Transferring Representations of Logical Connectives

Aaron Traylor, Ellie Pavlick, & Roman Feiman

BROWN

SuperGLUE

(Wang et al., 2019)

GLUE Leaderboard

Rank	Name	Model	RTE
1	SuperGLUE Human Baselines	SuperGLUE Human Baselines	93.6
2	T5 Team - Google	T5	92.5
3	Zhuiyi Technology	RoBERTa-mtl-adv	88.1
4	Facebook AI	RoBERTa	88.2
5	IBM Research AI	BERT-mtl	84.1

(Wang et al., 2019)

GLUE Leaderboard

“Recognizing Textual Entailment”

Rank	Name	Model	RTE
1	SuperGLUE Human Baselines	SuperGLUE Human Baselines	93.6
2	T5 Team - Google	T5	92.5
3	Zhuiyi Technology	RoBERTa-mtl-adv	88.1
4	Facebook AI	RoBERTa	88.2
5	IBM Research AI	BERT-mtl	84.1

(Wang et al., 2019)

Natural Language Inference Example

Premise

Natural Language Inference Example

Premise

Either he has a blind trust or he has a conflict of interest.

Natural Language Inference Example

Premise

Either he has a blind trust or he has a conflict of interest.

Hypothesis

Natural Language Inference Example

Premise

Either he has a blind trust or he has a conflict of interest.

Hypothesis

He has a conflict of interest.

Natural Language Inference Example

Disjunction

Premise

Either he has a blind trust or he has a conflict of interest.

Hypothesis

He has a conflict of interest.

GLUE Inference Diagnostics

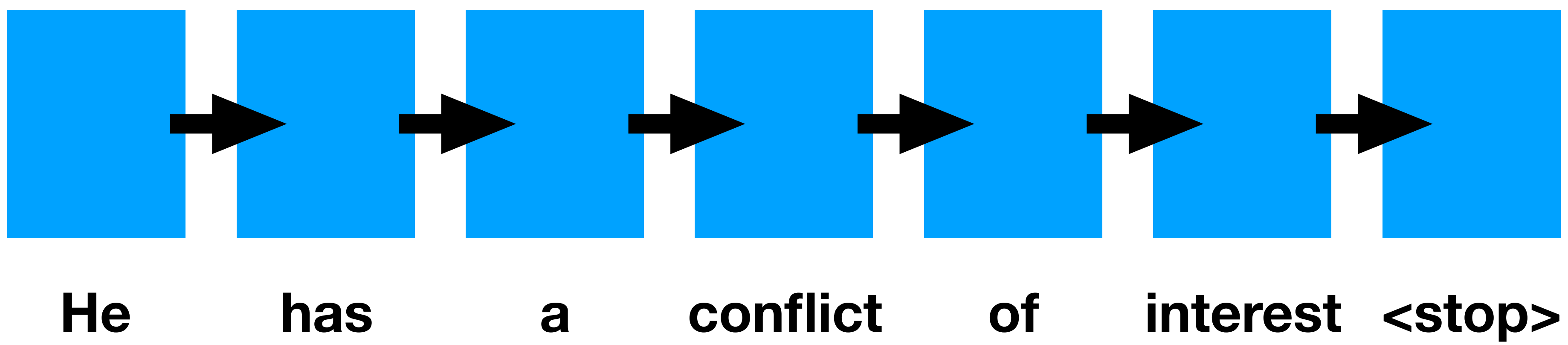
GLUE Inference Diagnostics

- Format
- Design
  - Standards for entailment
  - Handling Coreference
  - Definite Descriptions and Monotonicity
  - Background Knowledge
- Linguistic Categorization
  - Lexical Semantics
    - Lexical Entailment
    - Morphological Negation
    - Factivity
    - Symmetry/Collectivity
    - Redundancy
    - Named Entities
    - Quantifiers
  - Predicate-Argument Structure
    - Syntactic Ambiguity
    - Prepositional Phrases
    - Core Arguments
    - Alternations
    - Ellipsis/Implicits
    - Anaphora/Coreference
    - Intersectivity
    - Restrictivity
  - Logic
    - Propositional Structure
    - Quantification
    - Monotonicity
    - Richer Logical Structure
  - Knowledge and Common sense
    - World Knowledge
    - Common Sense

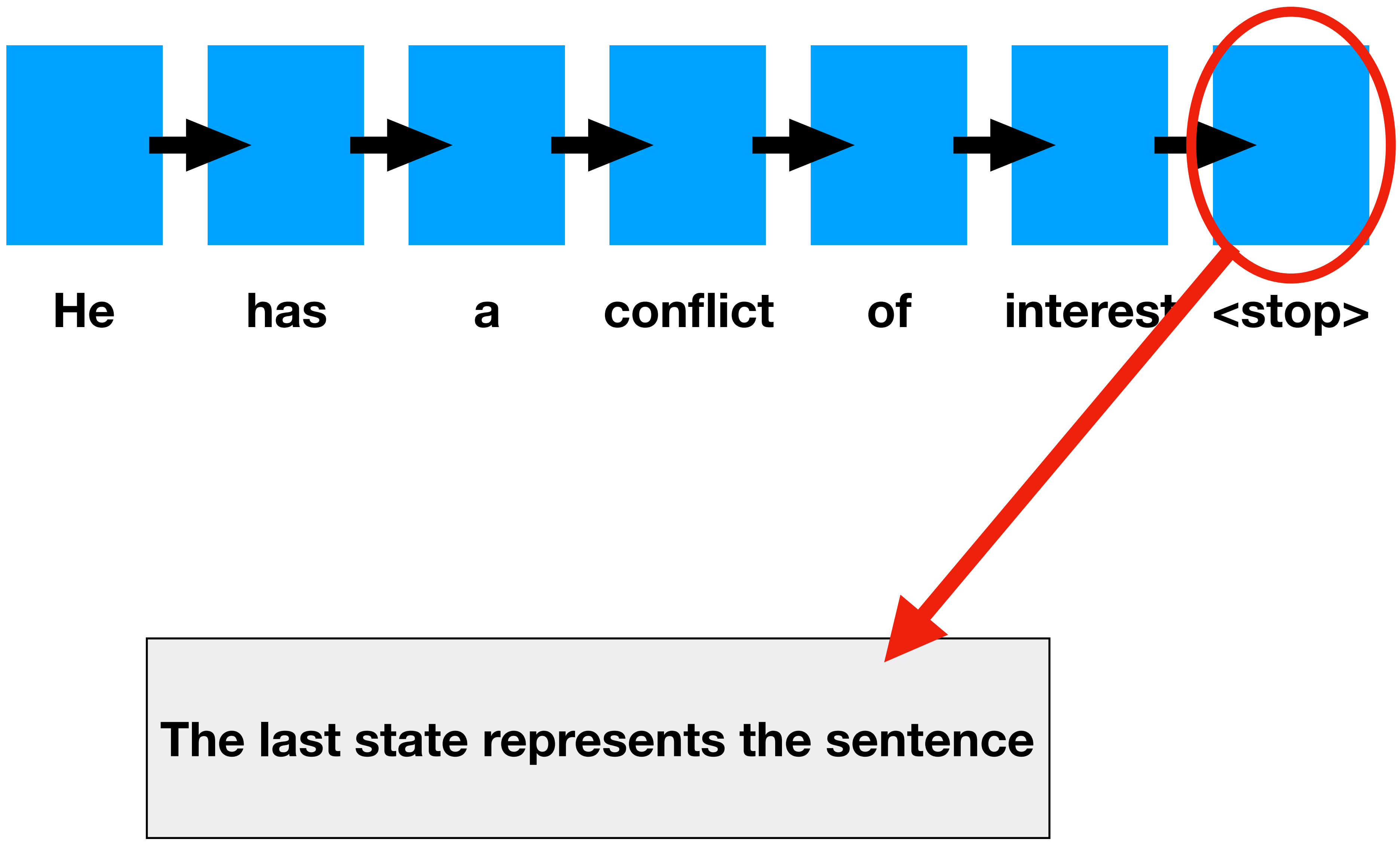
Sentence Representation

He has a conflict of interest <stop>

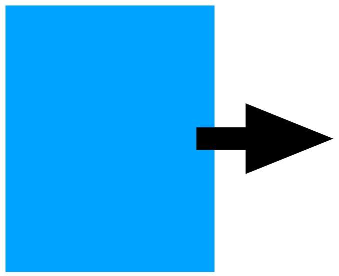
Sentence Representation



Sentence Representation

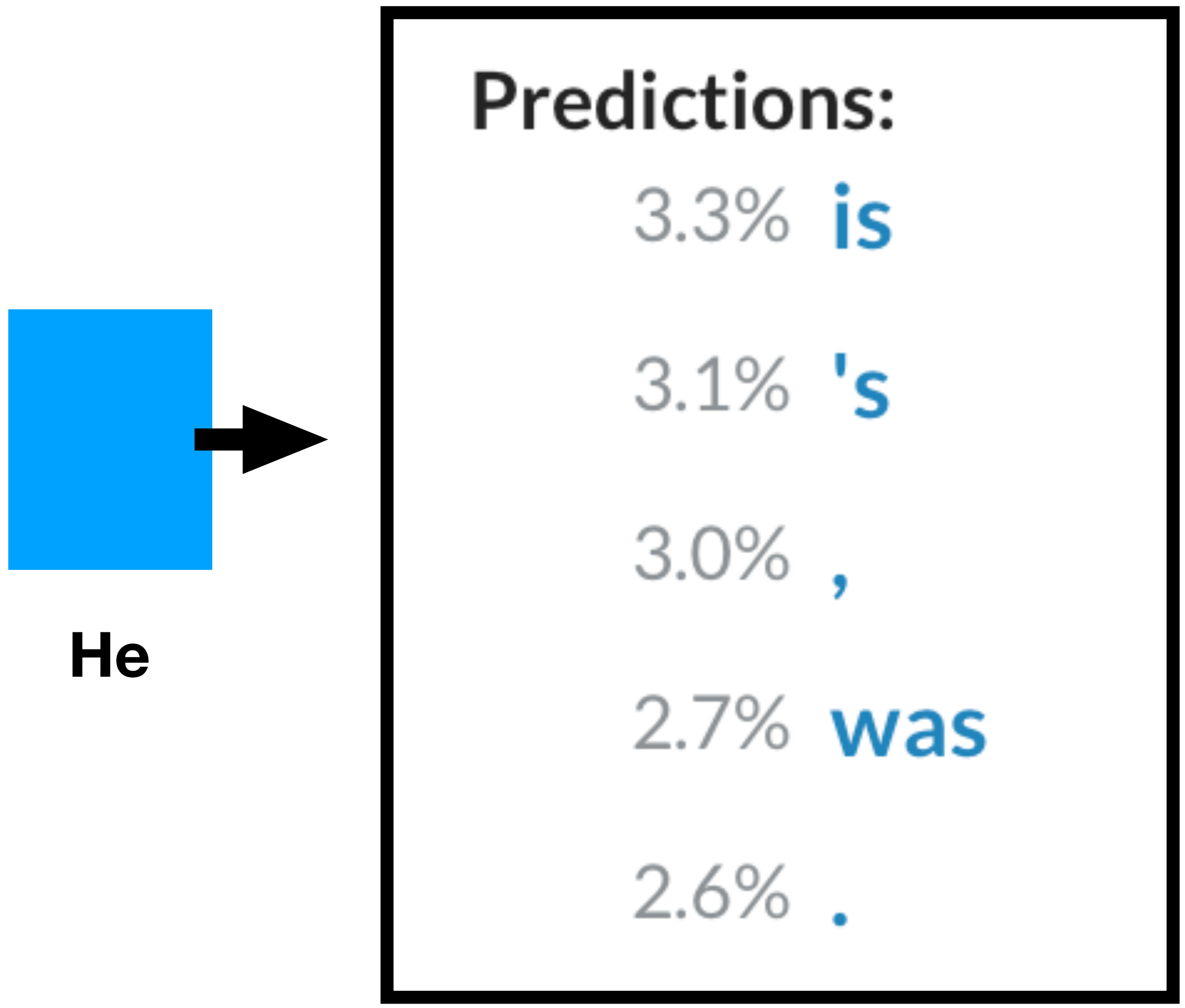


Language Modeling



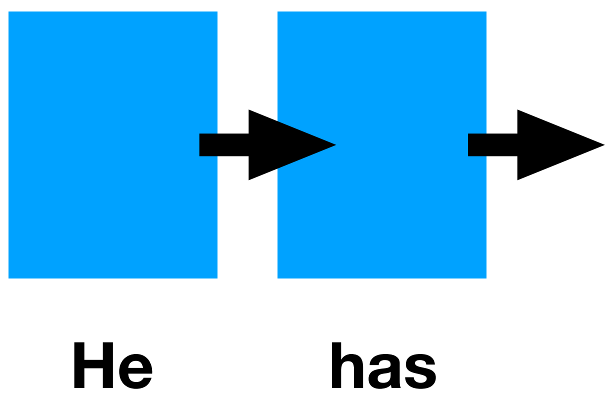
He

Language Modeling

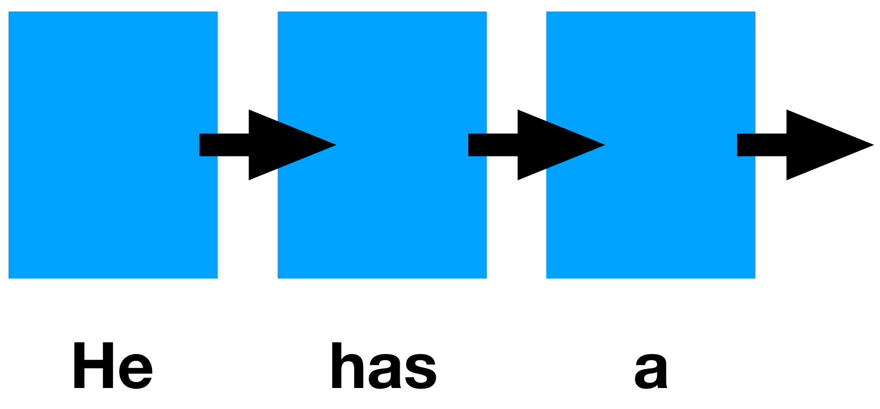


https://demo.allennlp.org/next-token-lm

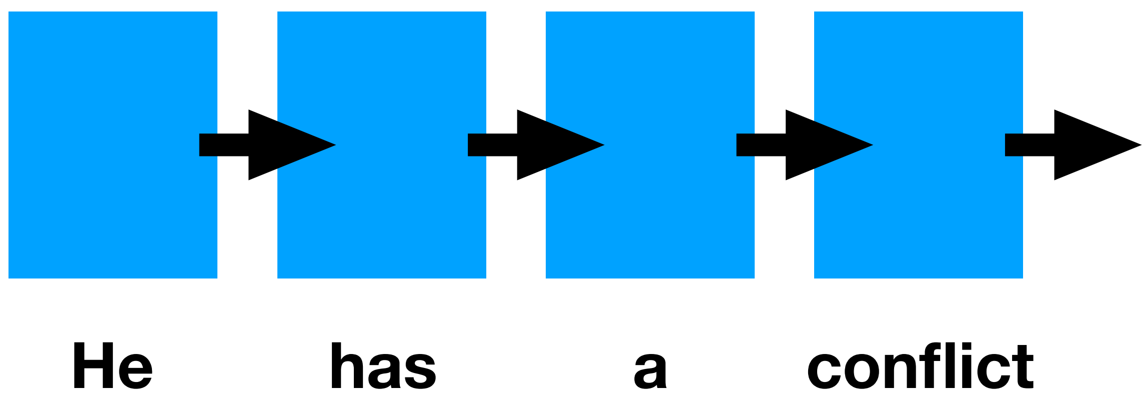
Language Modeling



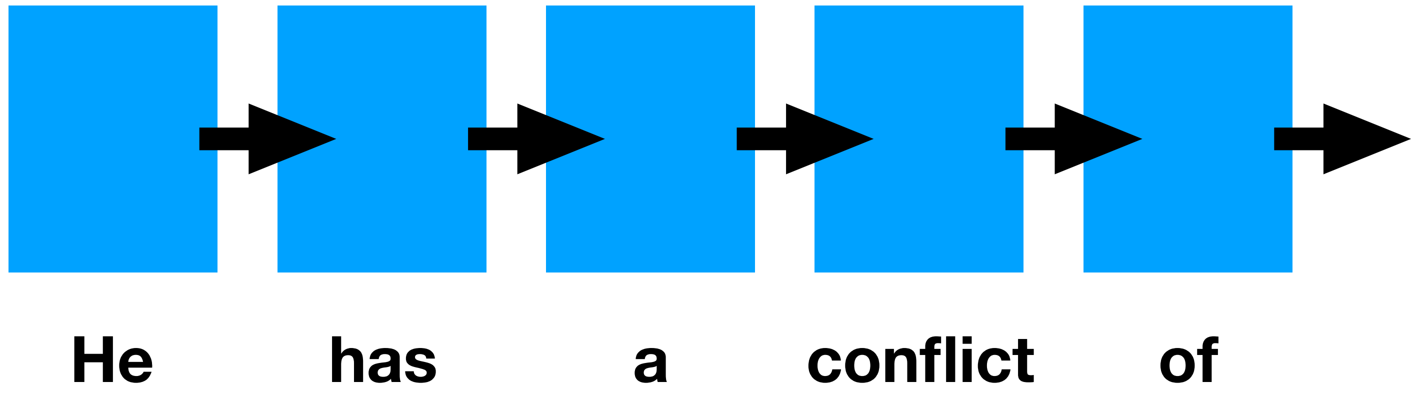
Language Modeling



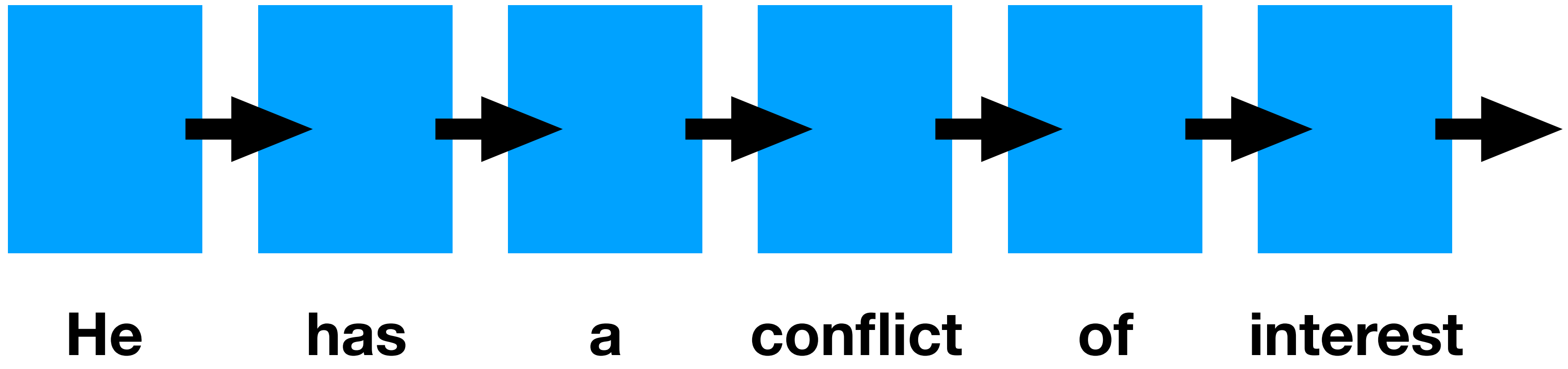
Language Modeling



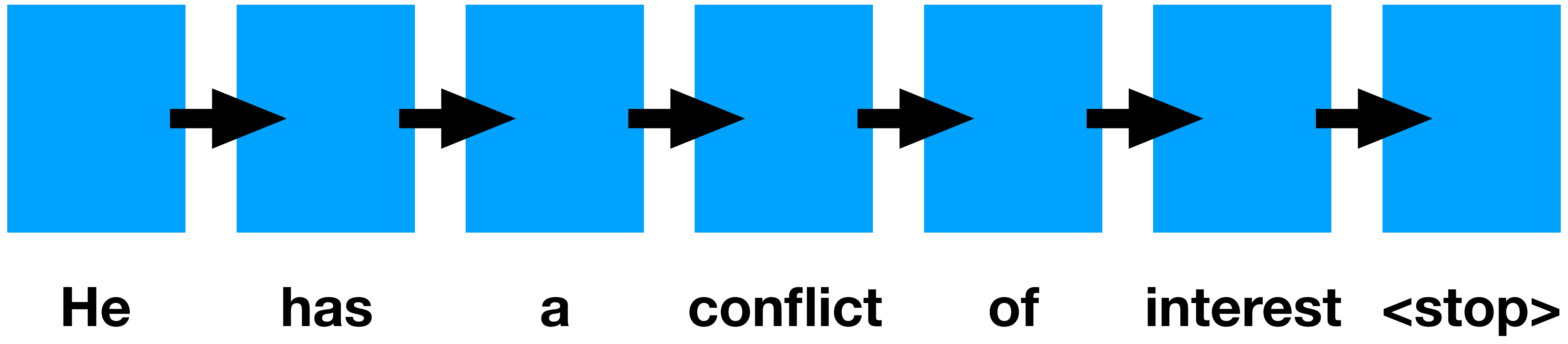
Language Modeling



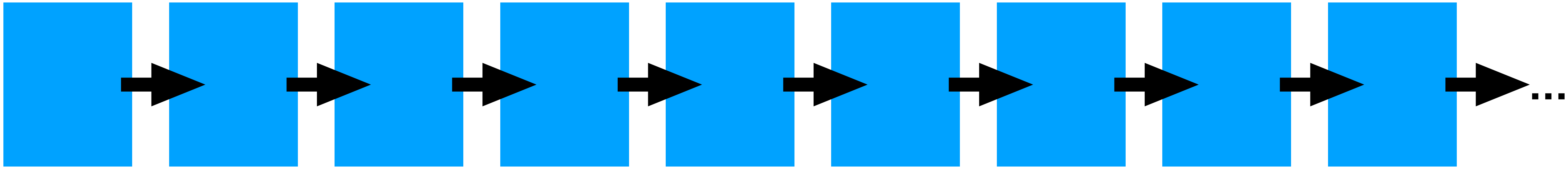
Language Modeling

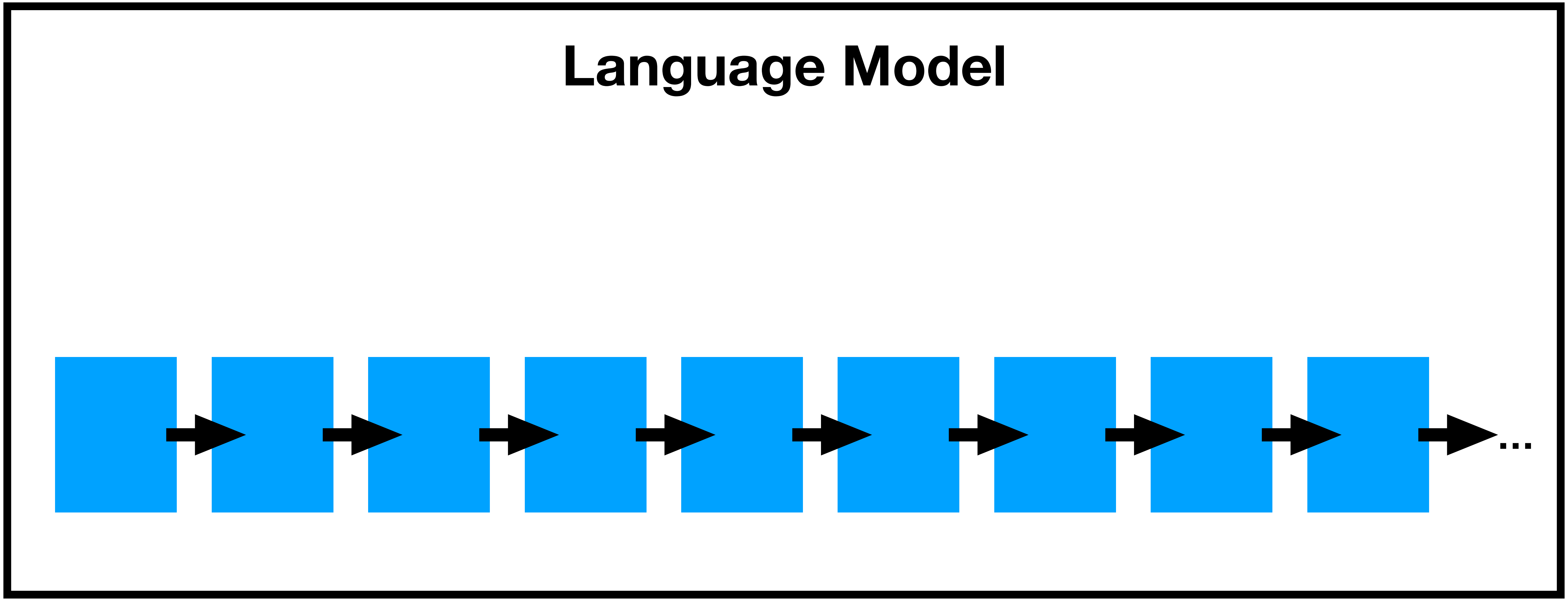


Language Modeling



Language Modeling




Language Model
...

