Calibrate Before Use: Improving Few-Shot Performance of Language Models

Tony Z. Zhao* Eric Wallace* Shi Feng
Dan Klein Sameer Singh

ICML 2021

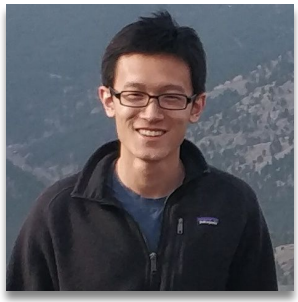
UC Berkeley

University of Maryland

UC Irvine

Tony Zhao
UC Berkeley

Eric Wallace
UC Berkeley



Shi Feng
UMD

Dan Klein
UC Berkeley

Sameer Singh
UC Irvine

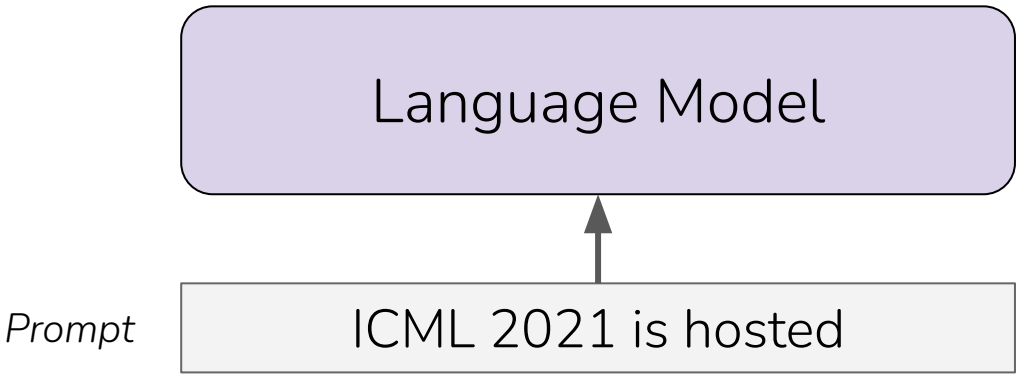
Slides, Blog, Code, and Video ericswallace.com/calibrate

Neural Language Models (LMs)

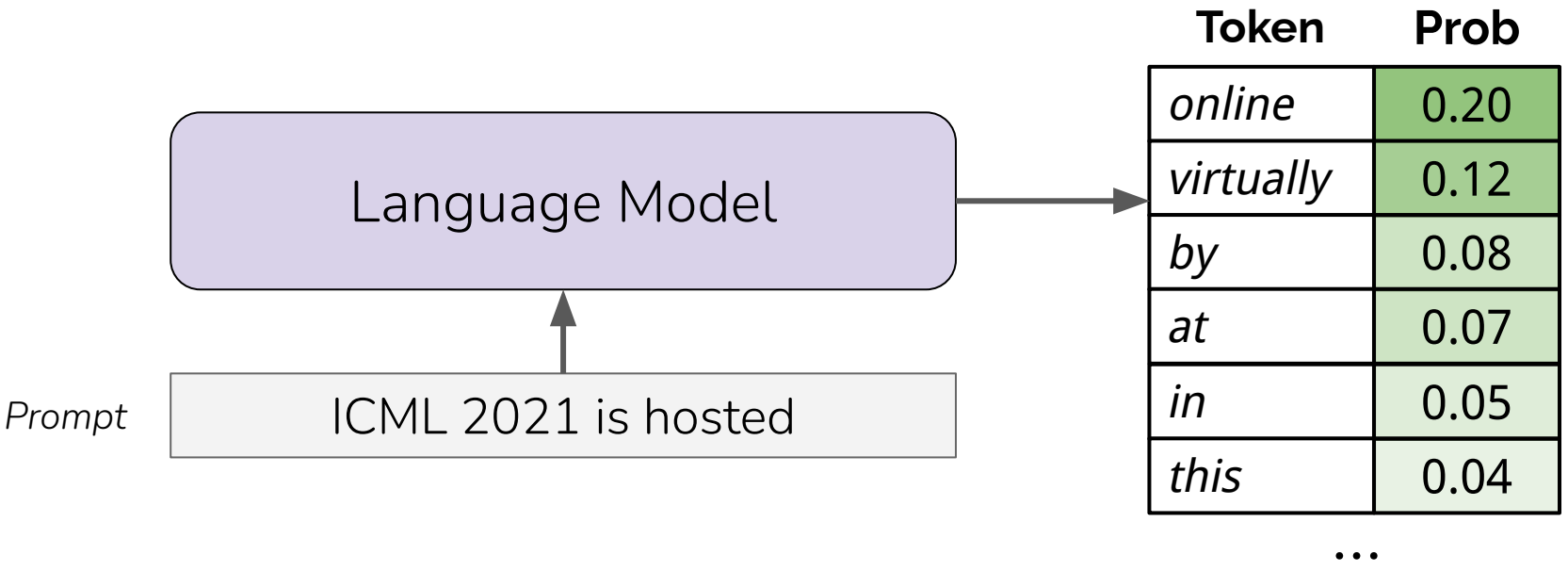
Neural Language Models (LMs)

Prompt ICML 2021 is hosted

Neural Language Models (LMs)



Neural Language Models (LMs)

