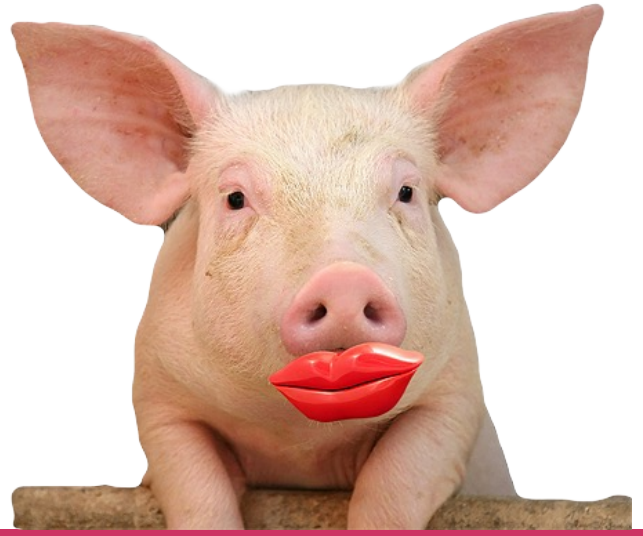
Lipstick on a Pig:

Using Language Models as Few-Shot Learners



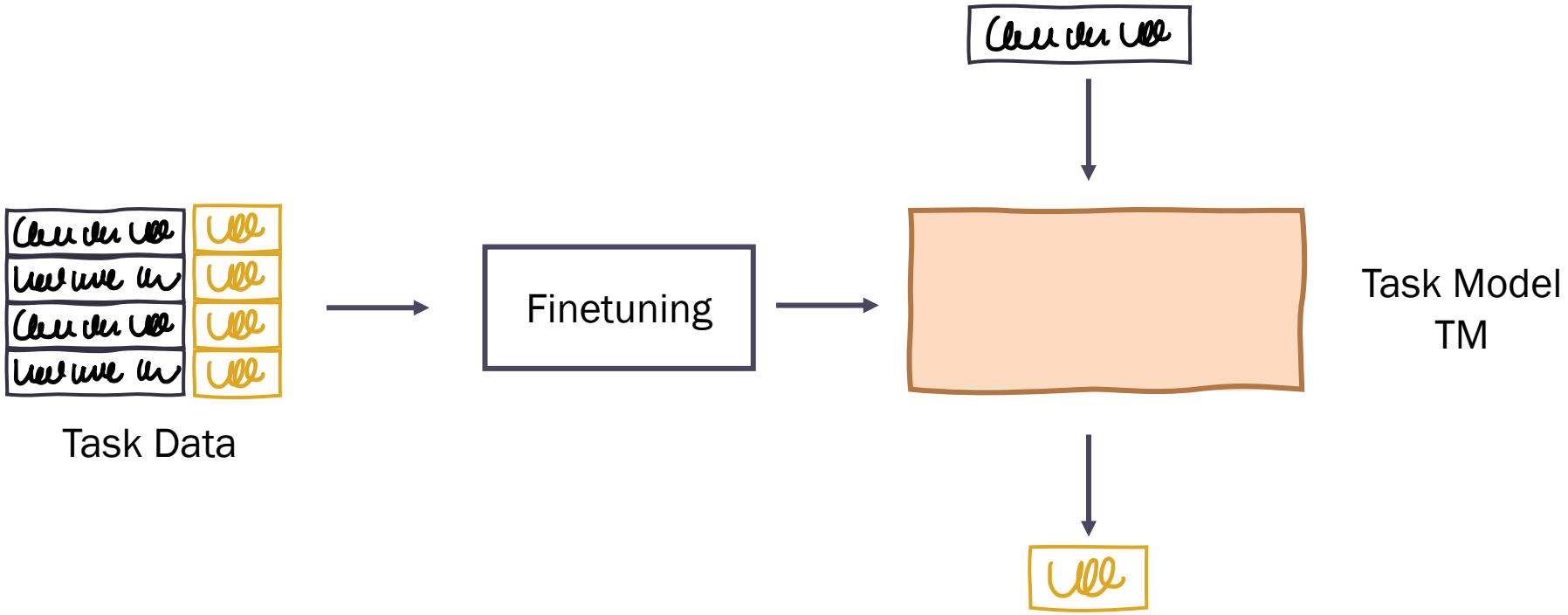
Sameer Singh

sameersingh.org

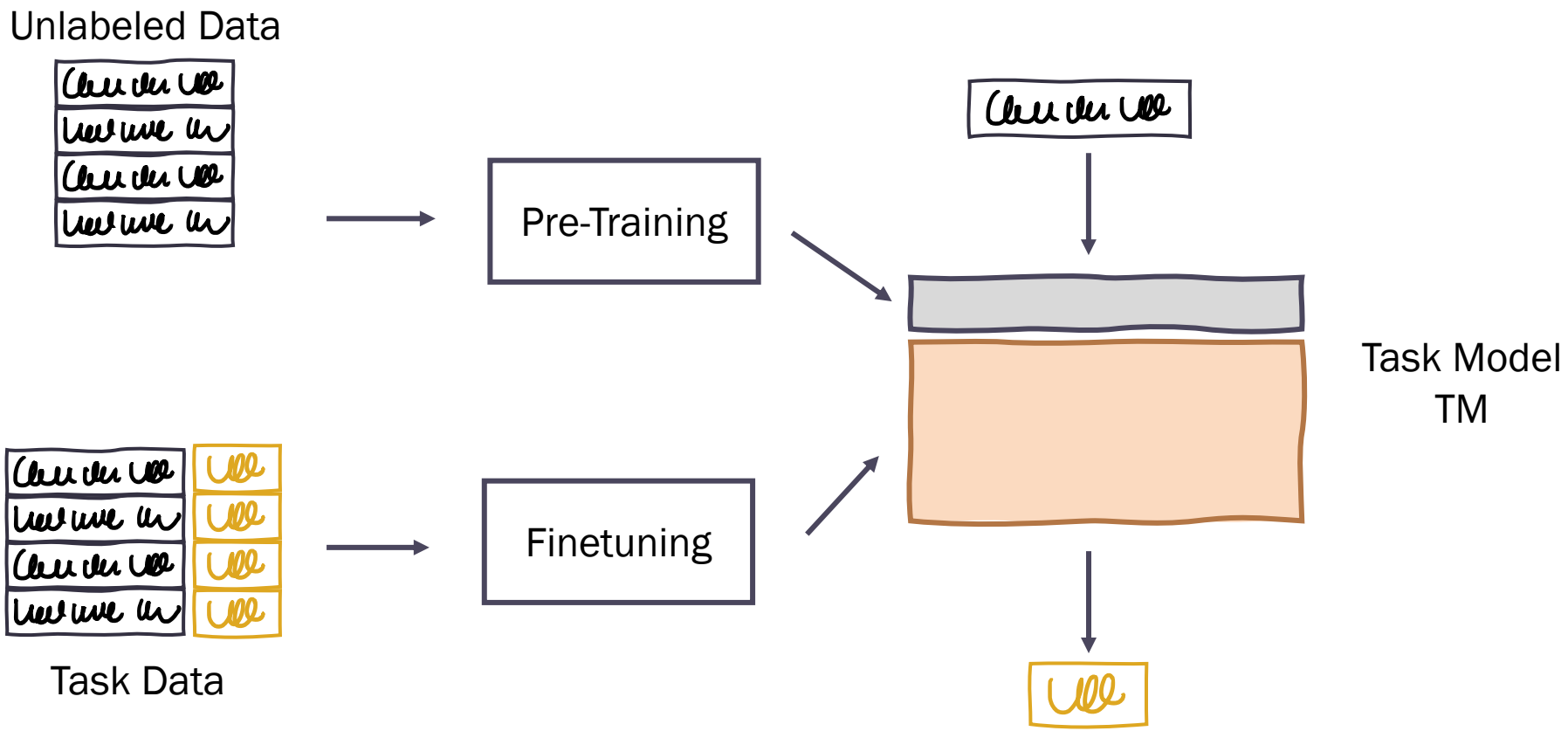
UCI nlp

AI2

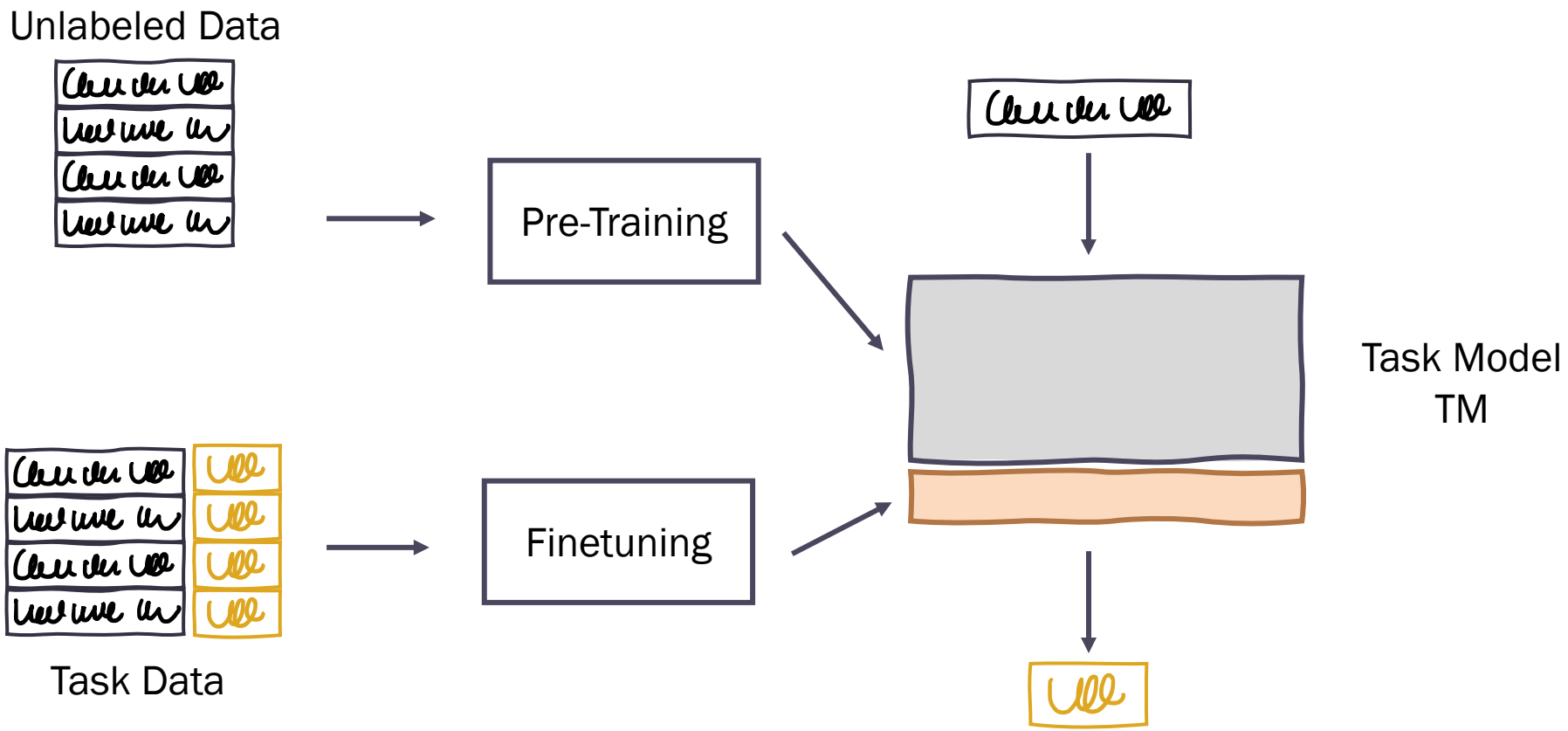
Natural Language Processing Pipeline



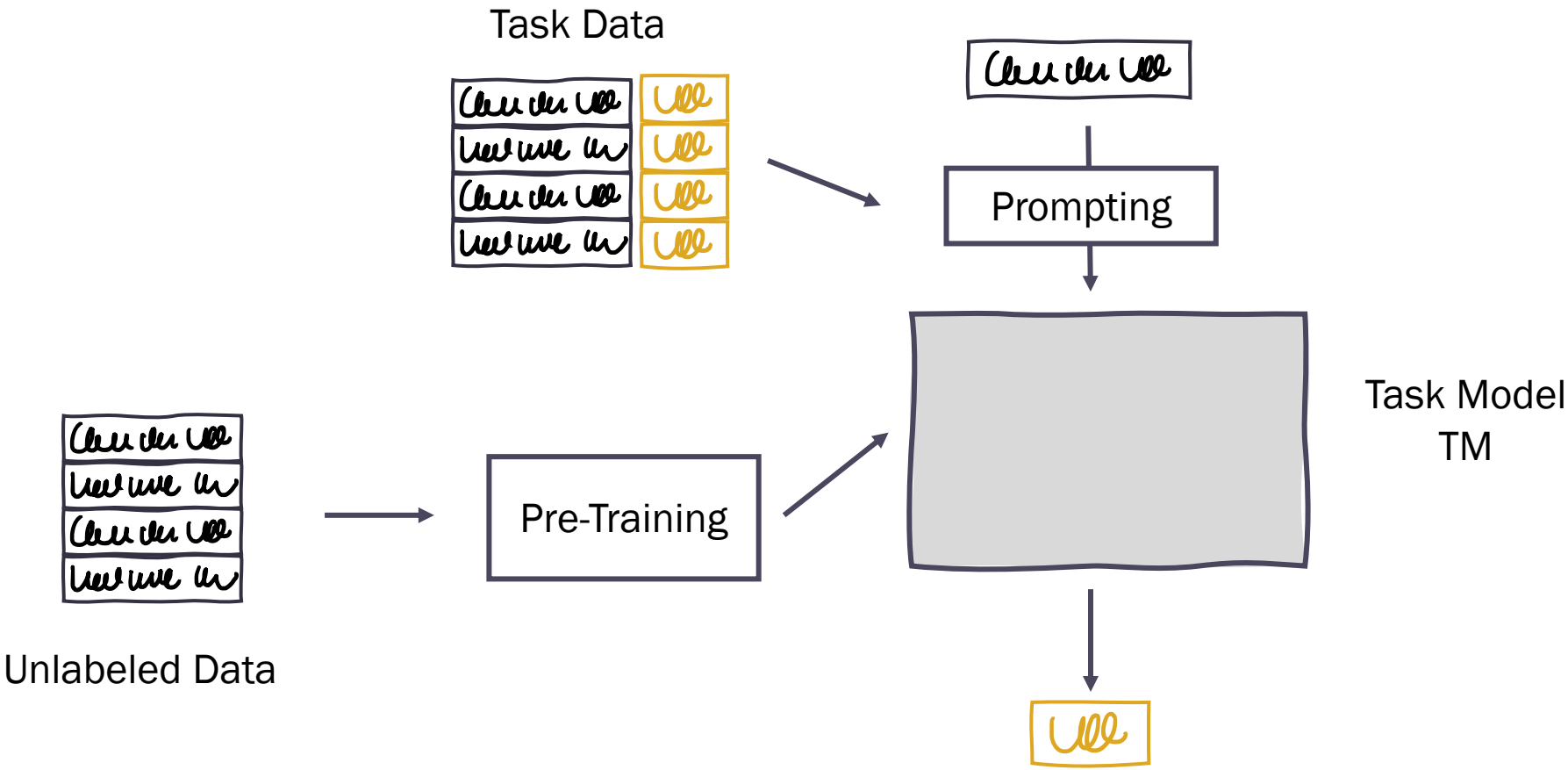
Natural Language Processing Pipeline



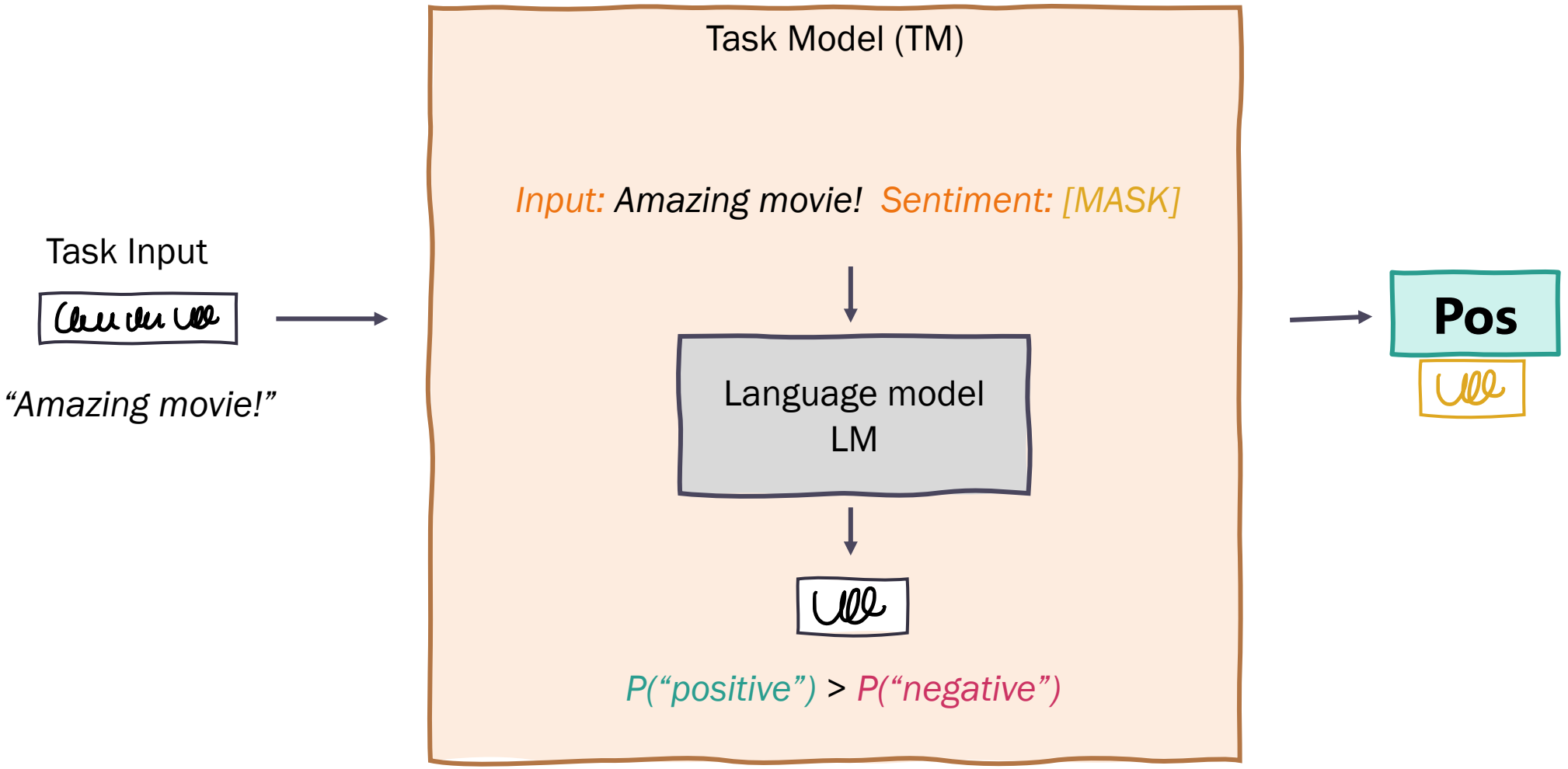
Natural Language Processing Pipeline



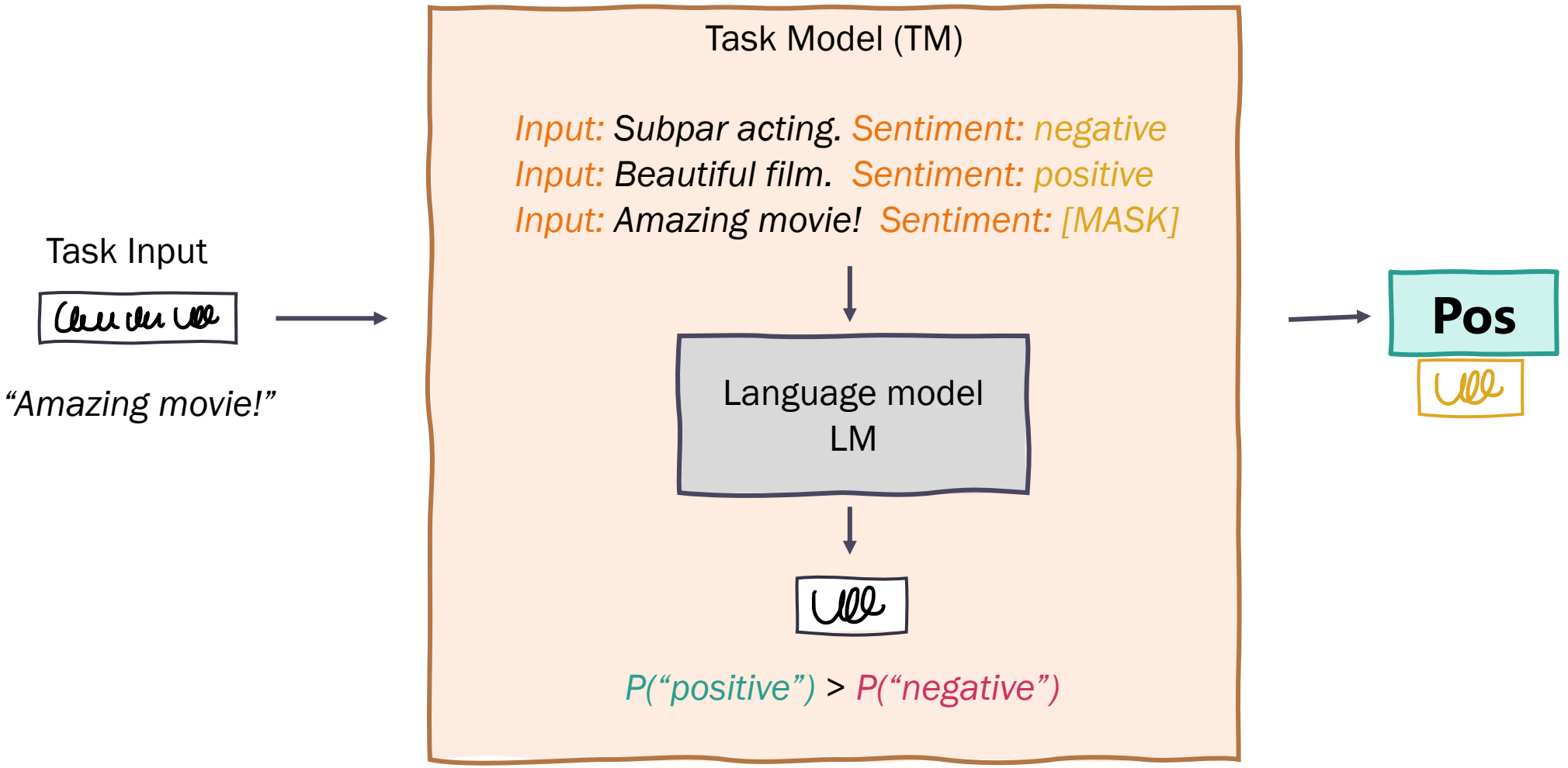
What's next? Get rid of finetuning!



Manual Prompts: Sentiment Analysis



In-Context Learning (Few-Shot Learning!)



