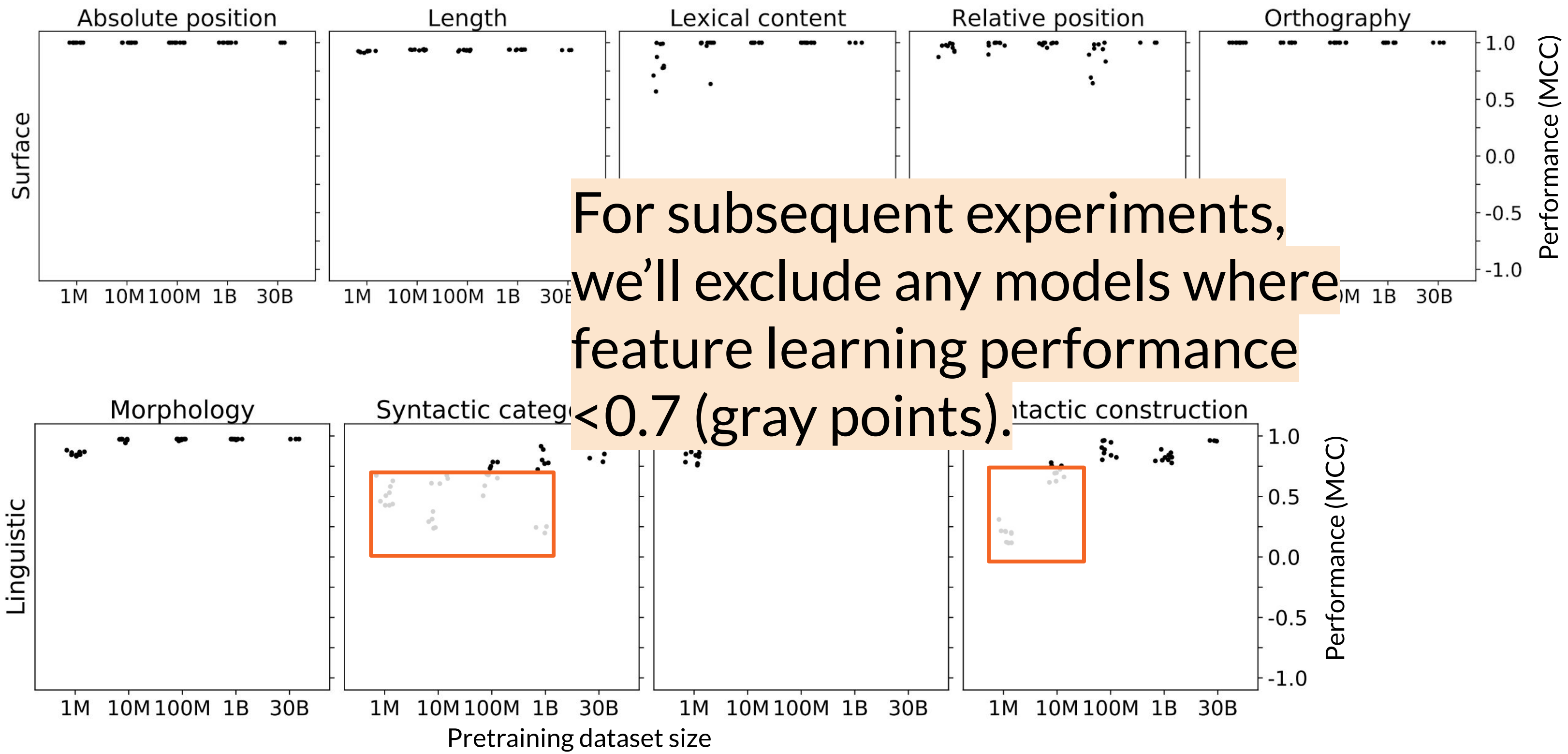
Surface features: Performance is at ceiling.