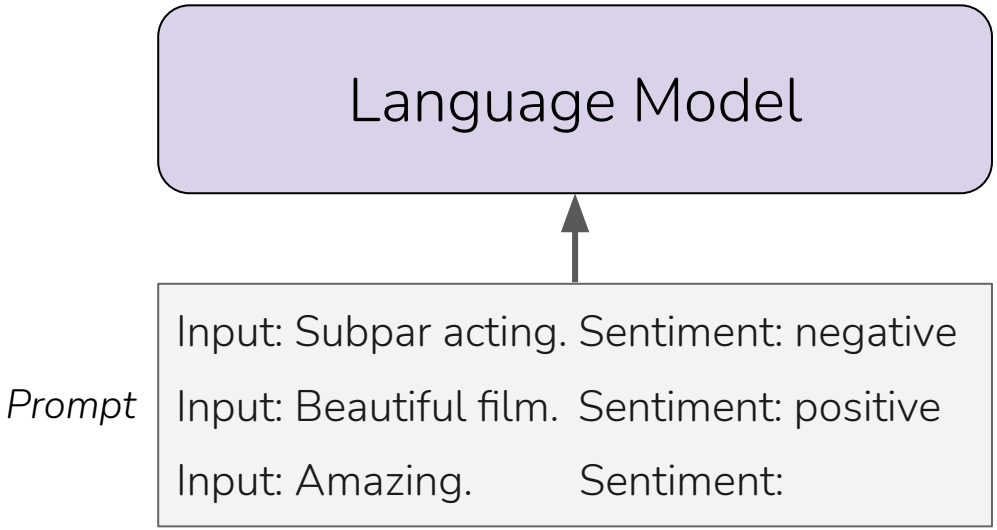


Few-shot Learning with LMs

Sentiment Analysis

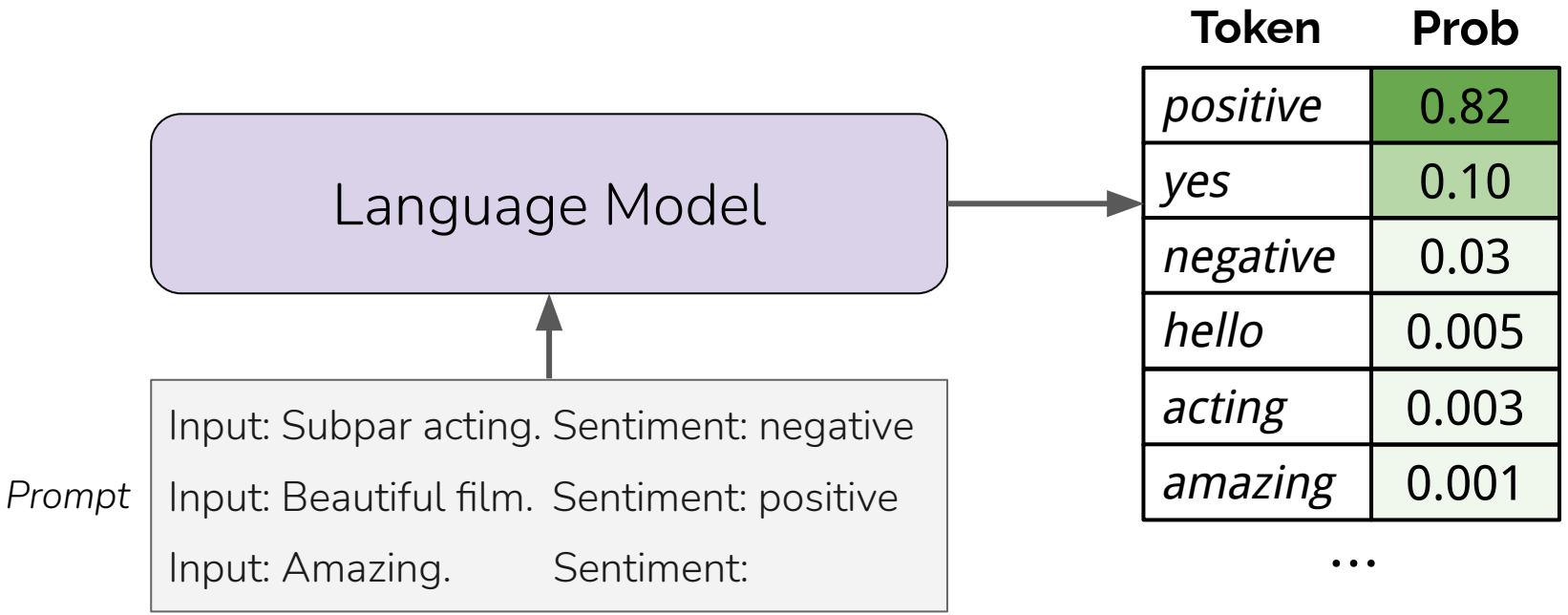
Few-shot Learning with LMs

Sentiment Analysis



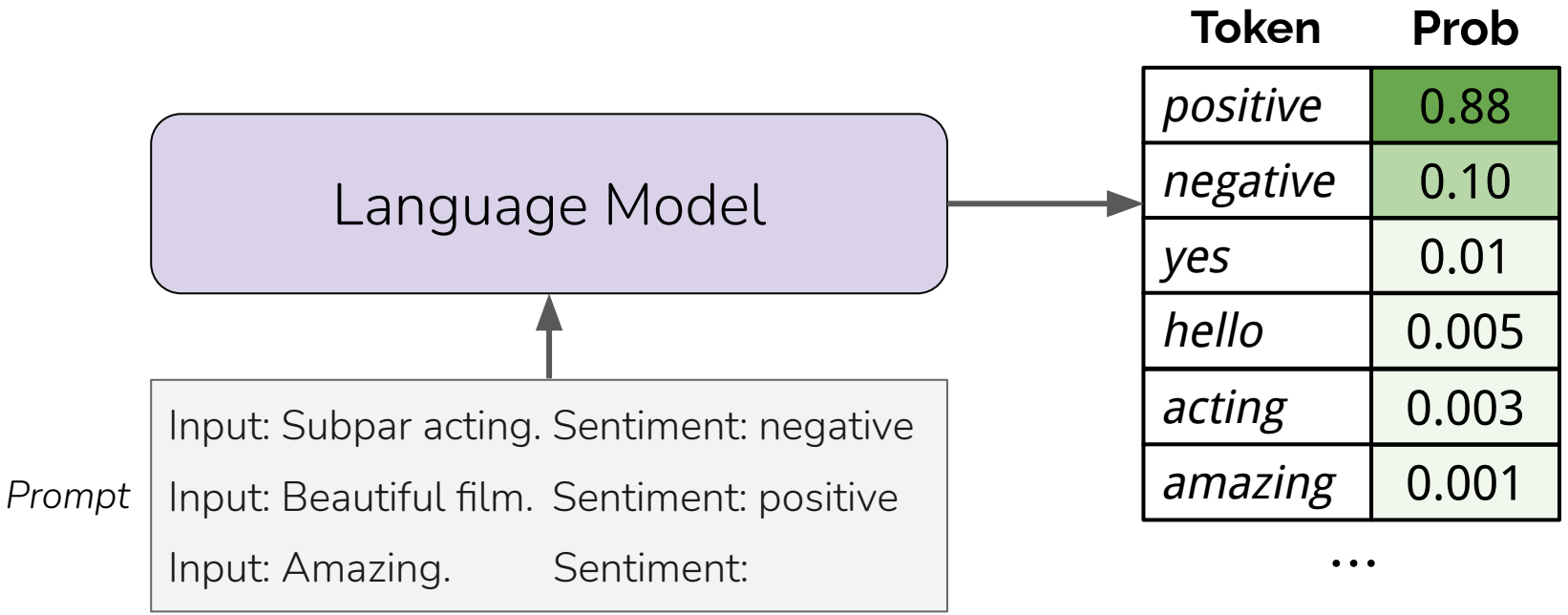
Few-shot Learning with LMs

Sentiment Analysis



Few-shot Learning with LMs

Sentiment Analysis



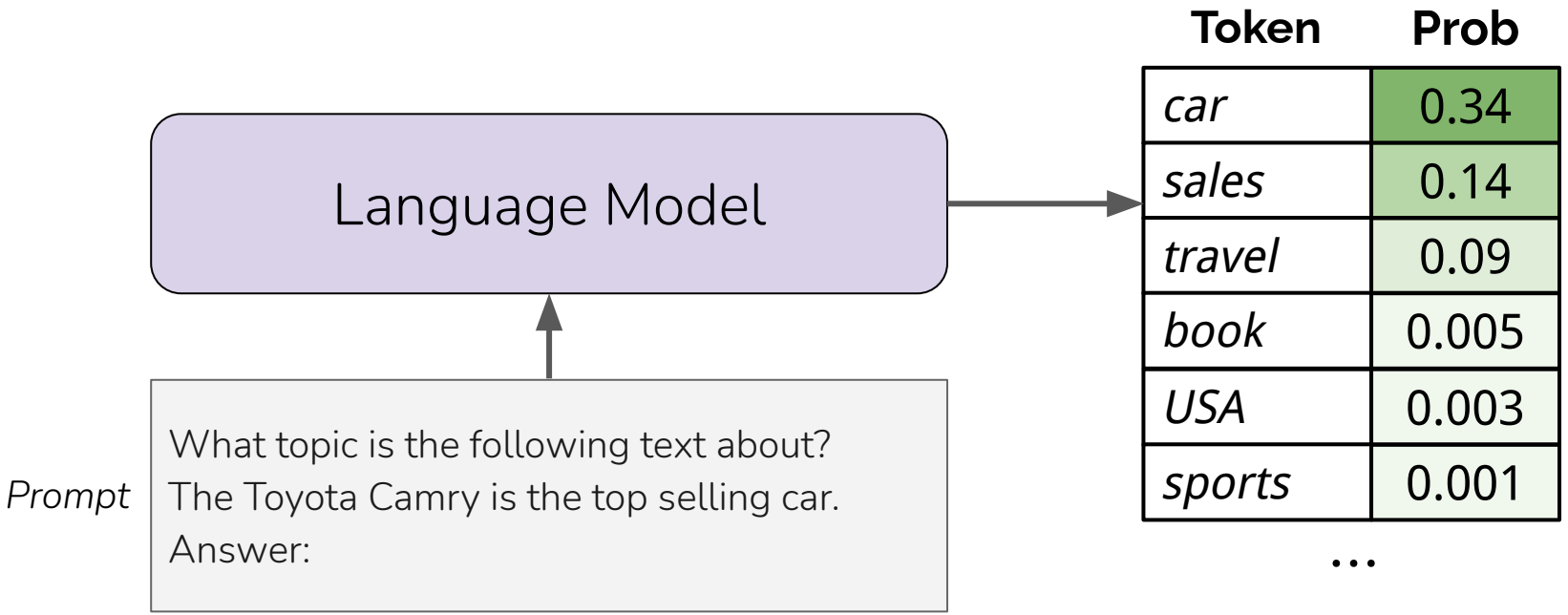
No model updates i.e., *in-context* learning

Few-shot Learning with LMs

Topic Classification

Few-shot Learning with LMs

Topic Classification

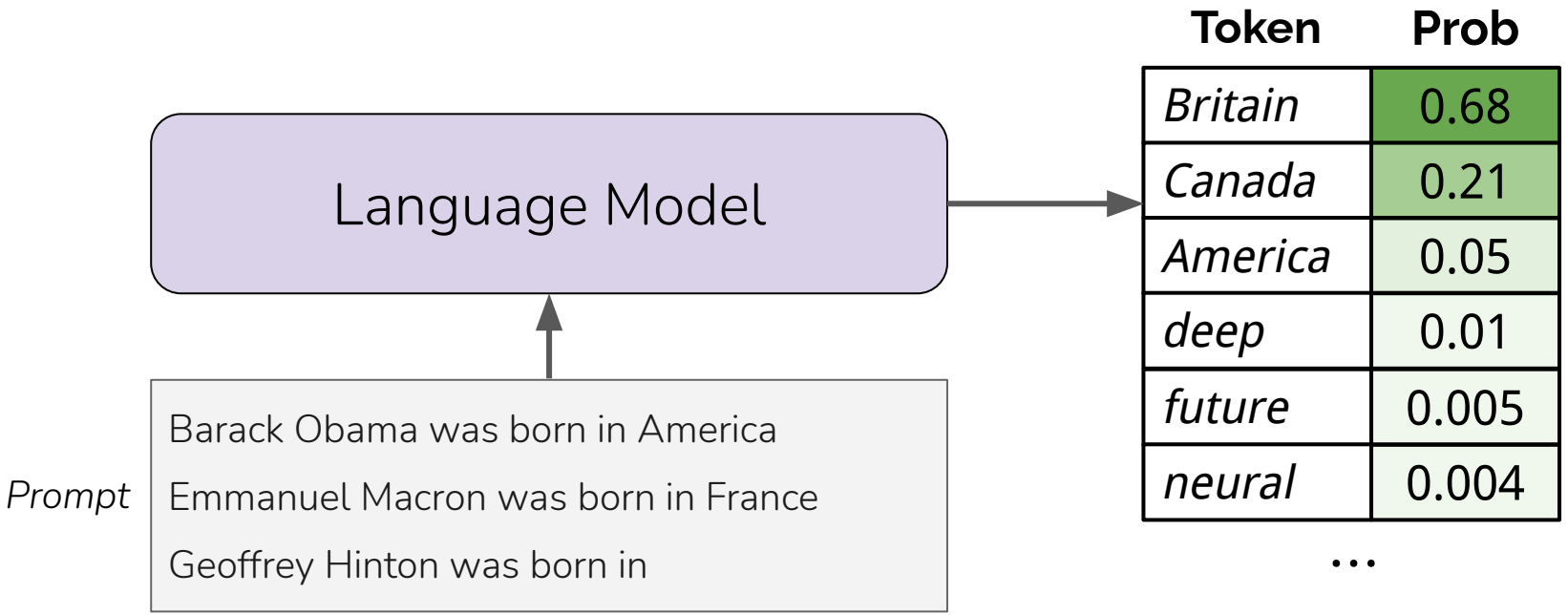


Few-shot Learning with LMs

Knowledge Base Completion

Few-shot Learning with LMs

Knowledge Base Completion



Why In-Context Learning?

- Academically interesting

- Practically relevant with GPT-3:
  - effective with ~0-16 examples
  - serve one model for many tasks
  - no ML expertise needed

Our paper's goal: analyze and improve in-context learning

Part 1:
The Impact of the Prompt

Components Of The Prompt

- Prompt format
- Training example selection
- Training example permutation

Components Of The Prompt

Prompt Format

Prompt

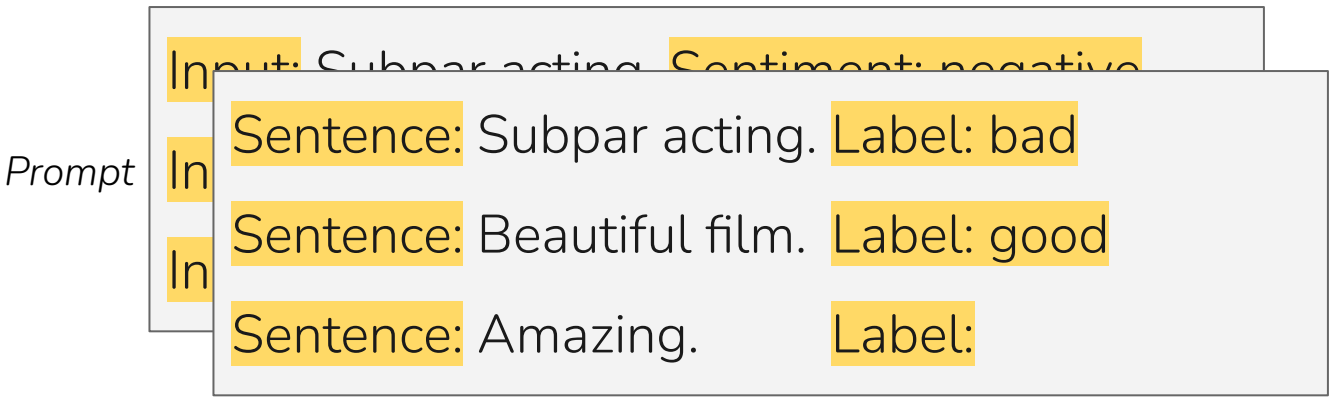
Input: Subpar acting. Sentiment: negative

Input: Beautiful film. Sentiment: positive

Input: Amazing. Sentiment:

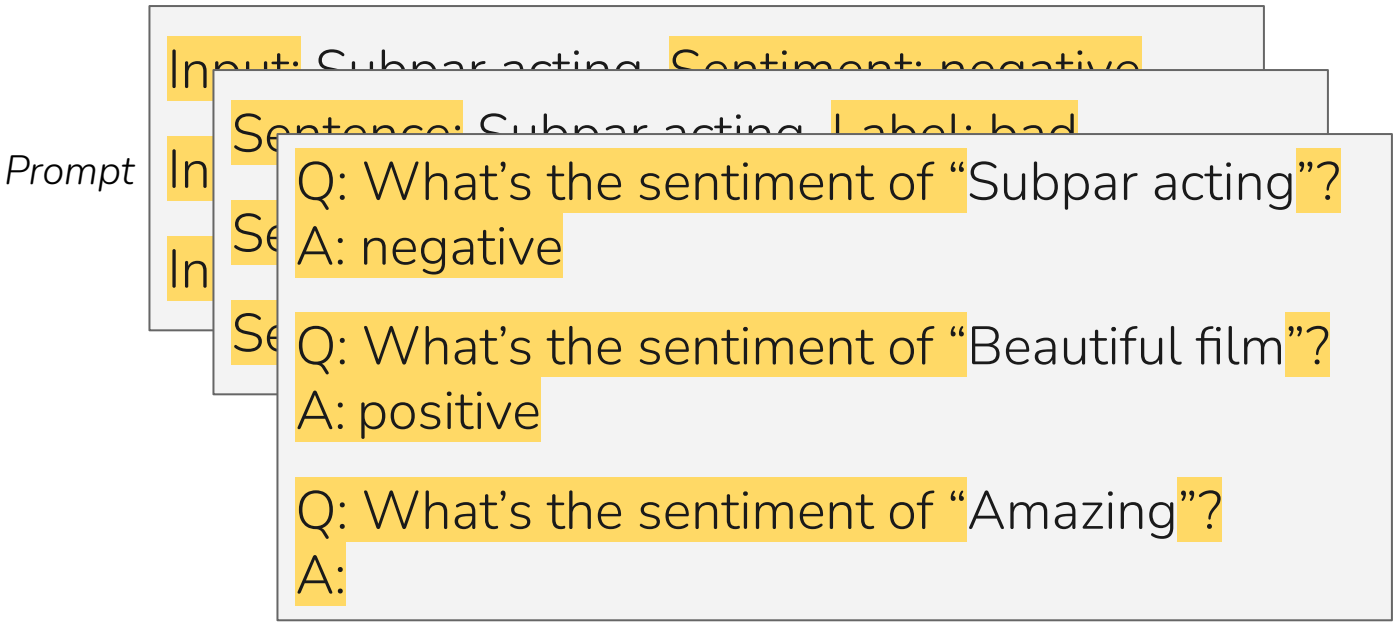
Components Of The Prompt

Prompt Format



Components Of The Prompt

Prompt Format



Components Of The Prompt

Training Example Selection

Prompt

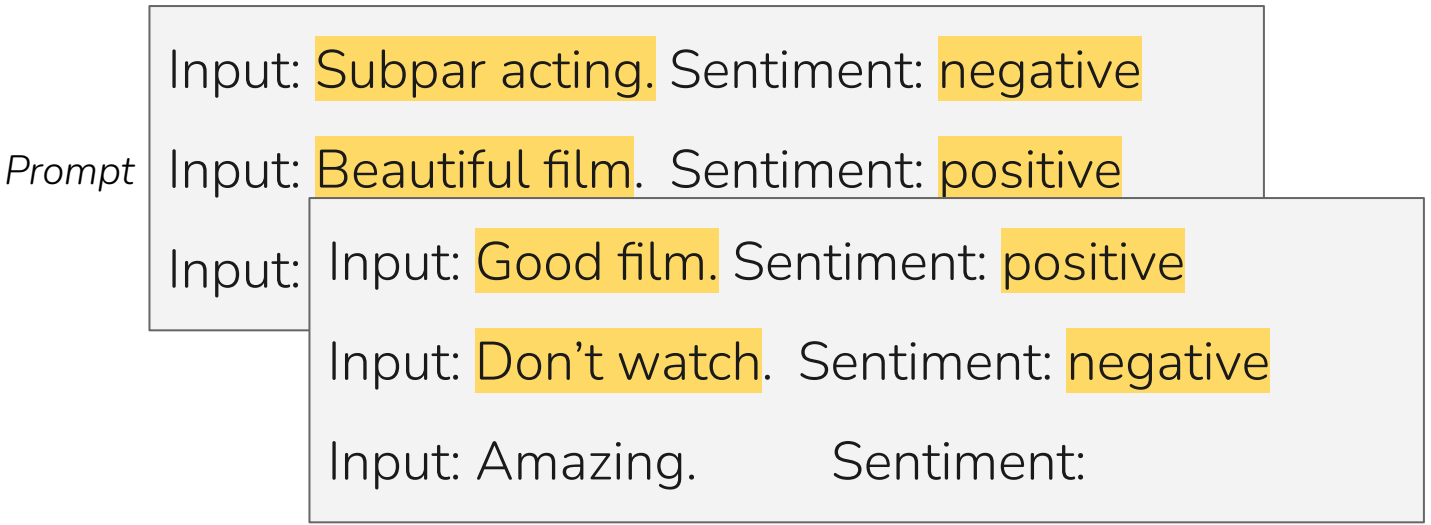
Input: Subpar acting. Sentiment: negative

Input: Beautiful film. Sentiment: positive

Input: Amazing. Sentiment:

Components Of The Prompt

Training Example Selection



Components Of The Prompt

Training Example Permutation

Prompt

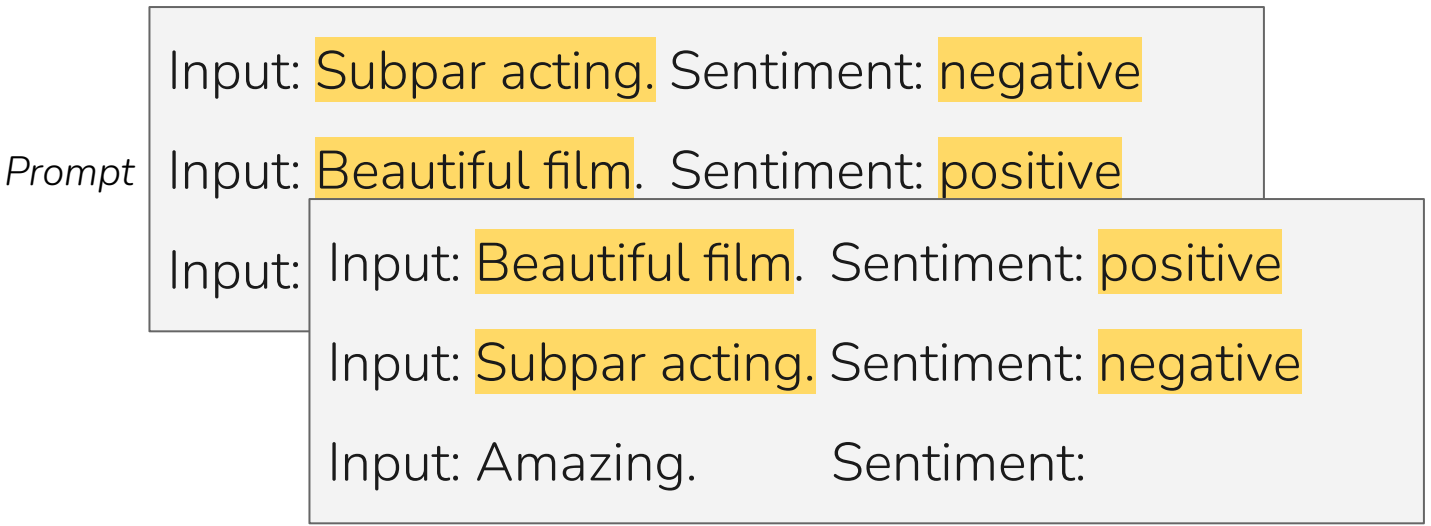
Input: Subpar acting. Sentiment: negative

Input: Beautiful film. Sentiment: positive

Input: Amazing. Sentiment:

Components Of The Prompt

Training Example Permutation



Does The Prompt Affect Accuracy?

- Prompt format
- Training example selection
- Training example permutation

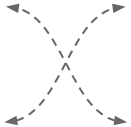
Does The Prompt Affect Accuracy?

- Prompt format
- Training example selection
- Training example permutation

Prompt 1

Review: the whole thing's fairly lame, making it par for the course for disney sequels.
Answer: Negative

Review: this quiet, introspective and entertaining independent is worth seeking.
Answer: Positive



Prompt 2

Review: this quiet, introspective and entertaining independent is worth seeking.
Answer: Positive

Review: the whole thing's fairly lame, making it par for the course for disney sequels .
Answer: Negative

Does The Prompt Affect Accuracy?

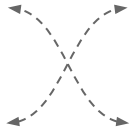
- Prompt format
- Training example selection
- Training example permutation

Prompt 1

Review: the whole thing's fairly lame, making it par for the course for disney sequels.
Answer: Negative

Review: this quiet, introspective and entertaining independent is worth seeking.
Answer: Positive

88.5% Acc.



Prompt 2

Review: this quiet, introspective and entertaining independent is worth seeking.
Answer: Positive

Review: the whole thing's fairly lame, making it par for the course for disney sequels .
Answer: Negative

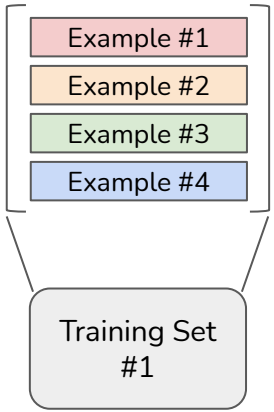
51.3% Acc.

Accuracy Is Highly Sensitive To Prompt Design

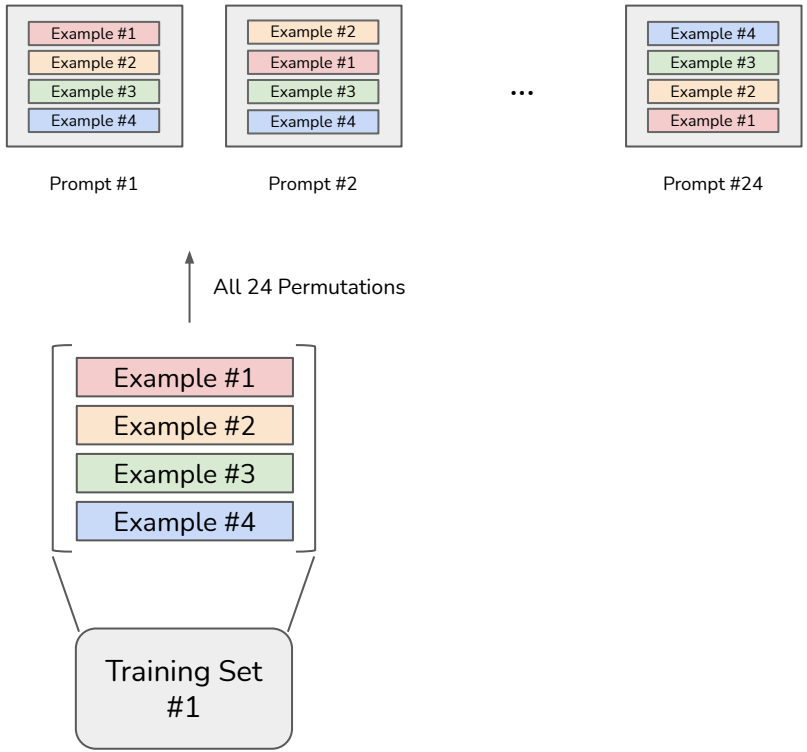
Accuracy Is Highly Sensitive To Prompt Design

Training Set
#1

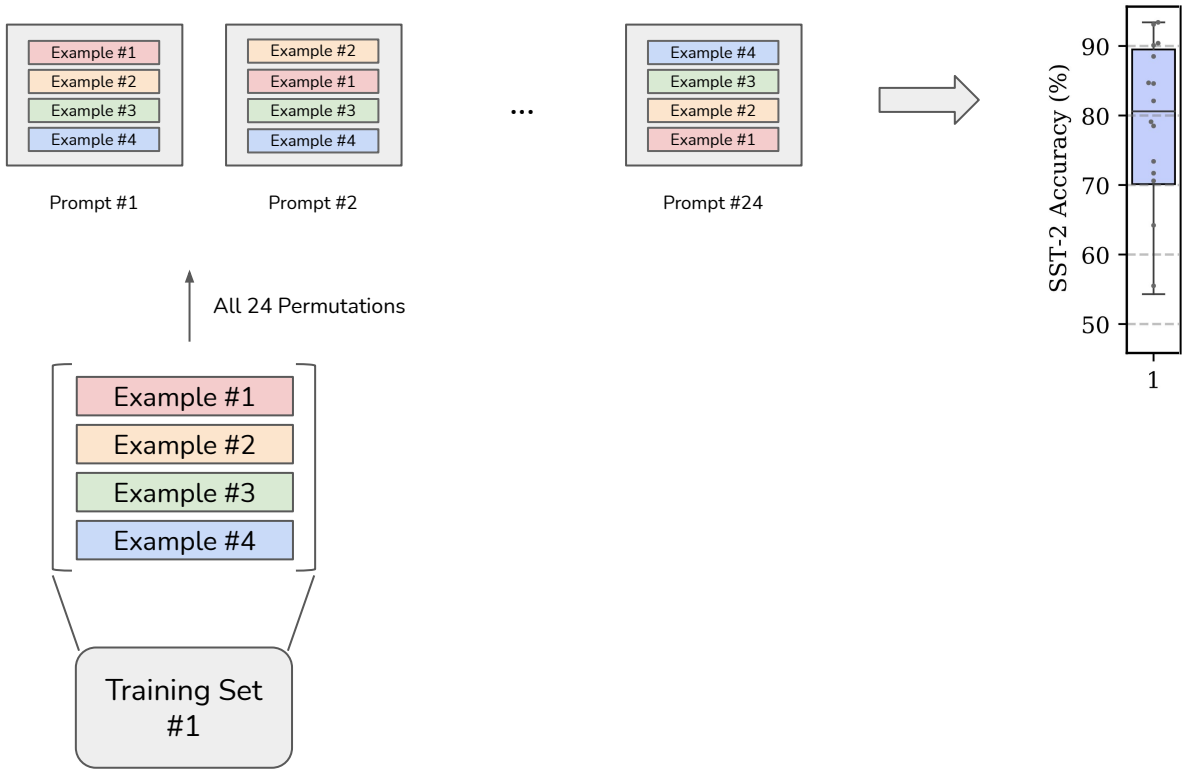
Accuracy Is Highly Sensitive To Prompt Design