Why is in-context learning interesting?

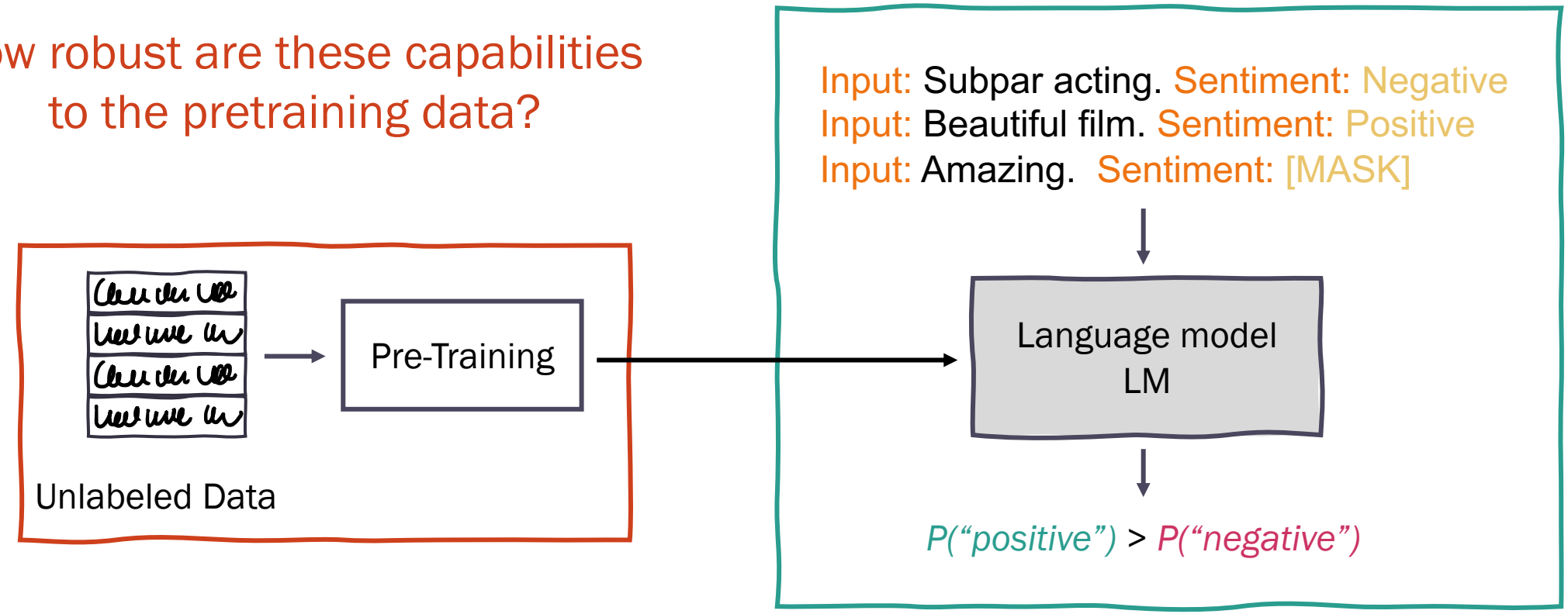
- Academically interesting
  - What do language models learn? How do we control them?
- Practically relevant (with GPT-3)
  - effective with ~0-16 examples
  - serve one model for many tasks
  - no ML expertise needed
- Related to other ways of adapting language models
  - AutoPrompt*: customized phrases to adapt LMs
  - Prompt/prefix tuning: continuous changes to input/weights
  - Increasingly more accurate and useful

* https://arxiv.org/abs/2010.15980

Today's Talk

What are the biases introduced by this format?

How robust are these capabilities to the pretraining data?