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

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

Results: Feature Learning Experiments

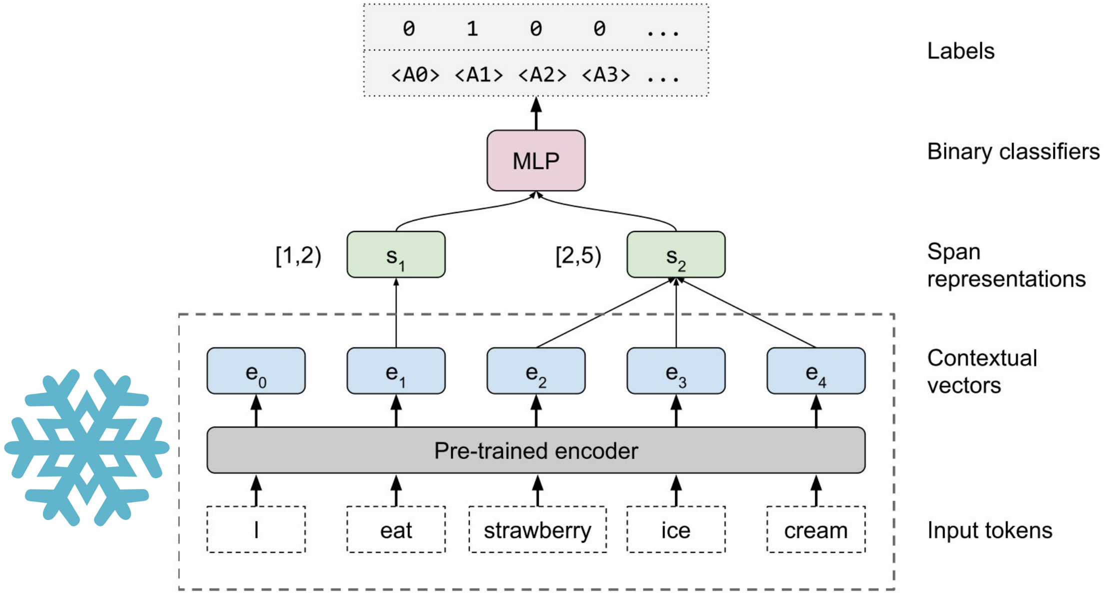


For subsequent experiments, we'll exclude any models where feature learning performance <0.7 (gray points).

Lessons for Language Acquisition

- The very idea that linguistic bias is learnable is controversial.
- We have earlier findings that BERT prefers linguistic generalizations in key empirical domains in this debate (in CogSci; Warstadt & Bowman, 2020)
- Focusing on data quantity is important: Humans are more efficient learners than Transformers.

1. “Standard” classifier probing



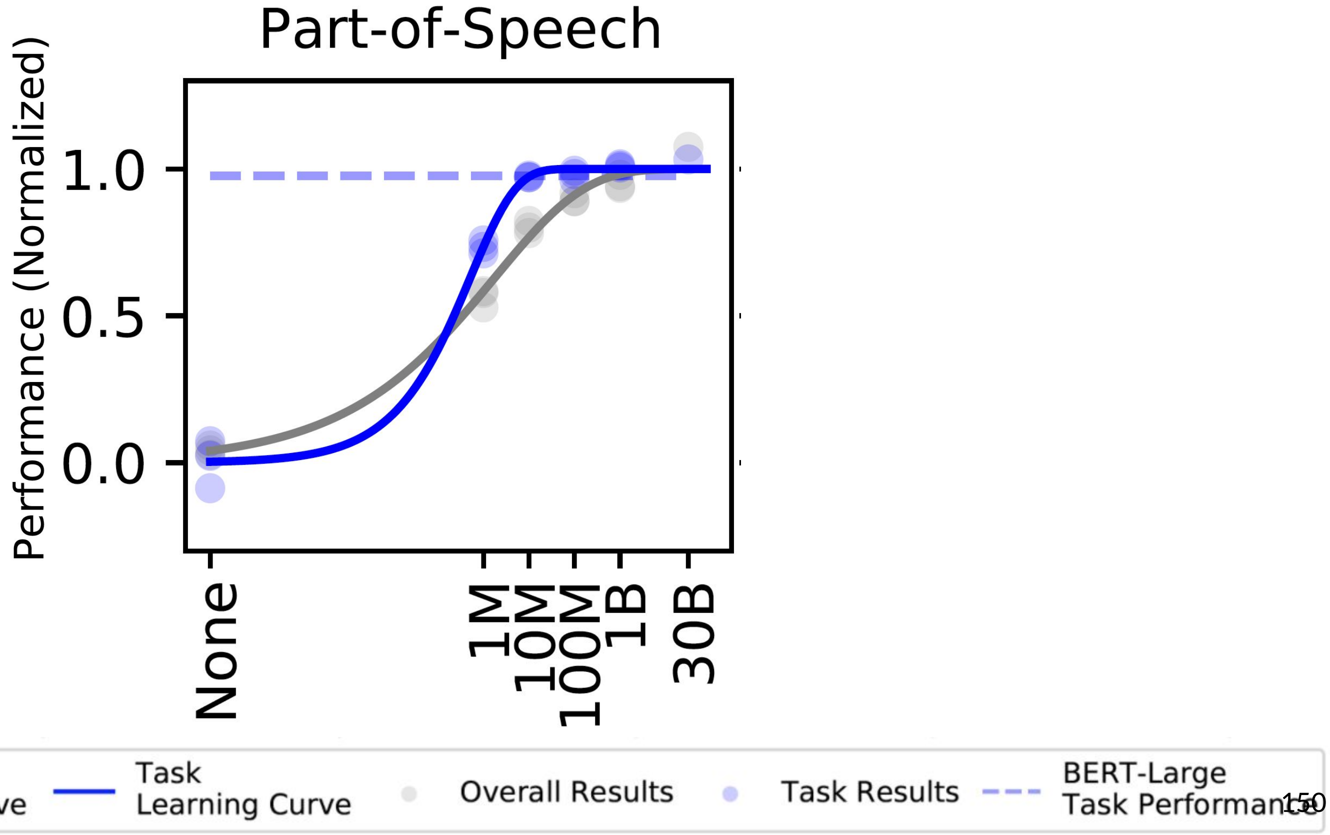
Source: Tenney et al. (2019)

