Large Language Models

Calibration of prompting LLMs

Sharif University of Technology

Soleymani

Fall 2023

Sensitivity of LLMs predictions

- LLMs are highly sensitive and even biased to:
  - the choice of templates
  - verbalizers or label spaces (such as yes/no, true/false, correct/incorrect)
  - demonstration examples and their permutations

- Calibration methods mitigate the effects of these biases while recovering LLM performance.

Prompt engineering difficulties

- Prompt engineering is an informal and difficult process.
  - Small changes to a prompt can cause massive changes to the model's output
    - highly sensitive and even biased to the choice of templates, verbalizers, and demonstrations
  - a harsh reality in creating applications with LLMs.

- Finding techniques that make LLMs more accurate and reliable

In-Context Learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  English translate to French:     ← task description
2  cheese =>  ......................  ← prompt
```

One-shot

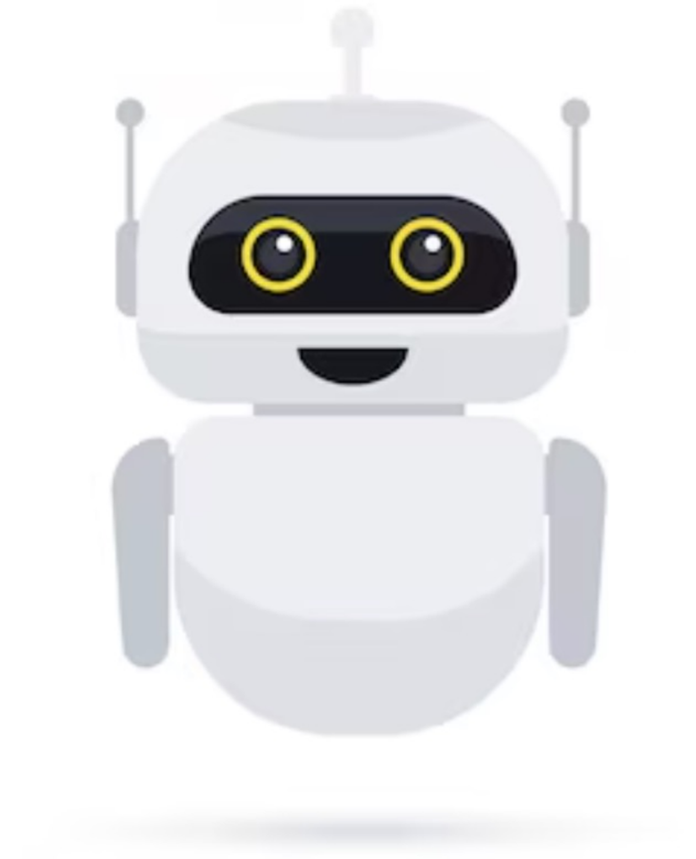
In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1  English translate to French:     ← task description
2  sea otter => loutre de mer       ← example
3  cheese =>  ......................  ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  English translate to French:     ← task description
2  sea otter => loutre de mer       ←
3  peppermint => menthe poivrée     ←  example
4  plush girafe => girafe peluche   ←
5  cheese =>  ......................  ← prompt
```