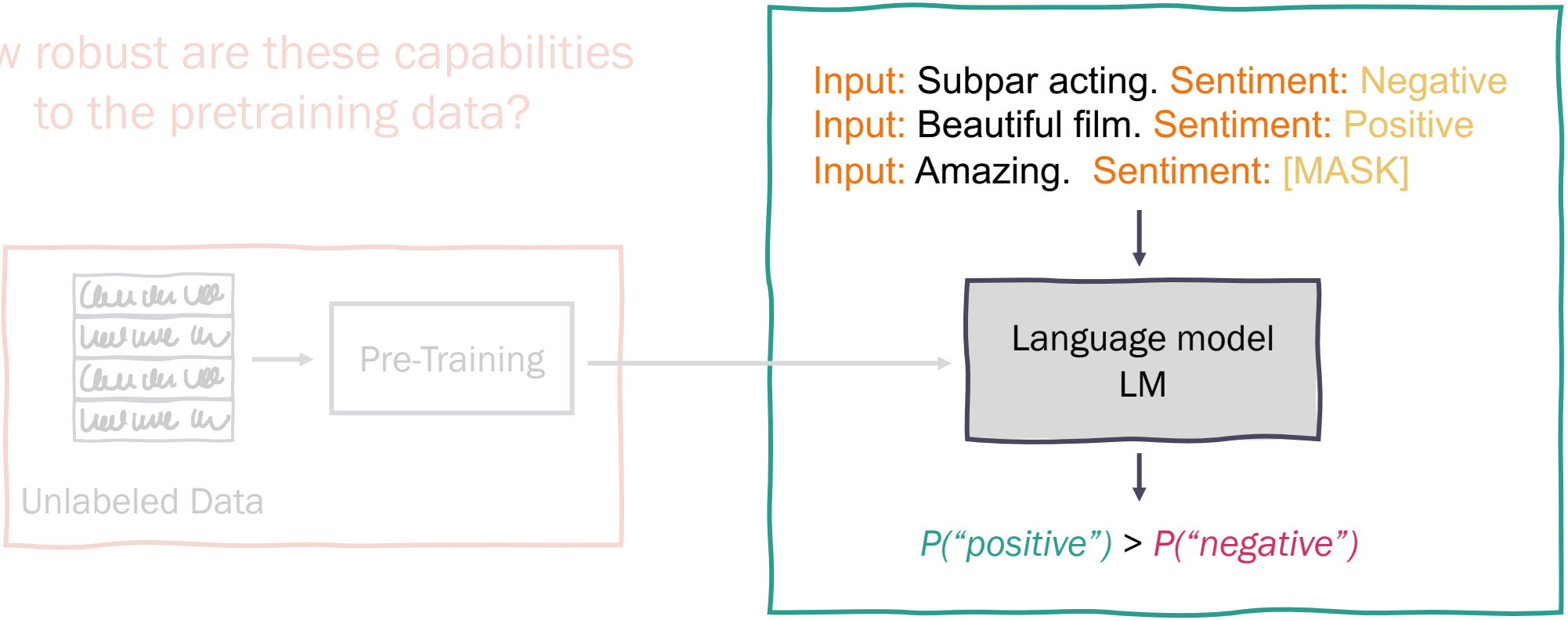


Today's Talk

Z. Zhao, E. Wallace, S. Feng, D. Klein, **S. Singh**.
Calibrate Before Use: Improving Few-shot Performance of Language Models.
International Conference on Machine Learning (ICML). 2021

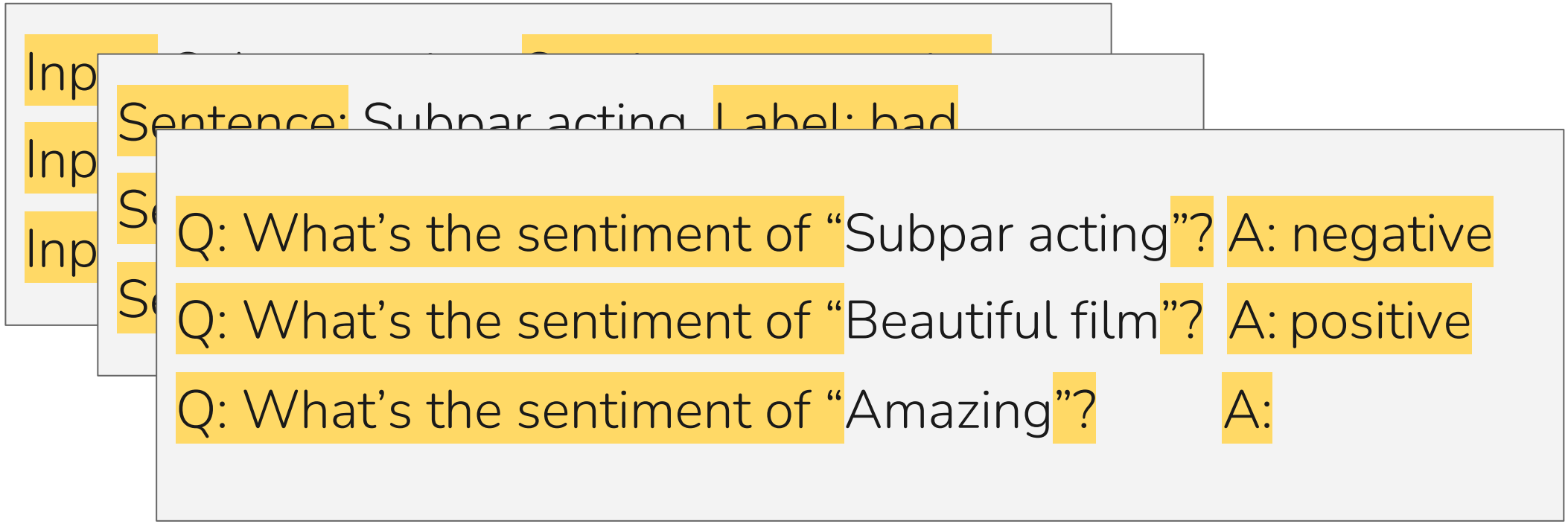
What are the biases introduced by this format?

How robust are these capabilities to the pretraining data?



Components Of The Prompt

Prompt Format



Components Of The Prompt

Training Example Selection

Input: Subpar acting. Sentiment: negative
Input: Beautiful film. Sentiment: positive
Input: Amazing. Sentiment:

Input: Good film. Sentiment: positive
Input: Don't watch. Sentiment: negative
Input: Amazing. Sentiment:

Components Of The Prompt

Training Example Perturbation

Input: Subpar acting. Sentiment: negative

Input: Beautiful film. Sentiment: positive

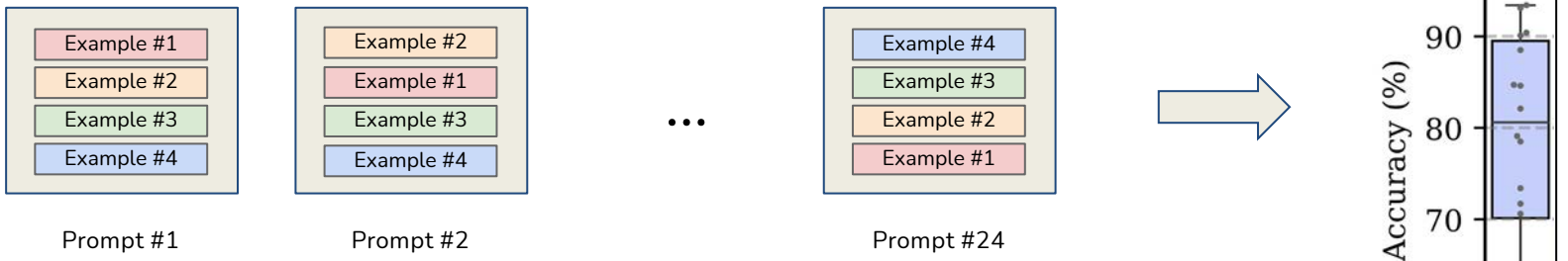
Input: Amazing. Sentiment:

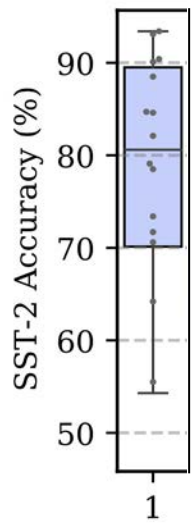
Input: Beautiful film. Sentiment: positive

Input: Subpar acting. Sentiment: negative

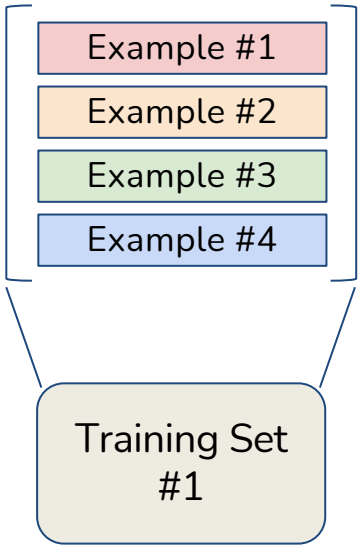
Input: Amazing. Sentiment:

Accuracy Is Highly Sensitive To Prompt Design



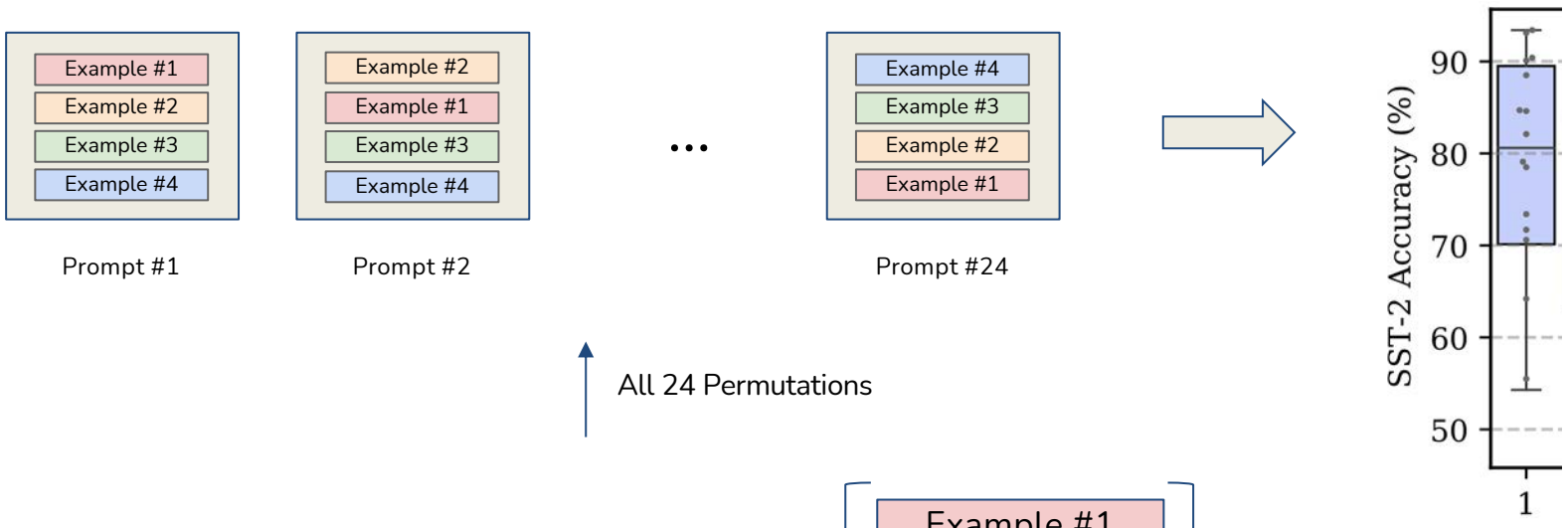


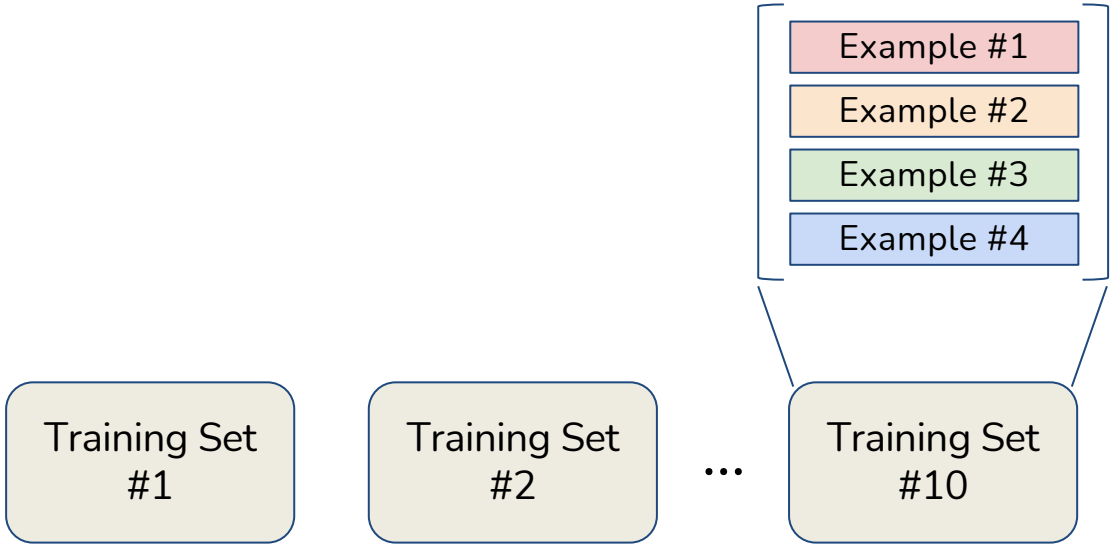
All 24 Permutations



Example Permutation Impacts Accuracy

Accuracy Is Highly Sensitive To Prompt Design





Example Selection Impacts Accuracy

Accuracy Is Highly Sensitive To Prompt Design

Format #1

Input: Subpar acting. Sentiment: negative
Input: Beautiful film. Sentiment: positive
Input: Amazing. Sentiment:

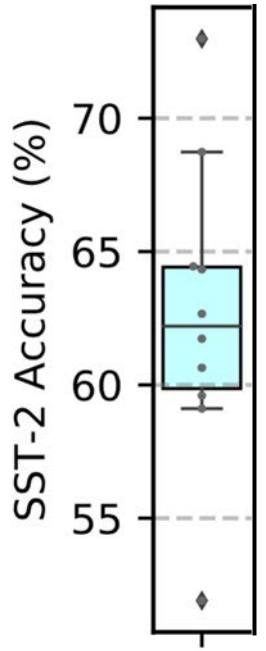
Format #2

Subpar acting. I hated the movie.
Beautiful film. I liked the movie.
Amazing.