1. “Standard” classifier probing

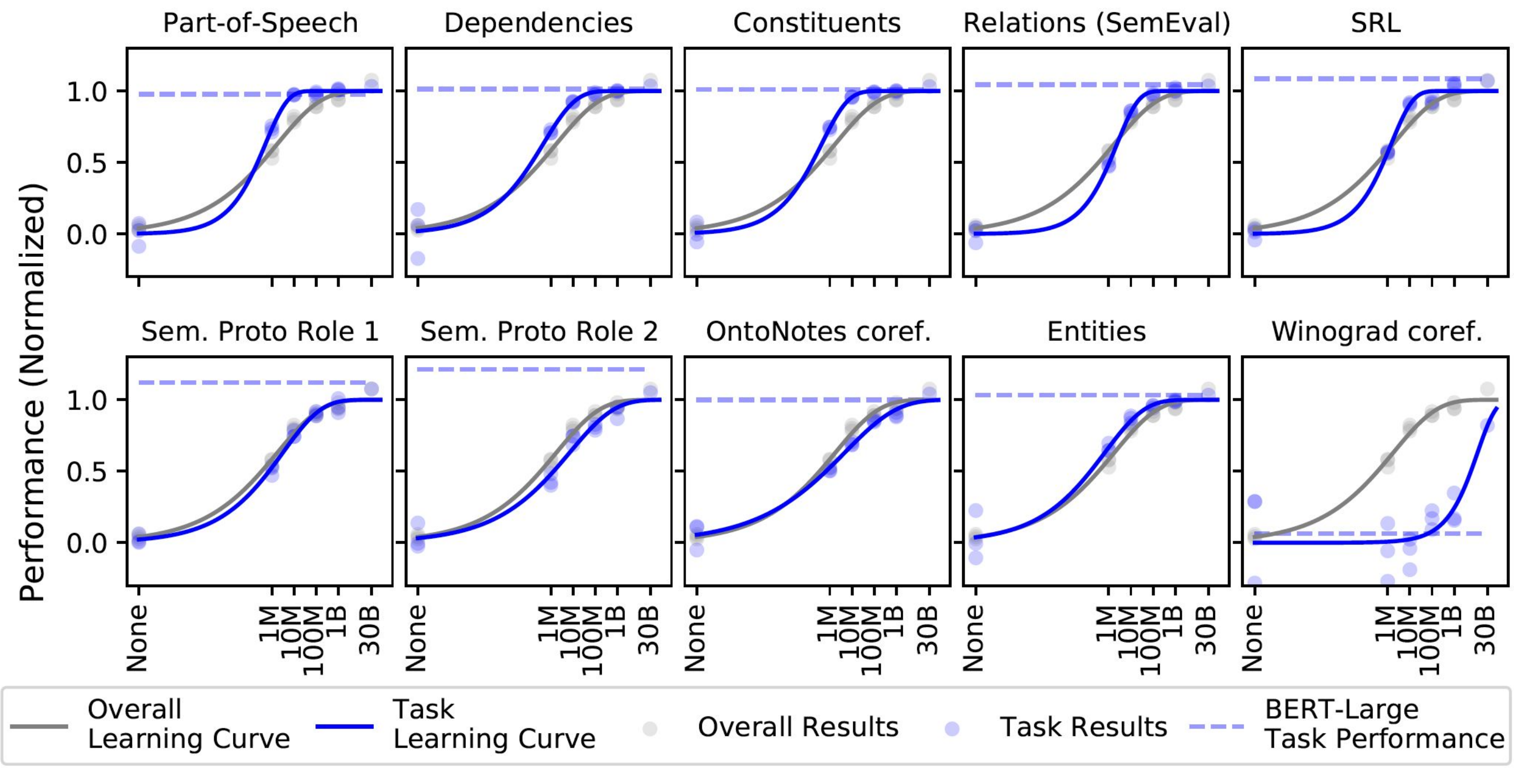
POS	The important thing about Disney is that it is a global $[\text{brand}]_1$. → NN (Noun)
Constit.	The important thing about Disney is that it $[\text{is a global brand}]_1$. → VP (Verb Phrase)
Depend.	$[\text{Atmosphere}]_1$ is always $[\text{fun}]_2$ → nsubj (nominal subject)
Entities	The important thing about $[\text{Disney}]_1$ is that it is a global brand. → Organization
SRL	$[\text{The important thing about Disney}]_2$ $[\text{is}]_1$ that it is a global brand. → Arg1 (Agent)
SPR	$[\text{It}]_1$ $[\text{endorsed}]_2$ the White House strategy. . . → {awareness, existed_after, . . . }
Coref.$^{\text{O}}$	The important thing about $[\text{Disney}]_1$ is that $[\text{it}]_2$ is a global brand. → True
Coref.$^{\text{W}}$	$[\text{Characters}]_2$ entertain audiences because $[\text{they}]_1$ want people to be happy. → True Characters entertain $[\text{audiences}]_2$ because $[\text{they}]_1$ want people to be happy. → False
Rel.	The $[\text{burst}]_1$ has been caused by water hammer $[\text{pressure}]_2$. → Cause-Effect(e_2, e_1)