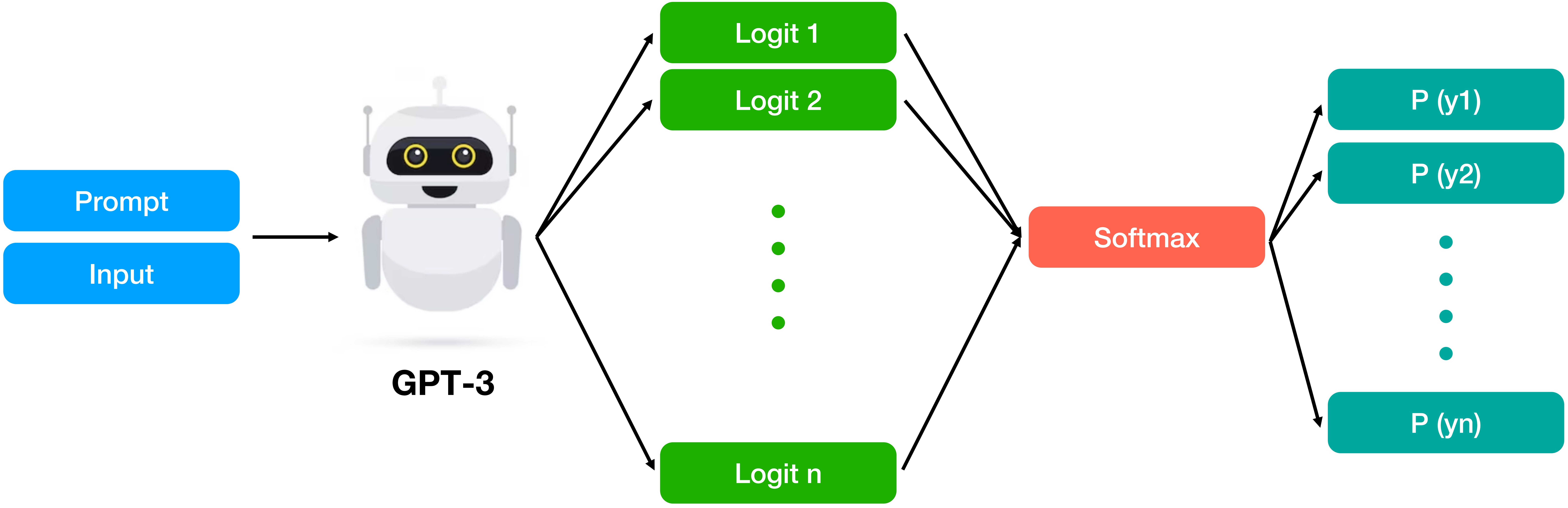


GPT-3

fromage

Brown et al., 2020

Language Modeling



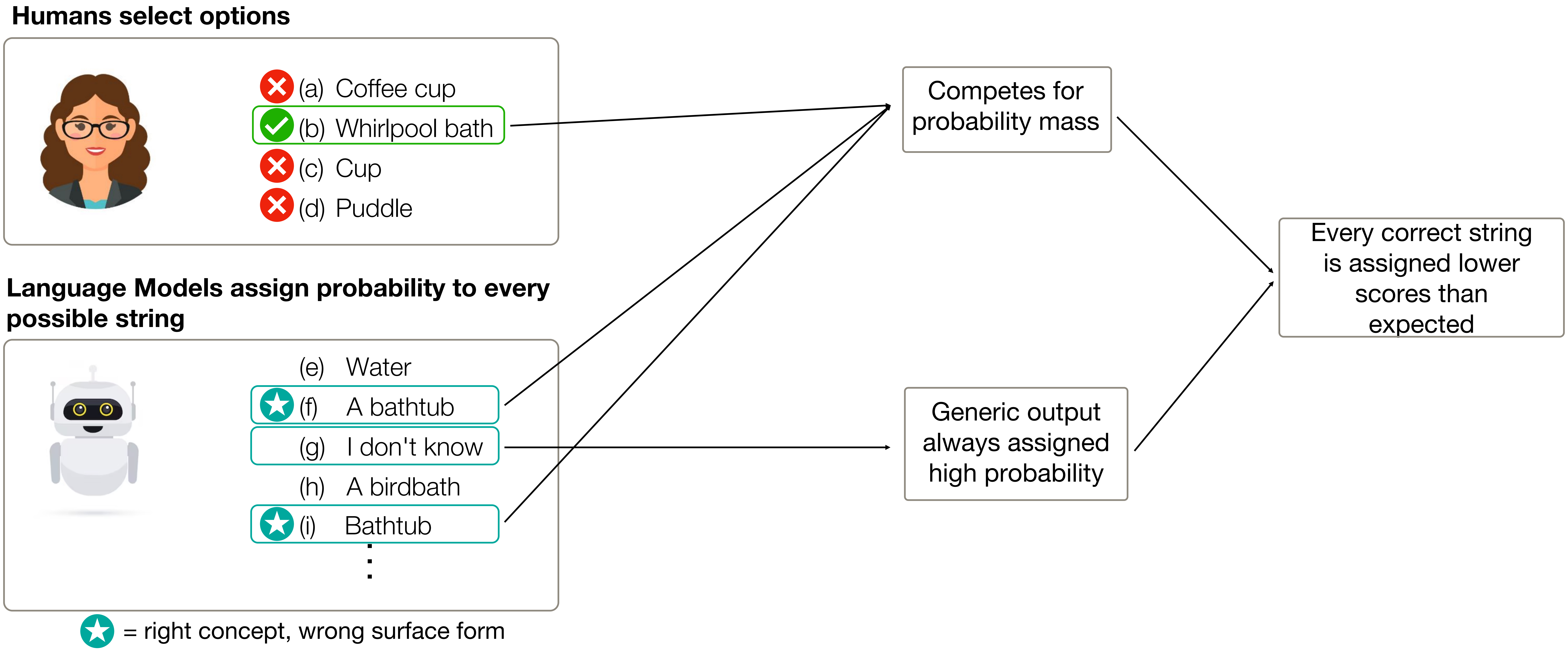
Question
What are some possible flaws?