Accuracy Is Highly Sensitive To Prompt Design

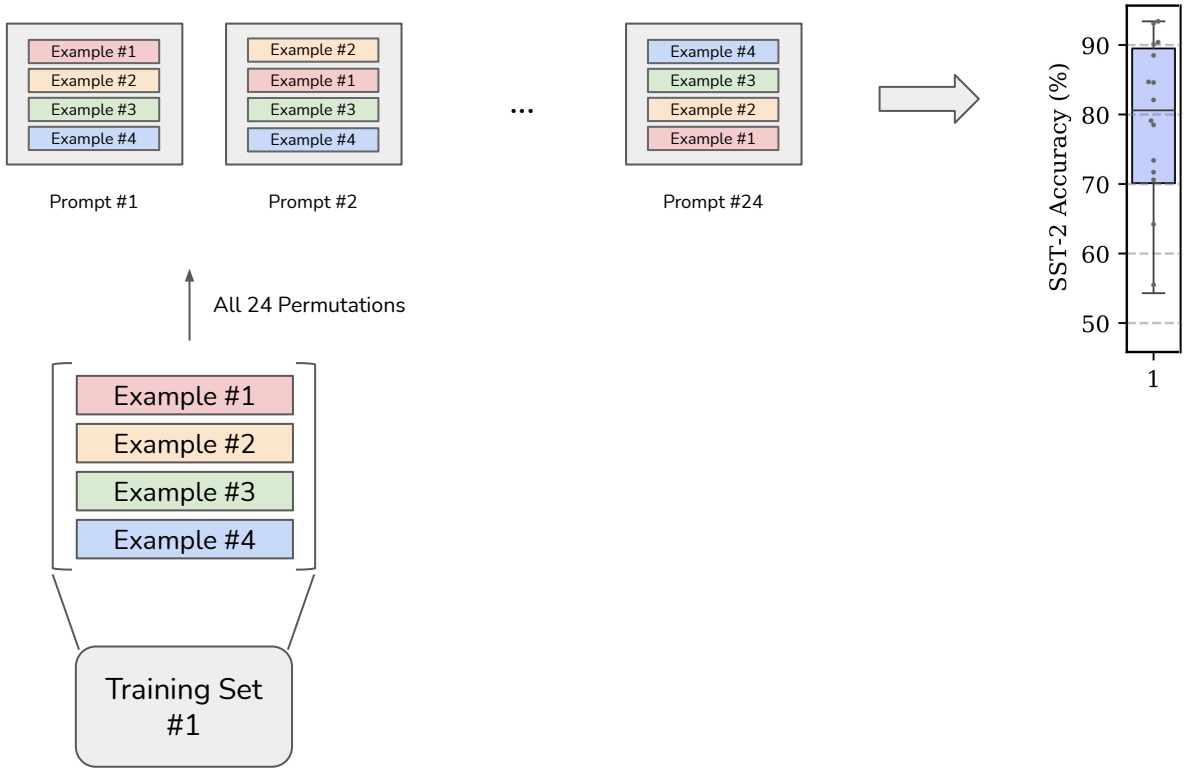


Accuracy Is Highly Sensitive To Prompt Design

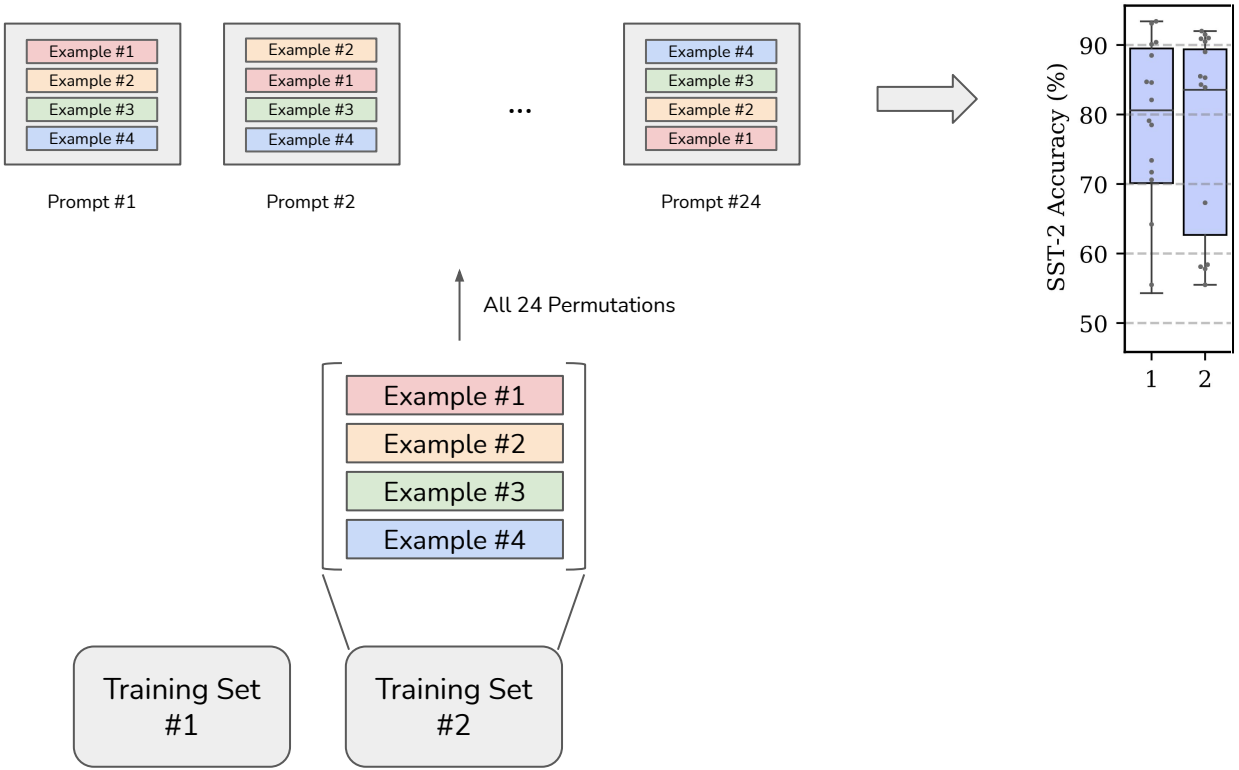


Accuracy Is Highly Sensitive To Prompt Design

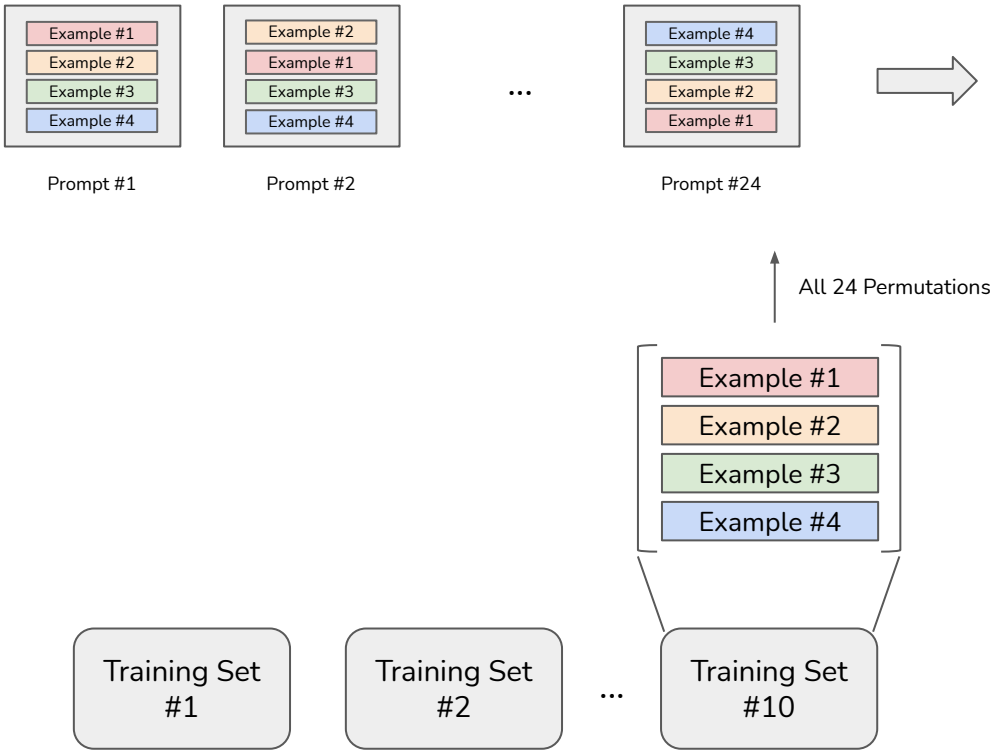
Example Permutation Impacts Accuracy

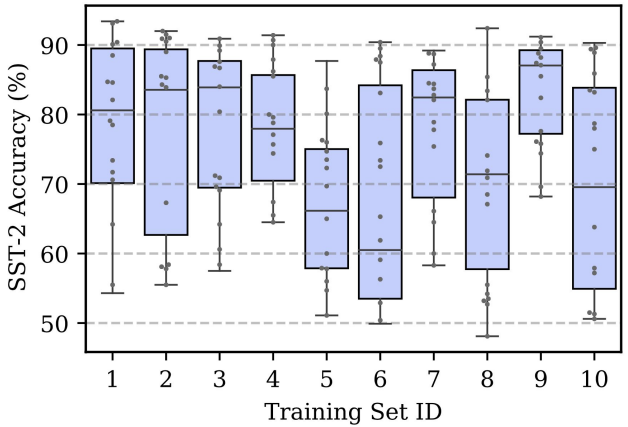


Accuracy Is Highly Sensitive To Prompt Design

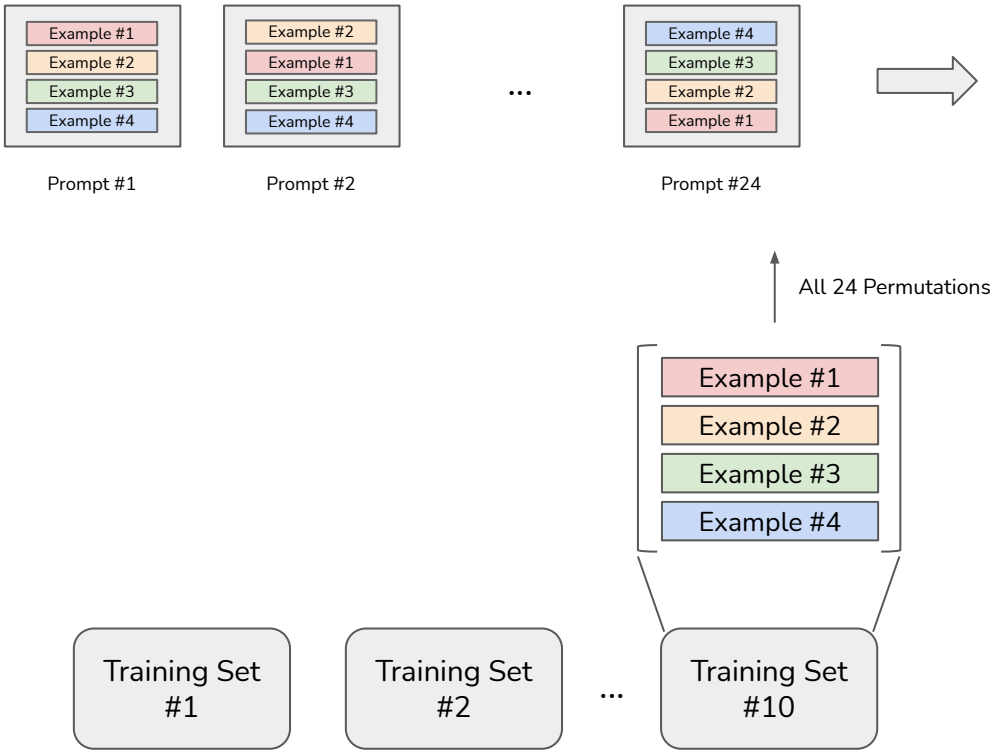


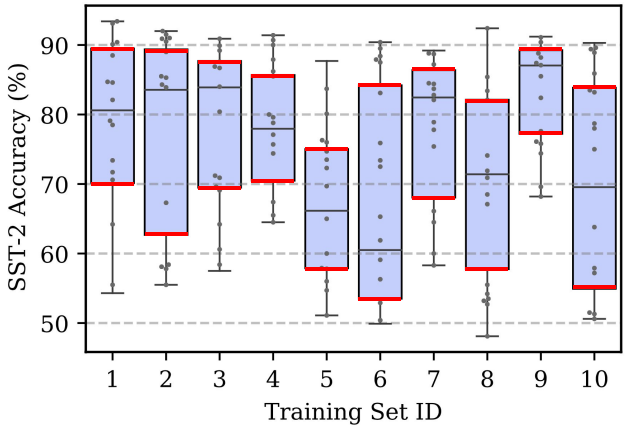
Accuracy Is Highly Sensitive To Prompt Design





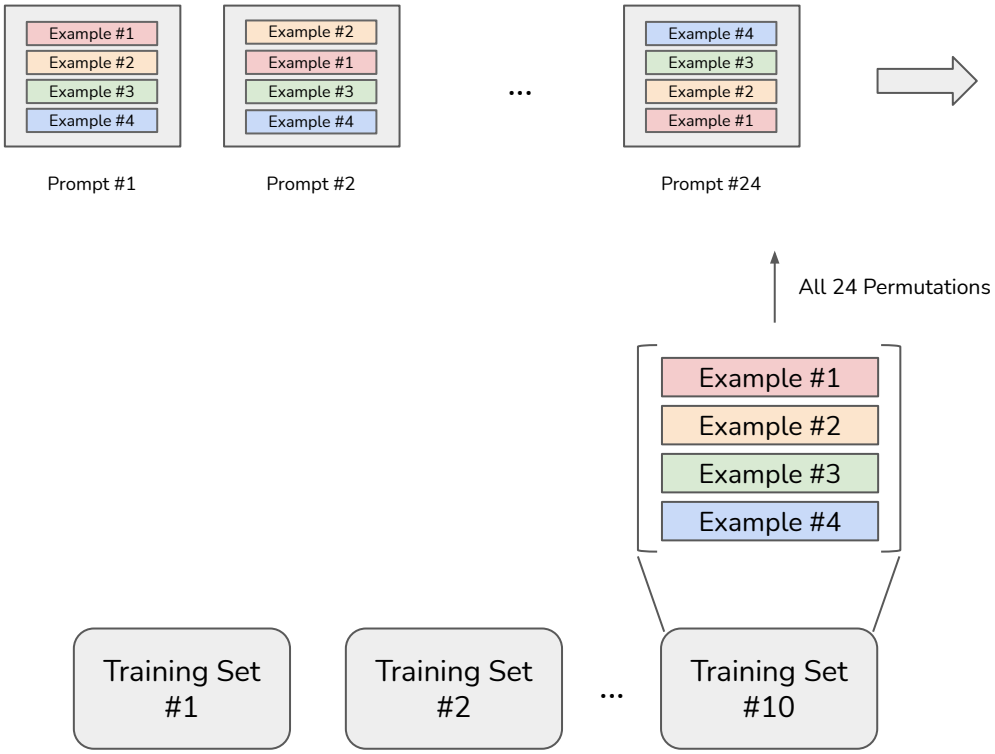
Accuracy Is Highly Sensitive To Prompt Design

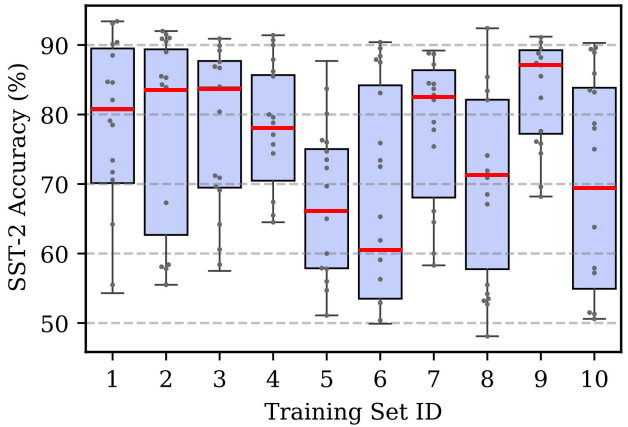




Accuracy Is Highly Sensitive To Prompt Design

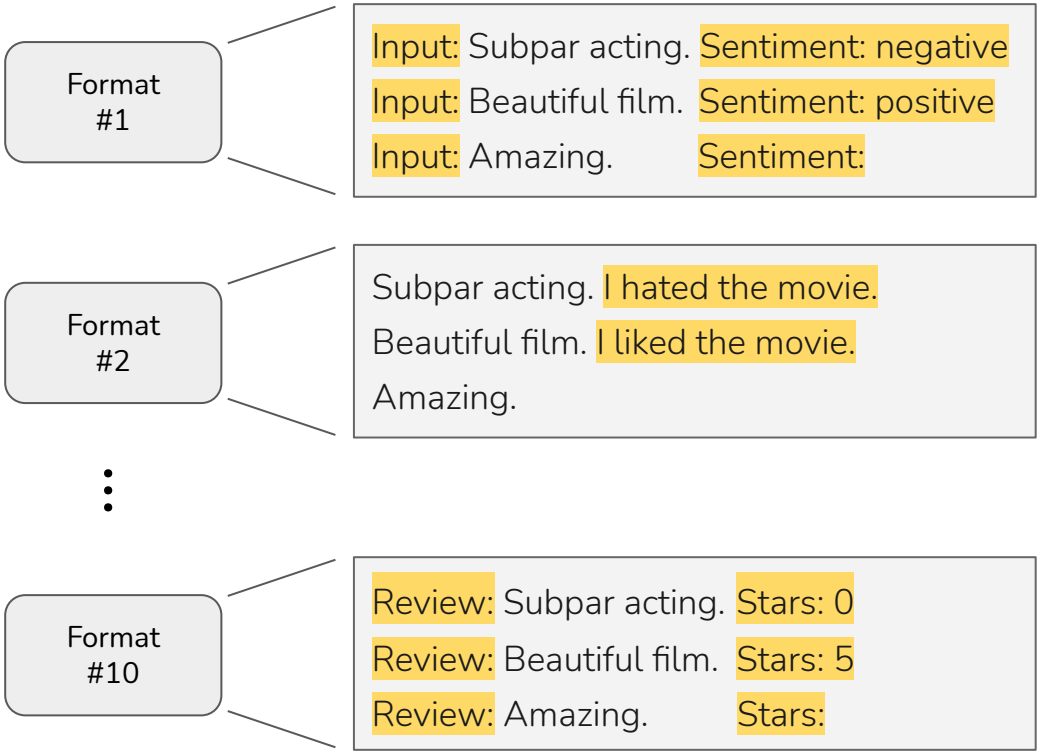
Example Selection Impacts Accuracy





Accuracy Is Highly Sensitive To Prompt Design

Accuracy Is Highly Sensitive To Prompt Design



Accuracy Is Highly Sensitive To Prompt Design

Prompt Format Impacts Accuracy

Format #1