Source: Tenney et al. (2019)

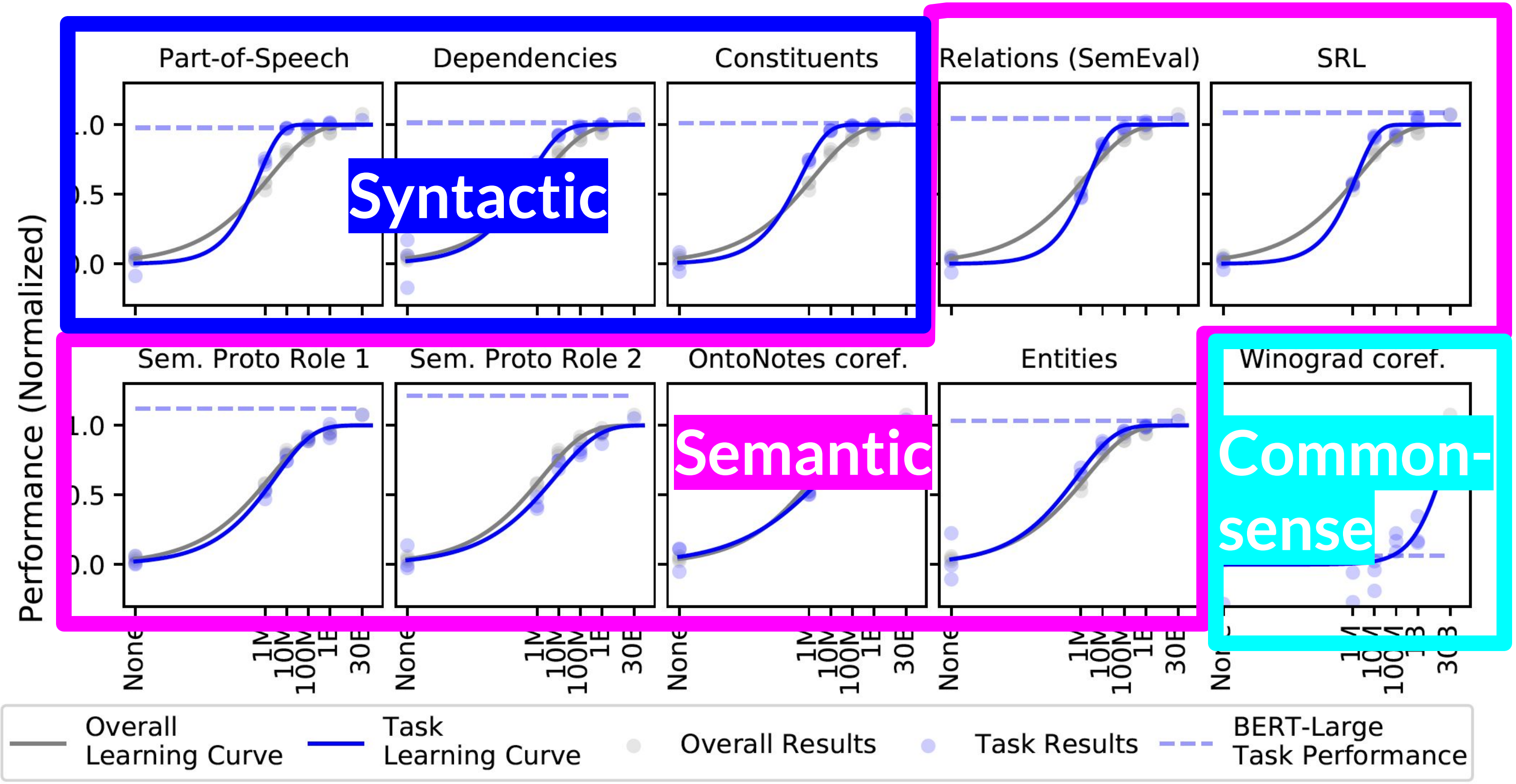
1. “Standard” classifier probing



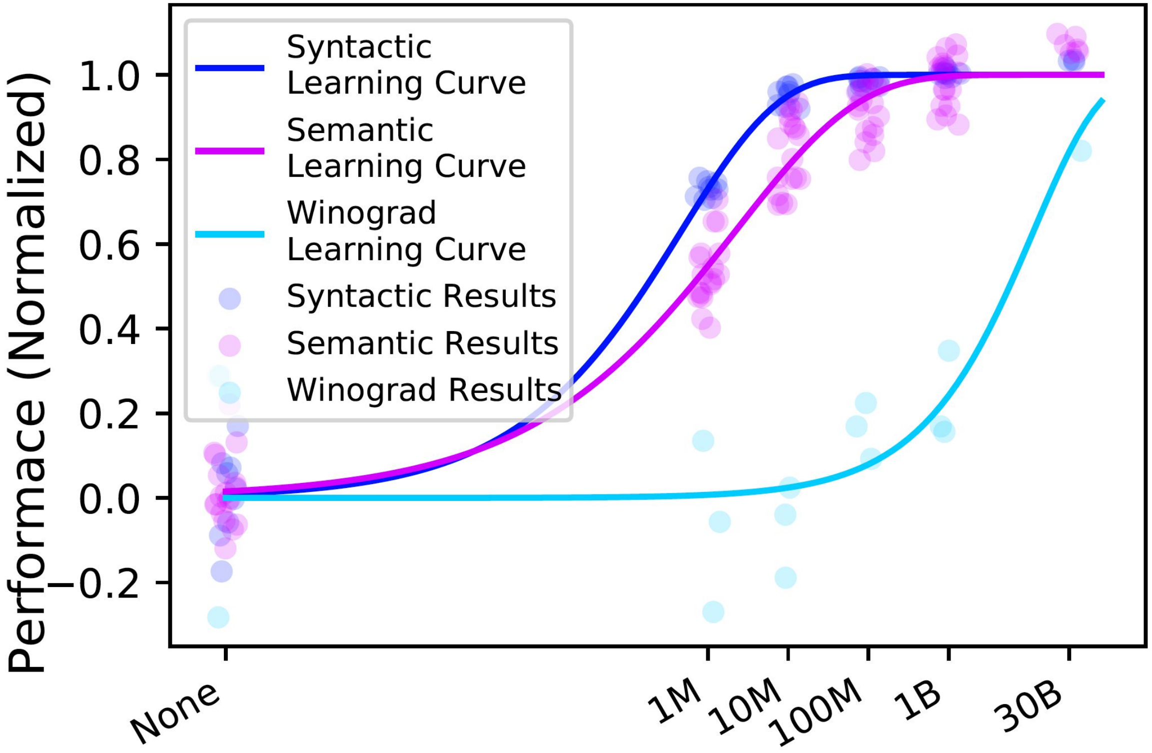