n = number of labels for close set classification tasks

n = number of words in the vocabulary for open set tasks

Surface Form Competition

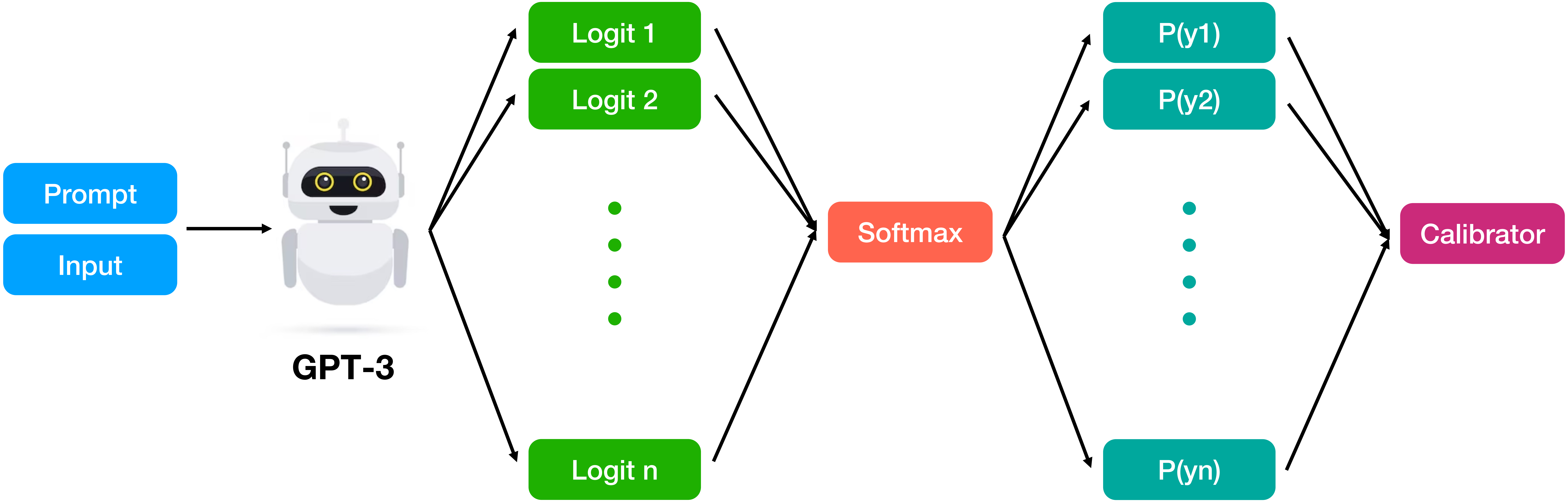
A human wants to submerge himself in water, what should he use?



Holtzman et al., Surface Form Competition: Why the Highest Probability Answer Isn't Always Right, EMNLP 2021

Calibration

calibration problem can be framed as an unsupervised decision (or few-shot) boundary learning problem



Question

How to calibrate?

n = number of labels for close set classification tasks

n = number of words in the vocabulary for open set tasks

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

Tony Z. Zhao [*1] **Eric Wallace** [*1] **Shi Feng** [2] **Dan Klein** [1] **Sameer Singh** [3]

ICML 2021

Some slides adapted from http://ericswallace.com/calibrate

How important is the structure of the prompt for in-context learning?

Components of a prompt:

1. **Prompt format**
2. Training example selection
3. Training example permutation

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

Q: What's the sentiment of "Subpar acting"?
A: negative
Q: What's the sentiment of "Beautiful film"?
A: positive
Q: What's the sentiment of "Amazing"?
A:

Zhao et al., Calibrate before Use: Improving Few-Shot Performance of Language Models, ICML 2021

How important is the structure of the prompt for in-context learning?