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

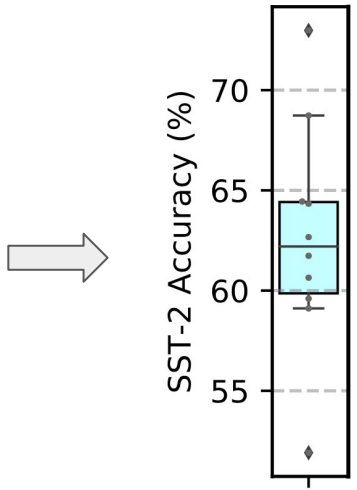
Format #2

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:



Does The Prompt Affect Accuracy?

- Prompt format - Yes
- Training example selection - Yes
- Training example permutation - Yes

Part 2:
Understanding Prompt Sensitivity

What Causes The High Variance?

We identify three biases in LMs

What Causes The High Variance?

We identify three biases in LMs

- **Majority label bias:** frequent training answers dominate predictions
  - helps explain variance across *example selections*

What Causes The High Variance?

We identify three biases in LMs

- **Majority label bias:** frequent training answers dominate predictions
  - helps explain variance across *example selections*

- **Recency bias:** examples near end of prompt dominate predictions
  - helps explain variance across *example permutations*

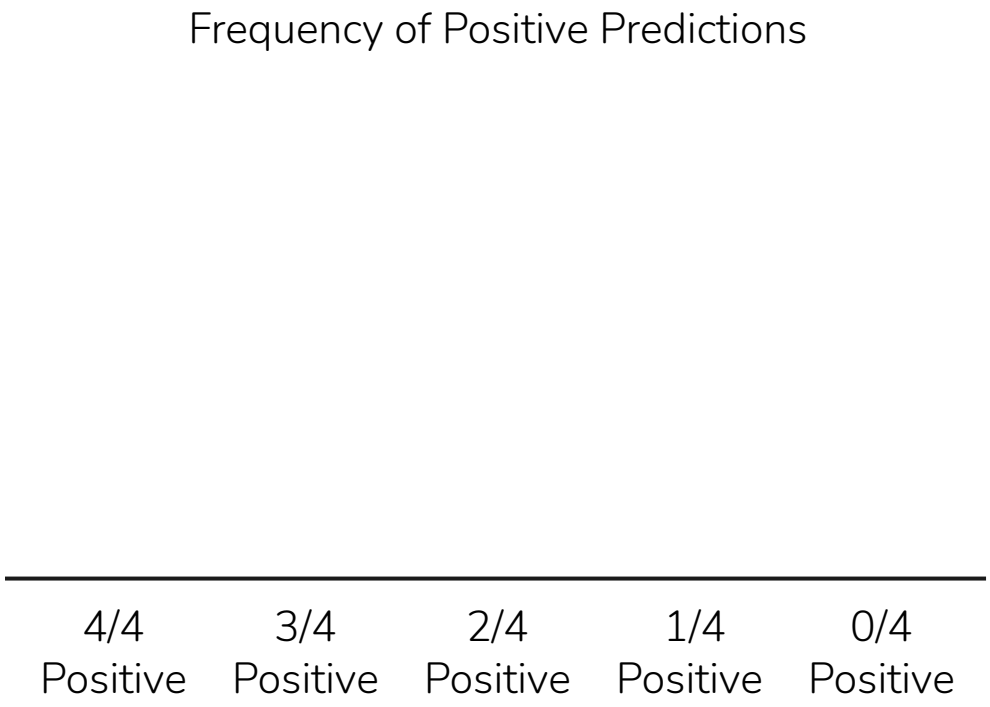
What Causes The High Variance?