1. "Standard" classifier probing



1. "Standard" classifier probing

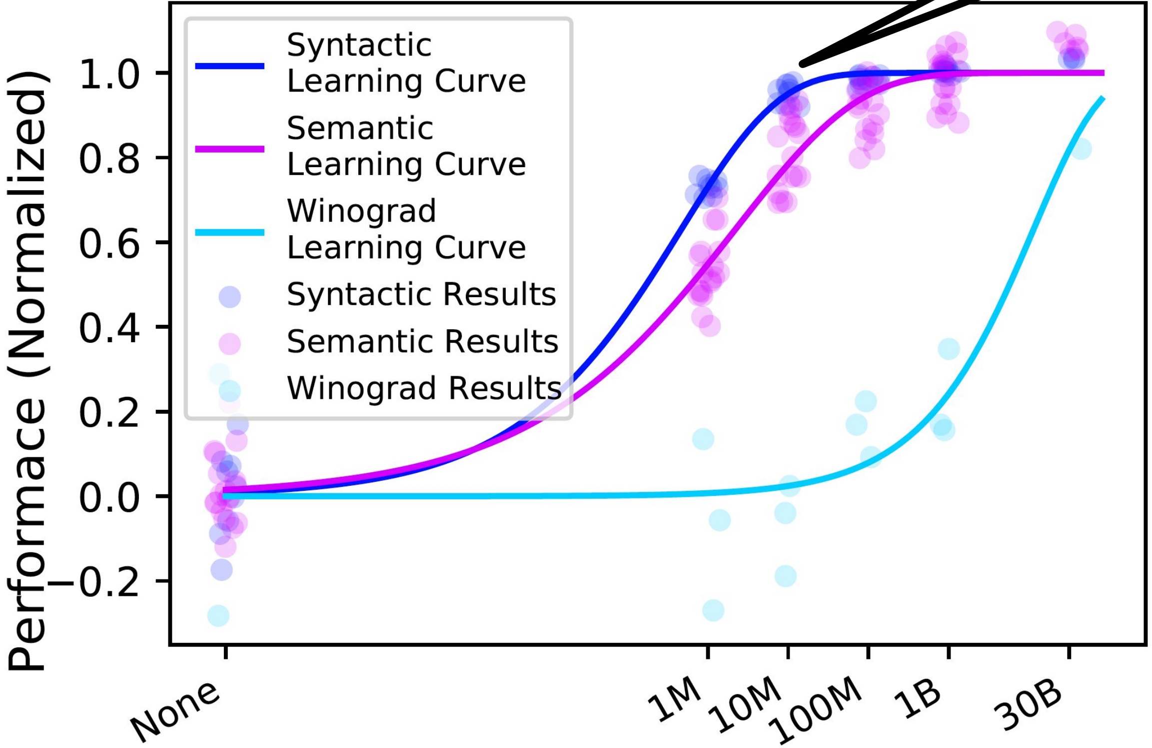


1. “Standard” classifier probing