Outline

Motivation

Experimental Design

Results

Discussion

Outline

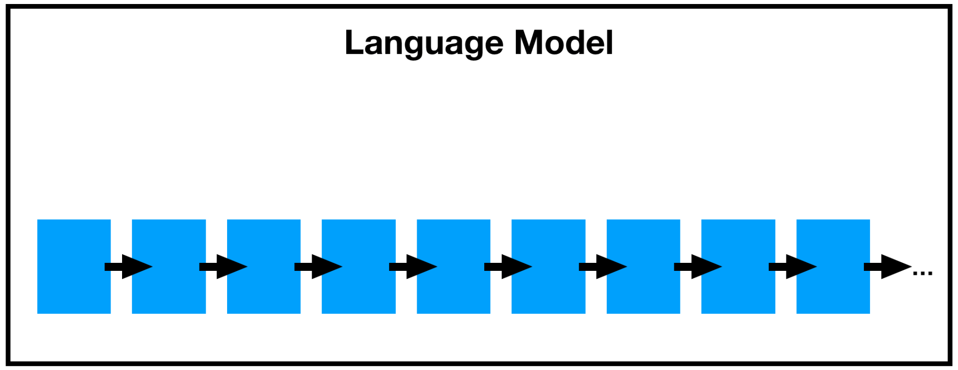
Motivation

Experimental Design

Results

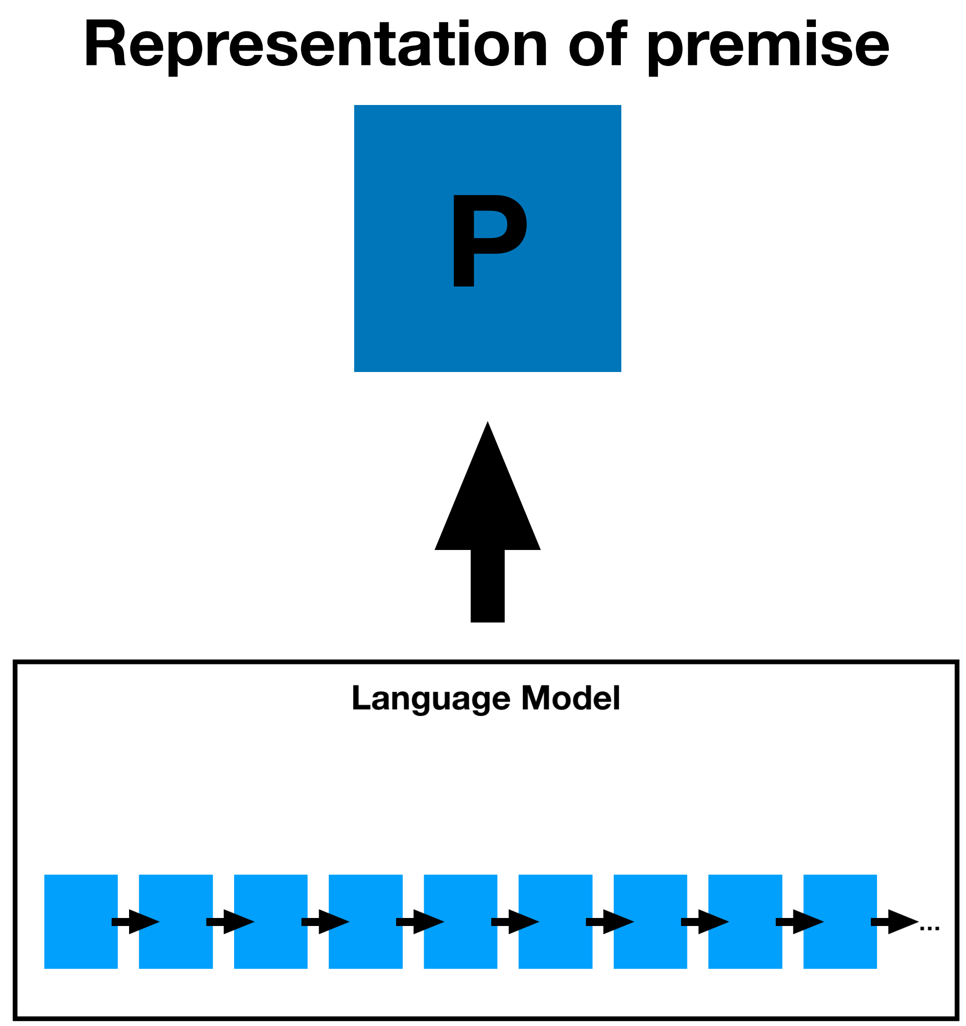
Discussion

Transferring Language Model Representations



A street filled with people walking and riding bikes.

Transferring Language Model Representations



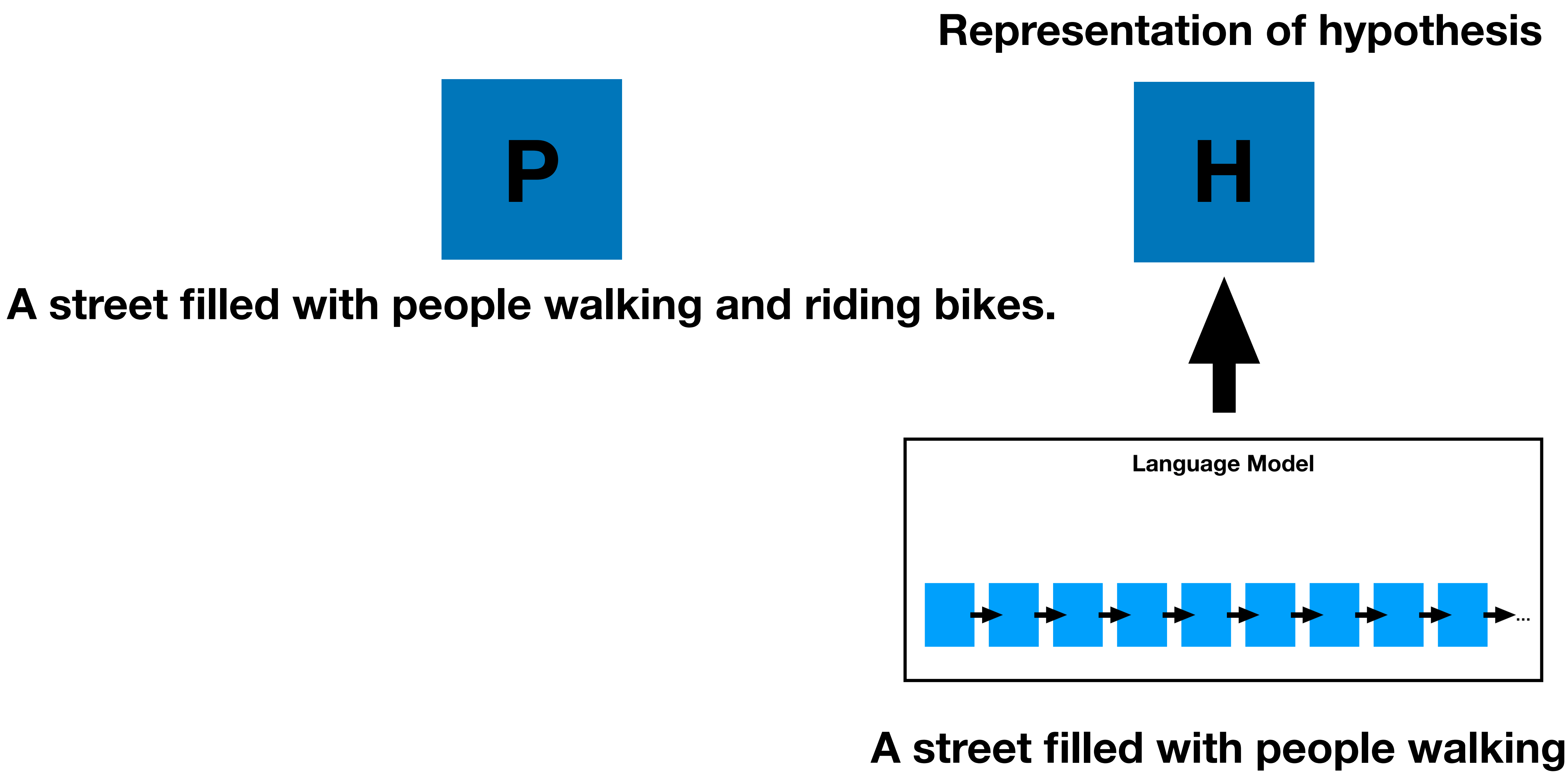
A street filled with people walking and riding bikes.

Transferring Language Model Representations

P

A street filled with people walking and riding bikes.

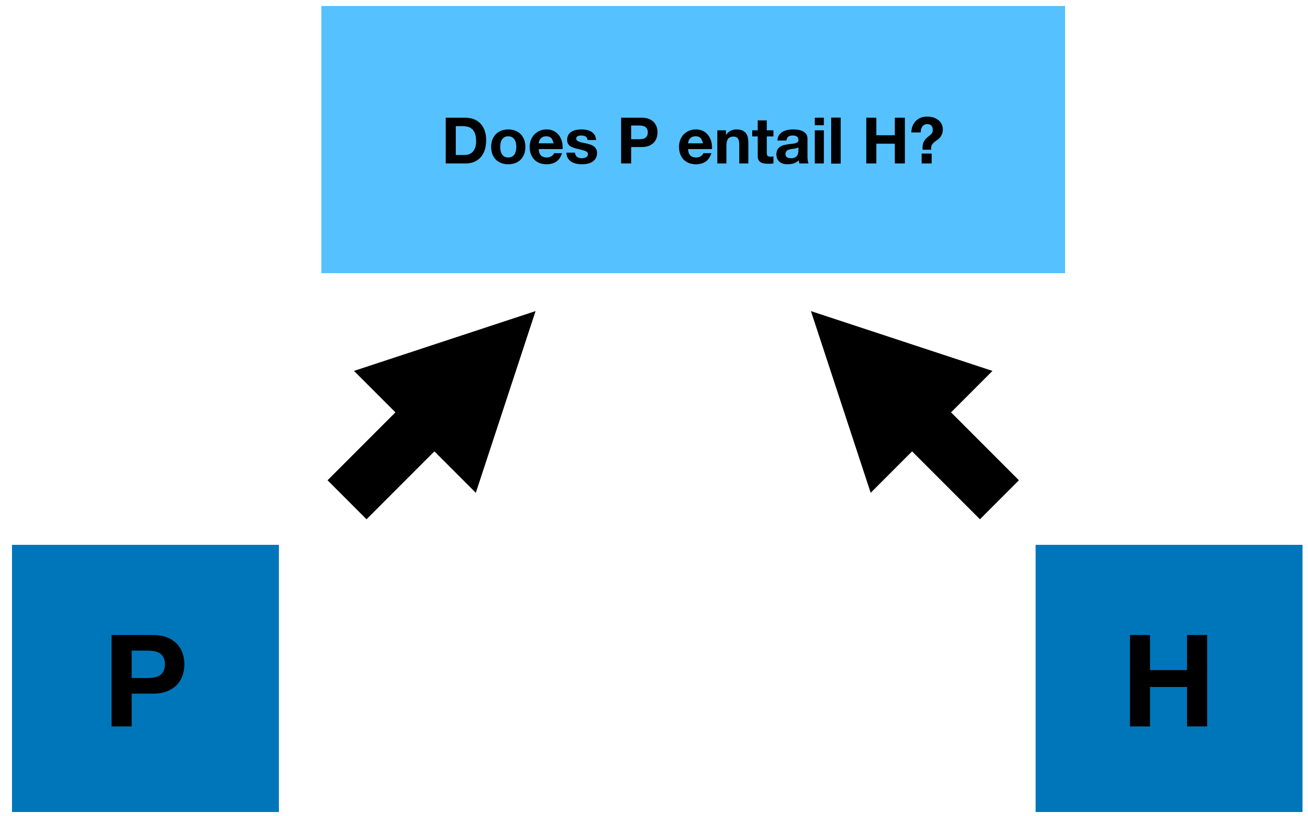
Transferring Language Model Representations



Transferring Language Model Representations

A street filled with people walking and riding bikes.

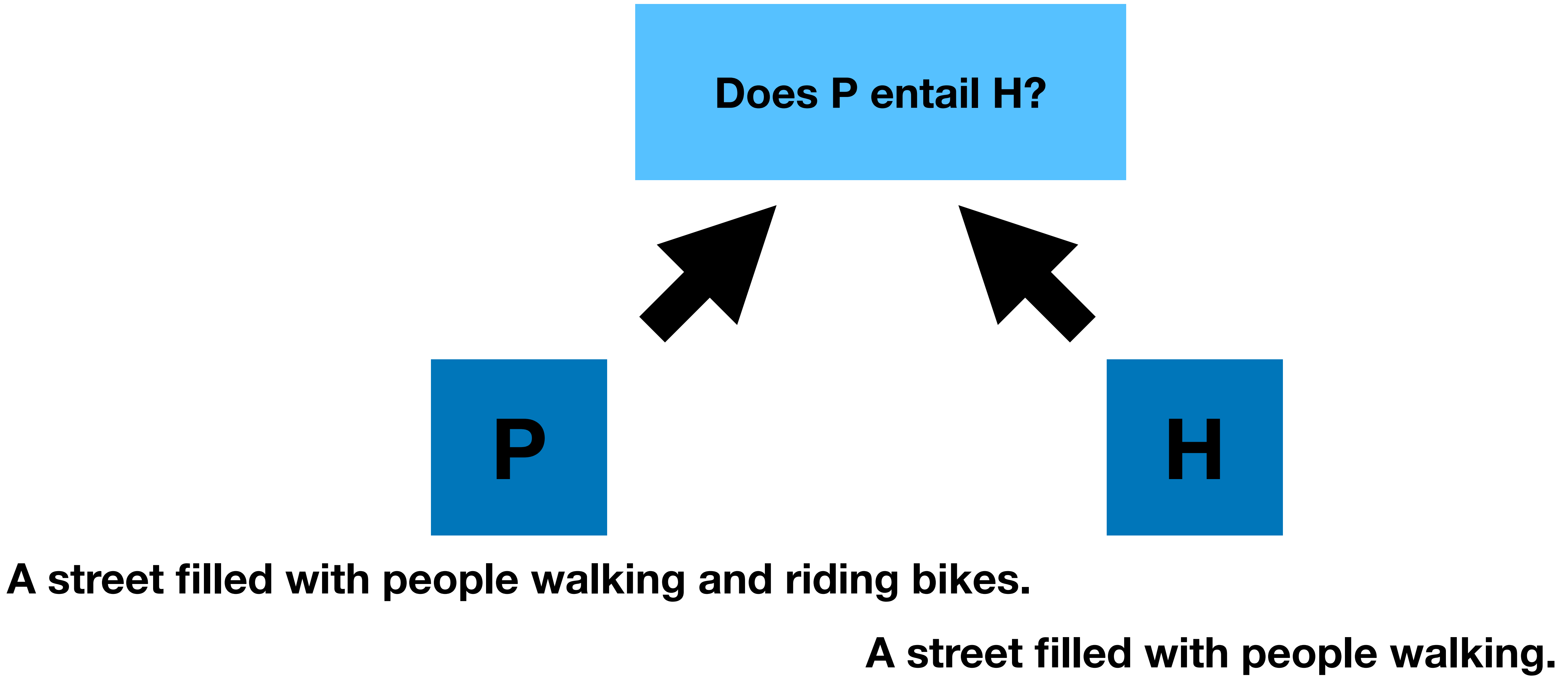
A street filled with people walking.

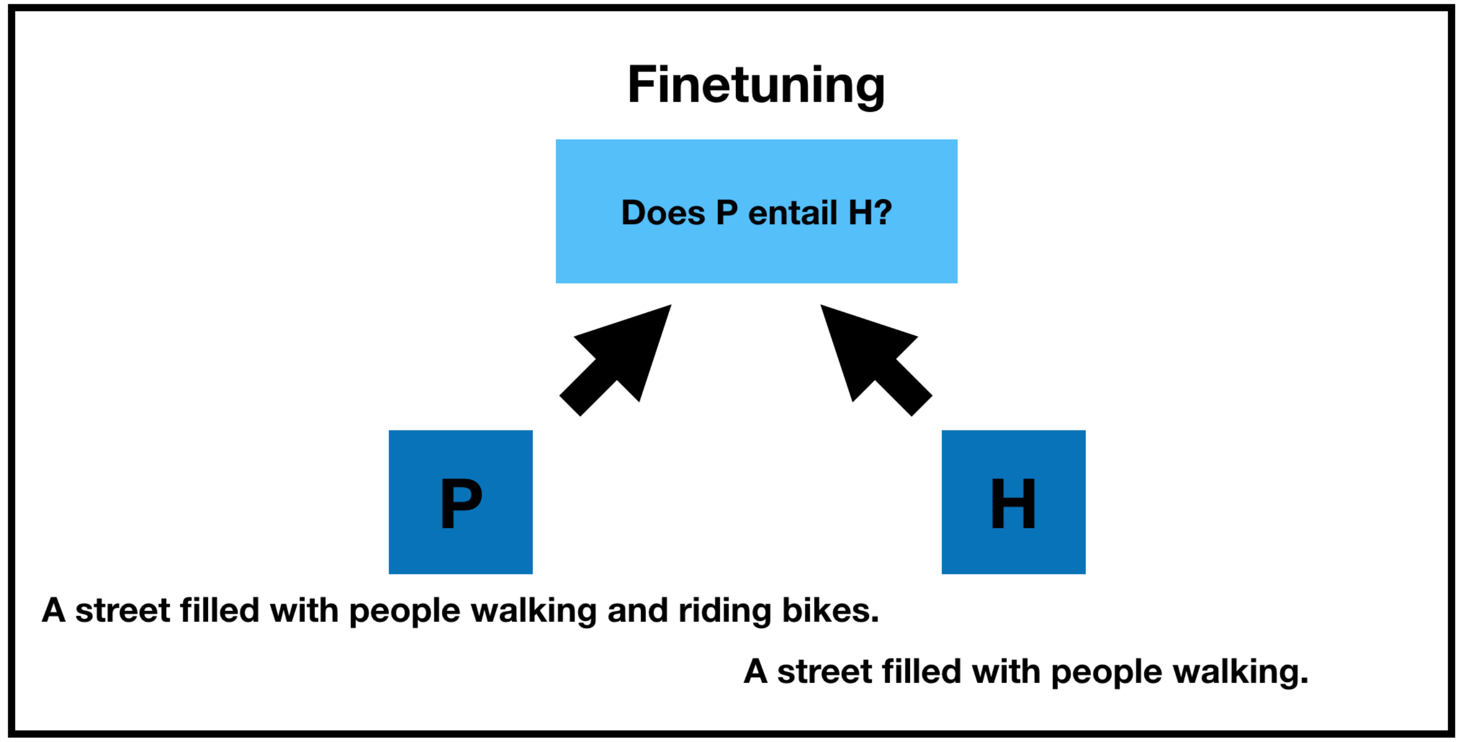


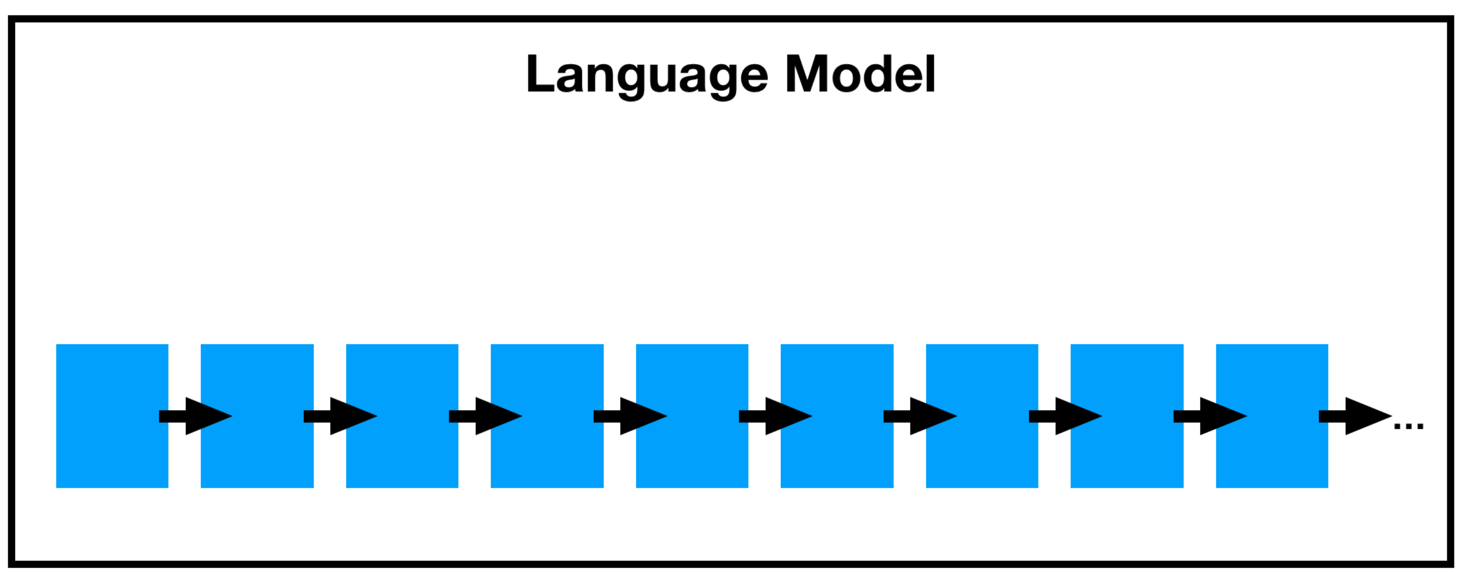
A street filled with people walking and riding bikes.

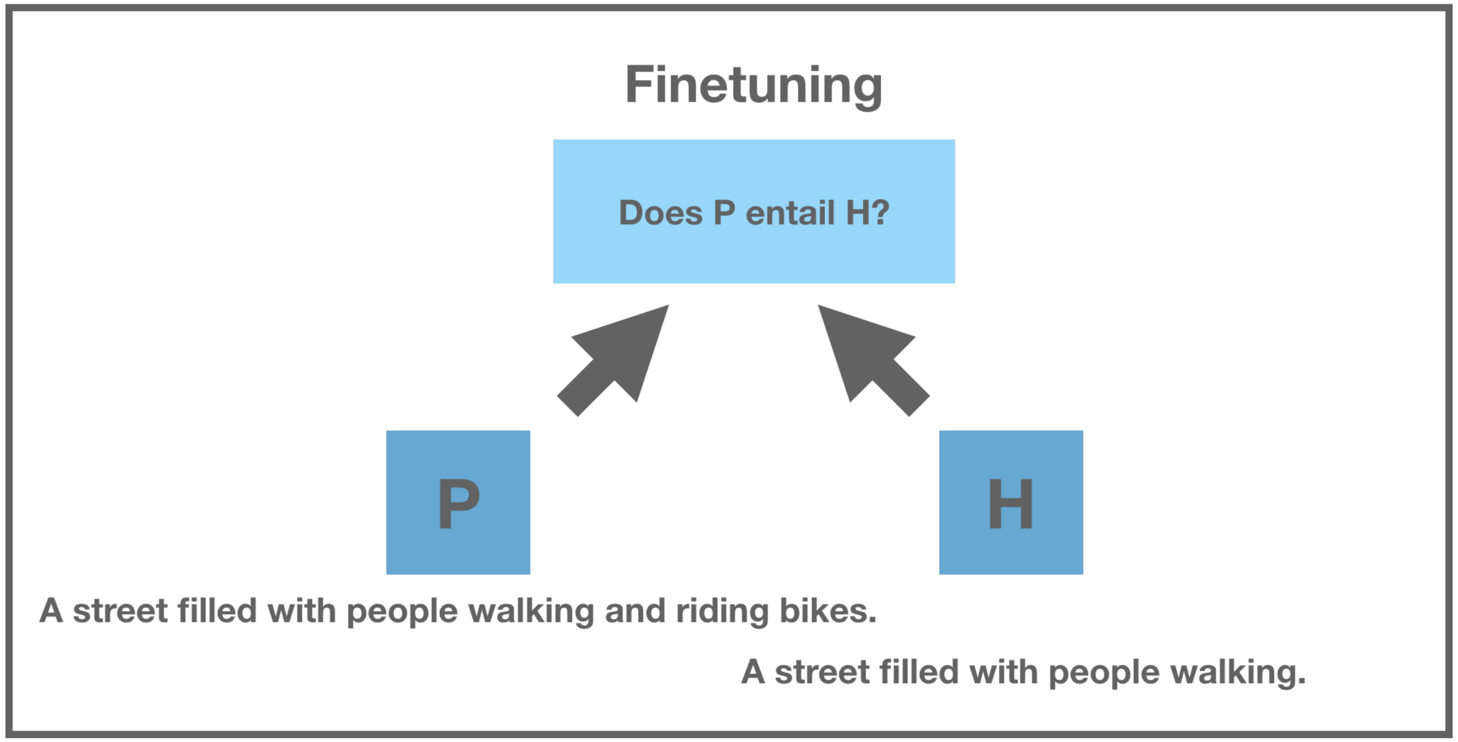
A street filled with people walking.