⋮

Format #10

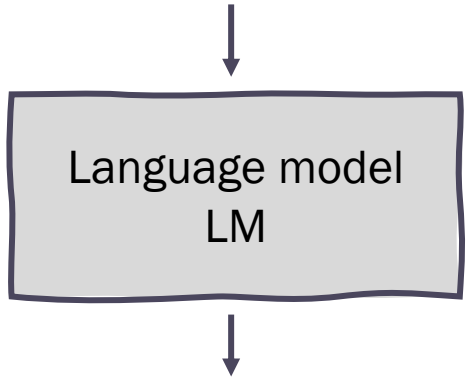
Review: Subpar acting. Stars: 0
Review: Beautiful film. Stars: 5
Review: Amazing. Stars:



Example Format Impacts Accuracy

In-Context Learning

Input: Meh movie. Sentiment: Negative
Input: Subpar acting. Sentiment: Negative
Input: Beautiful film. Sentiment: Positive
Input: Amazing. Sentiment: [MASK]



P("good") > *P("bad")*

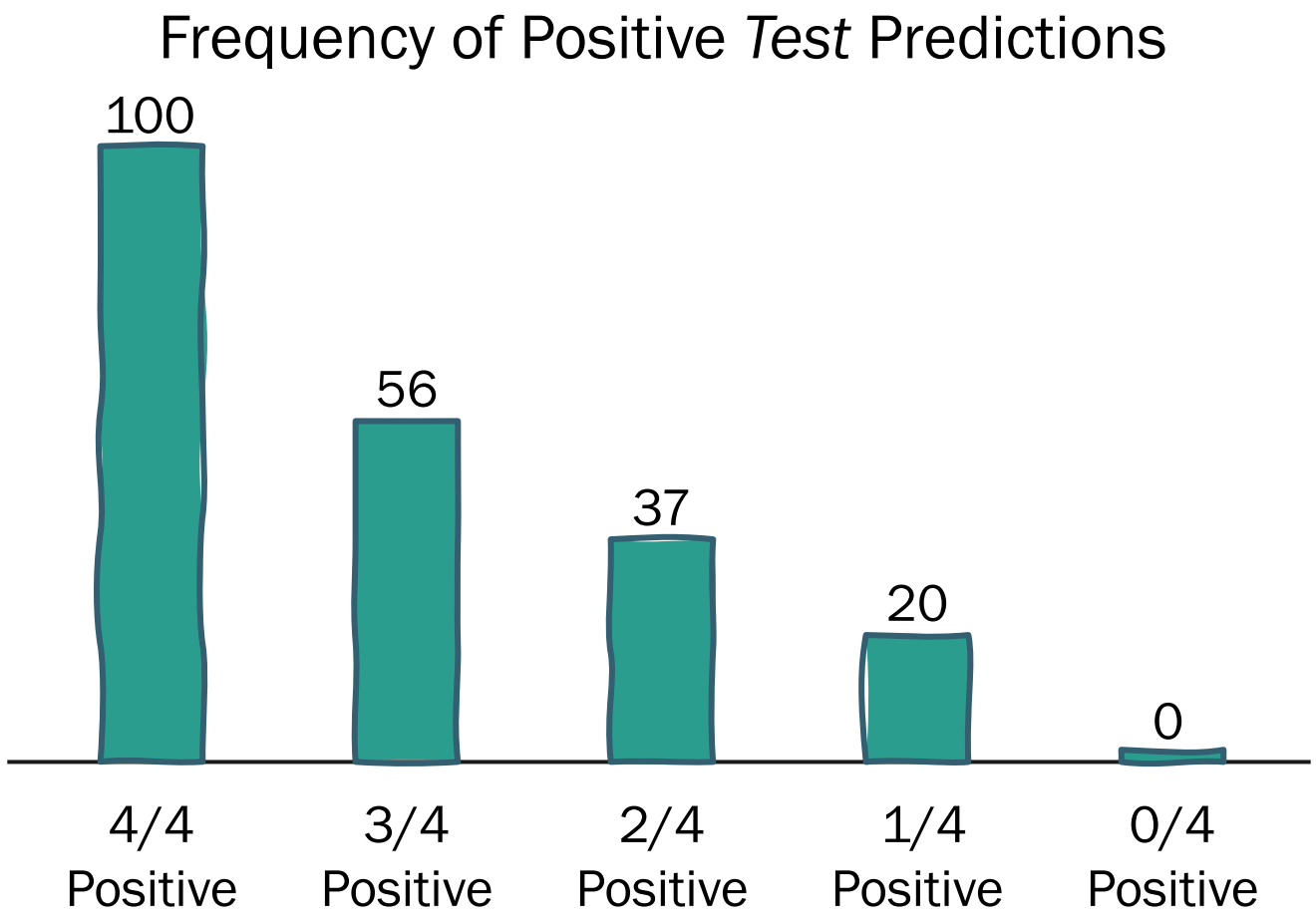
Format matters

Selection matters

Order matters

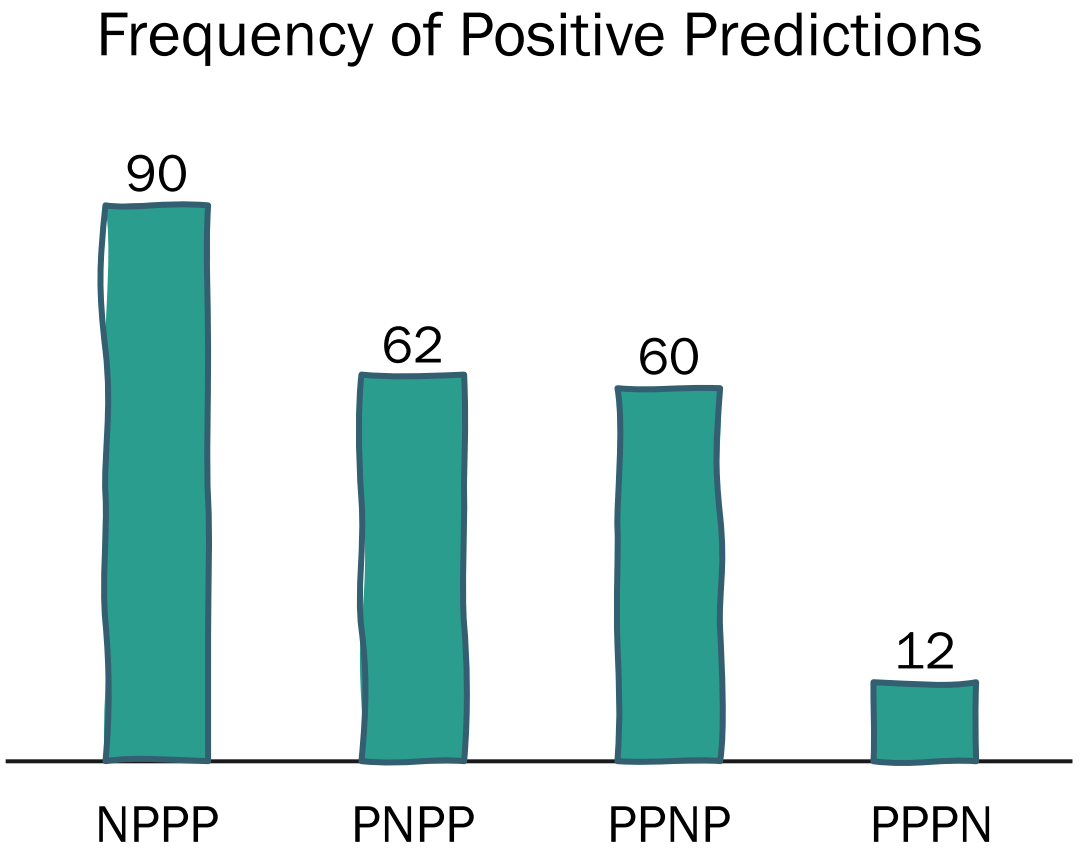
Word choice matters

Majority Label Bias



Frequent training answers dominate predictions

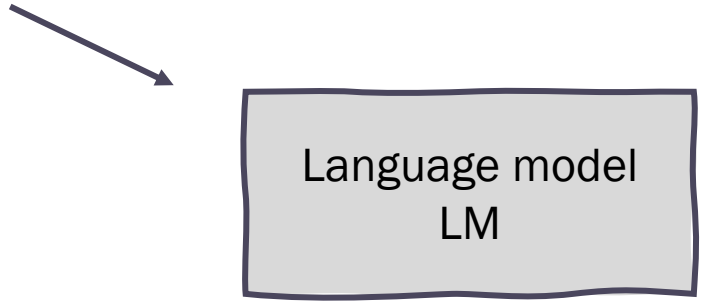
Recency Bias



Examples near end of prompt dominate predictions

Common Token Bias

What topic is the following text about?
The Model T was released by Ford in 1908.
Answer:



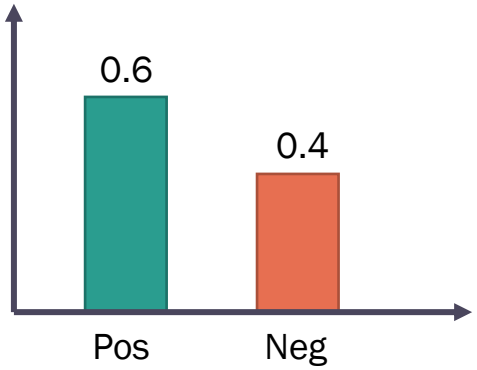
Token	Prob
book	0.35
transportation	0.23
school	0.11
village	0.03
company	0.02

Token	Web (%)	Label (%)	Prediction (%)
book	0.026	9	29
transportation	0.0000006	9	4

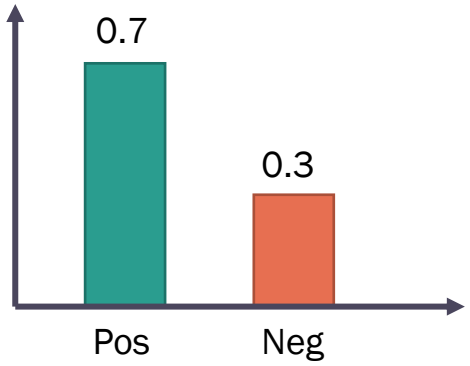
Common n-grams dominate predictions

Contextual Calibration of Language Models

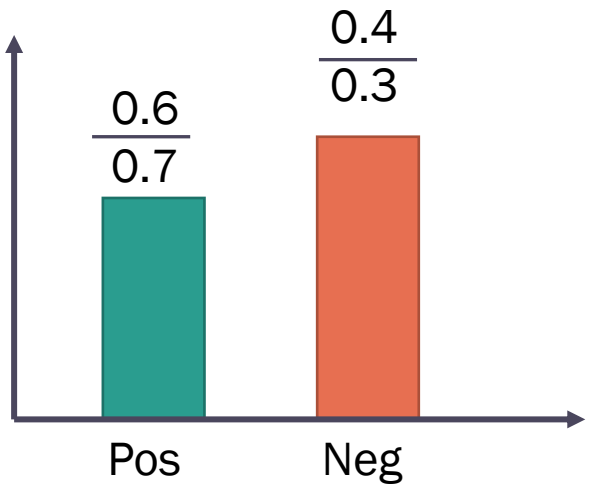
Input: Subpar acting. Sentiment: Negative
Input: Beautiful film. Sentiment: Positive
Input: Amazing. Sentiment: _____



Input: Subpar acting. Sentiment: Negative
Input: Beautiful film. Sentiment: Positive
Input: **N/A**. Sentiment: _____



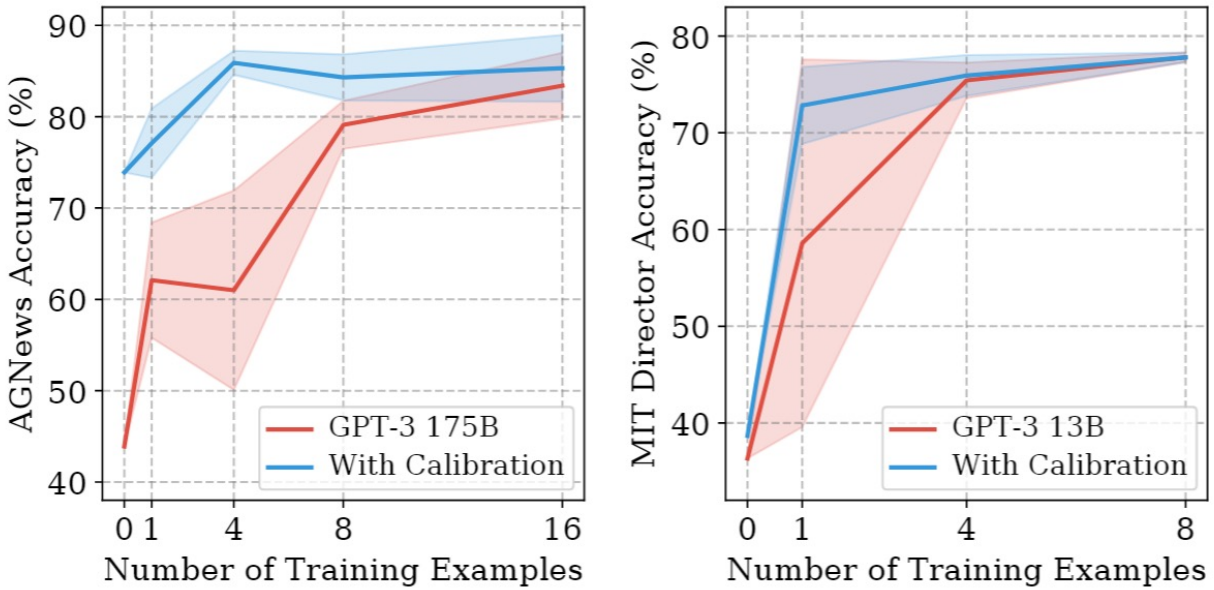
"meaningless" input,
but full context



More Accurate and Stable!