Components of a prompt:

1 Prompt format

2 Training example selection

3 Training example permutation

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:**

Zhao et al., Calibrate before Use: Improving Few-Shot Performance of Language Models, ICML 2021

How important is the structure of the prompt for in-context learning?

Components of a prompt:

1. Prompt format
2. Training example selection
3. **Training example permutation**

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

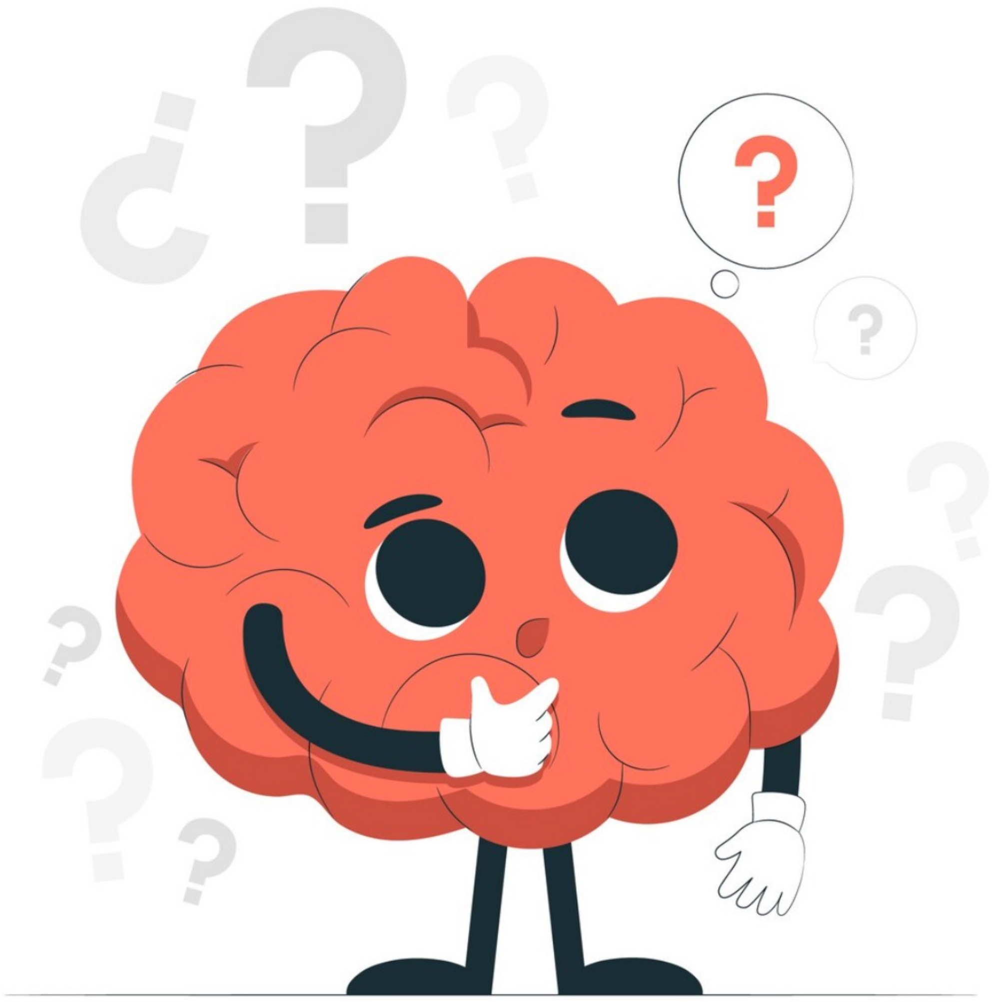
Input: Beautiful film.	**Sentiment:** positive
Input: Subpar acting.	**Sentiment:** negative
Input: Amazing.	**Sentiment:**

Zhao et al., Calibrate before Use: Improving Few-Shot Performance of Language Models, ICML 2021

How important is the structure of the prompt for in-context learning?