We identify three biases in LMs

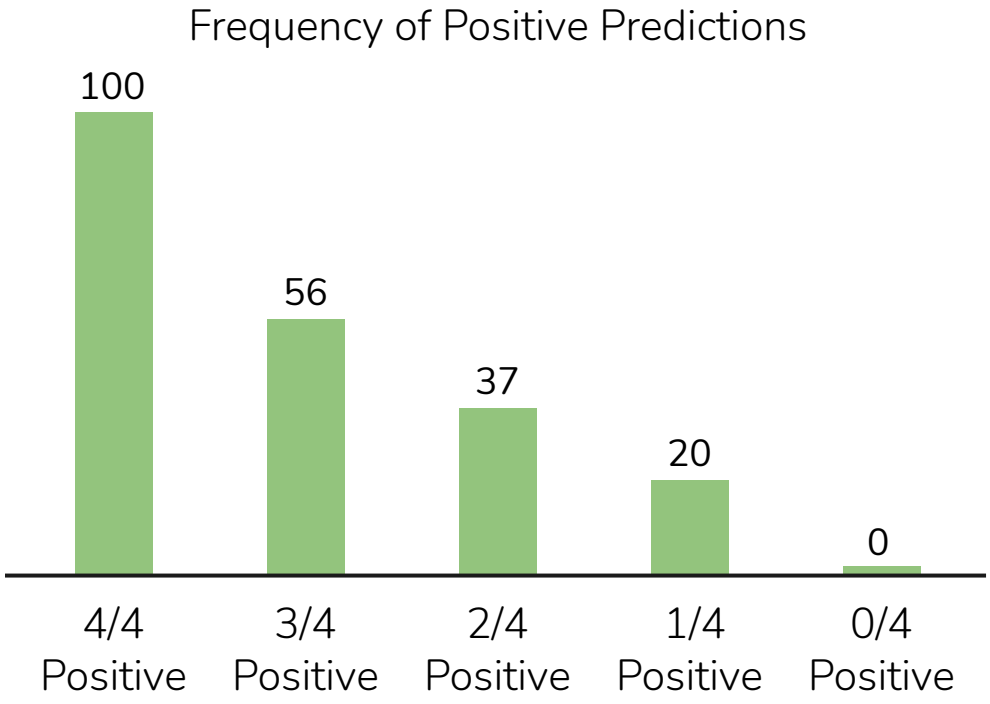
- **Majority label bias:** frequent training answers dominate predictions
  - helps explain variance across *example selections*
- **Recency bias:** examples near end of prompt dominate predictions
  - helps explain variance across *example permutations*
- **Common token bias:** common n-grams dominate predictions
  - helps explain variance across *prompt formats*

Majority Label Bias

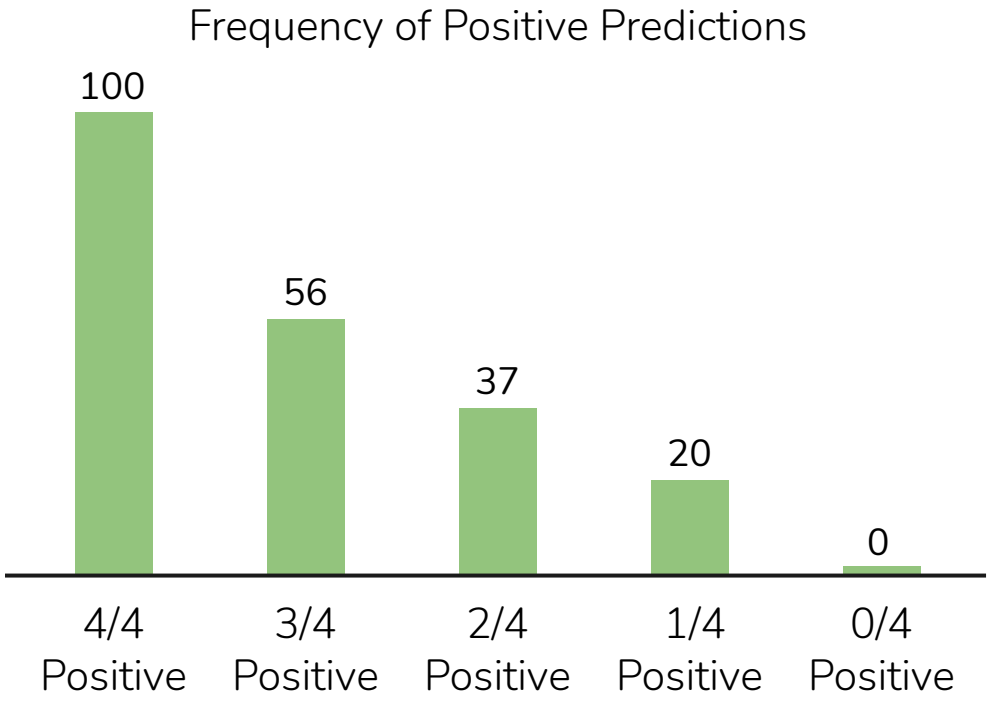
Majority Label Bias



Majority Label Bias



Majority Label Bias



Frequent training answers dominate predictions

Recency Bias

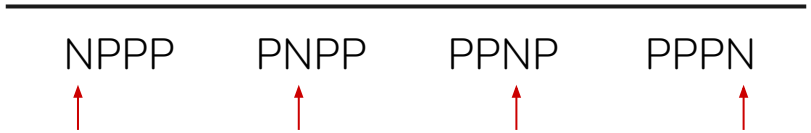
Recency Bias

Frequency of Positive Predictions

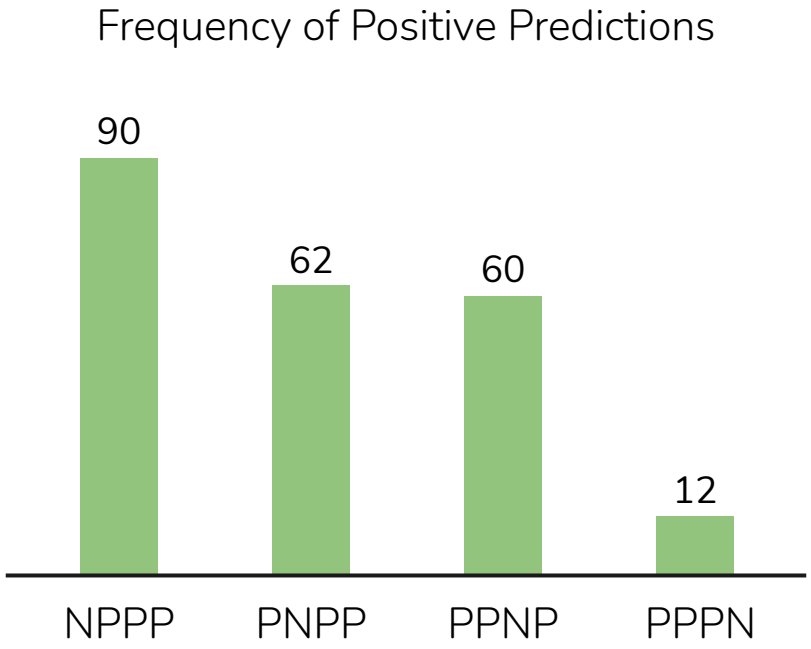
NPPP PNPP PPNP PPPN

Recency Bias

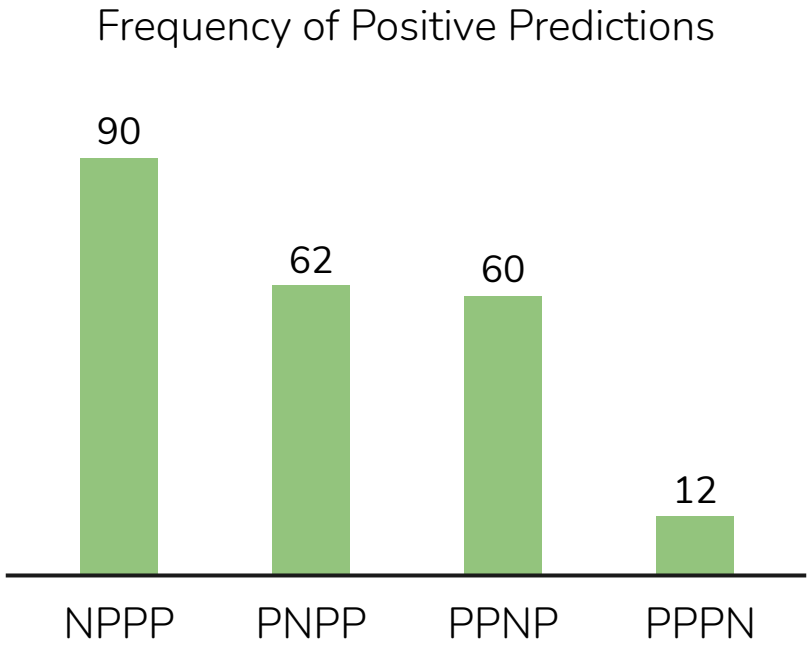
Frequency of Positive Predictions



Recency Bias



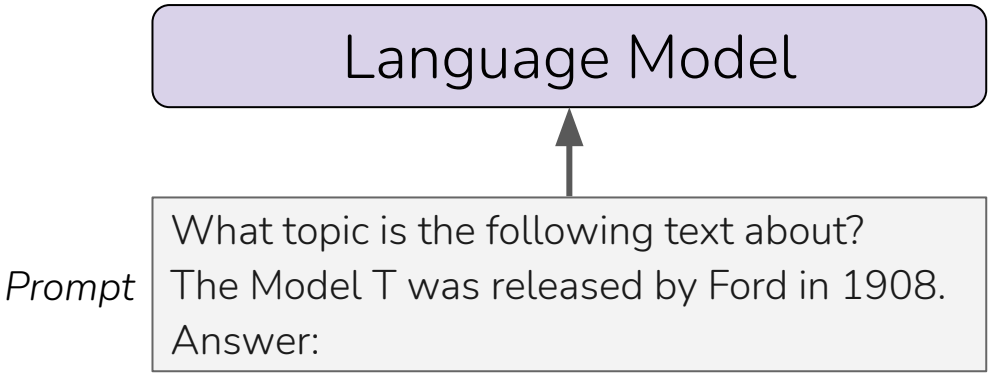
Recency Bias



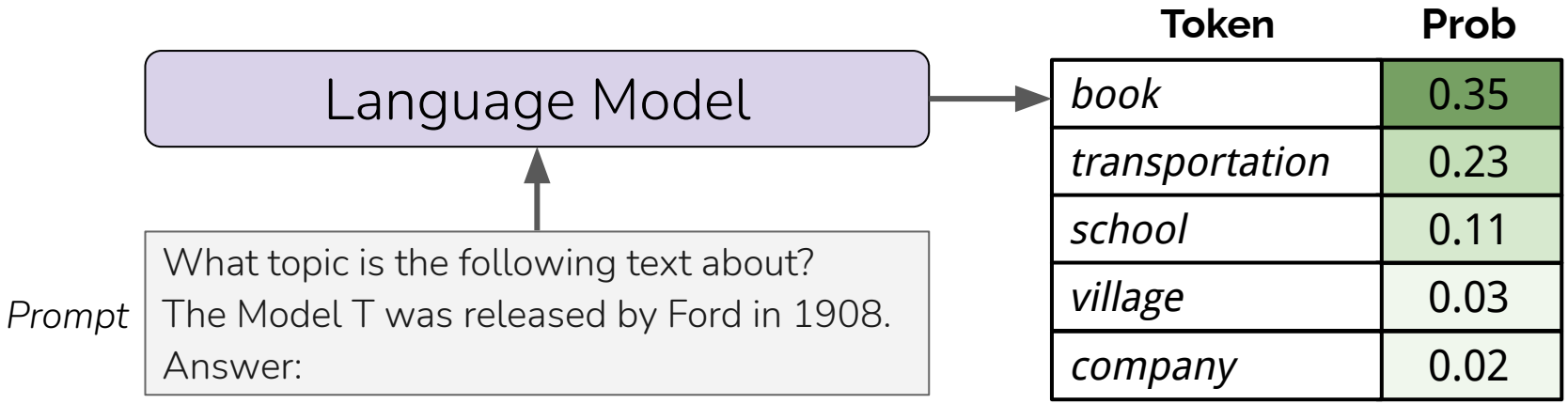
Examples near end of prompt dominate predictions