11 different datasets, 0-16 shots, GPT-2 and GPT-3 models

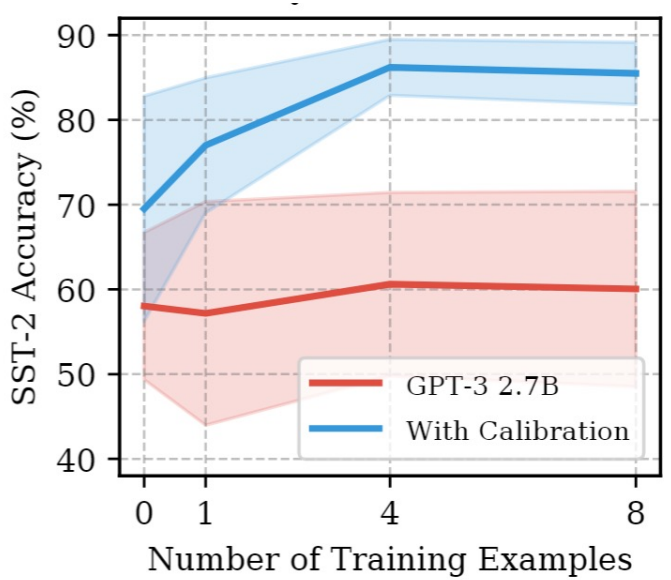
Different Training Examples



Improved mean and worst accuracy

Reduced variance for selection and ordering

Different Prompt Formats



Reduced variance for formats

Contextual Calibration for In-context Learning

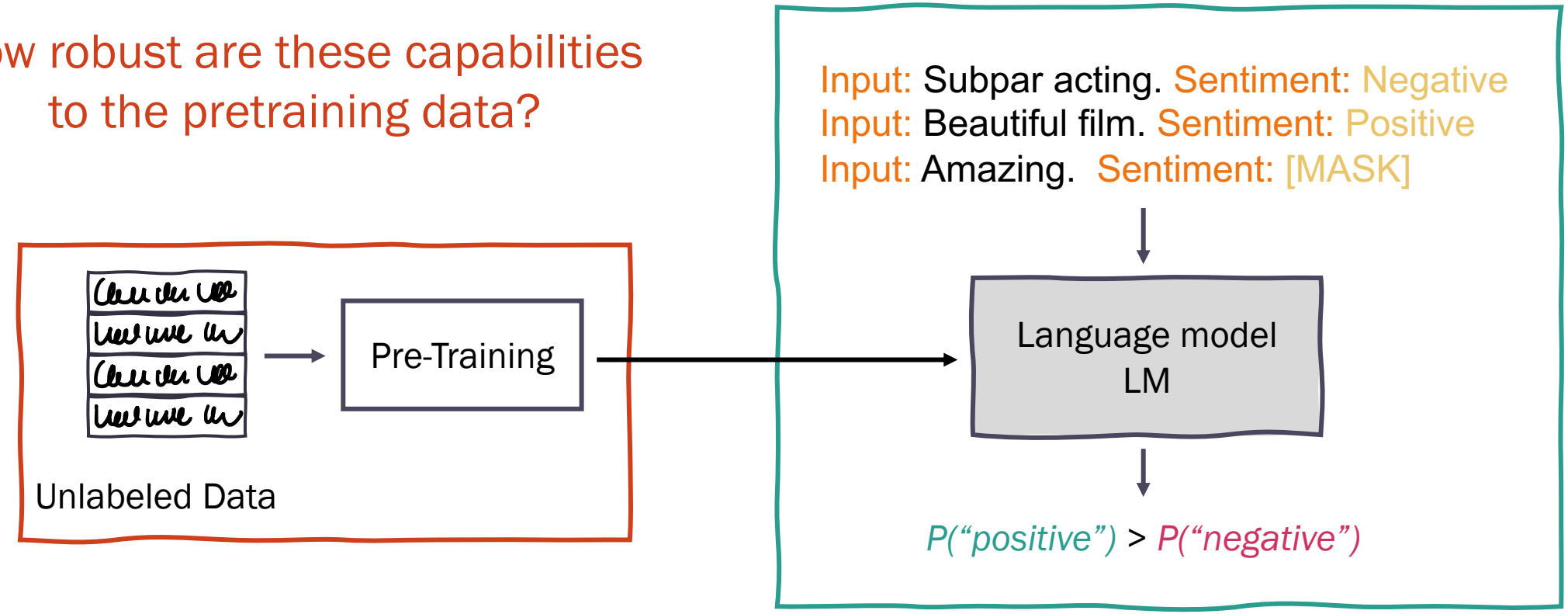
+ *extremely* simple fix
+ boosts accuracy, reduces variance

- Calibration doesn't completely solve brittleness
- Independent of the pretraining corpus

Today's Talk

What are the biases introduced by this format?

How robust are these capabilities to the pretraining data?

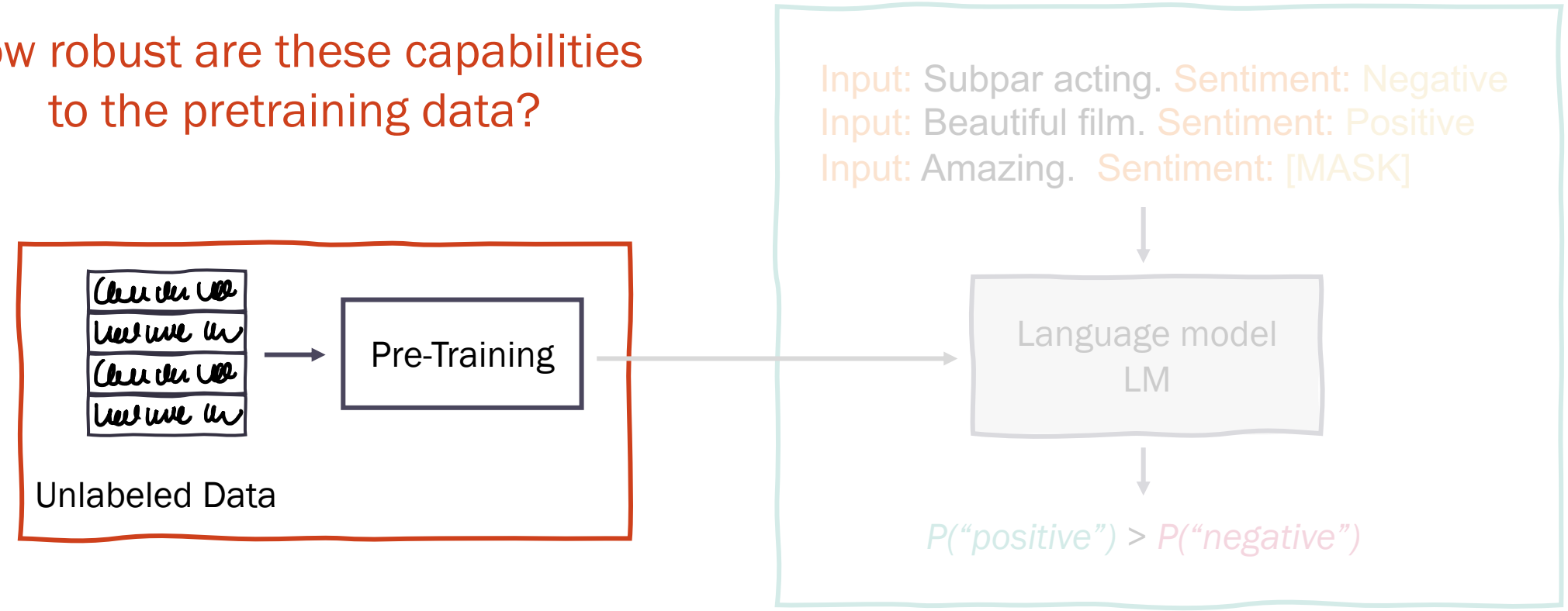


Today's Talk

Y. Razeghi, R. Logan, M. Gardner, **S. Singh**.
Impact of Pretraining Term Frequencies on Few-Shot Reasoning
ArXiV. 2022

What are the biases introduced by this format?

How robust are these capabilities to the pretraining data?



Reasoning and In-context Learning

- Instead of downstream classification, let's focus on Reasoning
  - Difficult to define precisely, but it's about inference
  - Go beyond regurgitation of what it has already seen
  - Feels different from memorization of facts
- Language Models need to perform reasoning

Went for a long lunch today, it lasted _____.

Alex loves chewing bones, which is not a surprise, given that he's a _____.

I wanted it in 10 days, but it took 2 weeks, which made me _____.

- And in-context few-shot reasoning is fairly accurate!
  - But how much of this performance is robust reasoning?

Numerical Reasoning

- One of the fundamental reasoning tasks
  - Version of common-sense reasoning
- Piece of the Neural vs Symbolic debate
  - Can LMs learn to multiply numbers?
- Good few-shot performance by big LMs
  - LMs are not explicitly trained for them

```
Prompt -------
What is 75*10?
--------------
Output:
750
What is -0.002 take away 72.75?
-72.752
Calculate -0.5 - 1039.
-1039.5
What is the difference between -1360 and 2?
1362
What is -27.95 less than -20?
7.95
Calculate -0.3 + -169.
-169.3
What is 0.7 minus 0.05?
0.65
Calculate -2 + 0.0899.
-1.9101
```

Example from GPT-J blog:

Motivating Example: Multiplication

- Good performance but not always correct

Q: What is 24 times 18? A: 432 ✓

Q: What is 23 times 18? A: 462 ✗

$\Omega(24) \simeq 10^7$

$\Omega(23) \simeq 10^6$

Why does the model perform differently on different instances?

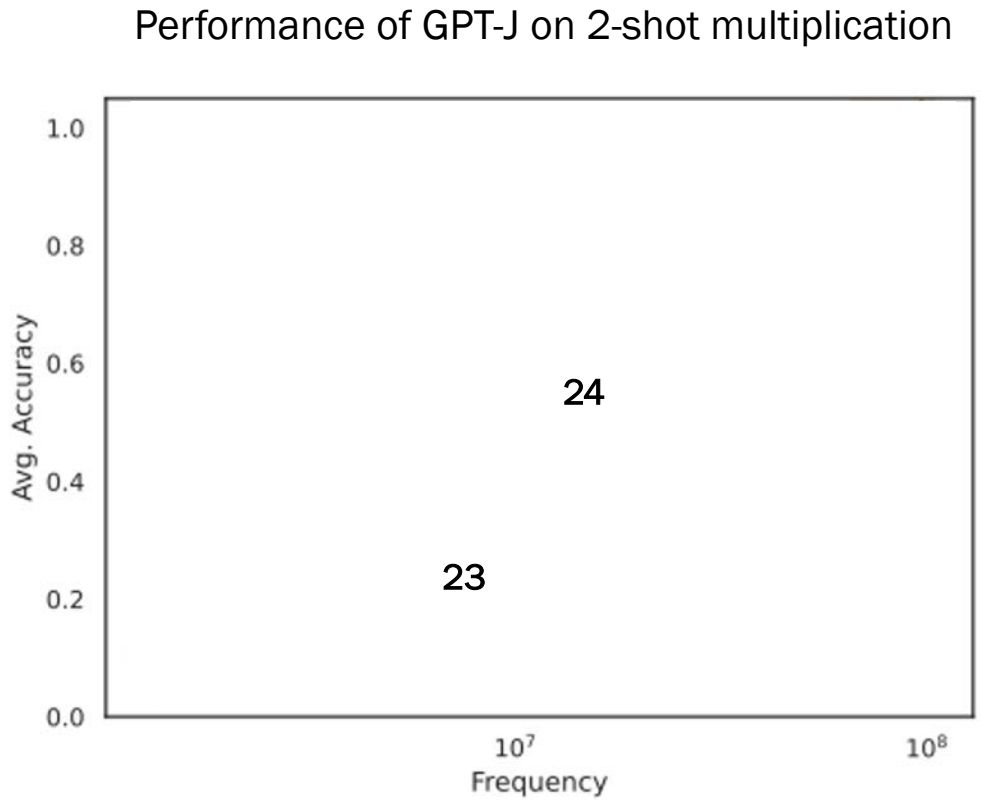
Hypothesis: maybe it depends on unigram statistics in pretraining?