1. “Standard” classifier probing

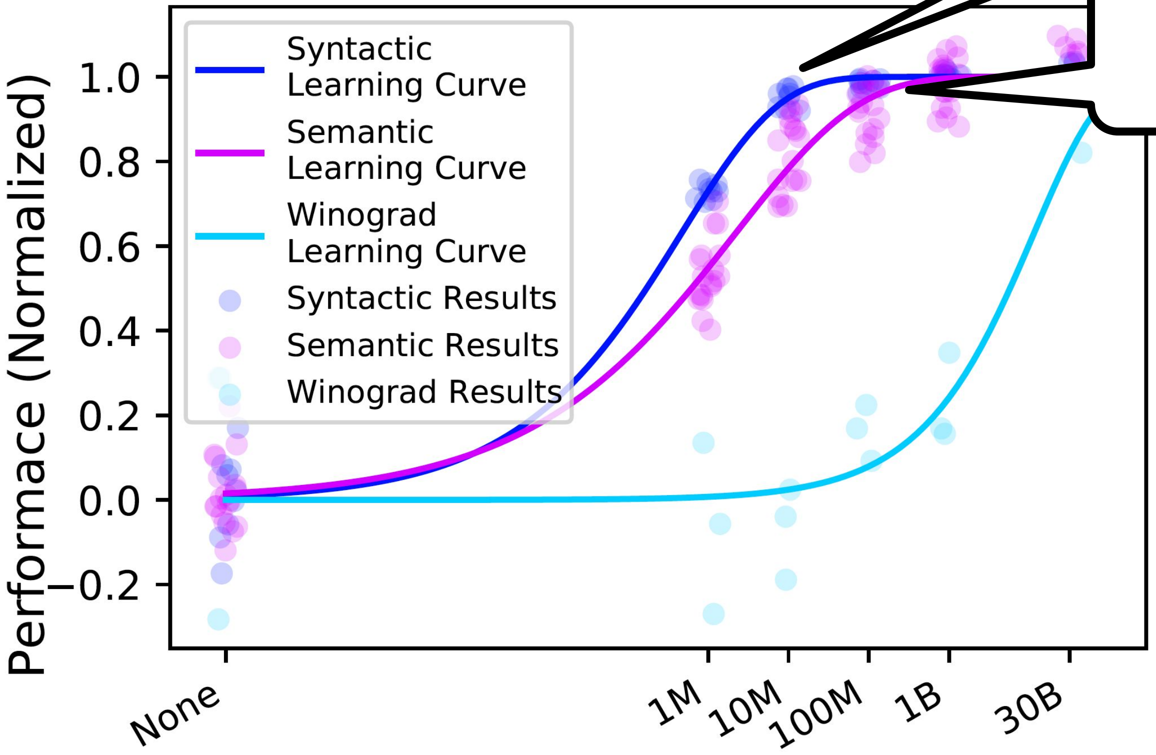
Syntactic feature learning converges ~10M words.