Common Token Bias

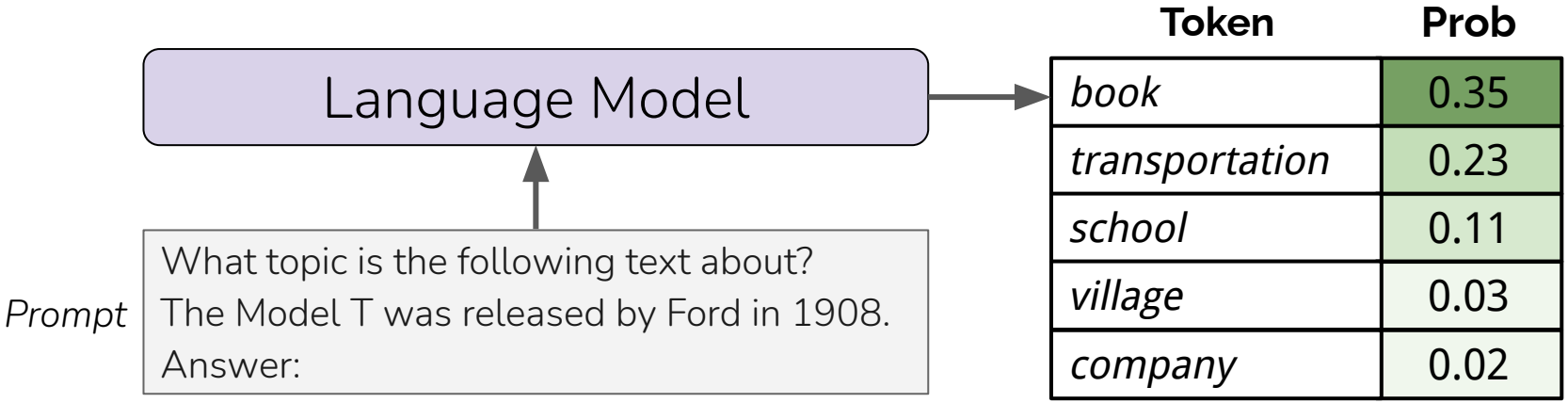
Common Token Bias



Common Token Bias

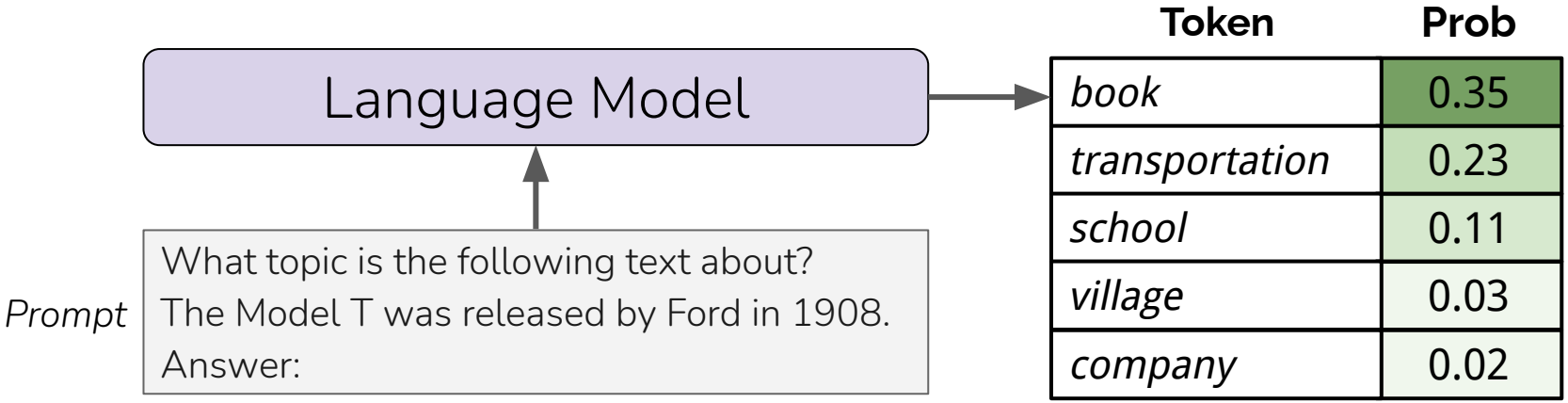


Common Token Bias



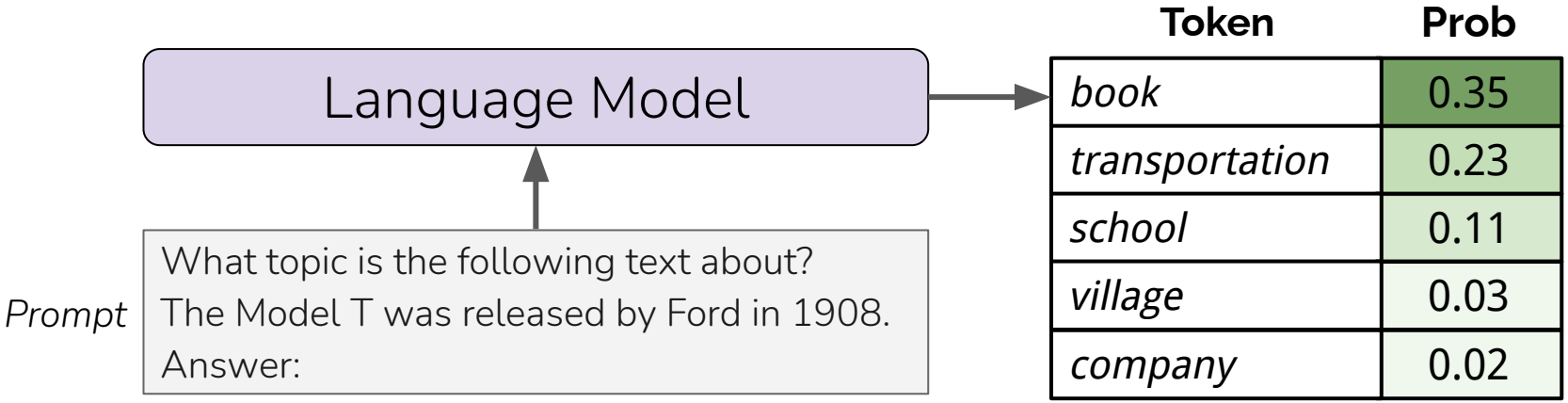
Token	Web (%)
book	0.026
transportation	0.0000006

Common Token Bias



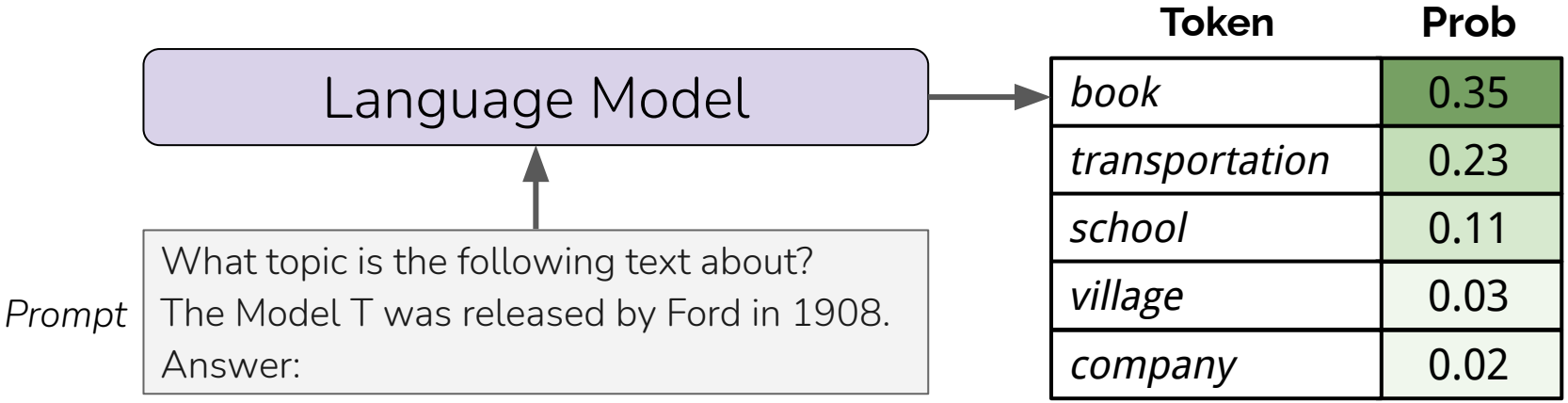
Token	Web (%)	Label (%)
book	0.026	9
transportation	0.0000006	9

Common Token Bias



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

Common Token Bias

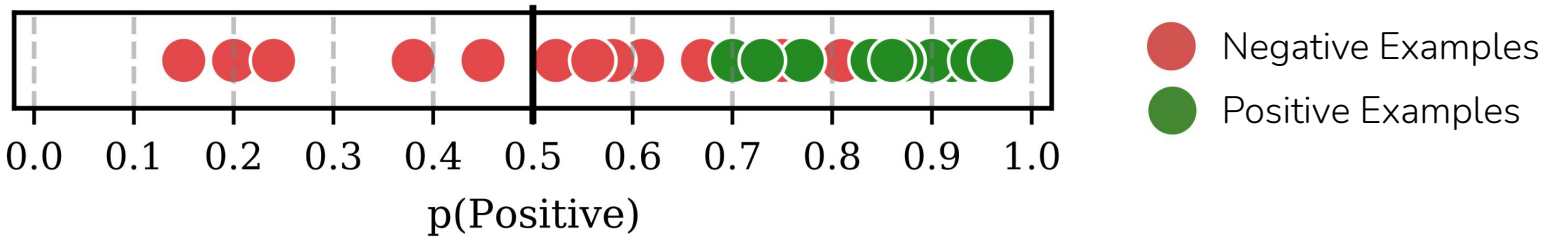


Token	Web (%)	Label (%)	Prediction (%)
book	0.026	9	29
transportation	0.0000006	9	4

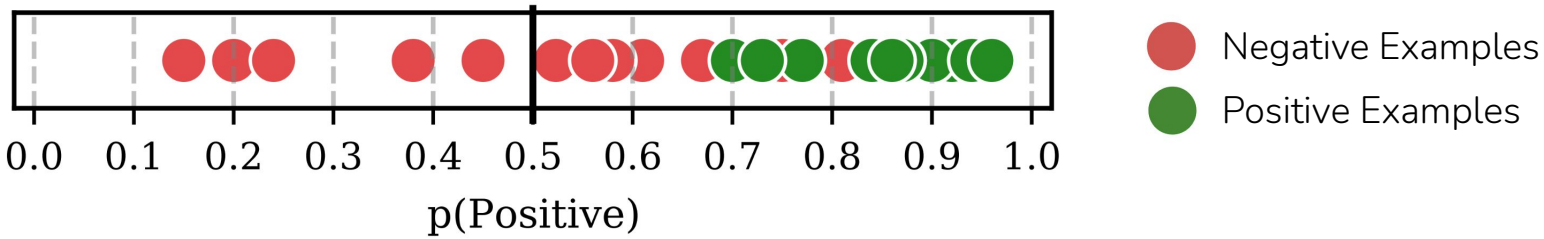
Common n-grams dominate predictions

What Is The Impact of These Biases?