Motivating Example: Multiplication

- First operand: numbers between 0-99
- Accuracy averaged over:
  - 5 choices of training instances
  - second operand: numbers in 1-50

Q: What is 24 times [x]? A: __

Q: What is 23 times [x]? A: __



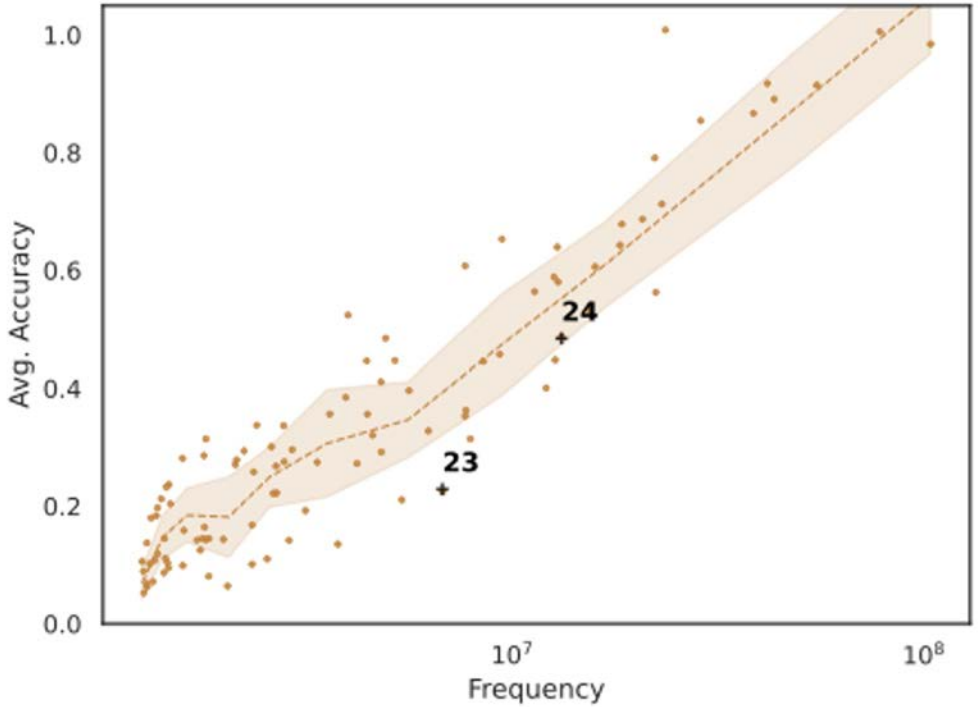
Motivating Example: Multiplication

- First operand: numbers between 0-99
- Accuracy averaged over:
  - 5 choices of training instances
  - second operands as numbers in 1-50

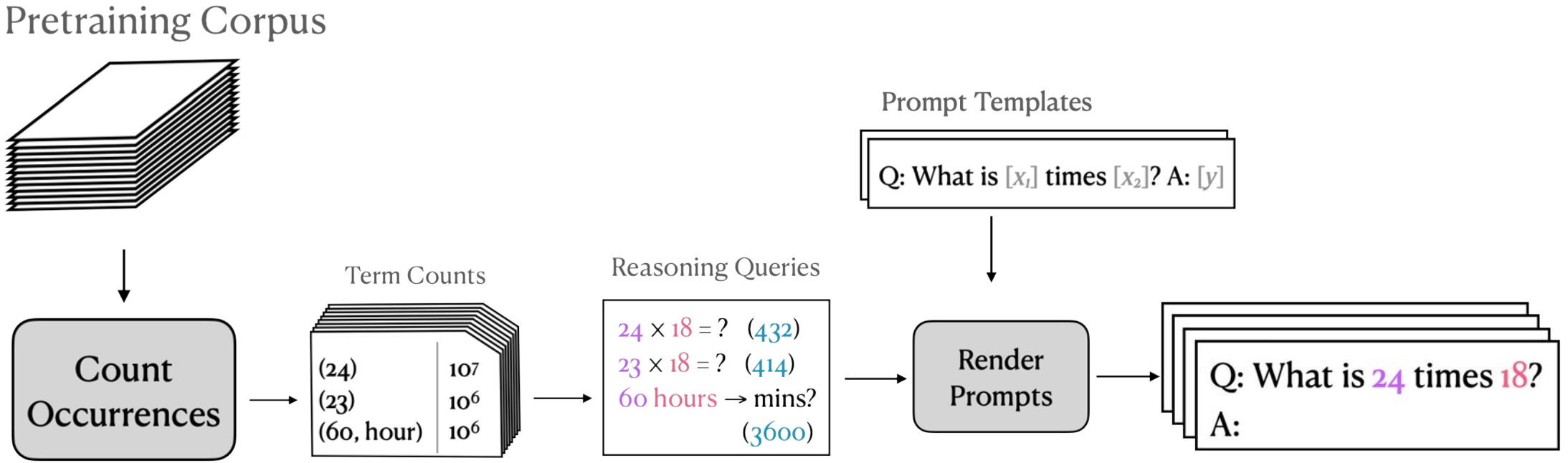
Q: What is 24 times [x]? A: __

Q: What is 23 times [x]? A: __

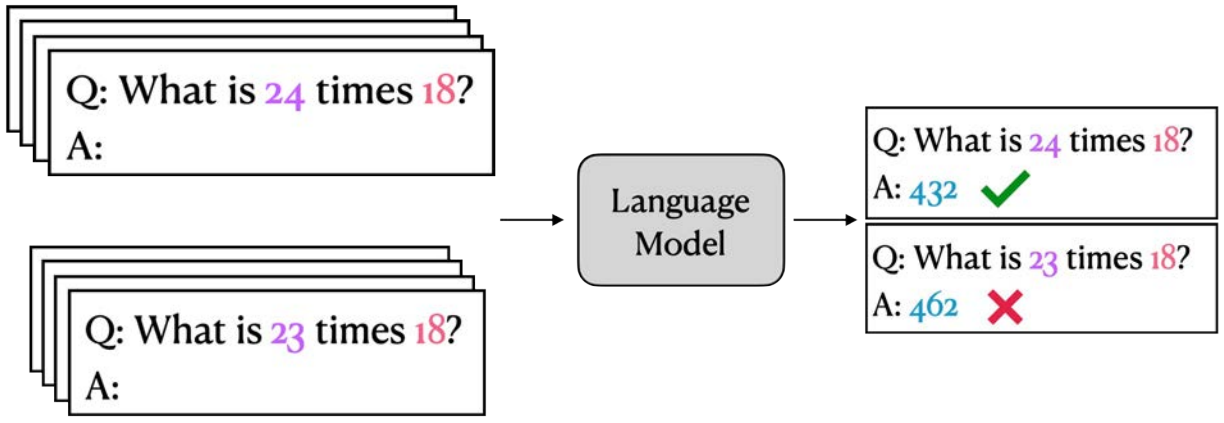
Performance of GPT-J on 2-shot multiplication

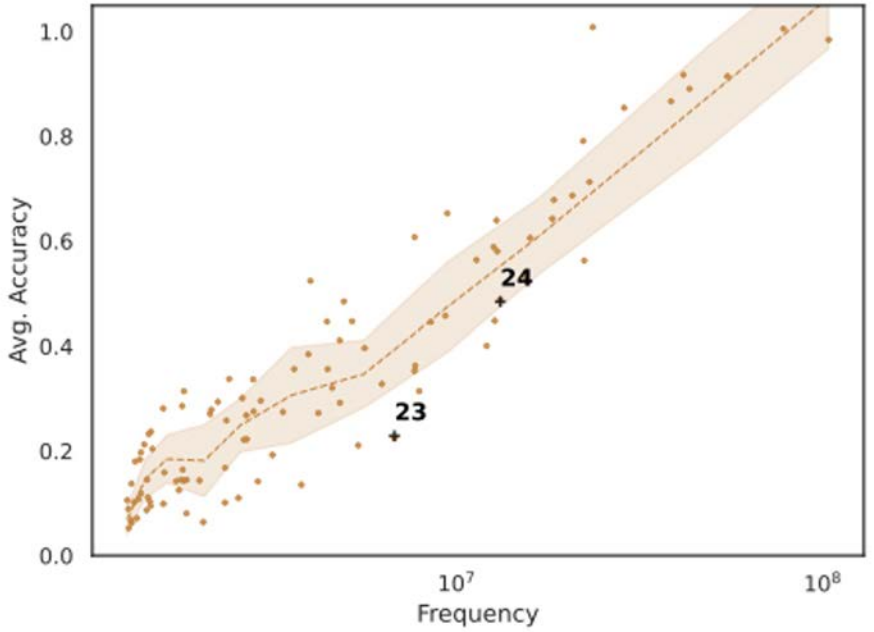


Pipeline for Evaluating this Effect



Analysis of Language Models





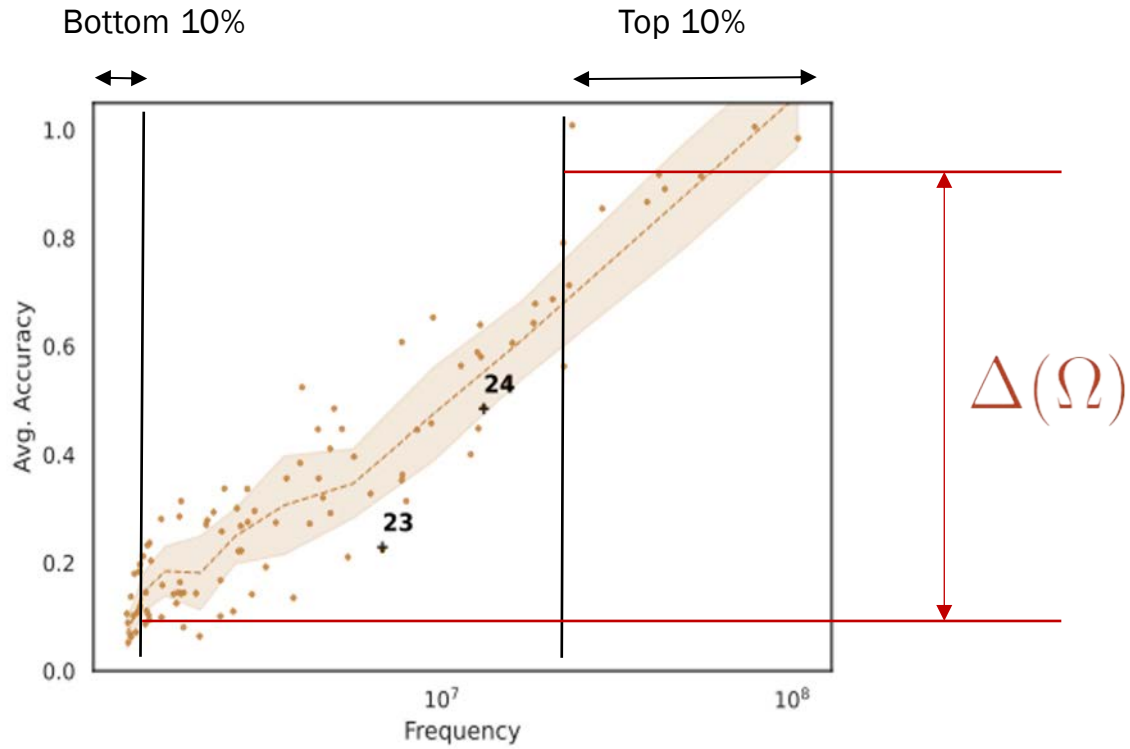
Term Counts

(24)	10^7
(23)	10^6
(60, hour)	10^6

Metric: Performance Gap

- Difference in average accuracy of the instances in the top and bottom quantiles of the distribution over term frequencies

$$\Delta(\Omega) = \mathrm{Acc}(\Omega_{>90\%}) - \mathrm{Acc}(\Omega_{<10\%})$$



Experiment Setup

- EleutherAI GPT-models
  - GPT-J-6B
  - GPT-Neo-2.7B
  - GPT-Neo-1.3B

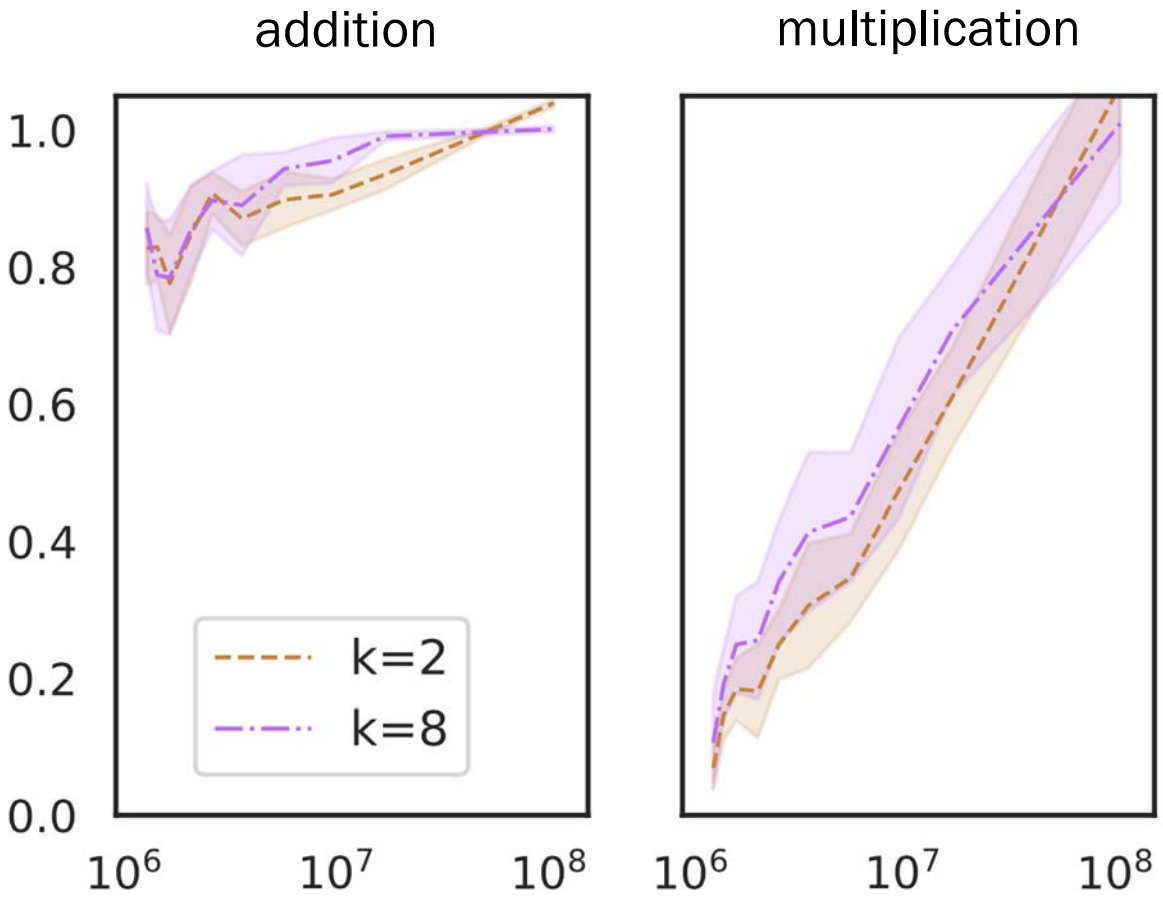
Pretrained on Pile Dataset

- 800GB pretraining corpus
- Publicly available!