Finetuning

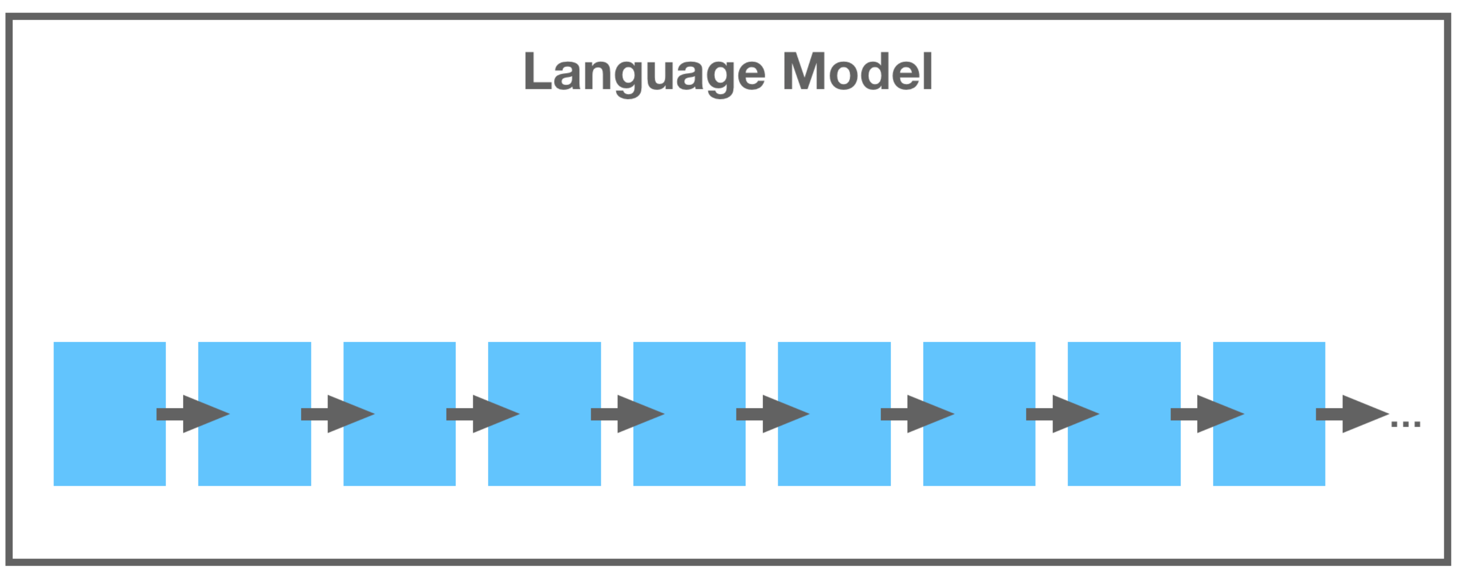

Does P entail H?
P
H
A street filled with people walking and riding bikes.
A street filled with people walking.


Finetuning
Does P entail H?
P
H
A street filled with people walking and riding bikes.
A street filled with people walking.

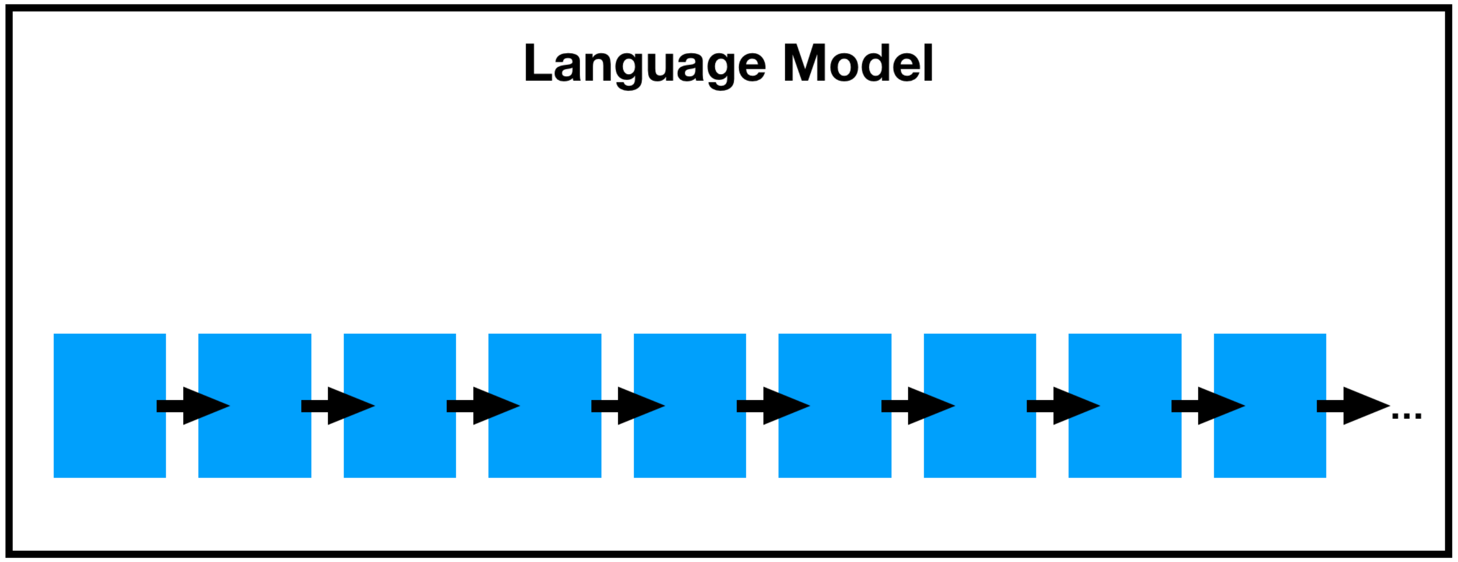

Language Model
...



Why would logical reasoning emerge from this process?

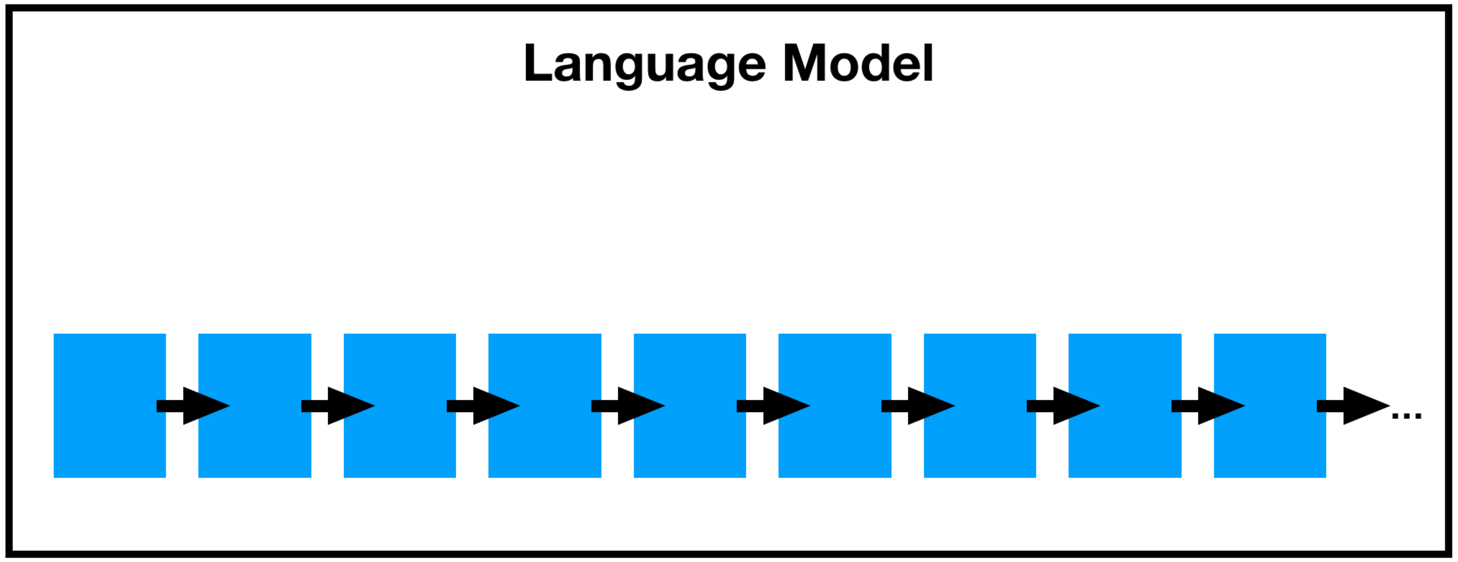


Reasons Logic Might Not Emerge From Pretraining-Finetuning



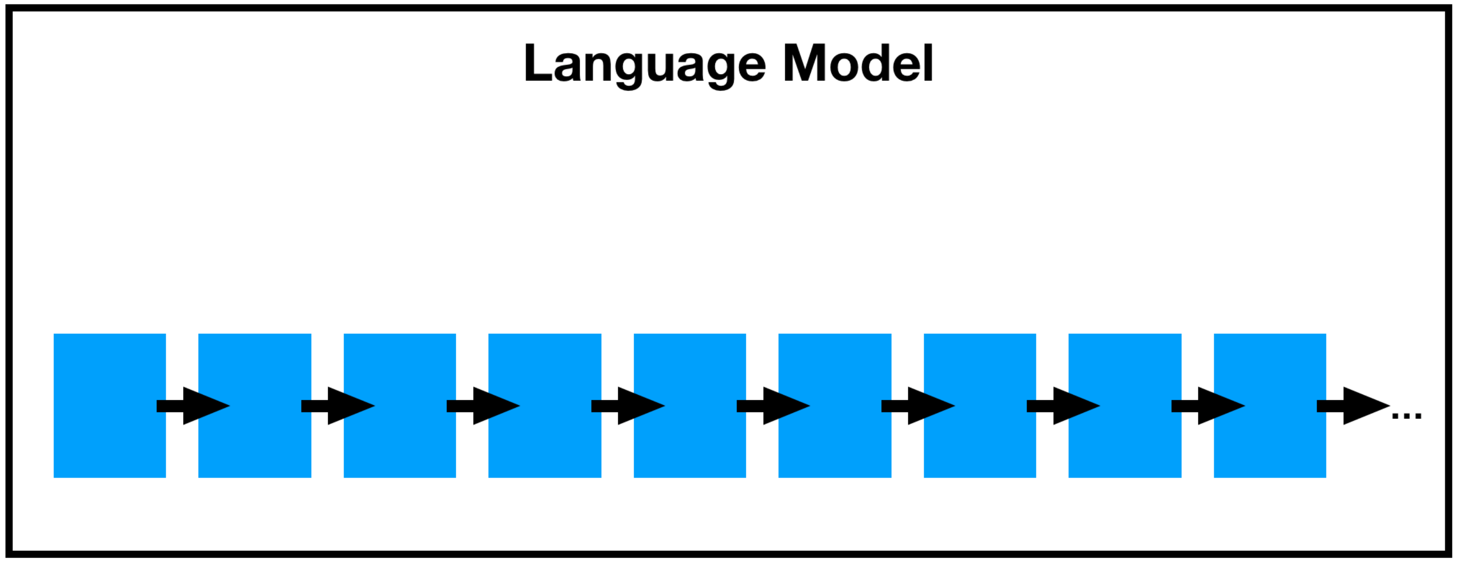
1.

Reasons Logic Might Not Emerge From Pretraining-Finetuning



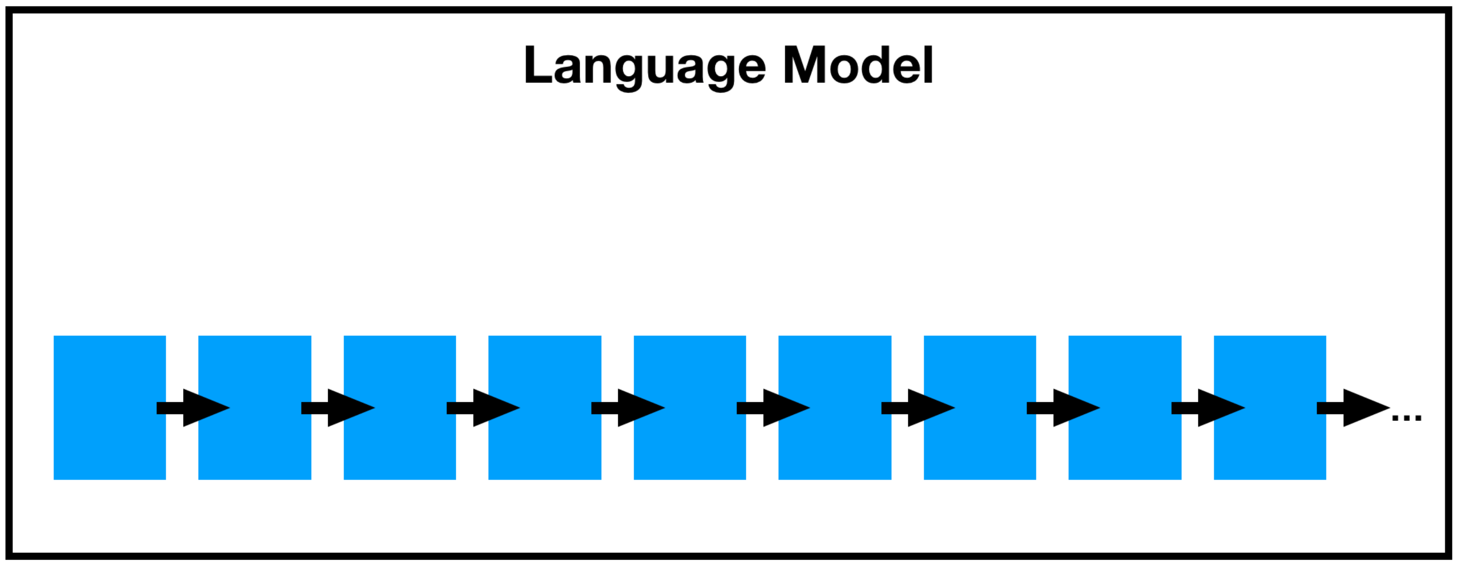
1. Language models do not access truth assignments.

Reasons Logic Might Not Emerge From Pretraining-Finetuning



A street filled with people walking and riding bikes.

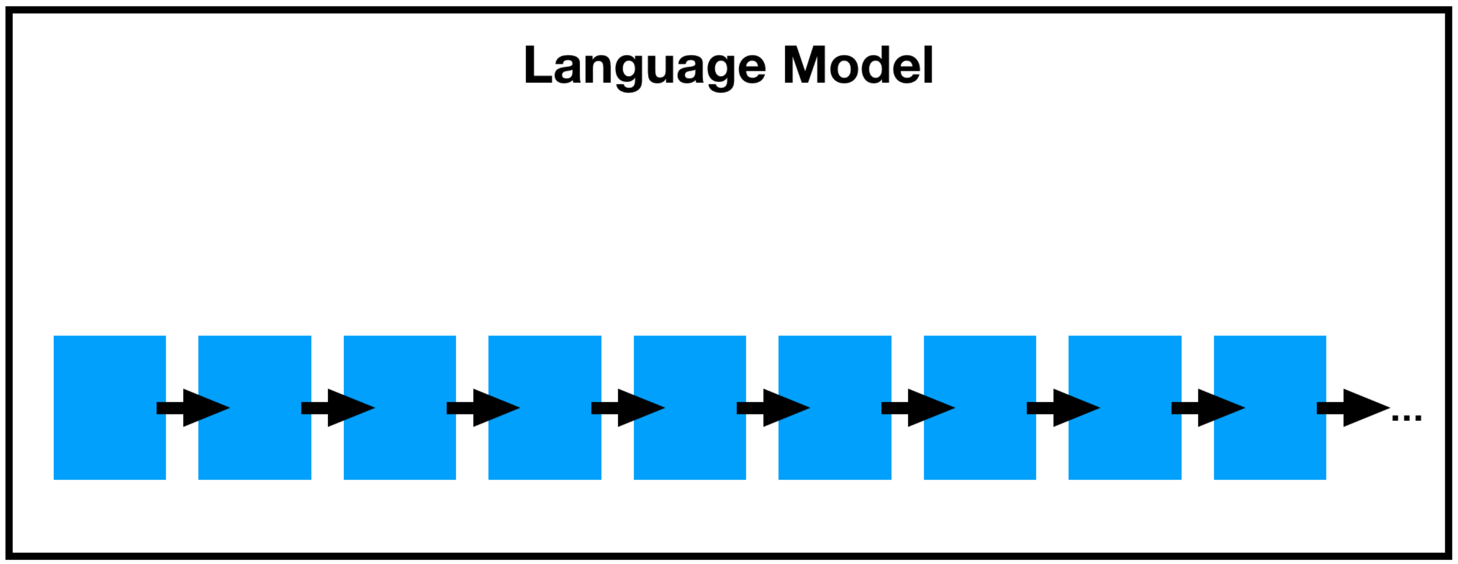
Reasons Logic Might Not Emerge From Pretraining-Finetuning



A street filled with people walking and riding bikes.

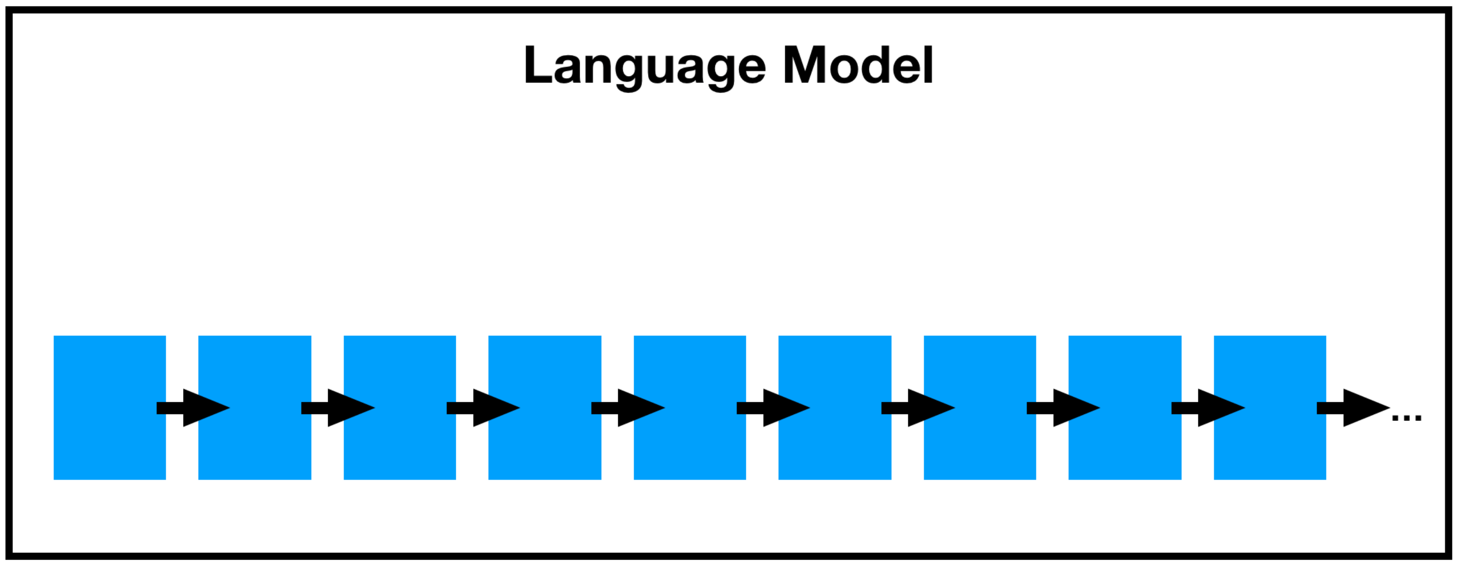
There is a street.	True
There are shops.	True
There are people.	True
There is a giraffe.	False

Reasons Logic Might Not Emerge From Pretraining-Finetuning



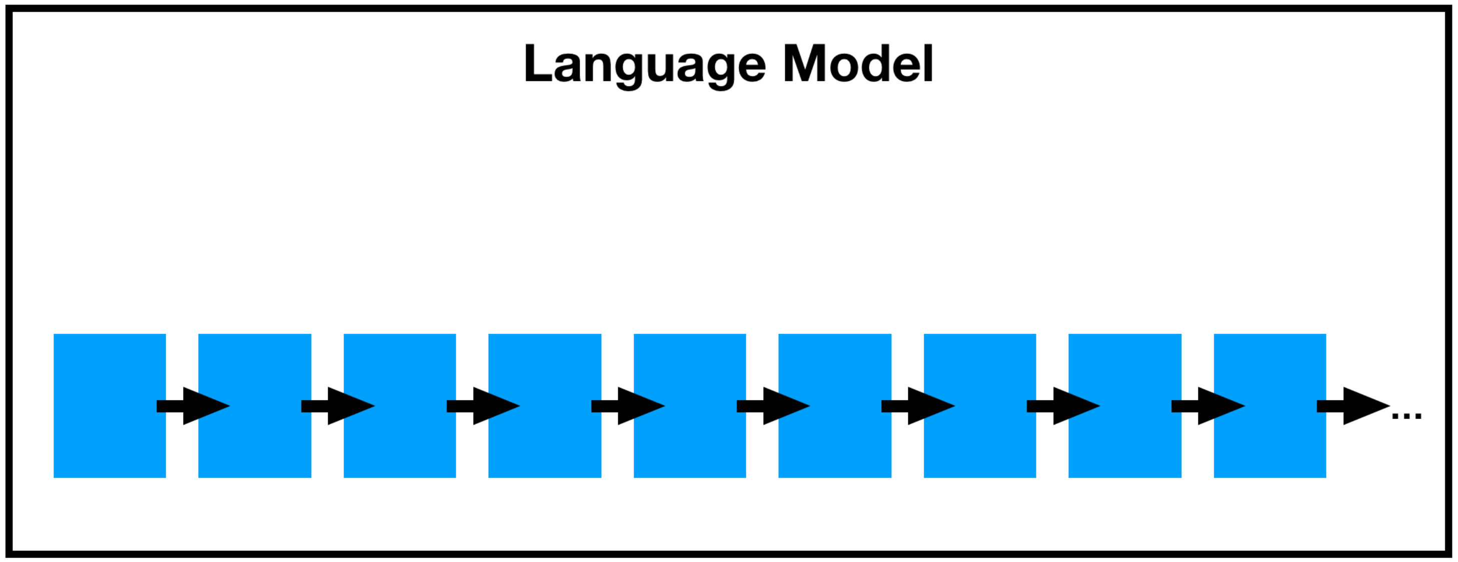
1. Language models do not access truth assignments.
2.

Reasons Logic Might Not Emerge From Pretraining-Finetuning



1. Language models do not access truth assignments.
2. Language models only ever observe “positive” examples.

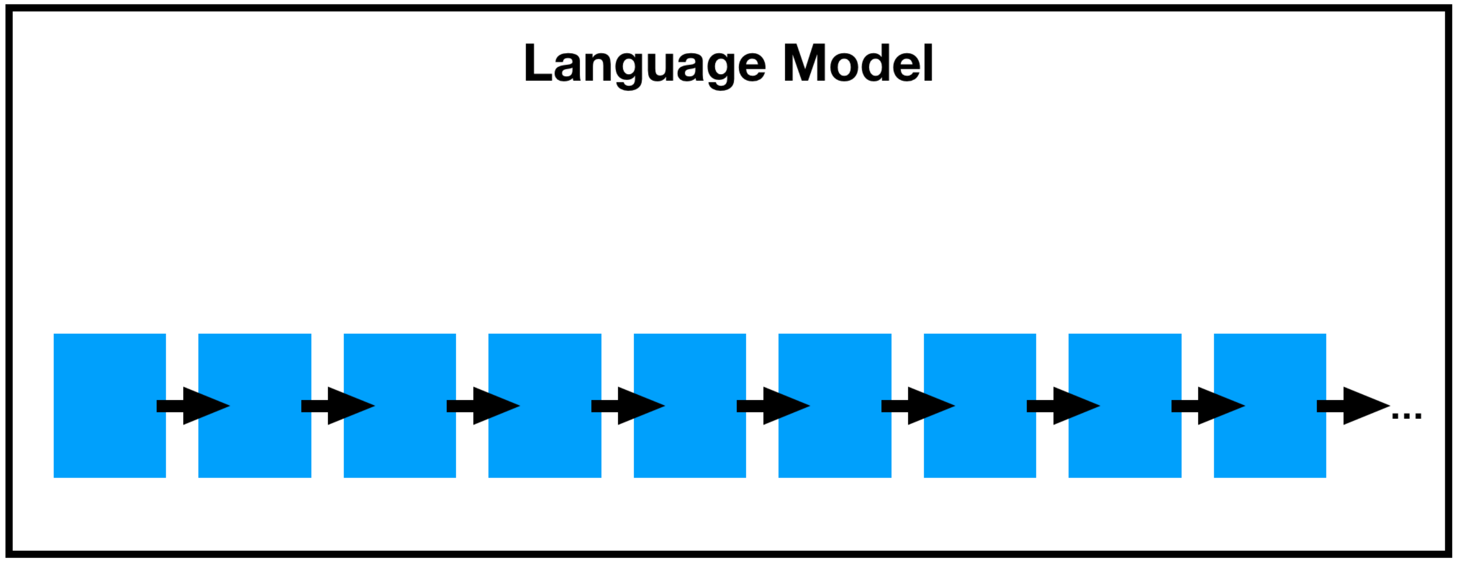
Reasons Logic Might Not Emerge From Pretraining-Finetuning



2. Language models only ever observe "positive" examples.

A street filled with people walking and riding bikes.

Reasons Logic Might Not Emerge From Pretraining-Finetuning



2. Language models only ever observe “positive” examples.

A street filled with people walking and riding bikes.

~~The street is full of people and not full of people.~~

Research Questions

Research Questions

- Can language modeling allow representations of logical reasoning to emerge?

Research Questions

- Can language modeling allow representations of logical reasoning to emerge?
- Because we expect the answer to be no, what modifications would allow emergence?

Research Questions

- Can language modeling allow representations of logical reasoning to emerge?
- Because we expect the answer to be no, what modifications would allow emergence?
  - Observing truth assignments?