What Is The Impact of These Biases?



What Is The Impact of These Biases?



Biases cause a shift in output distribution

Part 3:
Improving Few-Shot Accuracy

Contextual Calibration

Contextual Calibration

Step 1: Estimate the bias

Contextual Calibration

Step 1: Estimate the bias

Insert "*content-free*" test input

Input: Subpar acting. Sentiment: negative
Input: Beautiful film. Sentiment: positive
Input: **N/A** Sentiment:

Contextual Calibration

Step 1: Estimate the bias

Insert "*content-free*" test input

Input: Subpar acting. Sentiment: negative
Input: Beautiful film. Sentiment: positive
Input: **N/A** Sentiment:

Get model's prediction

positive	0.65
negative	0.35

Contextual Calibration

Step 1: Estimate the bias

Insert "*content-free*" test input

Input: Subpar acting. Sentiment: negative
Input: Beautiful film. Sentiment: positive
Input: **N/A** Sentiment:

Get model's prediction

positive	0.65
negative	0.35

Step 2: Counter the bias

Contextual Calibration

Step 1: Estimate the bias

Insert "*content-free*" test input

Input: Subpar acting. Sentiment: negative
Input: Beautiful film. Sentiment: positive
Input: **N/A** Sentiment:

Get model's prediction

positive	0.65
negative	0.35

Step 2: Counter the bias

"Calibrate" predictions with affine transformation

$$\hat{\mathbf{q}} = \text{softmax}(\mathbf{W}\hat{\mathbf{p}} + \mathbf{b})$$

Contextual Calibration

Step 1: Estimate the bias

Insert "*content-free*" test input

Input: Subpar acting. Sentiment: negative
Input: Beautiful film. Sentiment: positive
Input: **N/A** Sentiment:

Get model's prediction

positive	0.65
negative	0.35

Step 2: Counter the bias

"Calibrate" predictions with affine transformation

$$\hat{\mathbf{q}} = \text{softmax}(\mathbf{W}\hat{\mathbf{p}} + \mathbf{b})$$

Calibrated probs ↑ Original probs ↑

Contextual Calibration

Step 1: Estimate the bias

Insert "*content-free*" test input

Input: Subpar acting. Sentiment: negative
Input: Beautiful film. Sentiment: positive
Input: **N/A** Sentiment:

Get model's prediction

positive	0.65
negative	0.35

Step 2: Counter the bias

"Calibrate" predictions with affine transformation

$$\hat{\mathbf{q}} = \text{softmax}(\mathbf{W}\hat{\mathbf{p}} + \mathbf{b})$$

Calibrated probs — Original probs

Fit $\mathbf{W}$ and $\mathbf{b}$ to cause uniform prediction for "N/A"

Contextual Calibration

Step 1: Estimate the bias

Insert "*content-free*" test input into prompt

```
Input: Subpar acting. Sentiment: negative
Input: Beautiful film. Sentiment: positive
Input: N/A Sentiment:
```

Get model's prediction

positive	0.65
negative	0.35

Step 2: Counter the bias

"Calibrate" predictions with affine transformation

$$\hat{\mathbf{q}} = \text{softmax}(\mathbf{W}\hat{\mathbf{p}} + \mathbf{b})$$

Calibrated probs — Original probs

Fit $\mathbf{W}$ and $\mathbf{b}$ to cause uniform prediction for "N/A"

$$\mathbf{W} = \begin{pmatrix} \frac{1}{0.65} & 0 \\ 0 & \frac{1}{0.35} \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Effect of Contextual Calibration

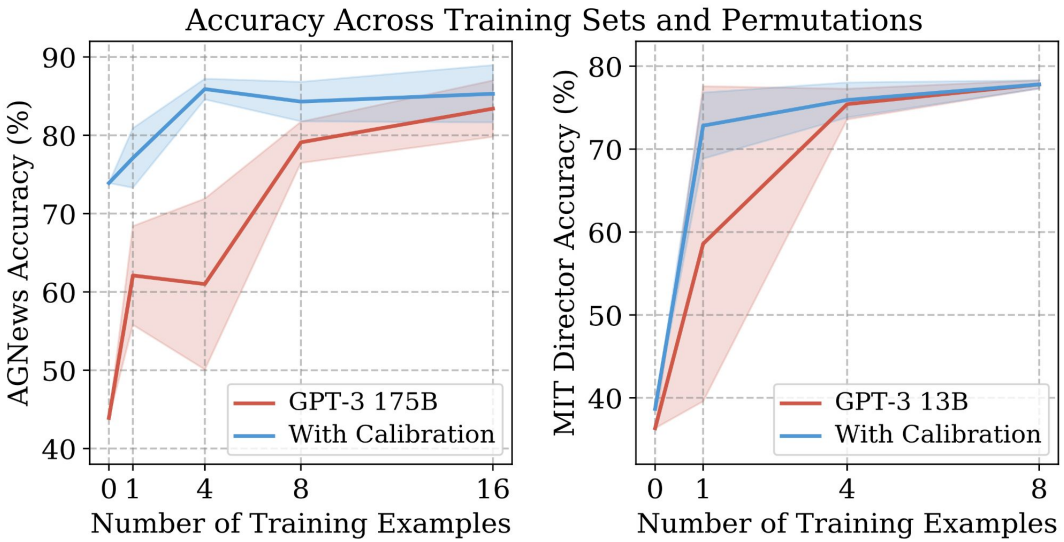
- Experiment with 11 different datasets
  - Classification
  - Knowledge base completion
  - Information extraction

- Consider 0, 1, 4, 8, and 16 training examples

- Different sizes of GPT-3 and GPT-2 language models

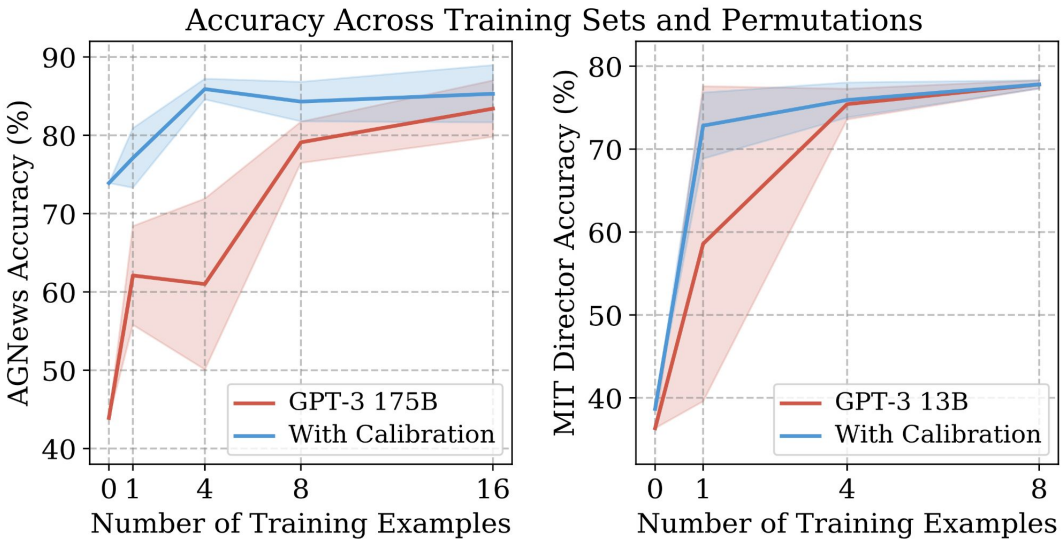
Effect of Contextual Calibration

- Improves mean and worst-case accuracy



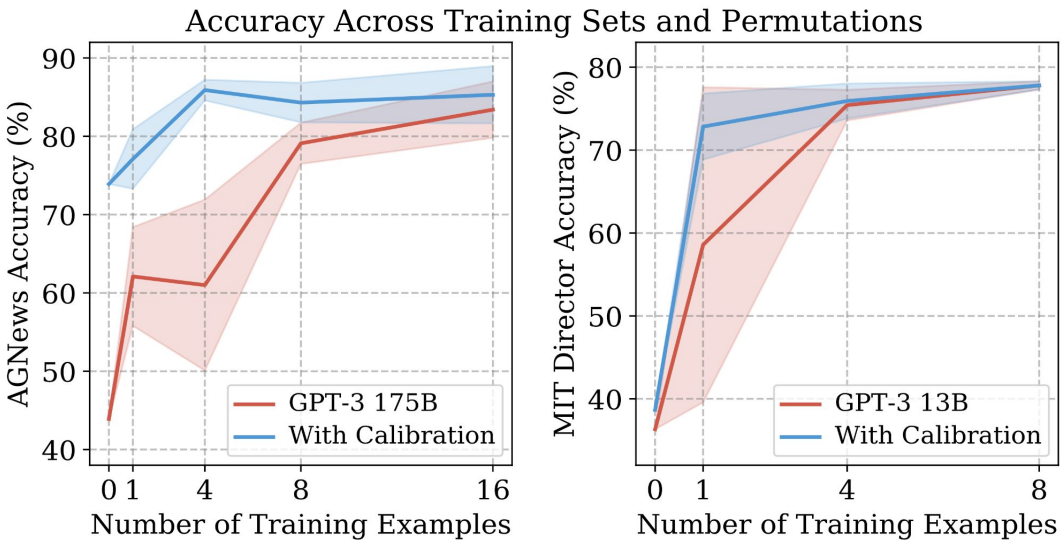
Effect of Contextual Calibration

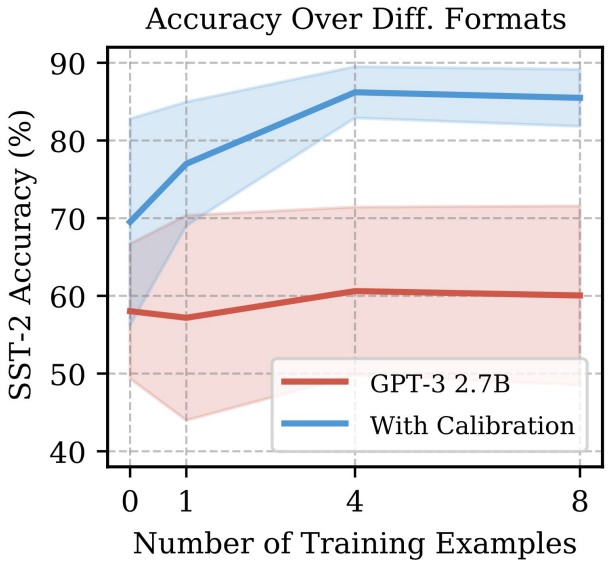
- Improves mean and worst-case accuracy
- Reduces variance across training sets and permutations



Effect of Contextual Calibration

- Improves mean and worst-case accuracy
- Reduces variance across training sets and permutations
- Reduces variance across prompt formats





Summary

1. GPT-3's accuracy has high variance across different prompts
2. LMs have biases that hurt few-shot learning
3. Contextual calibration improves accuracy and reduces variance

Paper and Code at ericswallace.com/calibrate