Training examples in the prompt:

- Randomly choose *k* examples
- 5 choice of random seeds

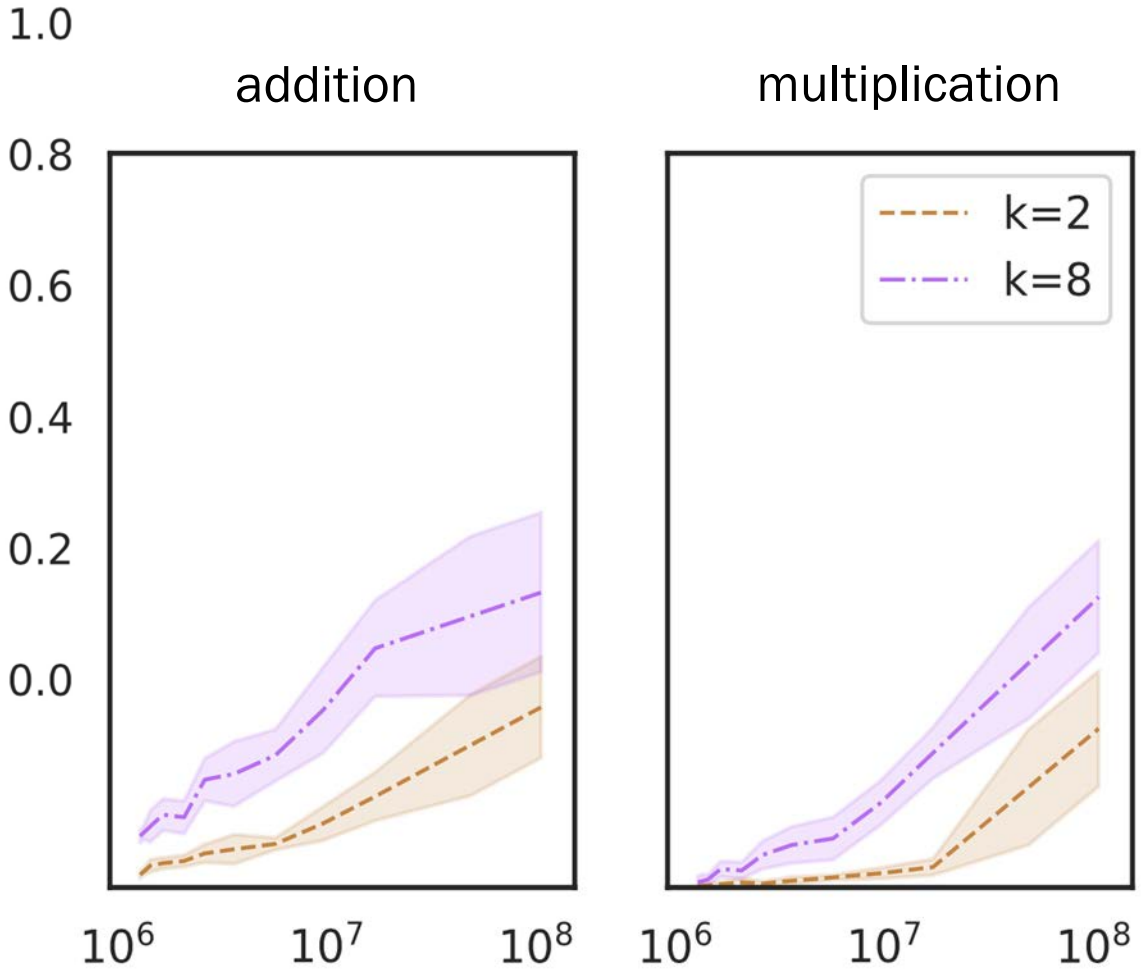
Arithmetic Reasoning



Q: What is 24 plus [x]? A: __
Q: What is 24 times [x]? A: __

k	Multiplication		Addition	
	Acc.	Δ_1	Acc.	Δ_1
0	5.4	18.0	1.6	8.4
2	35.9	77.6	88.2	16.8
4	39.2	70.8	91.4	15.0
8	42.9	74.6	89.6	16.3
16	40.9	73.3	88.6	16.4

Operation Inference



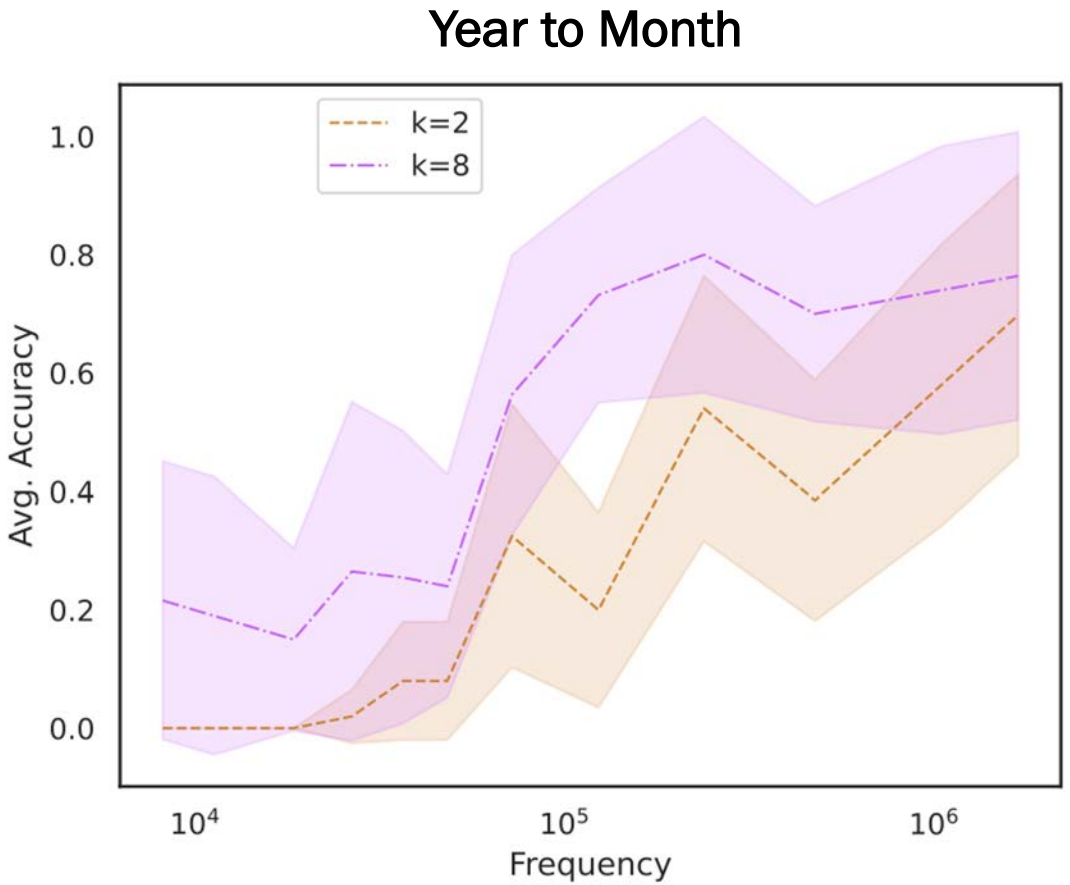
Q: What is 24 # [x]? A: __
Q: What is 24 # [x]? A: __

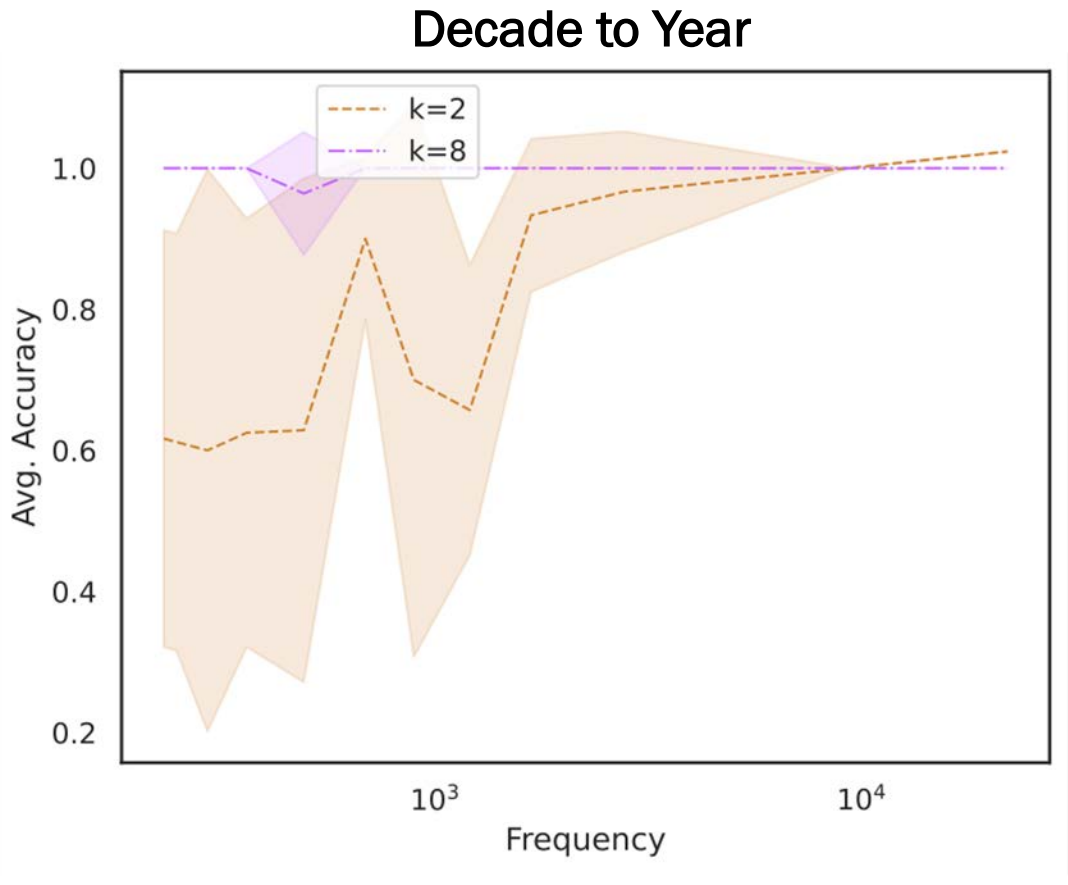
k	Multiplication (#)		Addition (#)	
	Acc.	Δ_1	Acc.	Δ_1
0	-	-	-	-
2	3.1	14.1	7.8	18.1
4	5.7	20.9	9.8	24.8
8	9.4	31.3	19.8	31.0
16	11.0	39.6	26.2	38.5

Time Unit Conversion

- Minute to Seconds — Q: What is 24 minutes in seconds? A: __
- Hour to Minutes — Q: What is 24 hours in minutes? A: __
- Day to Hour — Q: What is 24 days in hours? A: __
- Week to Day — Q: What is 24 weeks in days? A: __
- Month to Week — Q: What is 24 months in weeks? A: __
- Year to Month — Q: What is 24 years in months? A: __
- Decade to Year — Q: What is 24 decades in years? A: __