Research Questions

- Can language modeling allow representations of logical reasoning to emerge?
- Because we expect the answer to be no, what modifications would allow emergence?
  - Observing truth assignments?
  - Observing “negative” examples?

Outline

Motivation

Experimental Design

Results

Discussion

Our Framework

We create a dataset of propositional logic sentences

Our Framework

We create a dataset of propositional logic sentences because:

NLI datasets are biased

	Entailment		Neutral		Contradiction	
SNLI	outdoors	2.8%	tall	0.7%	nobody	0.1%
	least	0.2%	first	0.6%	sleeping	3.2%
	instrument	0.5%	competition	0.7%	no	1.2%
	outside	8.0%	sad	0.5%	tv	0.4%
	animal	0.7%	favorite	0.4%	cat	1.3%
MNLI	some	1.6%	also	1.4%	never	5.0%
	yes	0.1%	because	4.1%	no	7.6%
	something	0.9%	popular	0.7%	nothing	1.4%
	sometimes	0.2%	many	2.2%	any	4.1%
	various	0.1%	most	1.8%	none	0.1%

(Gururangan et al., 2018)

NLI datasets are biased

	Entailment		Neutral		Contradiction	
SNLI	outdoors	2.8%	tall	0.7%	nobody	0.1%
	least	0.2%	first	0.6%	sleeping	3.2%
	instrument	0.5%	competition	0.7%	no	1.2%
	outside	8.0%	sad	0.5%	tv	0.4%
	animal	0.7%	favorite	0.4%	cat	1.3%
MNLI	some	1.6%	also	1.4%	never	5.0%
	yes	0.1%	because	4.1%	no	7.6%
	something	0.9%	popular	0.7%	nothing	1.4%
	sometimes	0.2%	many	2.2%	any	4.1%
	various	0.1%	most	1.8%	none	0.1%

Strong indicators of contradiction

(Gururangan et al., 2018)

NLI datasets are biased

Hypothesis-only score: 67% SNLI, 53% MNLI

(Gururangan et al., 2018)

Our Framework

We create a dataset of propositional logic sentences because:

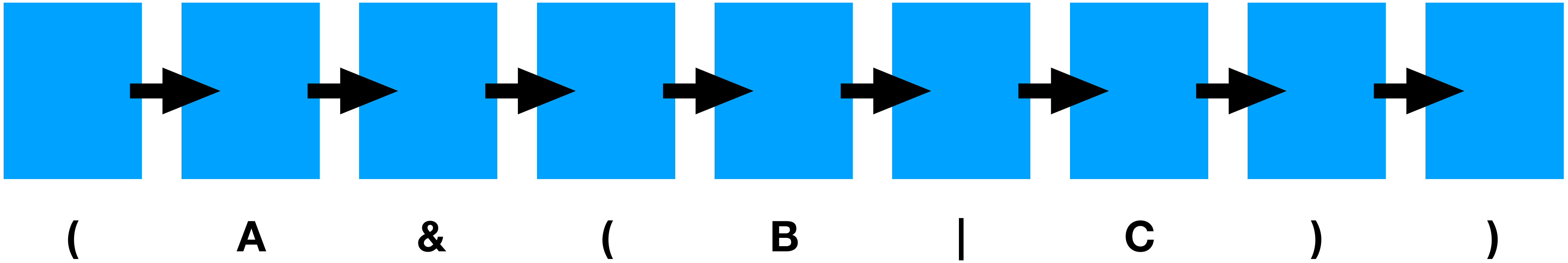
- Full control: minimize dataset bias

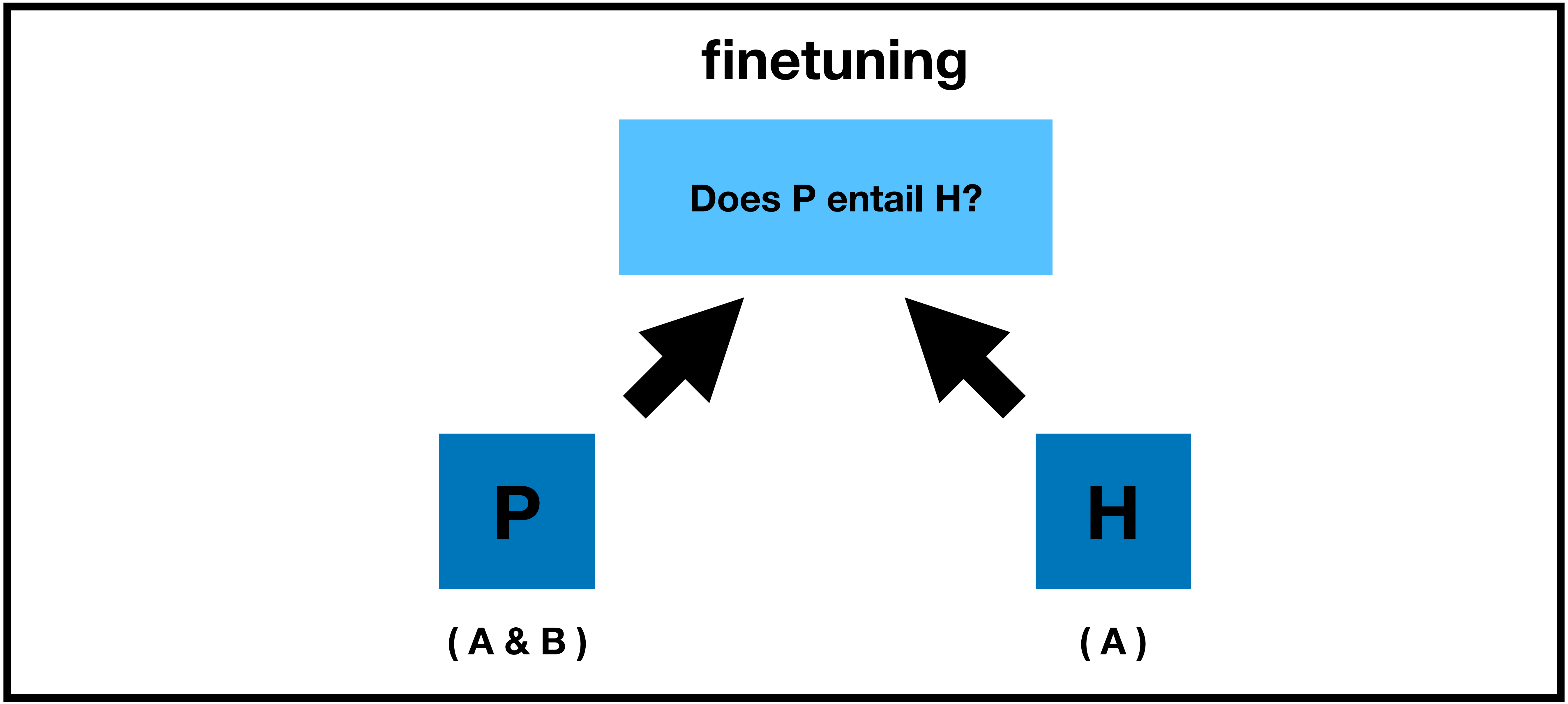
Our Framework

We create a dataset of propositional logic sentences because:

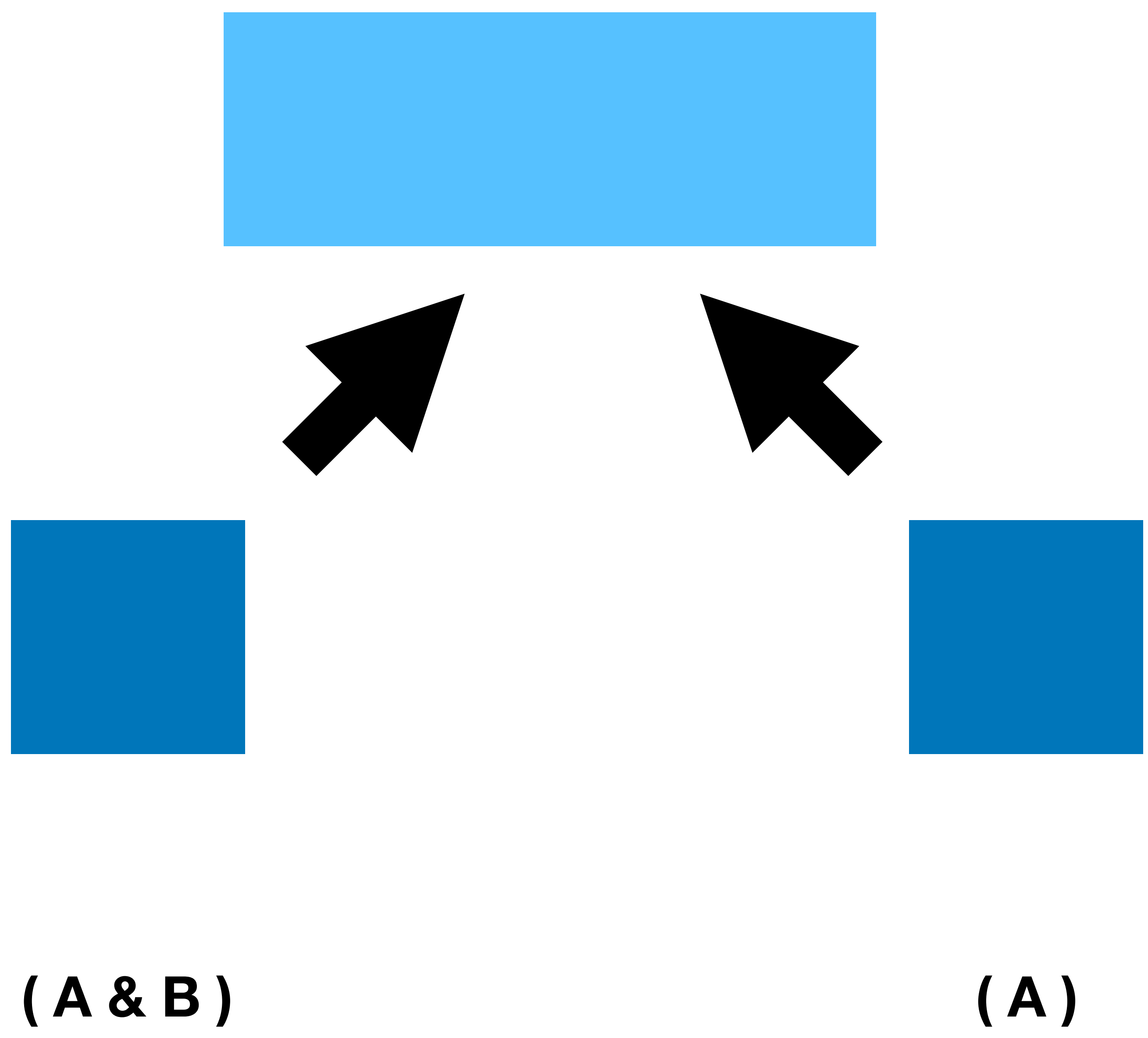
- Full control: minimize dataset bias
- No lexical priors / pragmatic effects to exploit: only logical information

language modeling

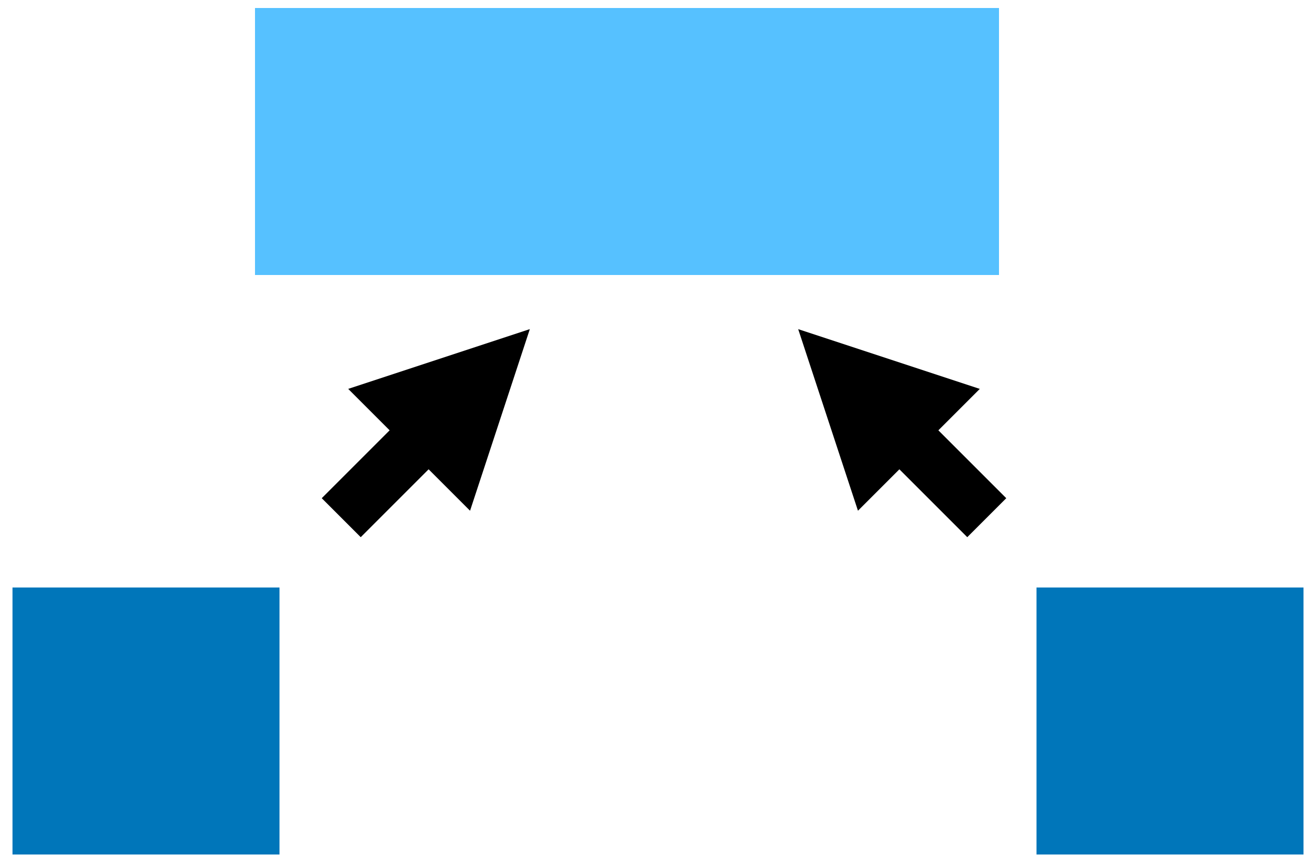



finetuning
Does P entail H?
P
H
(A & B)
(A)

Finetuning Architecture

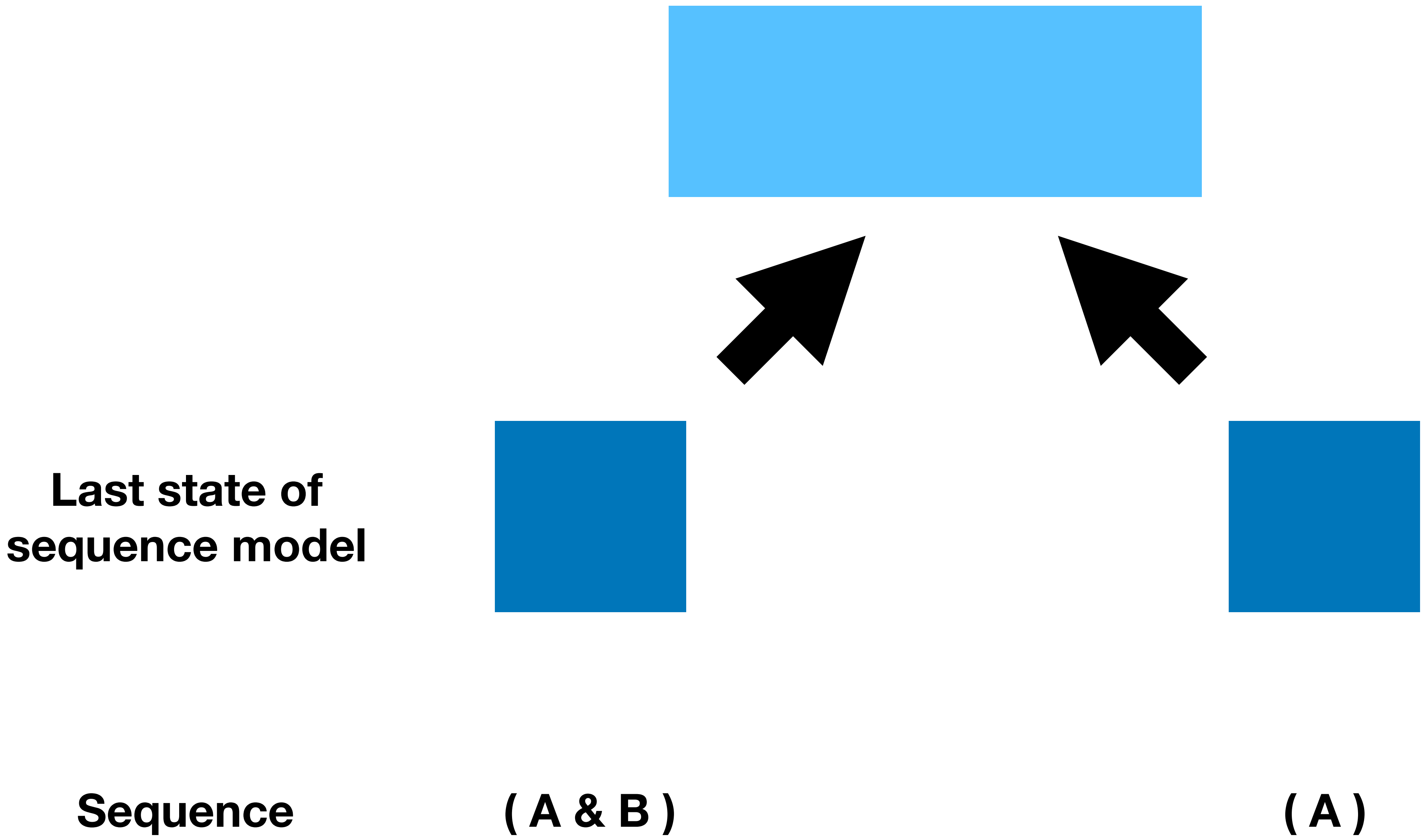


Finetuning Architecture

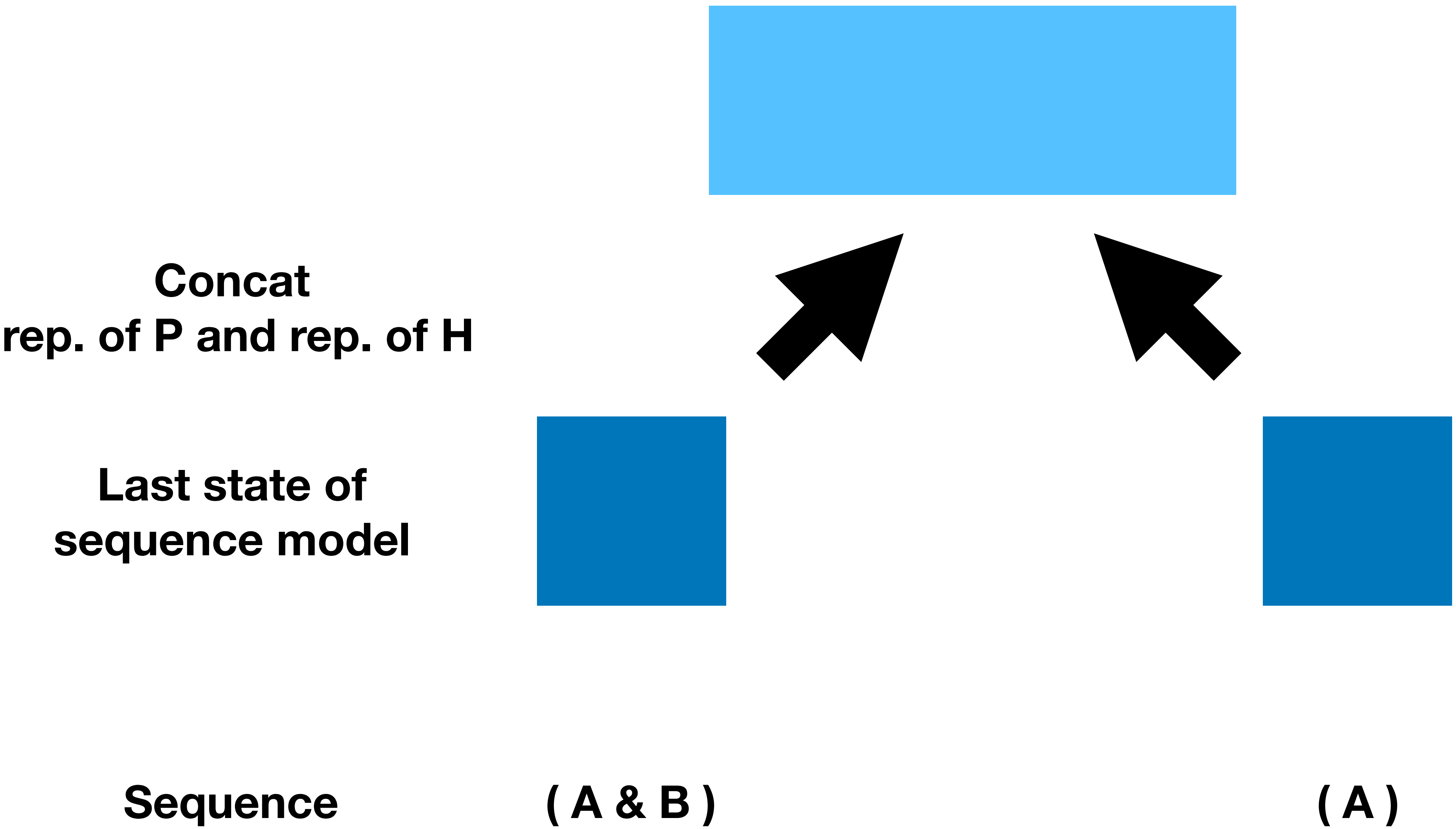


Sequence (A & B) (A)