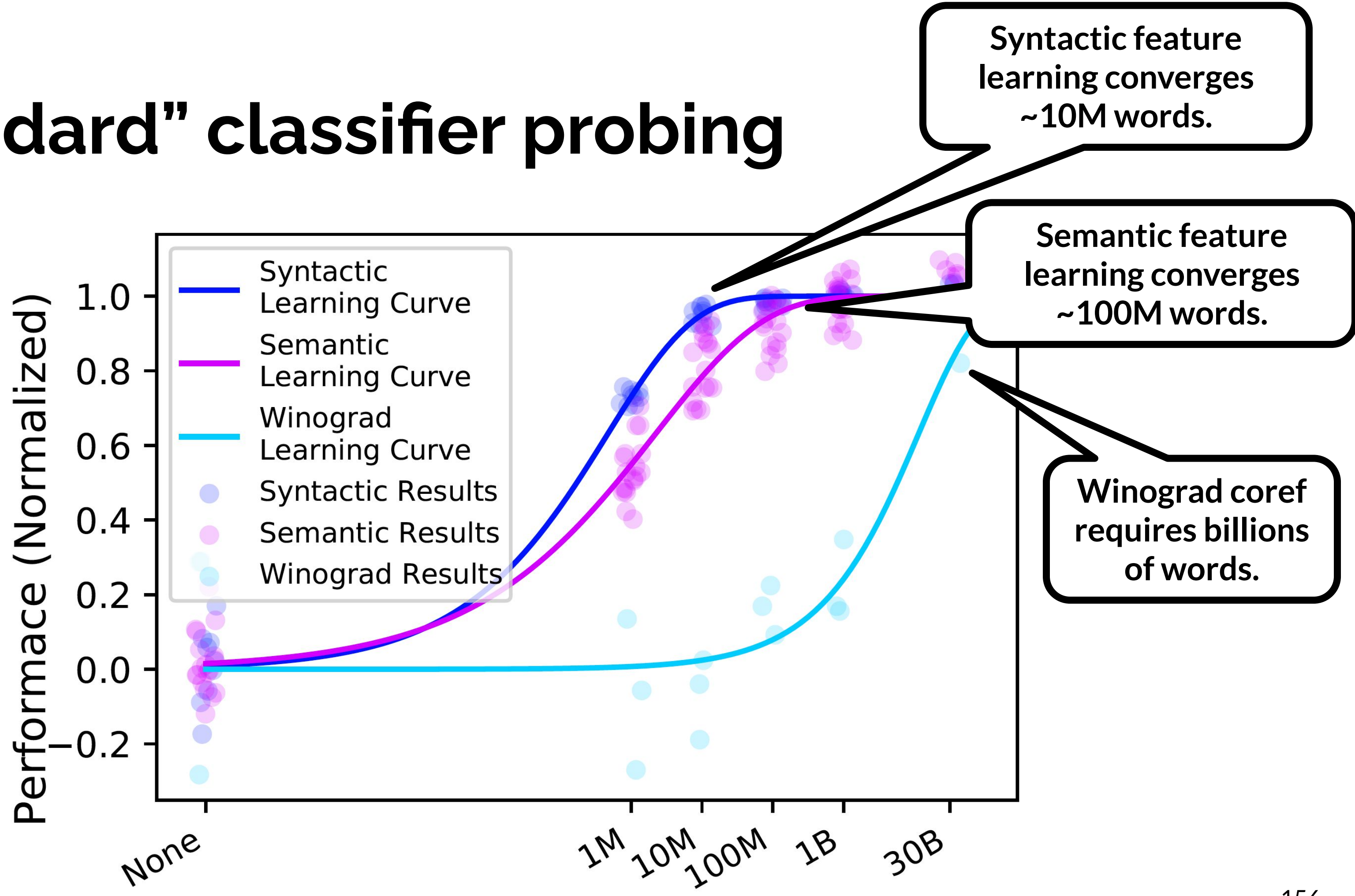
1. “Standard” classifier probing

Syntactic feature learning converges ~10M words.

Semantic feature learning converges ~100M words.



1. "Standard" classifier probing



3. BLiMP: Unsupervised Acceptability Judgments

The Benchmark of Linguistic Minimal Pairs for English

Warstadt et al. (2020)

- A collection of thousands of minimal pairs
- 67 types of contrasts, 1000 examples each
- 12 major phenomena in English **morphology**, **syntax**, and **semantics**.

Anaphor Agreement
Subject--Verb Agreement
Determiner--Noun Agreement
Ellipsis
NPIs
Argument Structure
Quantifiers
Filler--Gap Dependencies
Irregular Verb Forms
Binding
Control/Raising
Island Effects

3. BLiMP: Unsupervised Acceptability Judgments

Phenomenon	N	Acceptable example ✓	Unacceptable example ✗
Anaphor agreement	2	Many girls insulted themselves.	Many girls insulted herself.
Argument structure	9	Rose wasn't disturbing Mark.	Rose wasn't boasting Mark.
Binding	7	It's himself who Robert attacked.	It's himself who attacked Robert.
Control/Raising	5	Kevin isn't irritating to work with.	Kevin isn't bound to work with.
Determiner-N 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	Which bikes is John fixing?	Which is John fixing bikes?
NPI licensing	7	The truck has clearly tipped over.	The truck has ever tipped over.
Quantifiers	4	There was a cat annoying Alice.	There was each cat annoying Alice.
Subject-Verb agr.	6	These casseroles disgust Kayla.	These casseroles disgusts Kayla.

3. BLiMP: Unsupervised Acceptability Judgments

Phenomenon	N	Acceptable example ✓	Unacceptable example ✗
Anaphor agreement	2	Many girls insulted themselves.	Many girls insulted herself.
Argument structure	9	Rose wasn't disturbing Mark.	Rose wasn't boasting Mark.
Binding	7	It's himself who Robert attacked.	It's himself who attacked Robert.
Control/Raising		…with.	Kevin isn't bound to work with.
Determiner-N agr.		…air.	Rachelle had bought that chairs.
Ellipsis		…portant …w.	Anne's doctor cleans one book and Stacey cleans a few important.
Filler-gap		…rs find.	Brett knew that many waiters find.
Irregular forms		…ycle.	Aaron broken the unicycle.
Island effects		…fixing?	Which is John fixing bikes?
NPI licensing		…tipped over.	The truck has ever tipped over.
Quantifiers			
Subject-Verb agr.		…gust Kayla.	These casseroles disgusts Kayla.

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