Time Unit Conversion

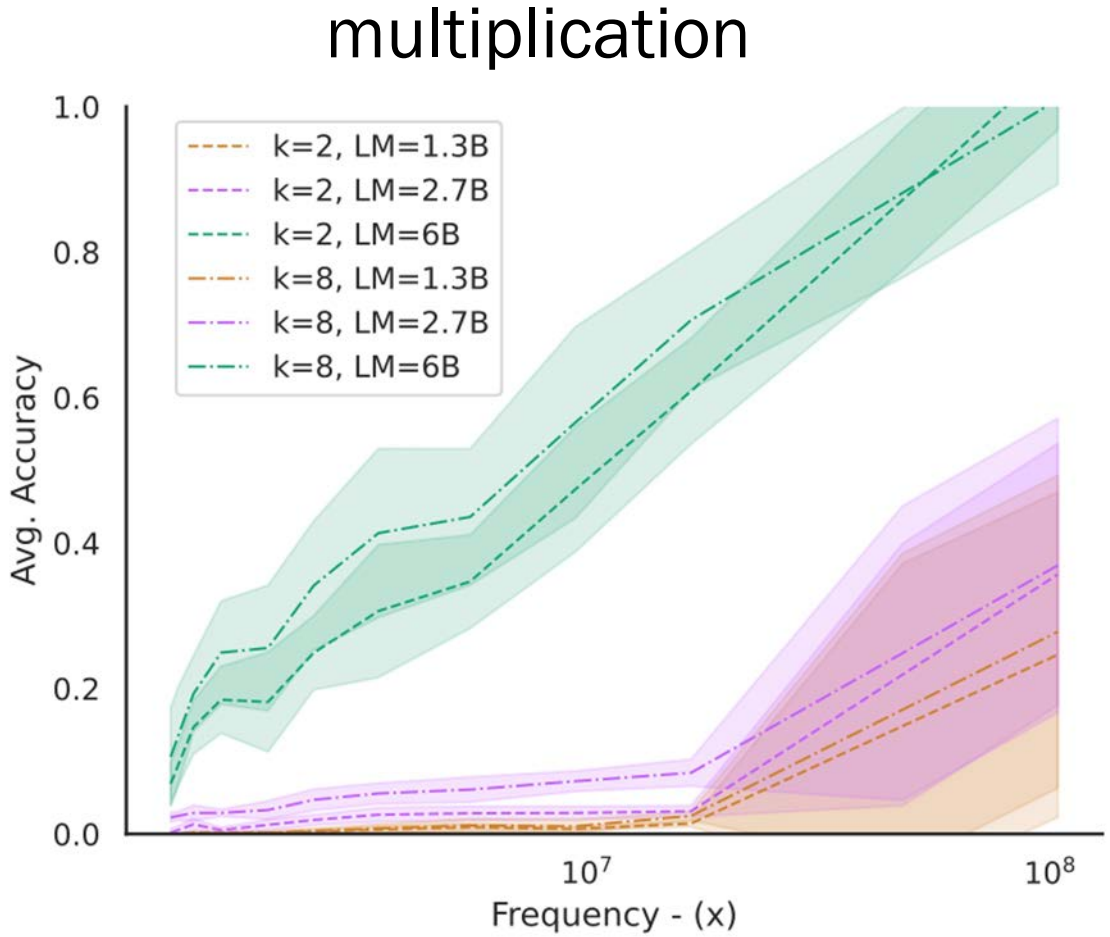




Time Unit Conversion

k	Min→Sec		Hour→Min		Day→Hour		Week→Day	
	Acc.	$\Delta_{1,2}$	Acc.	$\Delta_{1,2}$	Acc.	$\Delta_{1,2}$	Acc.	$\Delta_{1,2}$
0	1.3	0.0	1.0	0.0	1.0	0.0	1.0	0.0
2	25.5	62.5	19.4	58.0	12.1	28.9	13.1	43.5
4	35.5	60.0	29.1	76.4	22.7	46.4	19.2	40.9
8	49.9	72.1	36.3	74.6	31.0	59.1	28.6	70.6
16	58.4	82.7	42.8	80.1	43.3	62.8	28.0	22.1

Effect of Model Size



- As we increase size of model
  - Models get more accurate
  - But, more impacted by pretraining
- Number of shots is inconsistent
  - more training doesn't lead to robust reasoning by itself
- Difficult to detangle accuracy
  - By scale itself is not a solution

Effect of Pretraining on Reasoning

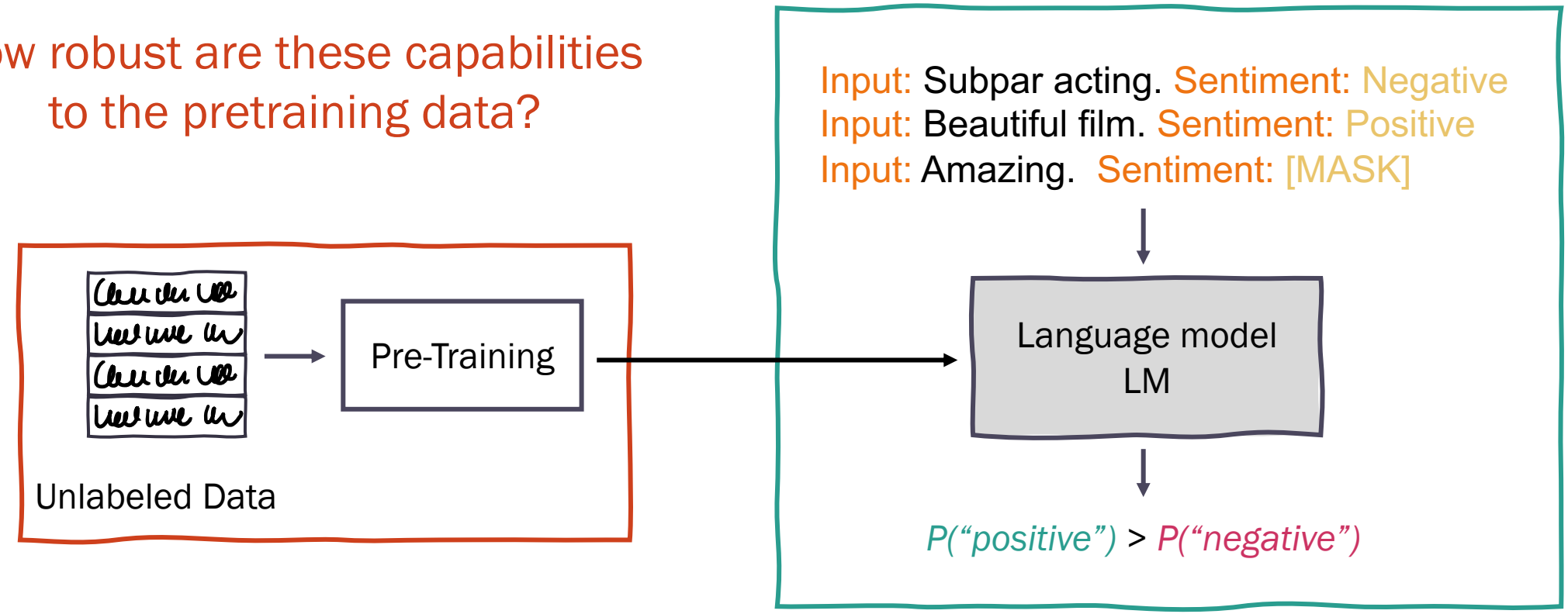
+ *high* impact on reasoning performance
+ raises questions about how to design, and evaluate, LMs

- we are not making a causal statement about reasoning
- only evaluated on numerical reasoning

Today's Talk

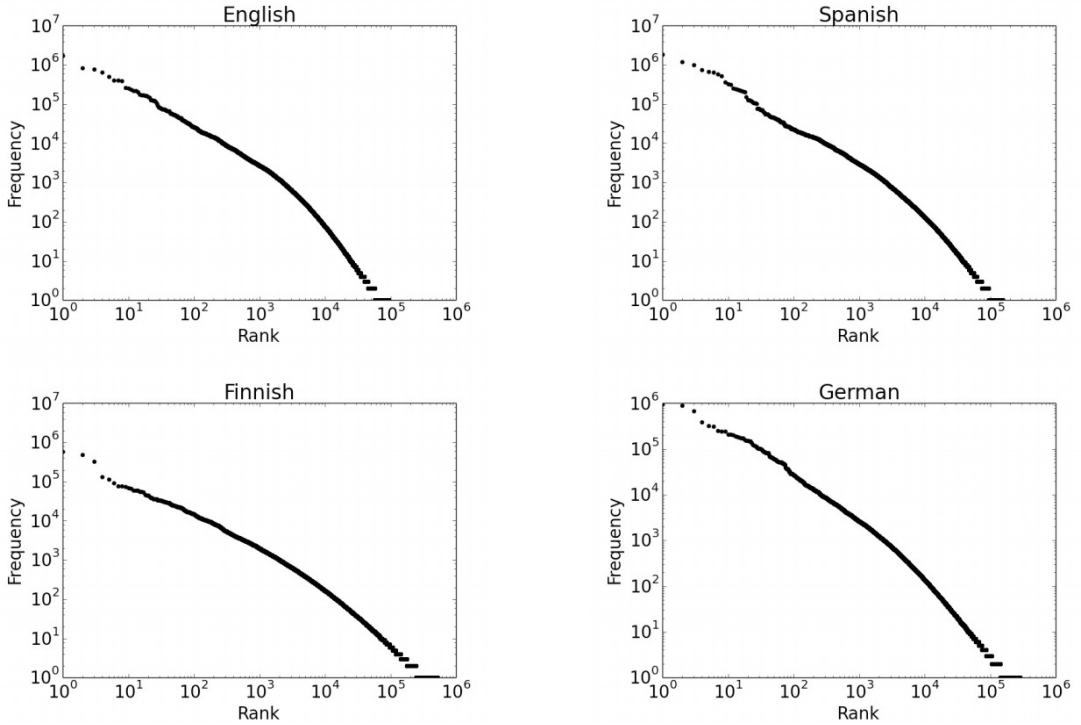
What are the biases introduced by this format?

How robust are these capabilities to the pretraining data?

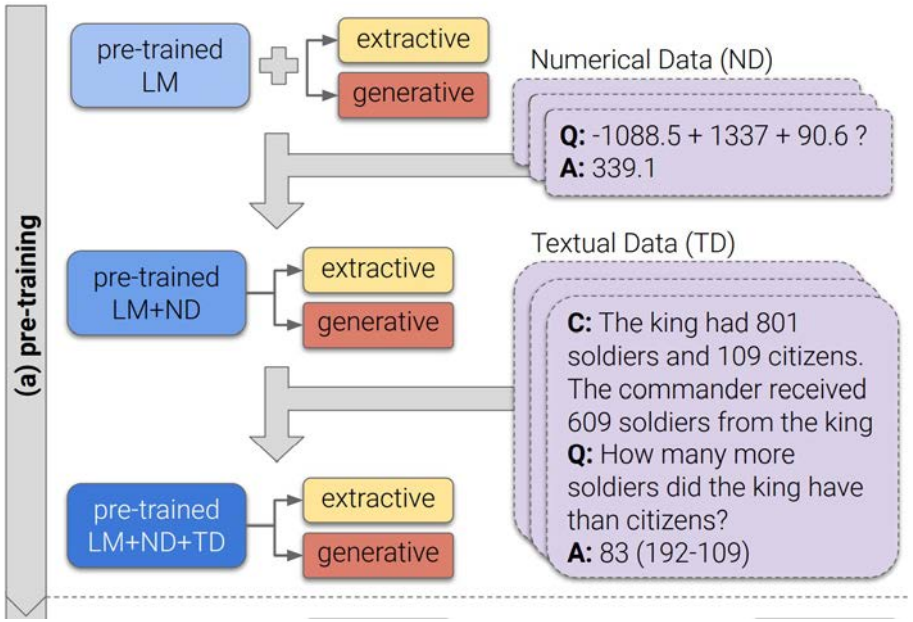


What Can We Do?

- More diverse data is better!
  - Will suffer from Zipf's Law
  - Future is more unique than the past



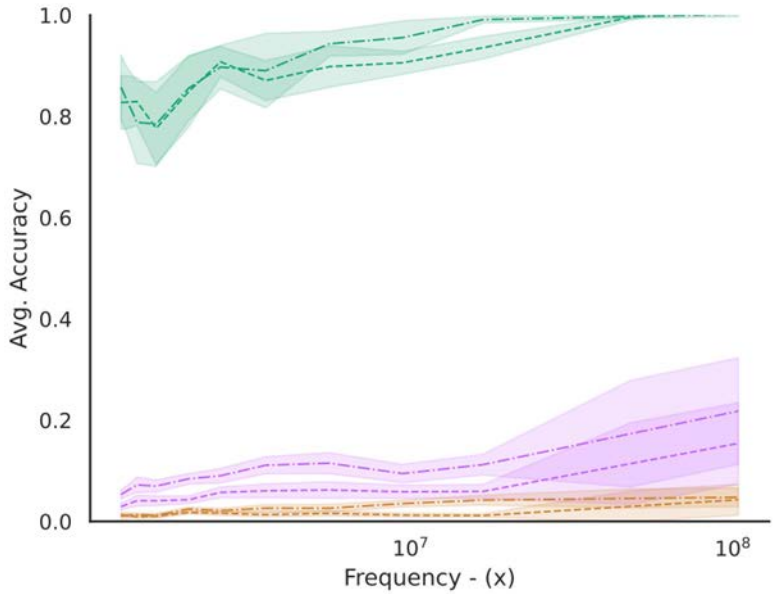
- Augmentation during pretraining?
  - Add data to address specific reasoning
  - Good for fixing the issues we have observed
  - Doesn't feel like the end goal



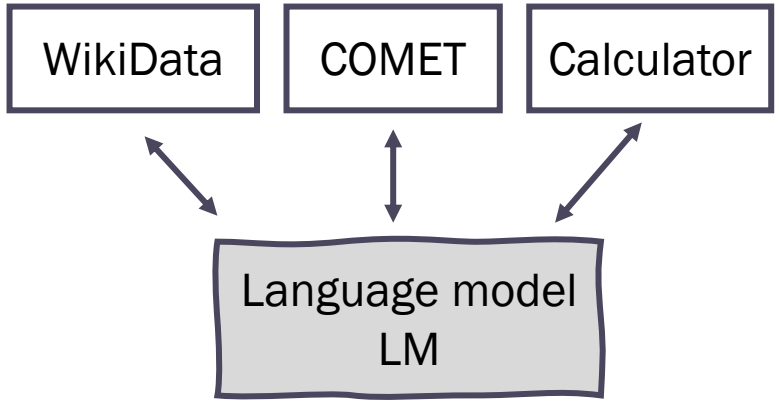
https://arxiv.org/abs/2004.04487

What Can We Do?

- Maybe scaling further will help?
  - Ultimately, they will just generalize perfectly?



- Neuro-symbolic language modeling?
  - Give LMS access to KGs, calculators, etc.
  - *Barack's Wife Hillary … [ACL 2019]* *



- Other losses for pretraining?
  - Should words really compete with each other?

* https://arxiv.org/abs/1906.07241

UCI nlp

AI2

Thank you!

@sameer_
sameer@uci.edu
sameersingh.org