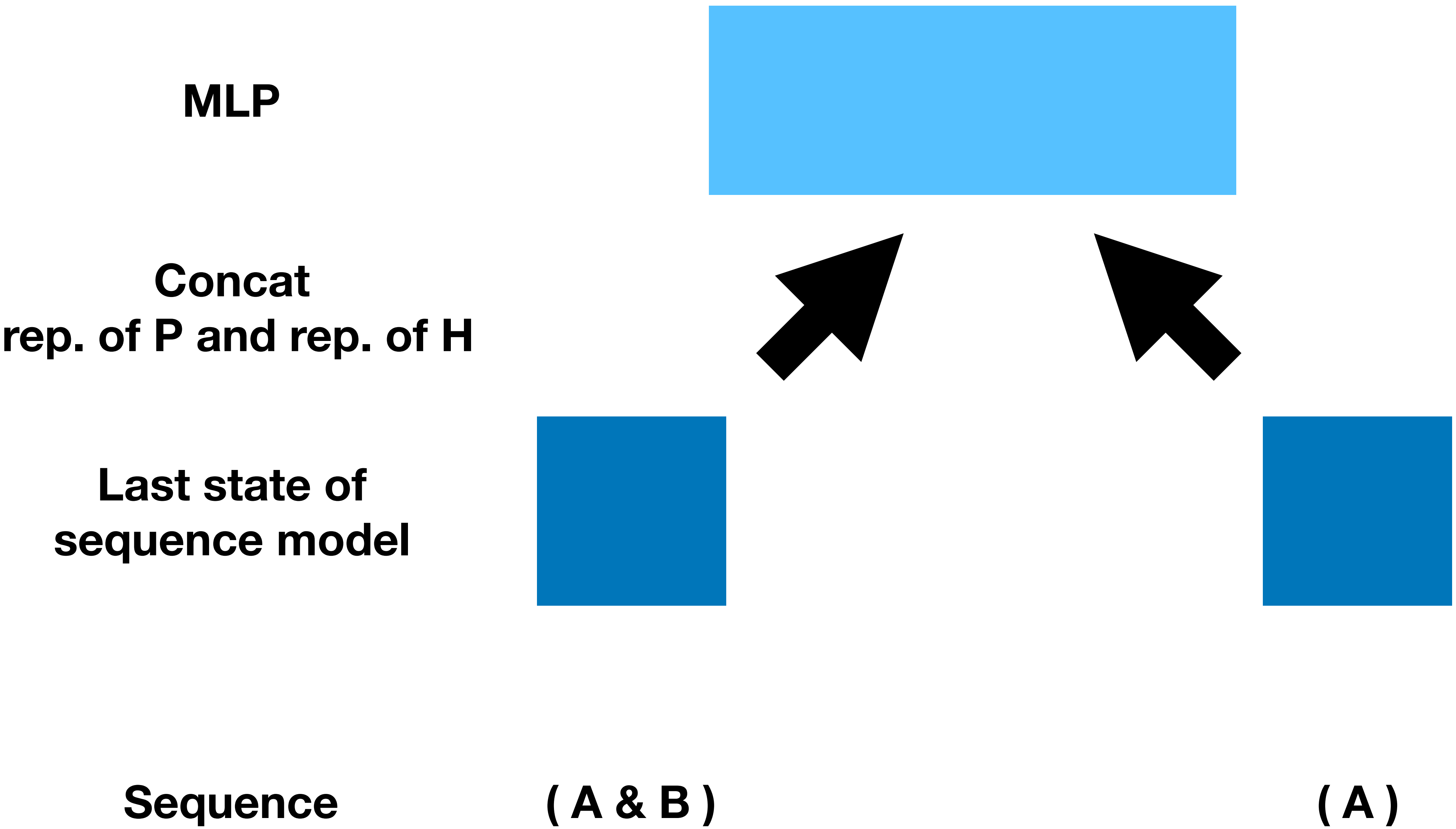
Finetuning Architecture



Finetuning Architecture



Finetuning Architecture



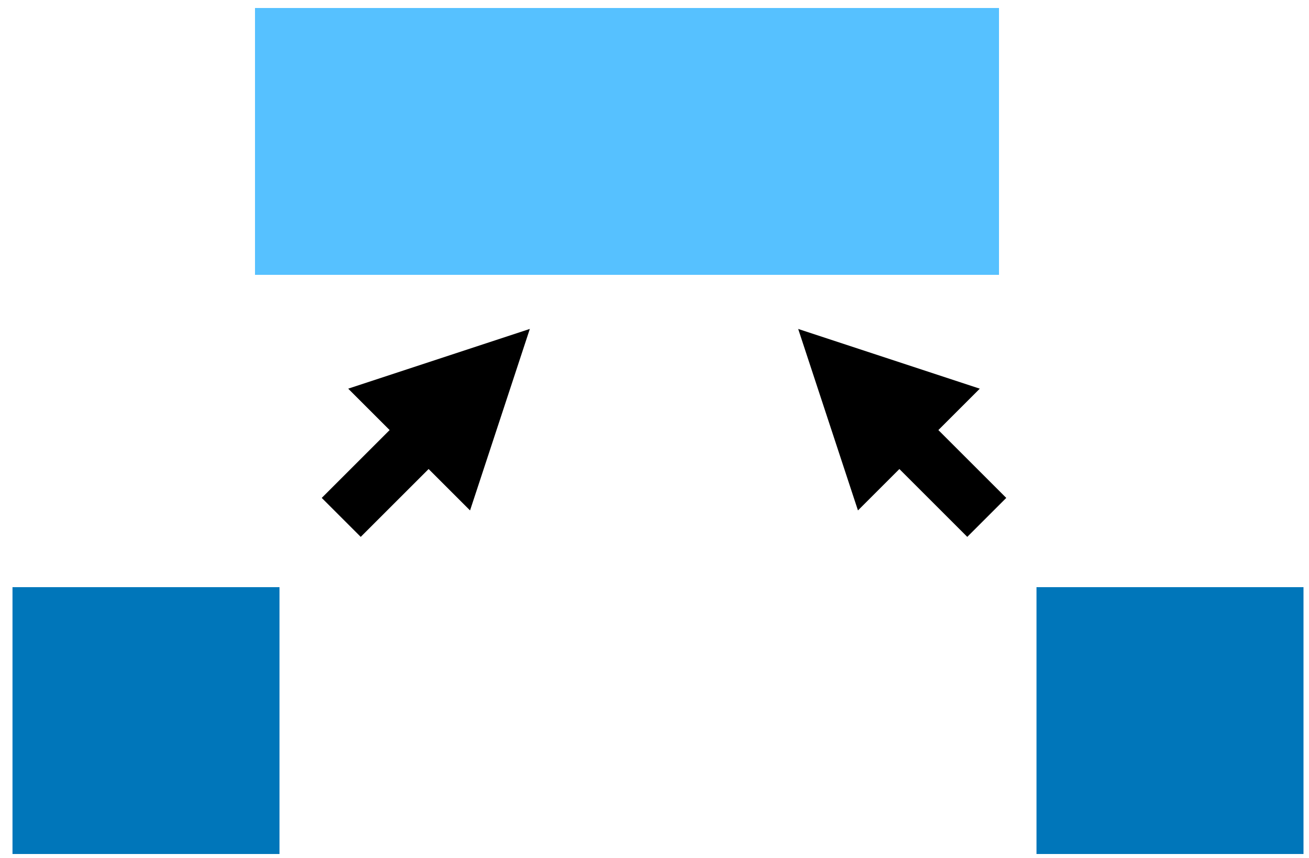
Finetuning Architecture

binary classification

MLP

Concat
rep. of P and rep. of H

Last state of
sequence model



Sequence

(A & B)

(A)

Dataset Statistics

Unary logical operators	Negation
Binary logical operators	Conjunction, disjunction, conditional

Dataset Statistics

Unary logical operators	Negation
Binary logical operators	Conjunction, disjunction, conditional
# of “variables”	30,000
Sentences in language modeling training dataset	500,000
Sentences in validation set	50,000

Dataset Statistics

Unary logical operators	Negation
Binary logical operators	Conjunction, disjunction, conditional
# of “variables”	30,000
Sentences in language modeling training dataset	500,000
Sentences in validation set	50,000
Sentences in entailment training dataset	100,000
Sentences in validation set	5,000

Our Framework

Entailment Dataset Balance

Generate premise,hypothesis pairs (a1,a2),(b1,b2) such that:

a1	entails	a2
b1	entails	b2
a1	does not entail	b2
b1	does not entail	a2

(Evans et al., 2018)

Our Framework

Entailment Dataset Balance

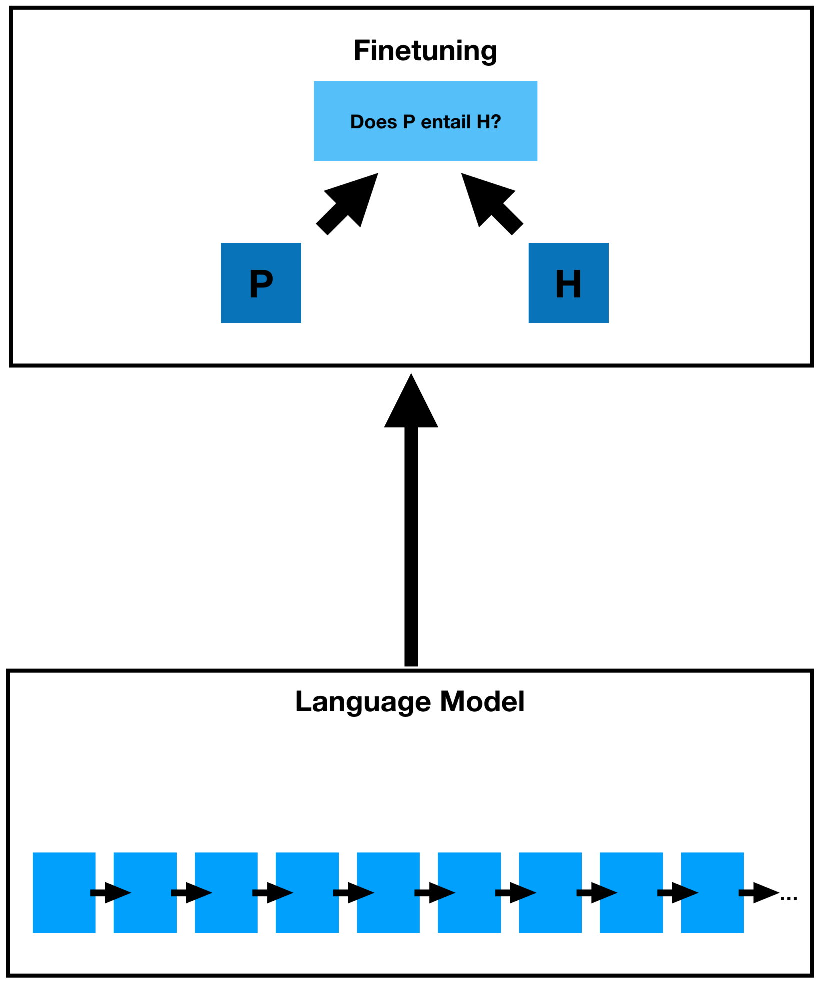
Generate premise,hypothesis pairs (a1,a2),(b1,b2) such that:

a1	entails	a2
b1	entails	b2
a1	does not entail	b2
b1	does not entail	a2

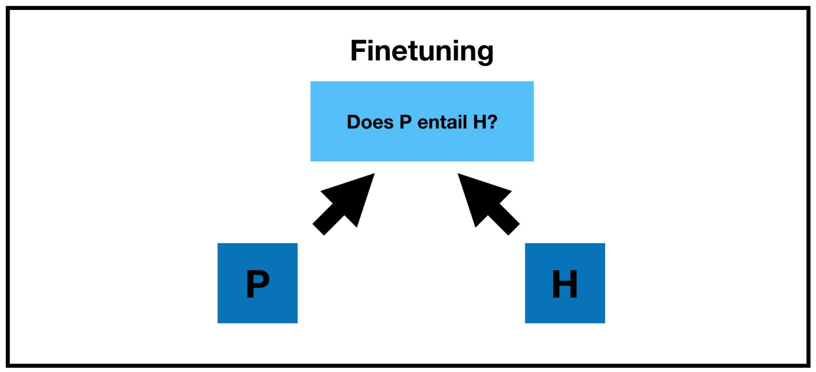
Thus both maximum-class and hypothesis-only accuracies are 50%

(Evans et al., 2018)

Our Framework

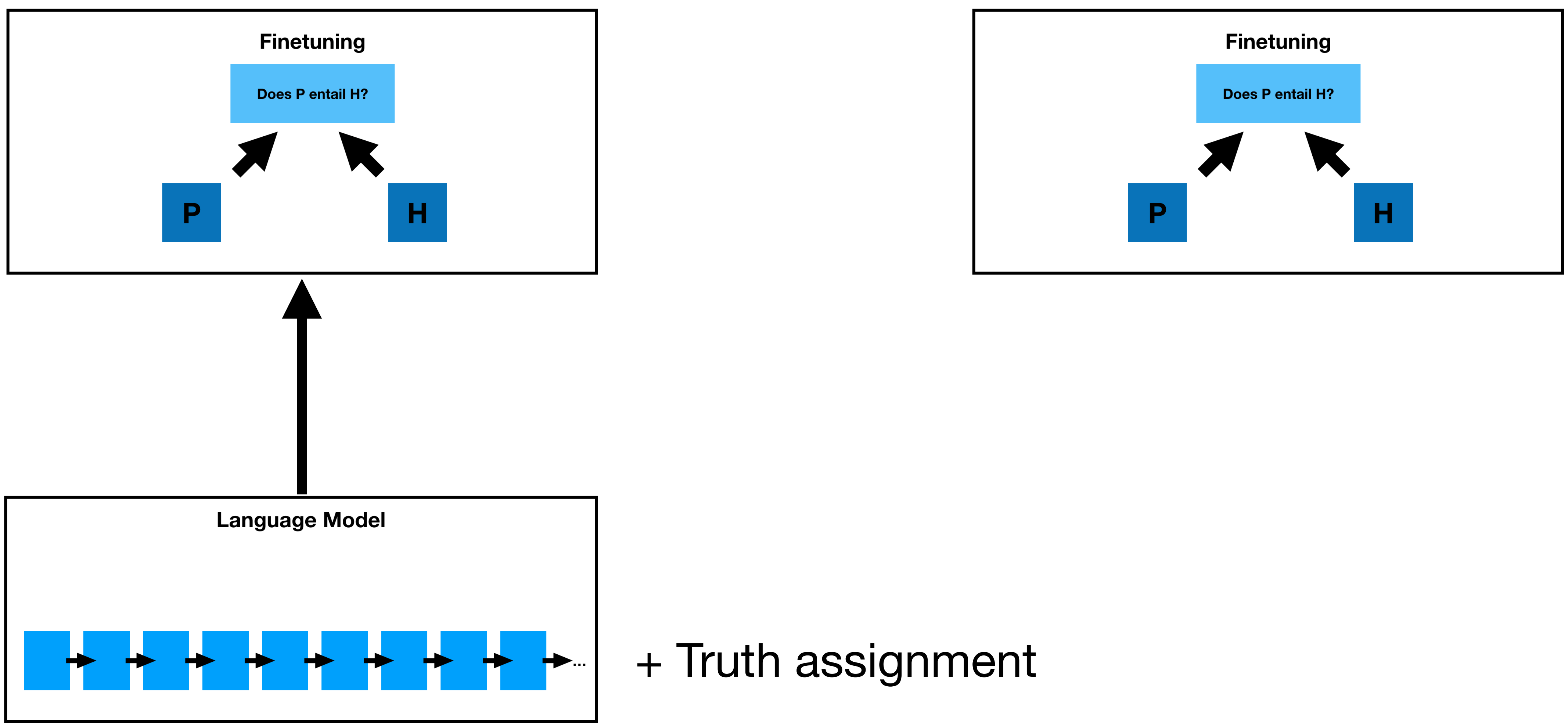


“from pretraining”



“from scratch”

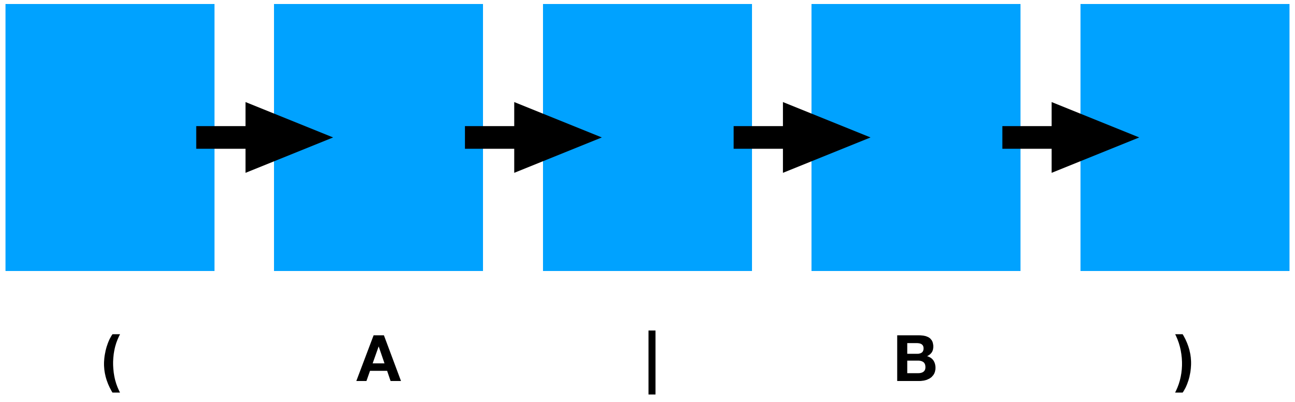
Our Framework



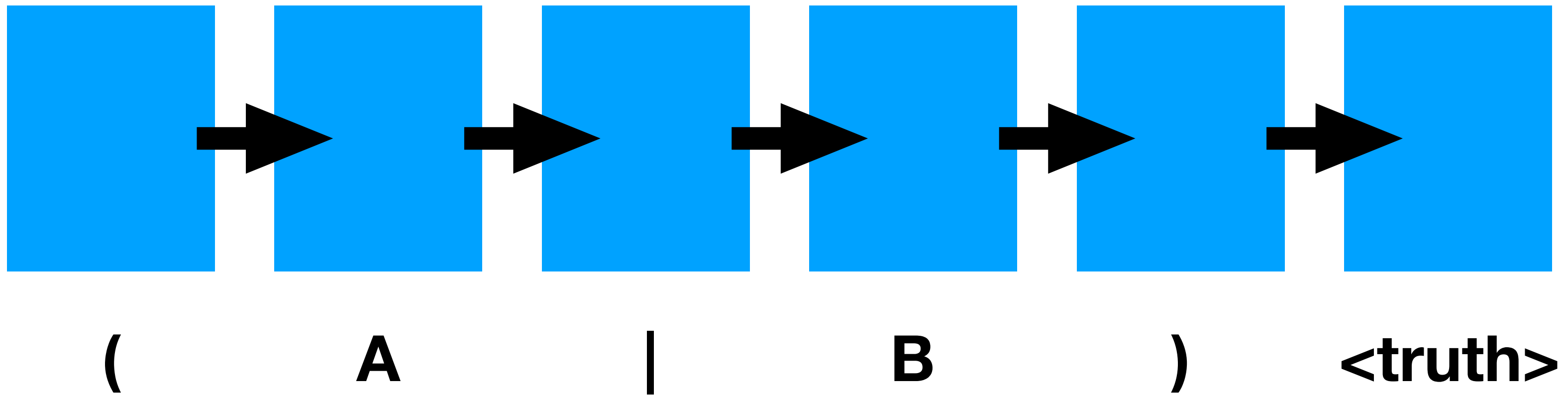
“from pretraining”

“from scratch”

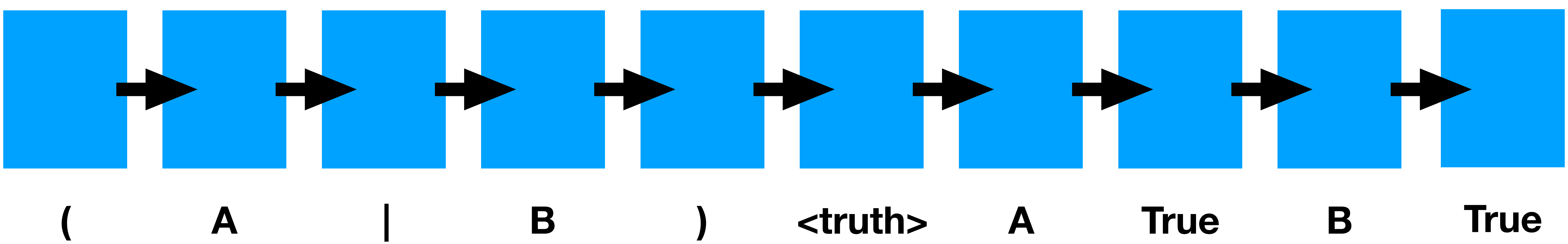
Truth Assignment Example



Truth Assignment Example



Truth Assignment Example



Inference Pattern Test Sets

<table>
<tr><th>Inference Pattern</th><th>Premise</th><th>Hypothesis</th></tr>
<tr><td>Double Negation</td><td>A</td><td>~~A</td></tr>
<tr><td>Conjunction Elimination</td><td>A & B</td><td>A</td></tr>
<tr><td>Disjunction Elimination</td><td>A</td><td>A|B</td></tr>
<tr><td>Disjunction Introduction</td><td>(A | B) & (A > C) & (B > C)</td><td>C</td></tr>
<tr><td>Modus Ponens</td><td>A & (A > B)</td><td>B</td></tr>
</table>

Inference Pattern Test Sets

Distractor Items

Inference Pattern Test Sets

Distractor Items

Inference Pattern	Premise	Hypothesis	Entailed?
Double Negation	A	~~A	Yes
	A	~A	No
	~~A	~A	No
	~A	~~A	No

Outline

Motivation

Experimental Design

Results

Discussion

Results

Inference Pattern Test Sets

Model	Validation Acc.
CBOW	51.136
LSTM	69.547
LSTM (pt)	68.079
LSTM (pt w/ TAs)	73.402
Transformer	63.917
Transformer (pt)	70.074
Transformer (pt w/ TAs)	75.949

Results

Inference Pattern Test Sets

Model	Validation Acc.
CBOW	51.136
LSTM	69.547
LSTM (pt)	68.079
LSTM (pt w/ TAs)	73.402
Transformer	63.917
Transformer (pt)	70.074
Transformer (pt w/ TAs)	75.949

CBOW baseline: Minimal dataset bias

Results

Inference Pattern Test Sets

Model	Validation Acc.
CBOW	51.136
LSTM	69.547
LSTM (pt)	68.079
LSTM (pt w/ TAs)	73.402
Transformer	63.917
Transformer (pt)	70.074
Transformer (pt w/ TAs)	75.949

LSTM: No benefit to pretraining

Results

Inference Pattern Test Sets

Model	Validation Acc.
CBOW	51.136
LSTM	69.547
LSTM (pt)	68.079
LSTM (pt w/ TAs)	73.402
Transformer	63.917
Transformer (pt)	70.074
Transformer (pt w/ TAs)	75.949

LSTM: benefit to truth assignment

Results

Inference Pattern Test Sets

Model	Validation Acc.
CBOW	51.136
LSTM	69.547
LSTM (pt)	68.079
LSTM (pt w/ TAs)	73.402
Transformer	63.917
Transformer (pt)	70.074
Transformer (pt w/ TAs)	75.949

Transformer: benefit to pretraining

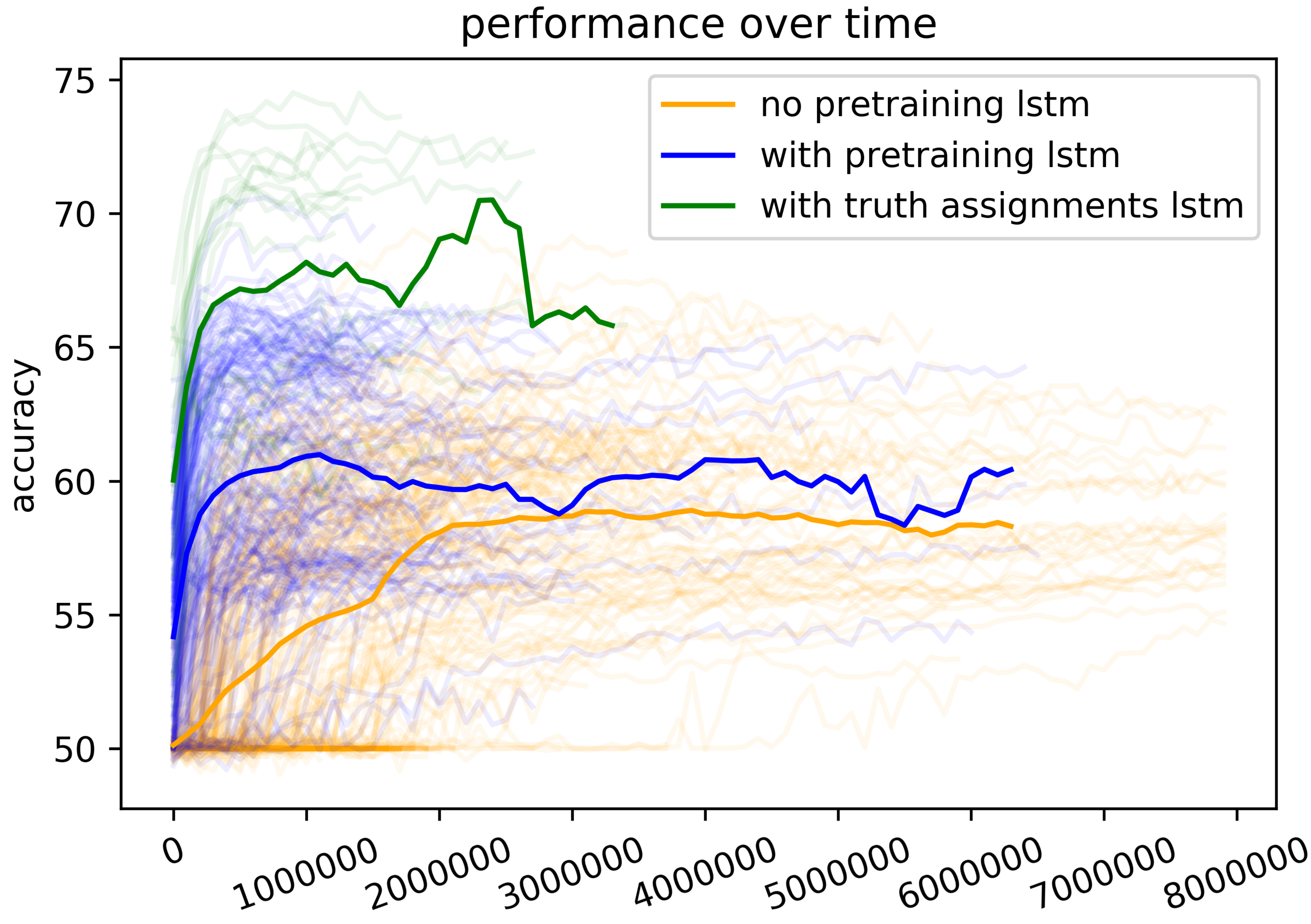
Results

Inference Pattern Test Sets

Model	Validation Acc.
CBOW	51.136
LSTM	69.547
LSTM (pt)	68.079
LSTM (pt w/ TAs)	73.402
Transformer	63.917
Transformer (pt)	70.074
Transformer (pt w/ TAs)	75.949

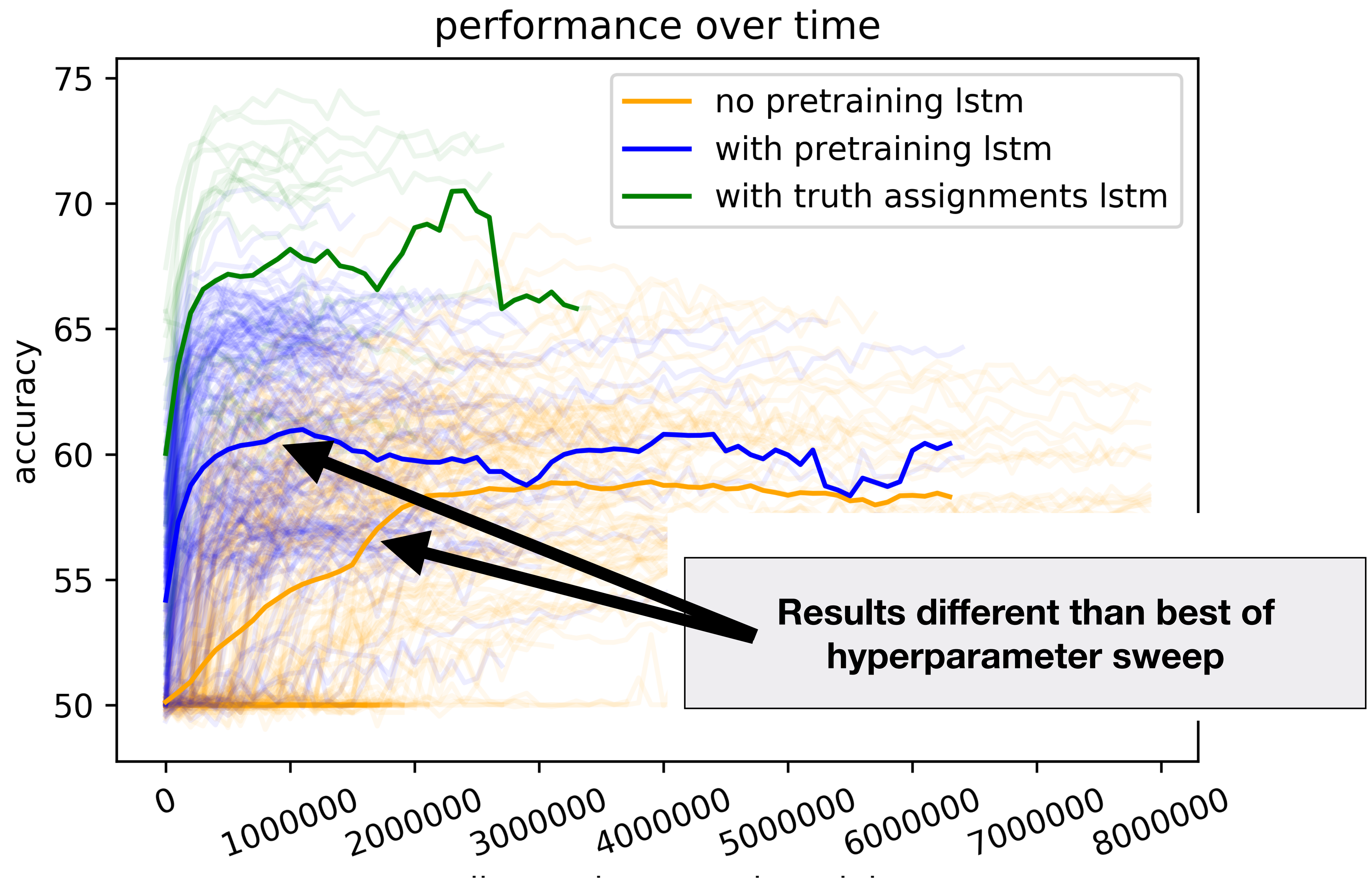
Transformer: benefit to truth assignment

Results

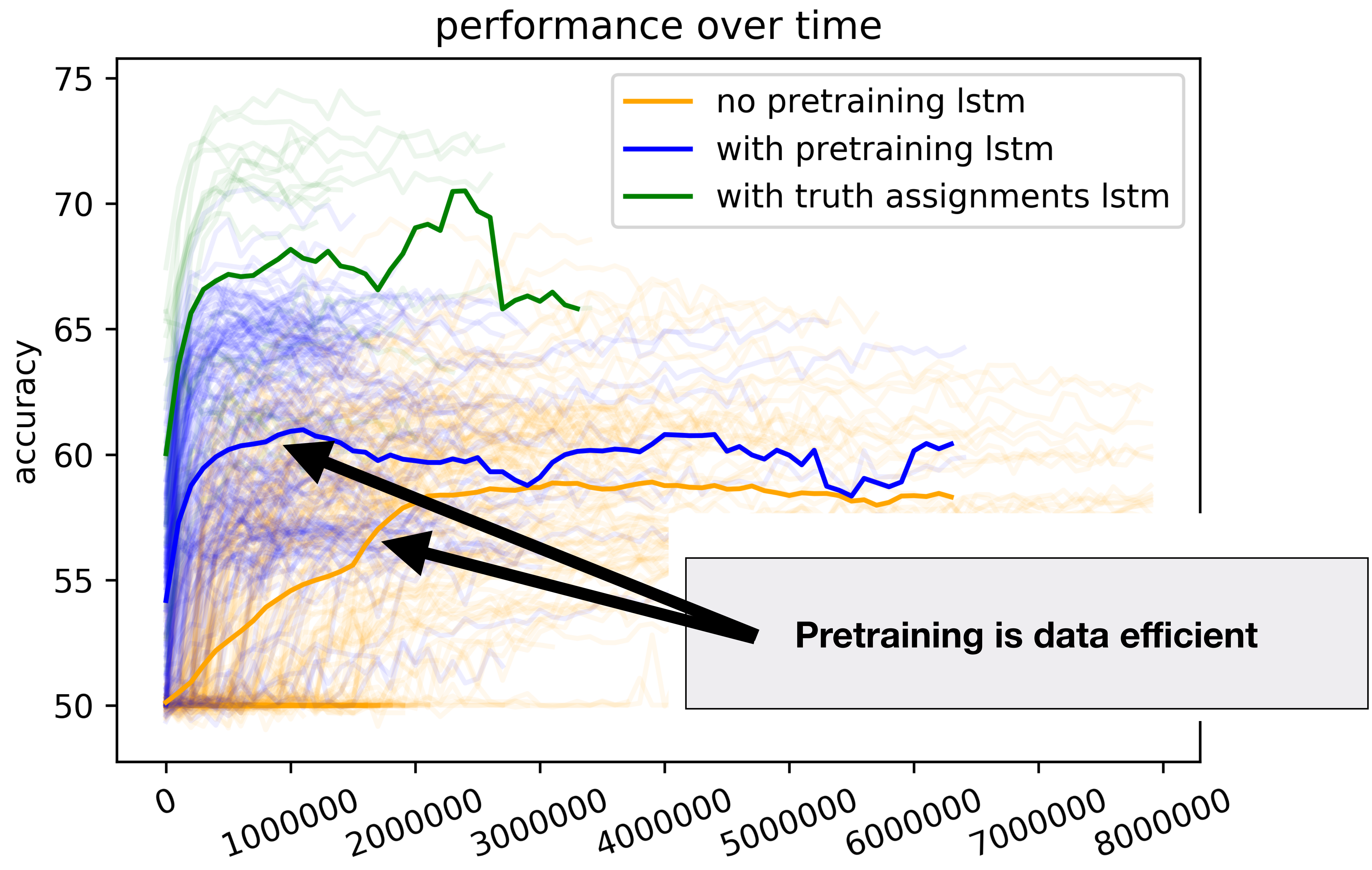


performance over time
accuracy
75
70
65
60
55
50
0
1000000
2000000
3000000
4000000
5000000
6000000
7000000
8000000
no pretraining lstm
with pretraining lstm
with truth assignments lstm

Results

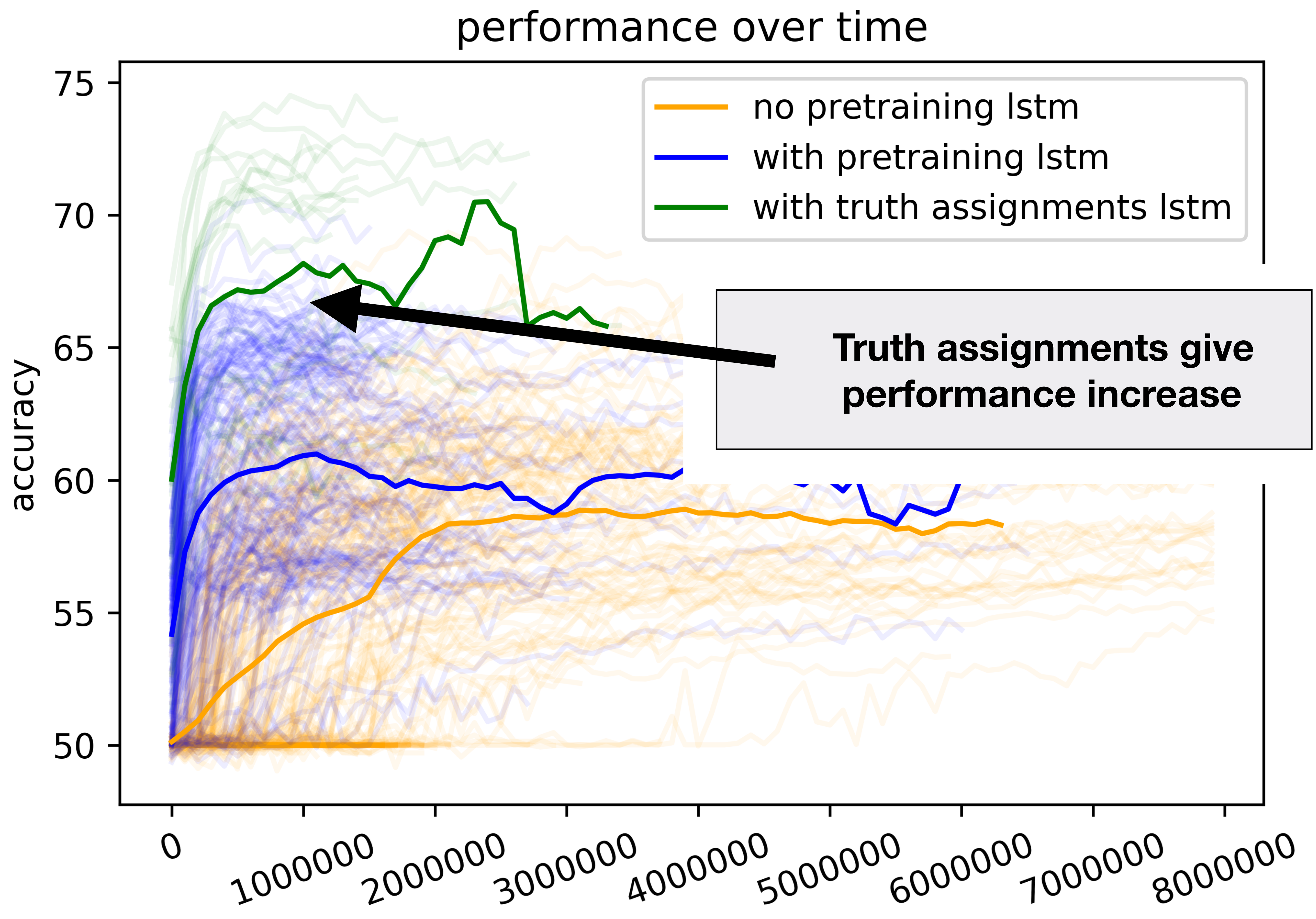

performance over time
no pretraining lstm
with pretraining lstm
with truth assignments lstm
accuracy
75
70
65
60
55
50
0
1000000
2000000
3000000
4000000
5000000
6000000
7000000
8000000
Results different than best of hyperparameter sweep

Results



performance over time
no pretraining lstm
with pretraining lstm
with truth assignments lstm
accuracy
75
70
65
60
55
50
0
1000000
2000000
3000000
4000000
5000000
6000000
7000000
8000000
Pretraining is data efficient

Results


performance over time
accuracy
75
70
65
60
55
50
0
1000000
2000000
3000000
4000000
5000000
6000000
7000000
8000000
no pretraining lstm
with pretraining lstm
with truth assignments lstm
Truth assignments give performance increase

Results

Inference Pattern Test Sets

Model	Validation Acc.	Inf. Pattern Acc.
CBOW	51.136	0.501
LSTM	69.547	0.683
LSTM (pt)	68.079	0.566
LSTM (pt w/ TAs)	73.402	0.531
Transformer	63.917	0.679
Transformer (pt)	70.074	0.701
Transformer (pt w/ TAs)	75.949	0.693

Results

Inference Pattern Test Sets

Model	Validation Acc.	Inf. Pattern Acc.
CBOW	51.136	0.501
LSTM	69.547	0.683
LSTM (pt)	68.079	0.566
LSTM (pt w/ TAs)	73.402	0.531
Transformer	63.917	0.679
Transformer (pt)	70.074	0.701
Transformer (pt w/ TAs)	75.949	0.693

LSTM: No benefit to pretraining

Results

Inference Pattern Test Sets

Model	Validation Acc.	Inf. Pattern Acc.
CBOW	51.136	0.501
LSTM	69.547	0.683
LSTM (pt)	68.079	0.566
LSTM (pt w/ TAs)	73.402	0.531
Transformer	63.917	0.67[illegible]
Transformer (pt)	70.074	0.701
Transformer (pt w/ TAs)	75.949	0.693

LSTM: truth assignments hinder performance

Results

Inference Pattern Test Sets

Transformer: small benefit to pretraining

Model	Validation Acc.	Inf. Pattern Acc.
CBOW	51.136	0.501
LSTM	69.547	0.683
LSTM (pt)	68.079	0.566
LSTM (pt w/ TAs)	73.402	0.531
Transformer	63.917	0.679
Transformer (pt)	70.074	0.701
Transformer (pt w/ TAs)	75.949	0.693

Results

Inference Pattern Test Sets

Transformer: no benefit to truth assignments

Model	Validation Acc.	Inf. Pattern Acc.
CBOW	51.136	0.501
LSTM	69.547	0.683
LSTM (pt)	68.079	0.566
LSTM (pt w/ TAs)	73.402	0.531
Transformer	63.917	0.679
Transformer (pt)	70.074	0.701
Transformer (pt w/ TAs)	75.949	0.693

Results

Inference Pattern Test Sets

Model	Validation Acc.	Inf. Pattern Acc.	Inf. Pattern P(A) Acc.	Inf. Pattern N(A) Acc.
CBOW	51.136	0.501	0.271	0.736
LSTM	69.547	0.683	0.768	0.480
LSTM (pt)	68.079	0.566	0.360	0.680
LSTM (pt w/ TAs)	73.402	0.531	0.145	0.881
Transformer	63.917	0.679	0.749	0.563
Transformer (pt)	70.074	0.701	0.983	0.441
Transformer (pt w/ TAs)	75.949	0.693	0.919	0.409

Results

Inference Pattern Test Sets

Model	Validation Acc.	Inf. Pattern Acc.	Inf. Pattern P(A) Acc.	Inf. Pattern N(A) Acc.
CBOW	51.136	0.501	0.271	0.736
LSTM	69.547	0.683	0.768	0.480
LSTM (pt)	68.079	0.566	0.360	0.680
LSTM (pt w/ TAs)	73.402	0.531	0.145	0.881
Transformer	63.917	0.679	0.749	0.563
Transformer (pt)	70.074	0.701	0.983	0.441
Transformer (pt w/ TAs)	75.949	0.693	0.919	0.409

Extreme skew!

Results

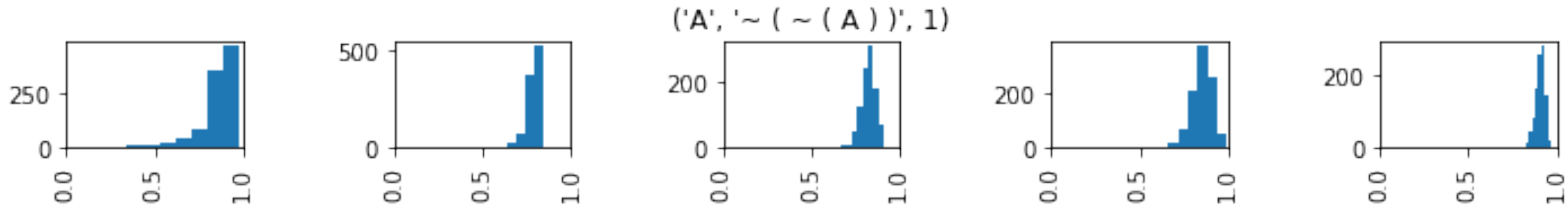
Inference Pattern Test Sets

Model	Validation Acc.	Inf. Pattern Acc.	Inf. Pattern P(A) Acc.	Inf. Pattern N(A) Acc.
CBOW	51.136	0.501	0.271	0.736
LSTM	69.547	0.683	0.768	0.480
LSTM (pt)	68.079	0.566	0.360	0.680
LSTM (pt w/ TAs)	73.402	0.531	0.145	0.881
Transformer	63.917	0.679	0.749	0.563
Transformer (pt)	70.074	0.701	0.983	0.441
Transformer (pt w/ TAs)	75.949	0.693	0.919	0.409

Great job on positive, still very skewed

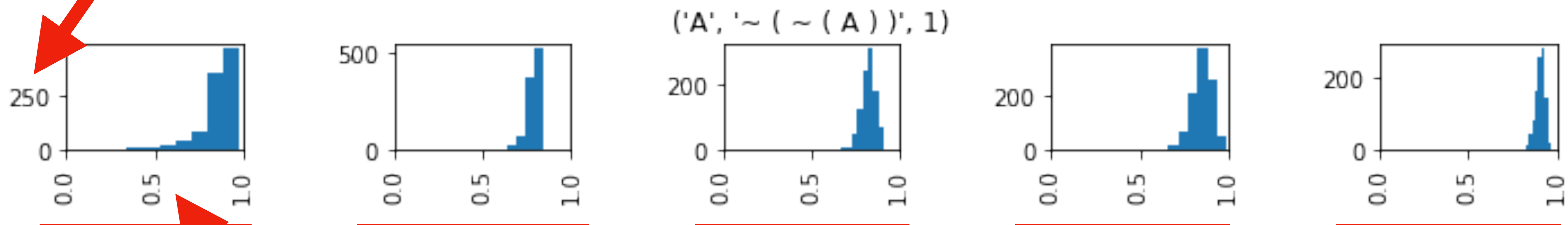
Results

Inference Pattern Test Sets



Results

of examples

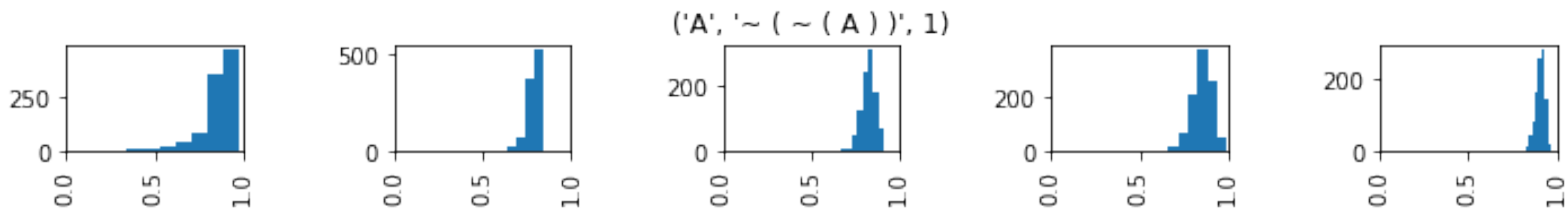


"Entailed" label Activation

Five different runs of same hyperparameters

Results

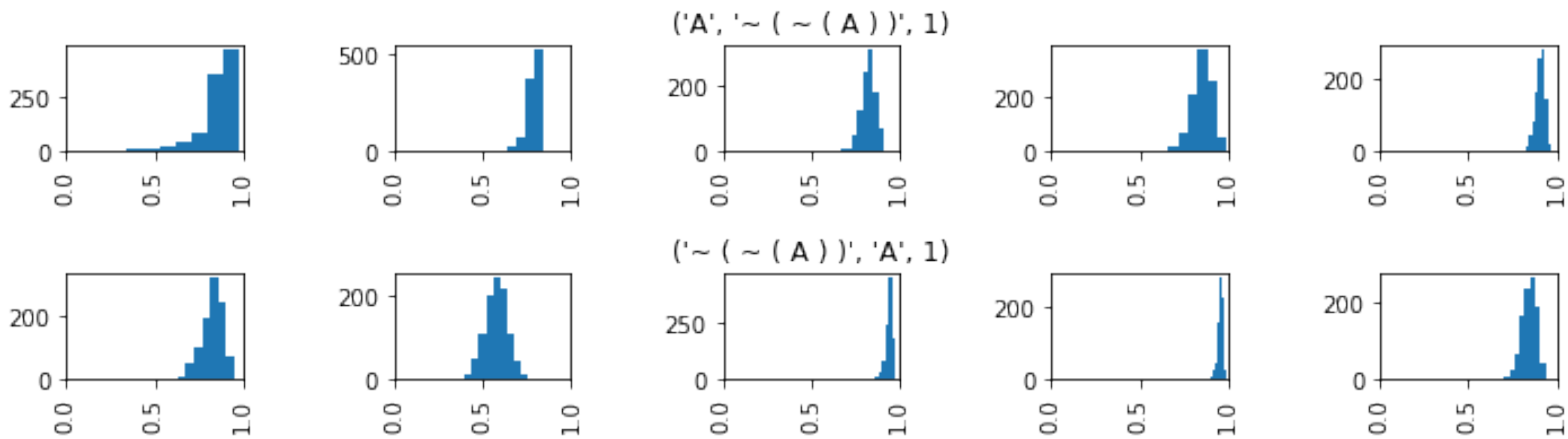
Inference Pattern Test Sets



Good

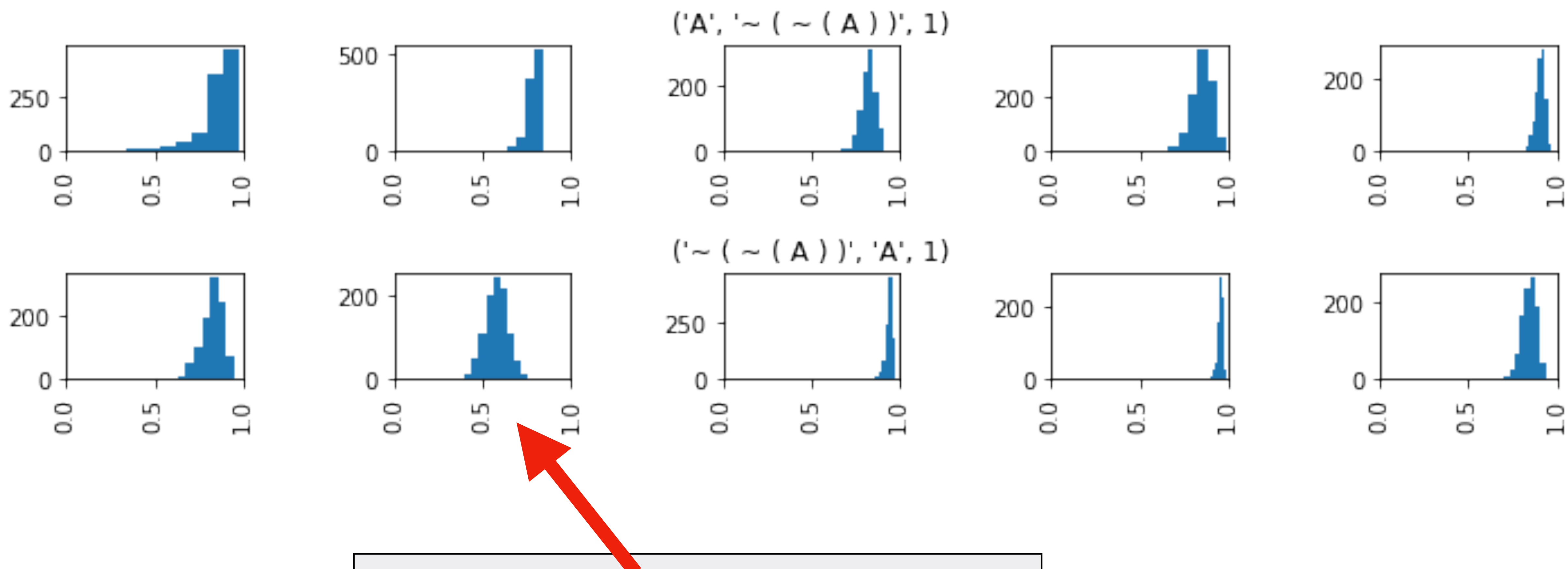
Results

Inference Pattern Test Sets



Results

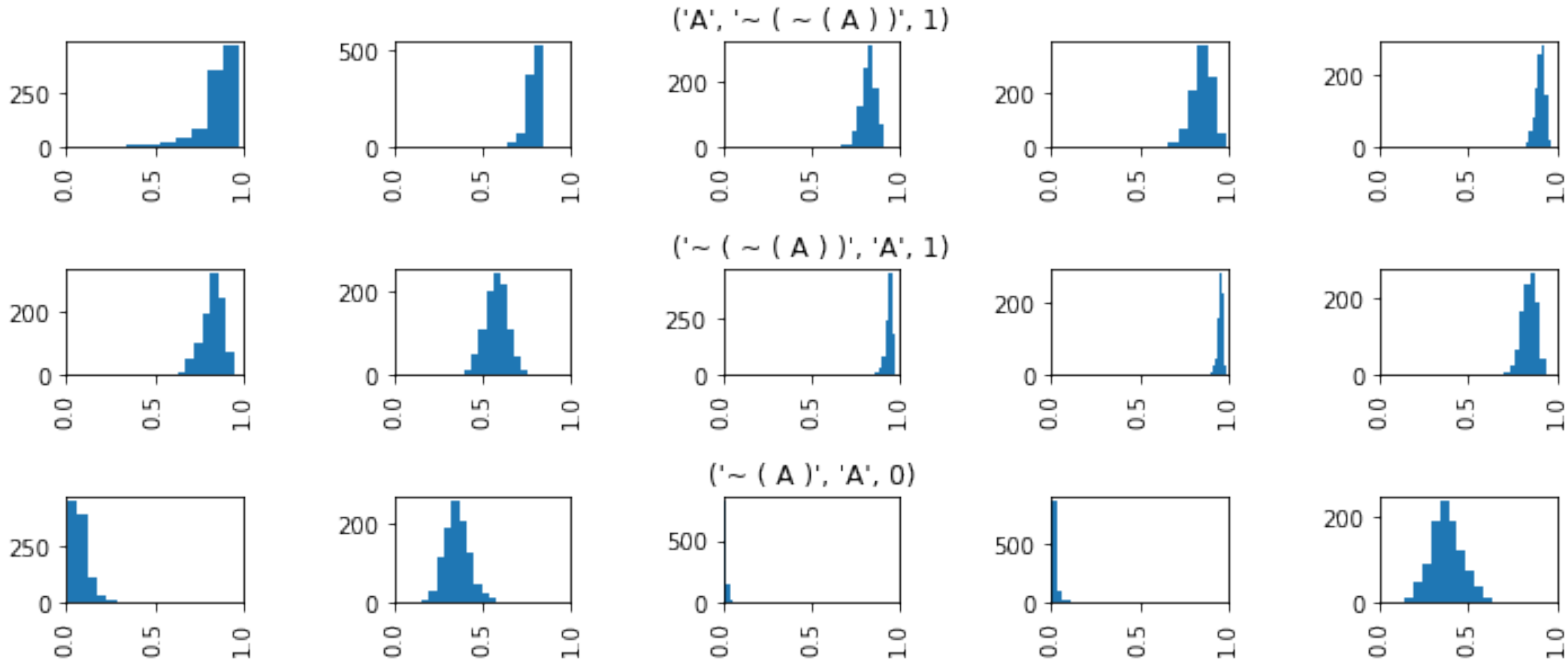
Inference Pattern Test Sets



Not very good

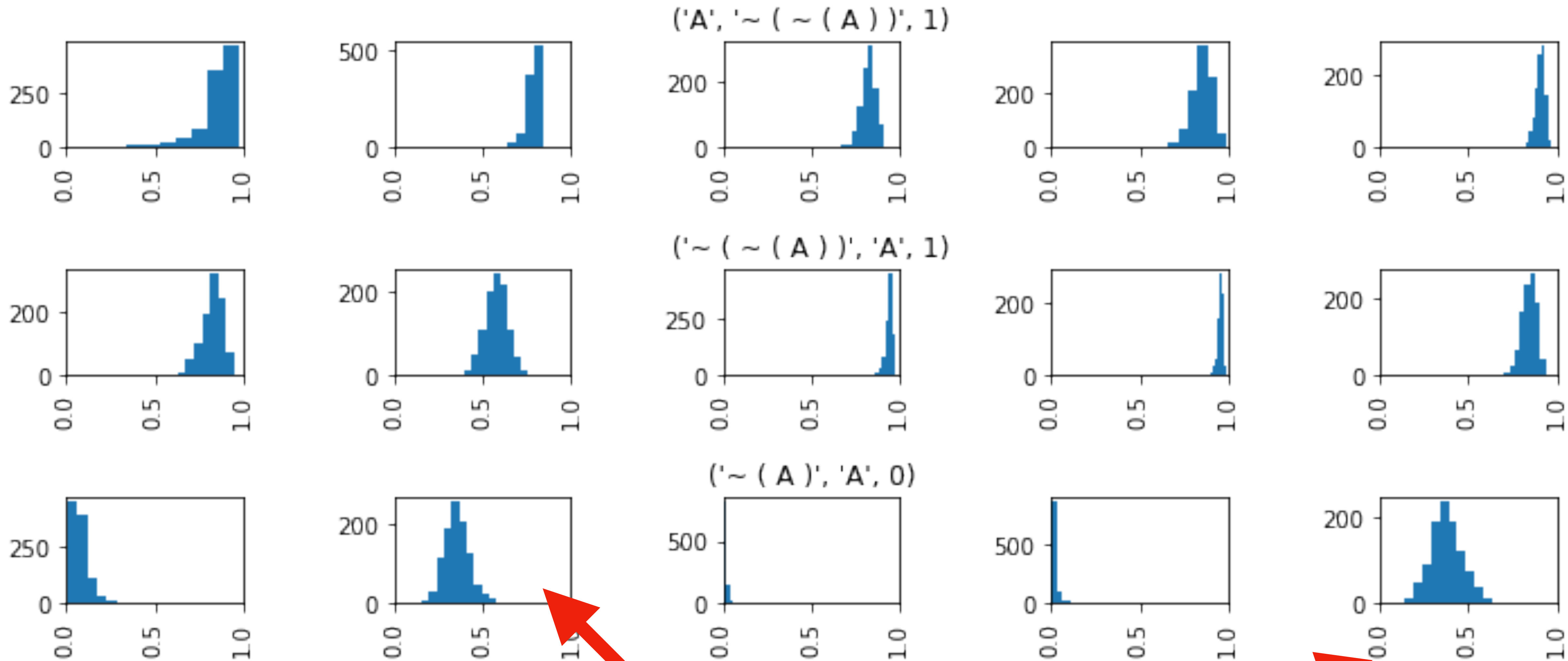
Results

Inference Pattern Test Sets



Results

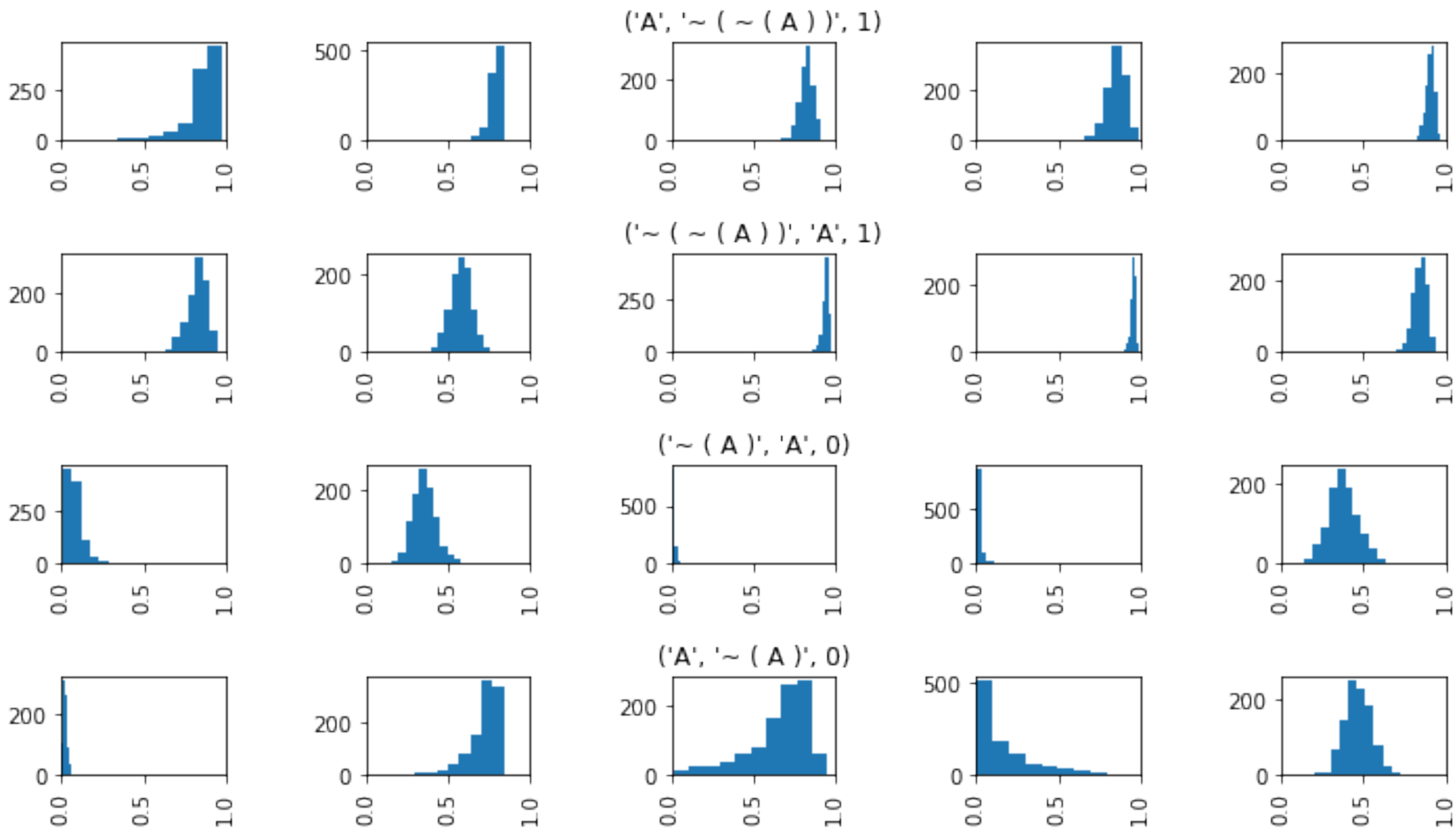
Inference Pattern Test Sets



Not very good

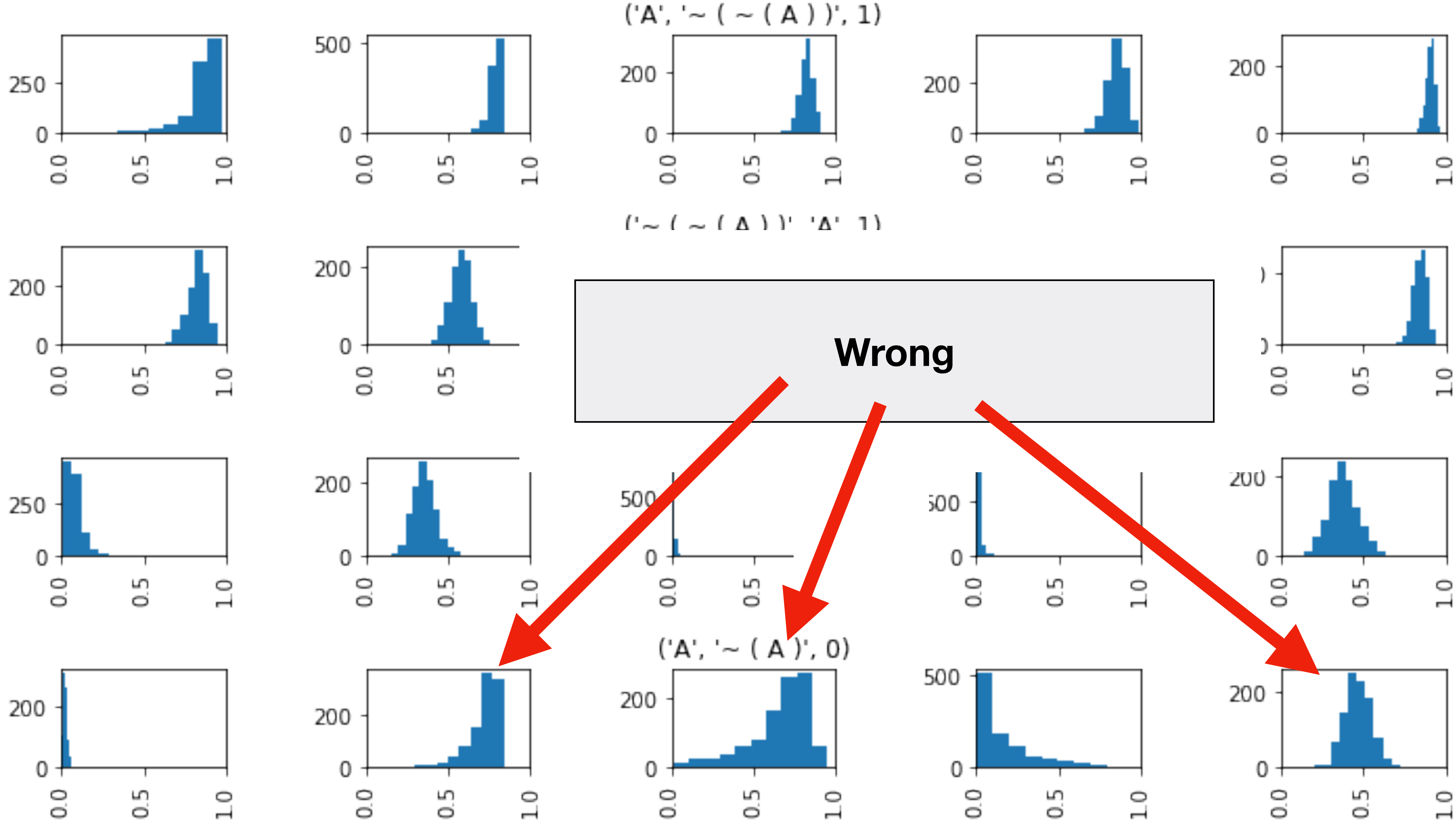
Results

Inference Pattern Test Sets



Results

Inference Pattern Test Sets



Outline

Motivation

Experimental Design

Results

Discussion

Conclusion

Conclusion

- Results negative, inconclusive, dependent on sequence model type

Conclusion

- Results negative, inconclusive, dependent on sequence model type
- Language model pretraining helps only with data efficiency

Conclusion

- Results negative, inconclusive, dependent on sequence model type

- Language model pretraining helps only with data efficiency

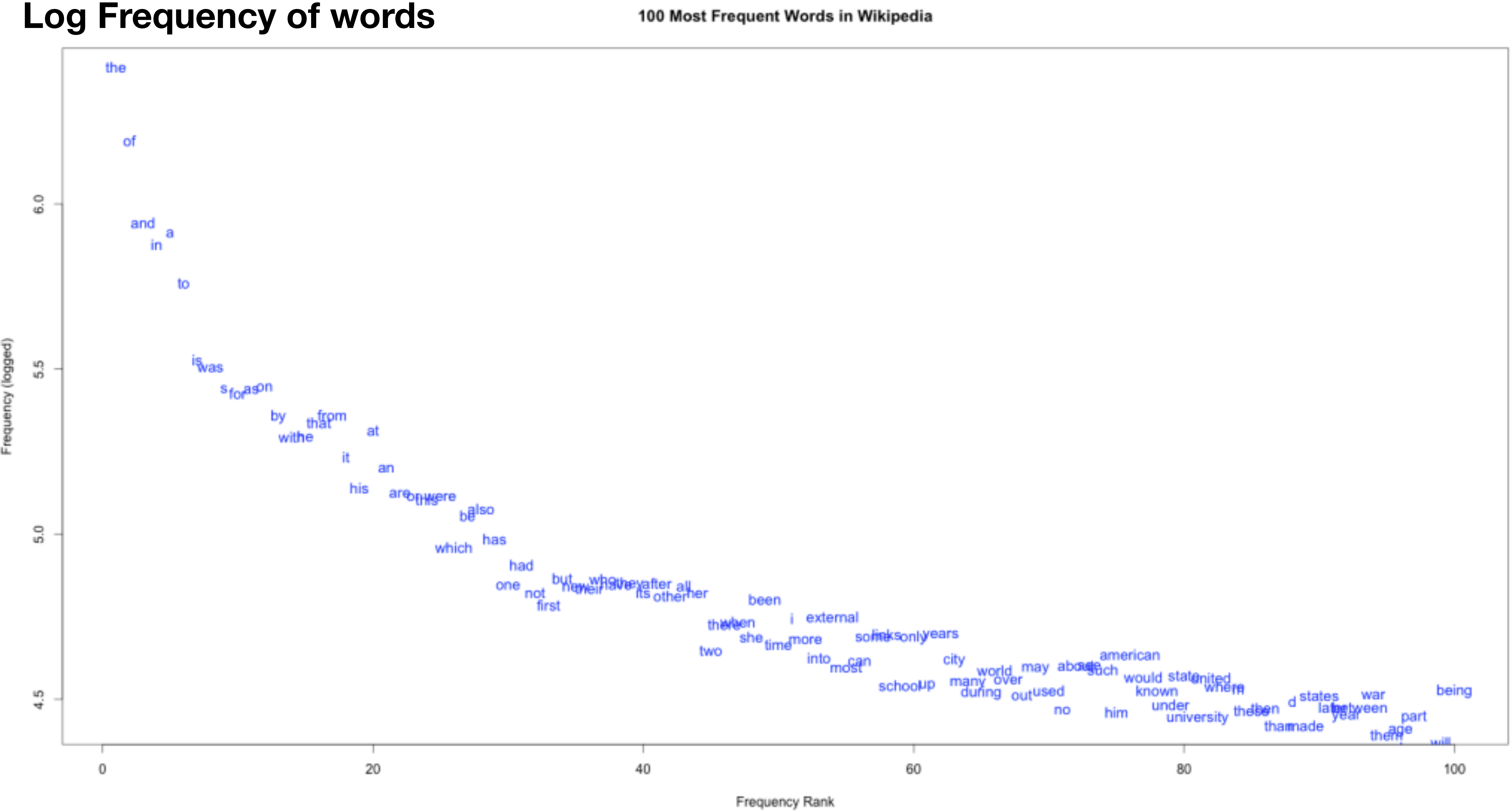
- All models struggle with inference pattern test sets

Next Steps

- Can success observed on natural logic datasets be explained by exploitation of cooccurence and complex lexical heuristics?

Zipfian distribution

Log Frequency of words

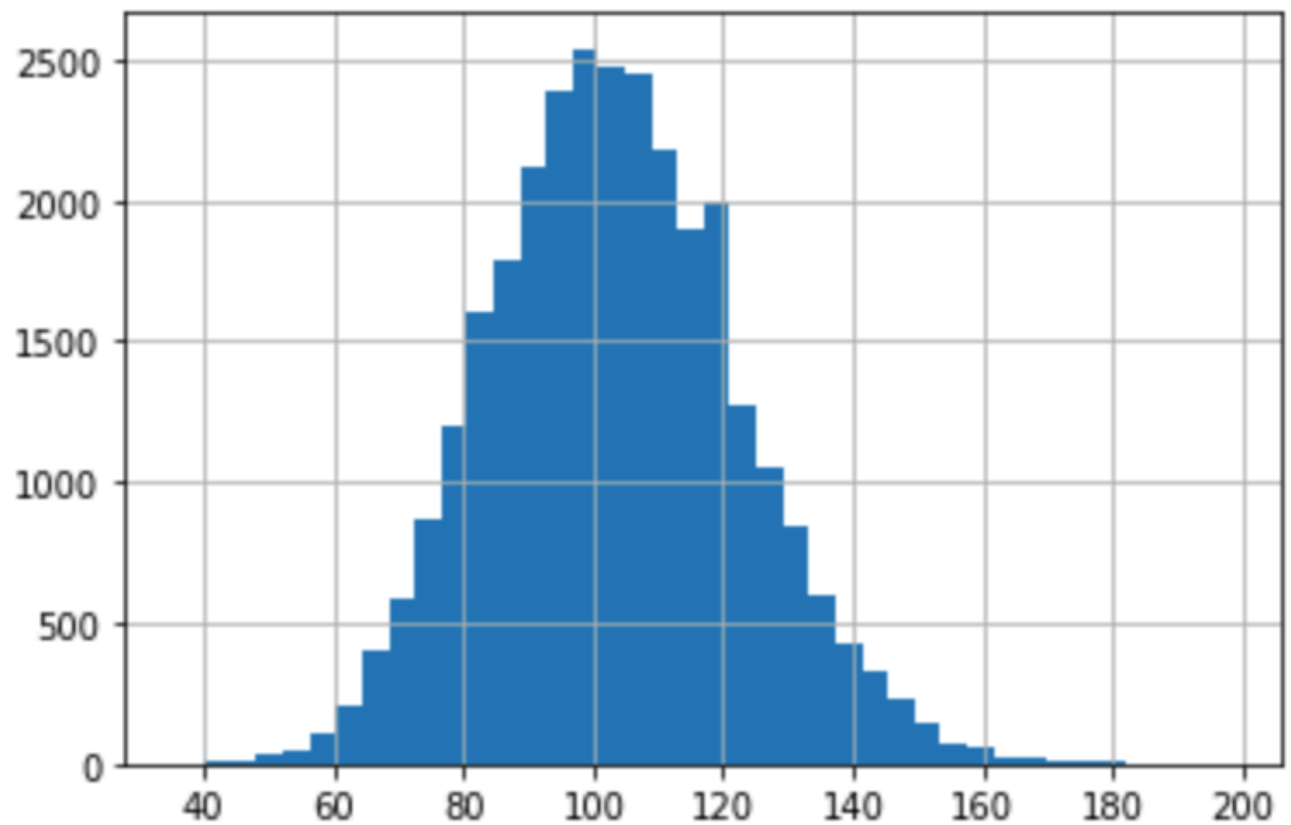


Frequency Rank of words

http://wugology.com/zipfs-law/

Symbol distribution in our datasets

of symbols in
bucket



# of times symbol appears in dataset

Next Steps

- Can success observed on natural logic datasets be explained by exploitation of cooccurence and complex lexical heuristics?
  - Skew frequency of symbols in our dataset

Next Steps

- Can success observed on natural logic datasets be explained by exploitation of cooccurence and complex lexical heuristics?
    - Skew frequency of symbols in our dataset
- Would including “unattested” sentences assist the model in learning logical properties?

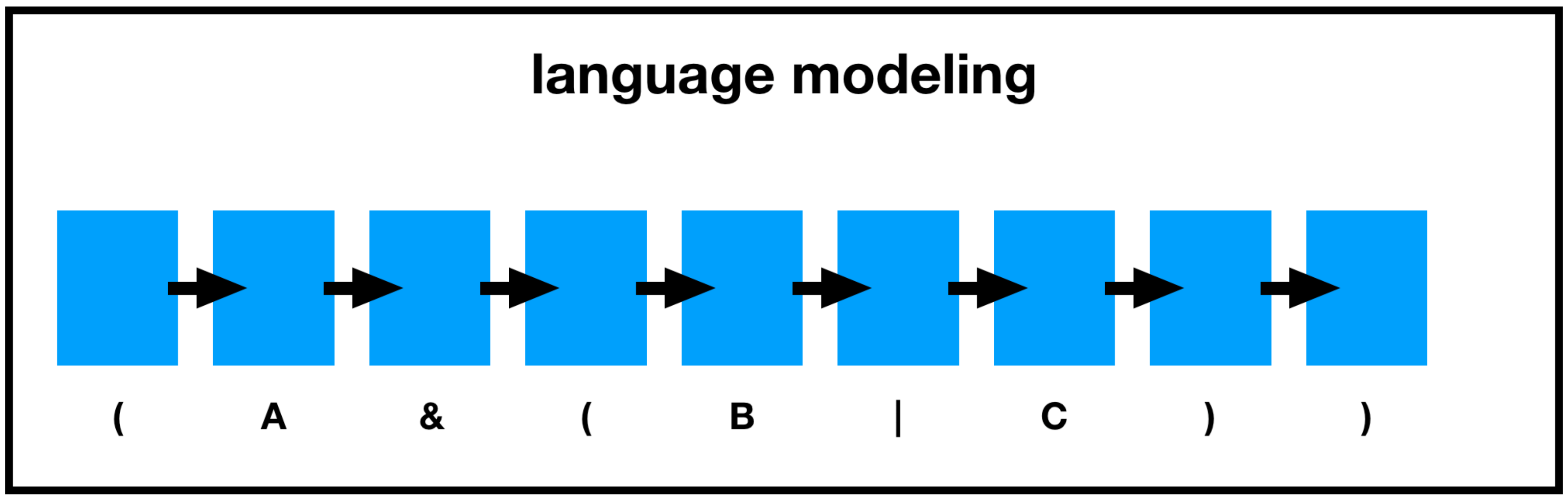
Next Steps

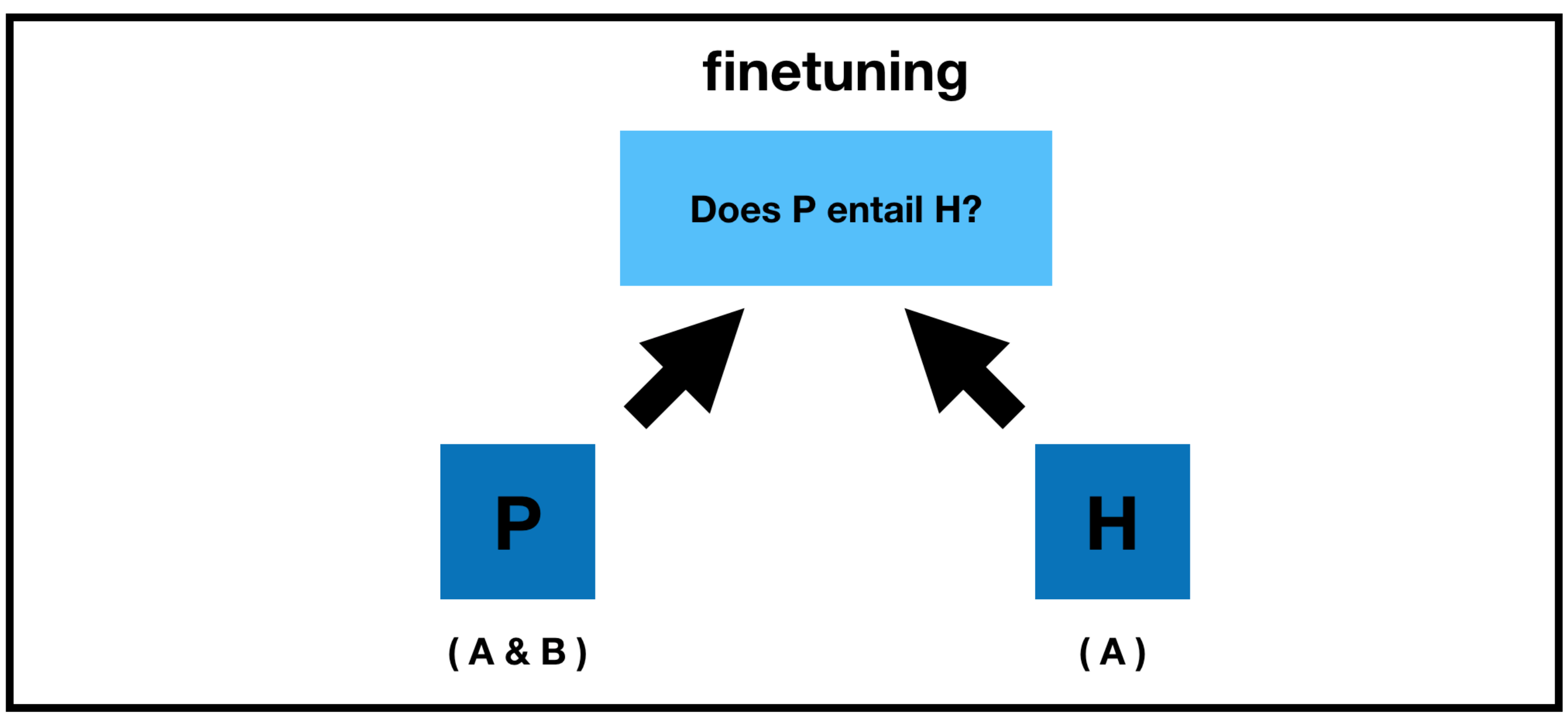
- Can success observed on natural logic datasets be explained by exploitation of cooccurence and complex lexical heuristics?
    - Skew frequency of symbols in our dataset
- Would including “unattested” sentences assist the model in learning logical properties?
    - Include sentences that evaluate to False at pretraining

thanks! :)

bonus slides!

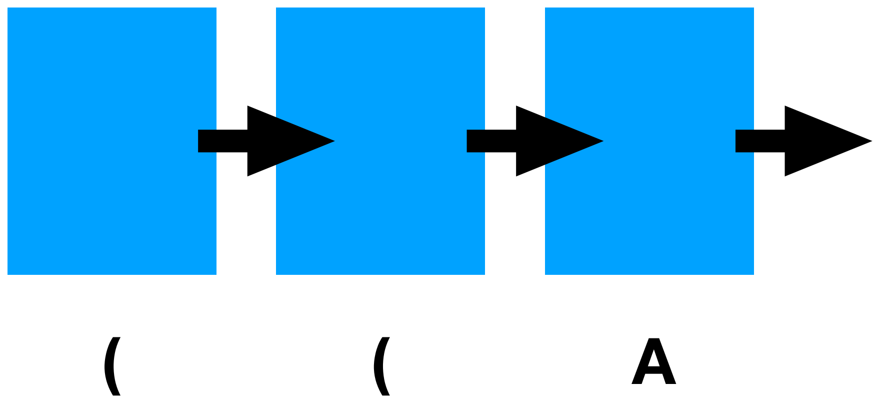
Paradigm





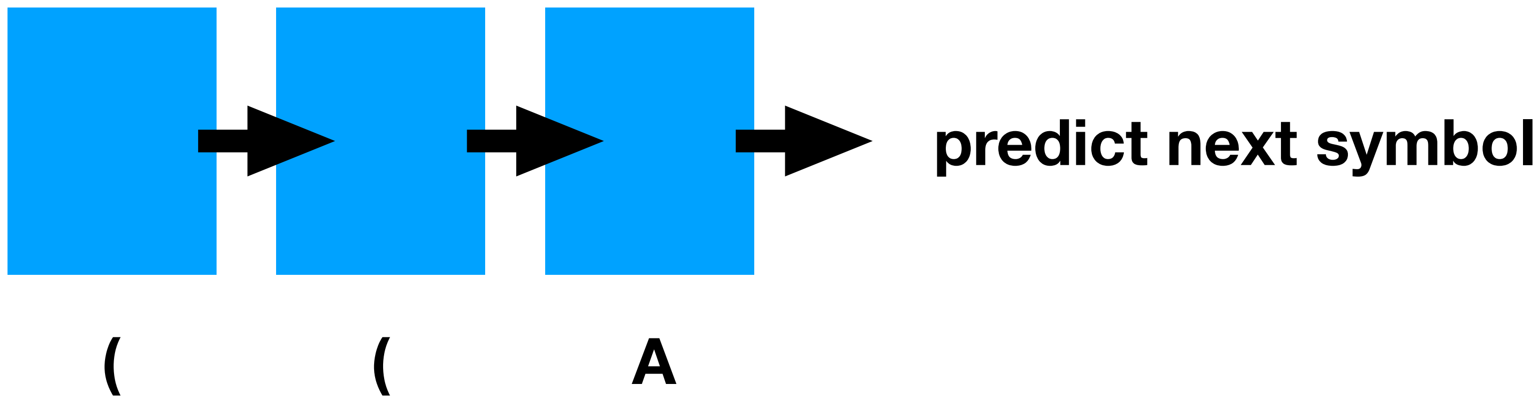
Sanity Check!

Experiment 1



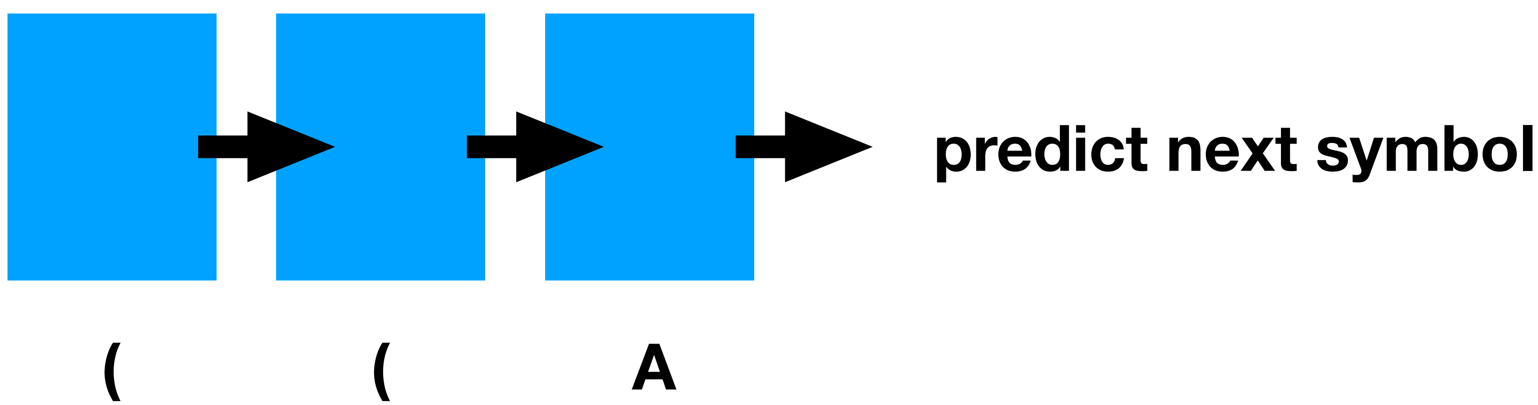
Sanity Check!

Experiment 1



Sanity Check!

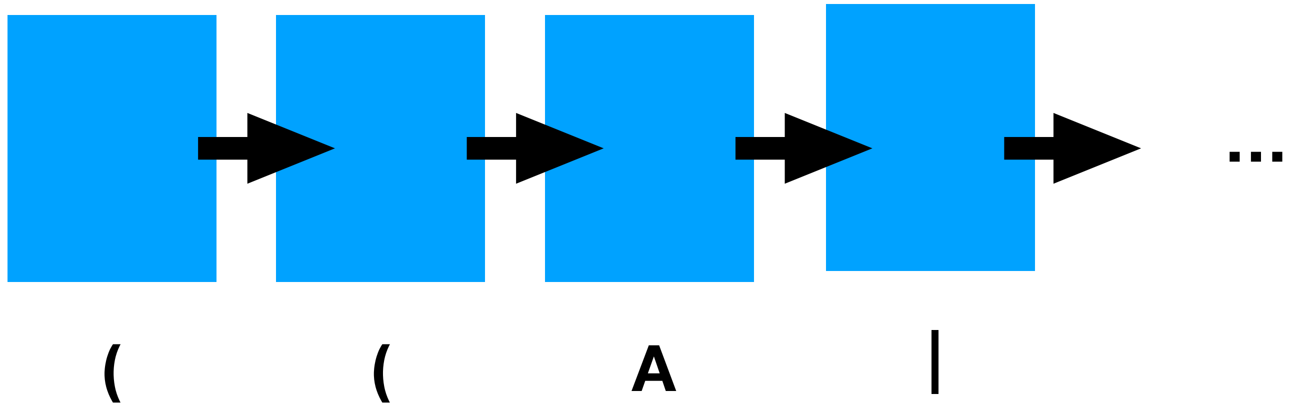
Experiment 1



|	0.281
&	0.123
A	0.0001
...	...

Sanity Check!

Experiment 1



Sanity Check!

Experiment 1 model sampling

Output type	Example	% of data
Consistent	(A | A)	84.270%
Inconsistent	(A | ~ (A))	1.037%
Syntax invalid	(A | & A)	0.443%
Unfinished	(A | ((((…	14.25%

bonus slides!

How do we know how many symbols to include?

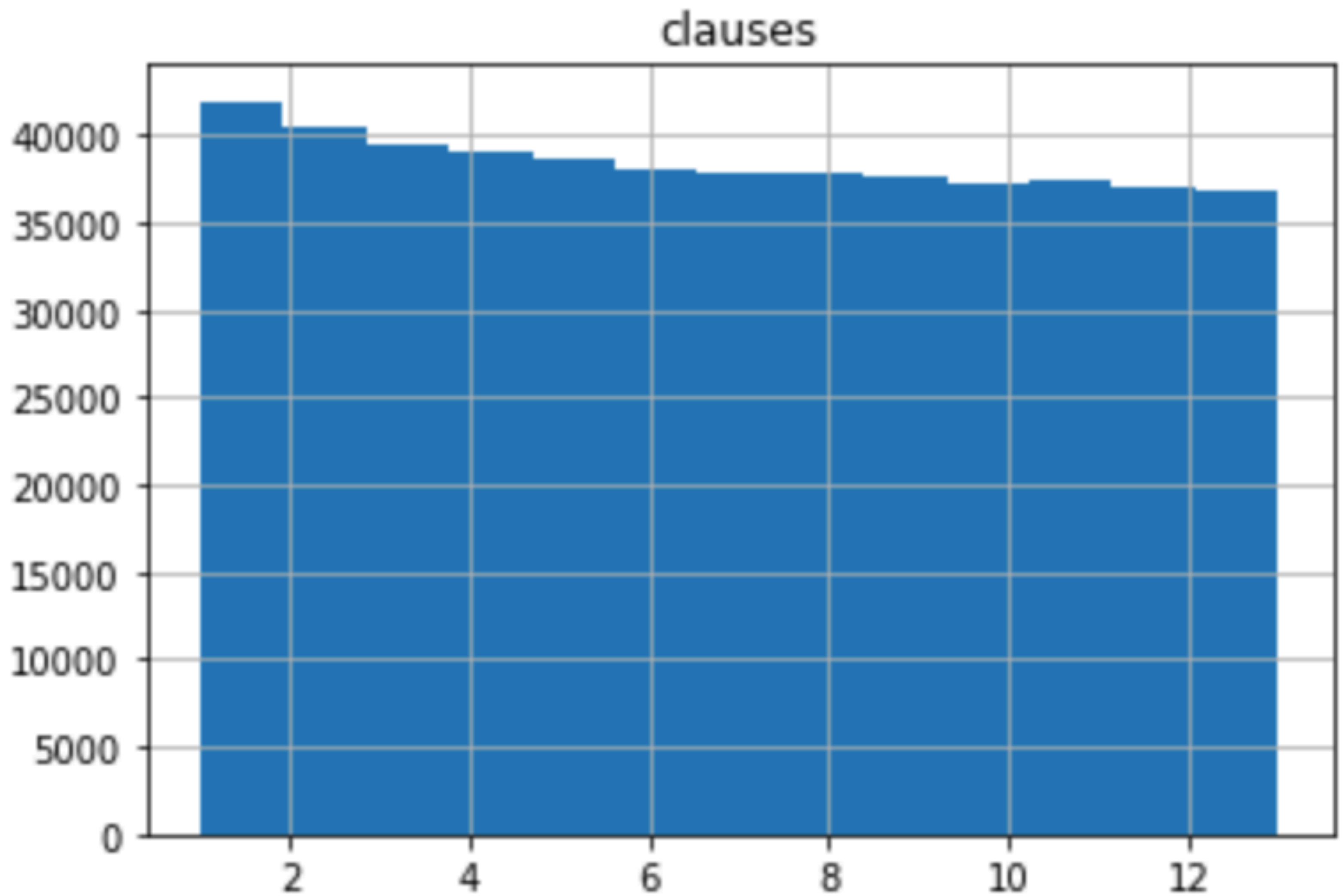
Train examples	# Symbols	Heldout patterns, training symbols	Heldout patterns, novel symbols
10K	25	0.957	0.5
100K	25	0.957	0.8125
10K	5K	0.511	0.511
100K	5K	1	0.969
1M	50K	1	0.98

bonus slides!

Algorithm 1 Generating Satisfiable Sentences

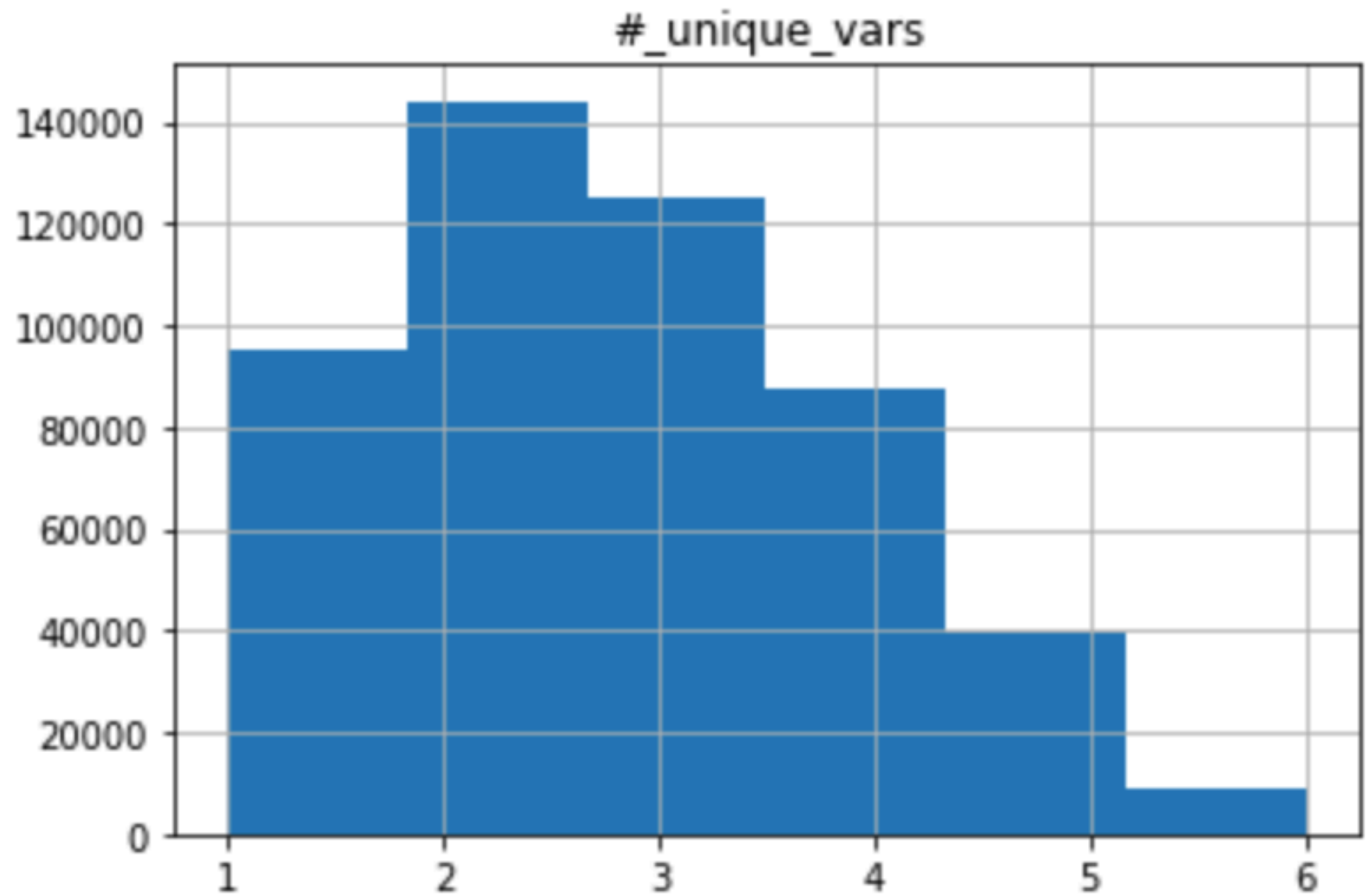
procedure GENERATE_SENTENCE(X)
    Randomly pick *num_clauses* between 1 and 13
    Randomly pick *maximum_unique_variables* between 1 and 5
    vocab = *maximum_unique_variables* symbols from X, sampling via uniform distribution
    clauses_in_sentence = *num_clauses* samples from &, |, ⊨, ¬
    final_sentence = clauses[0]
    open_indices = indices of clause[0] where variable or clause could be inserted (for (&, |, ⊨, add indices 0 and 2, and for ¬, add index 1)
    for all clauses in *clauses_in_sentence*[1:] **do**
        Randomly nest clause in *final_sentence* by inserting it at random index from *open_indices*
        Update *open_indices* by removing chosen index and adding indices of clause modulated by current position in *final_sentence*
    for all index in *open_indices* **do**
        final_sentence[index] = randomly sampled variable from *vocab*
    if *final_sentence* is satisfiable **then**
     return *final_sentence*
    else
     return Generate_Sentence(X)

bonus slides!



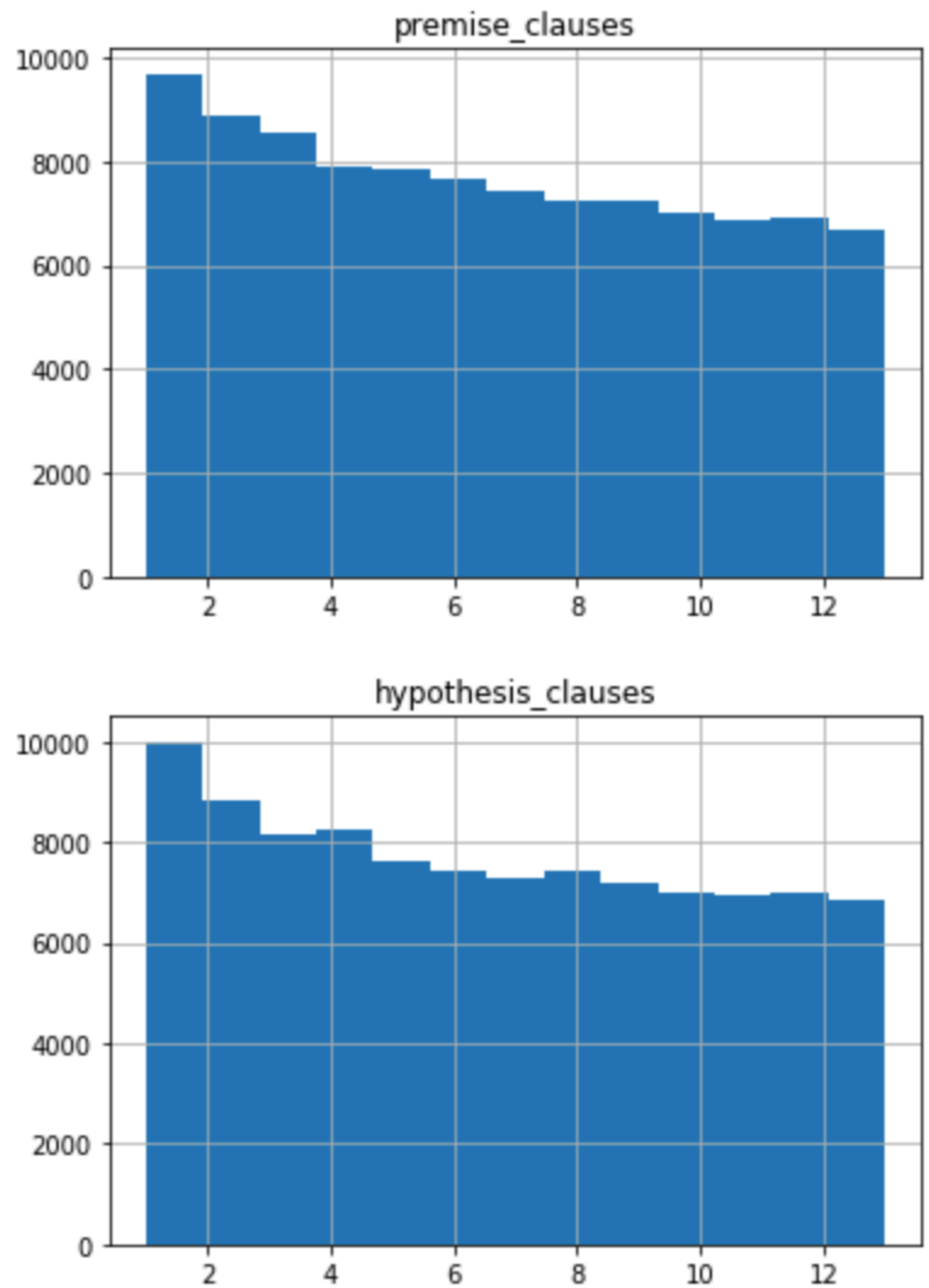
pretraining

bonus slides!



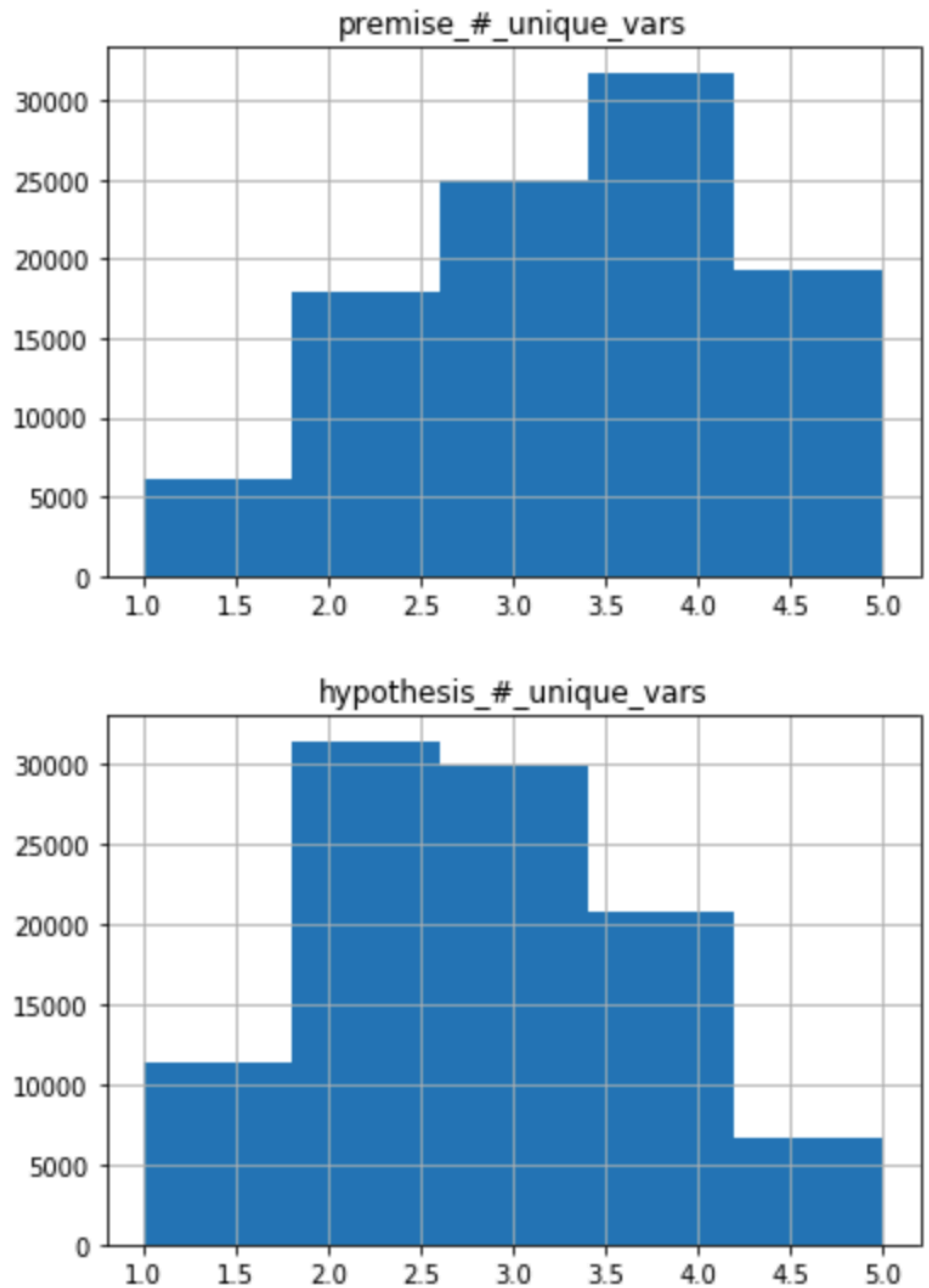
pretraining

bonus slides!



finetuning

bonus slides!



finetuning