Components of a prompt:

1. Prompt format
2. Training example selection
3. Training example permutation



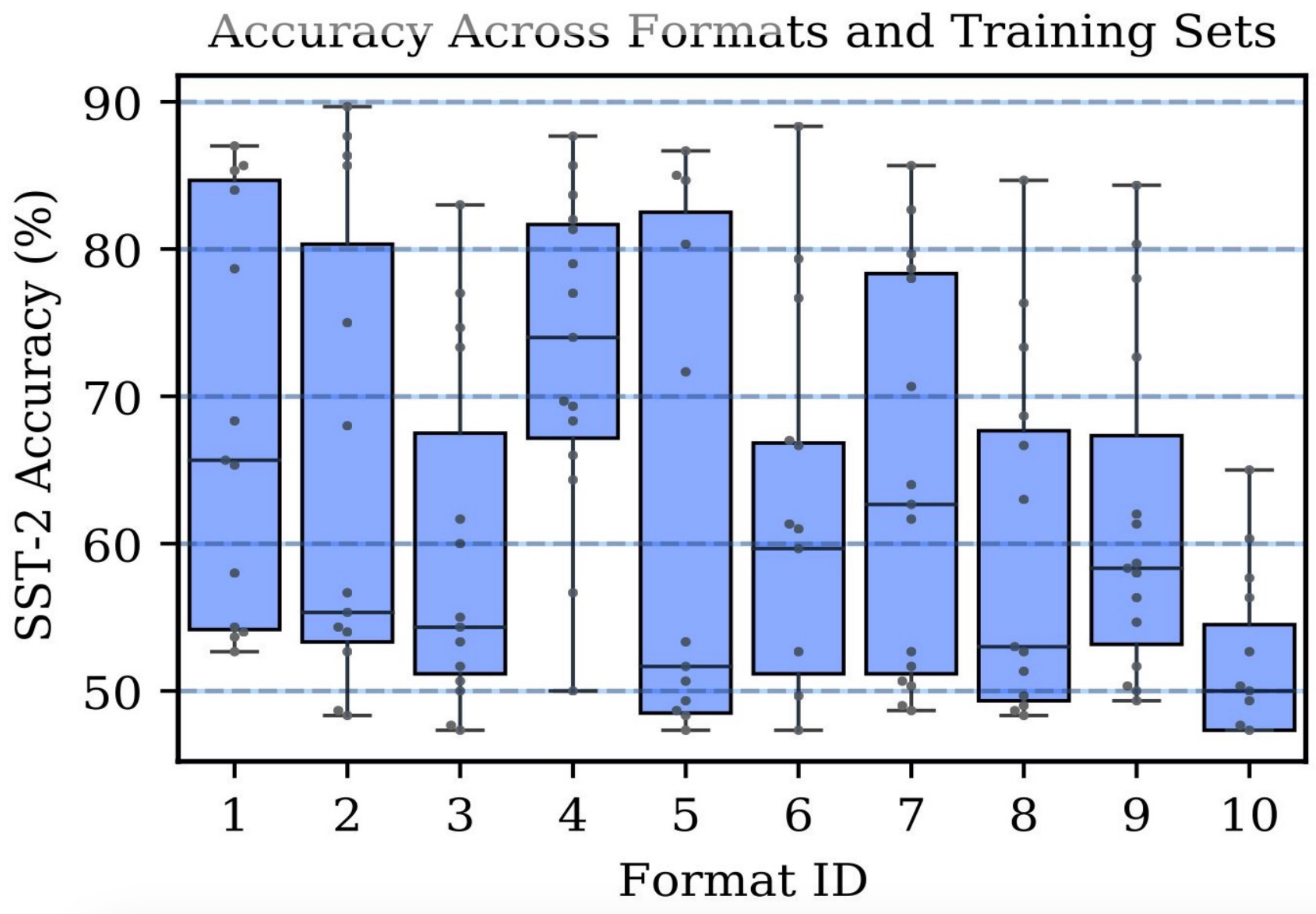
Let's try to ablate each component ...

Zhao et al., Calibrate before Use: Improving Few-Shot Performance of Language Models, ICML 2021

How important is the structure of the prompt for in-context learning?

Components of a prompt:

1. **Prompt format**
2. Training example selection
3. Training example permutation



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:**

Note

In-context learning is highly sensitive to prompt **format**

Zhao et al., Calibrate before Use: Improving Few-Shot Performance of Language Models, ICML 2021

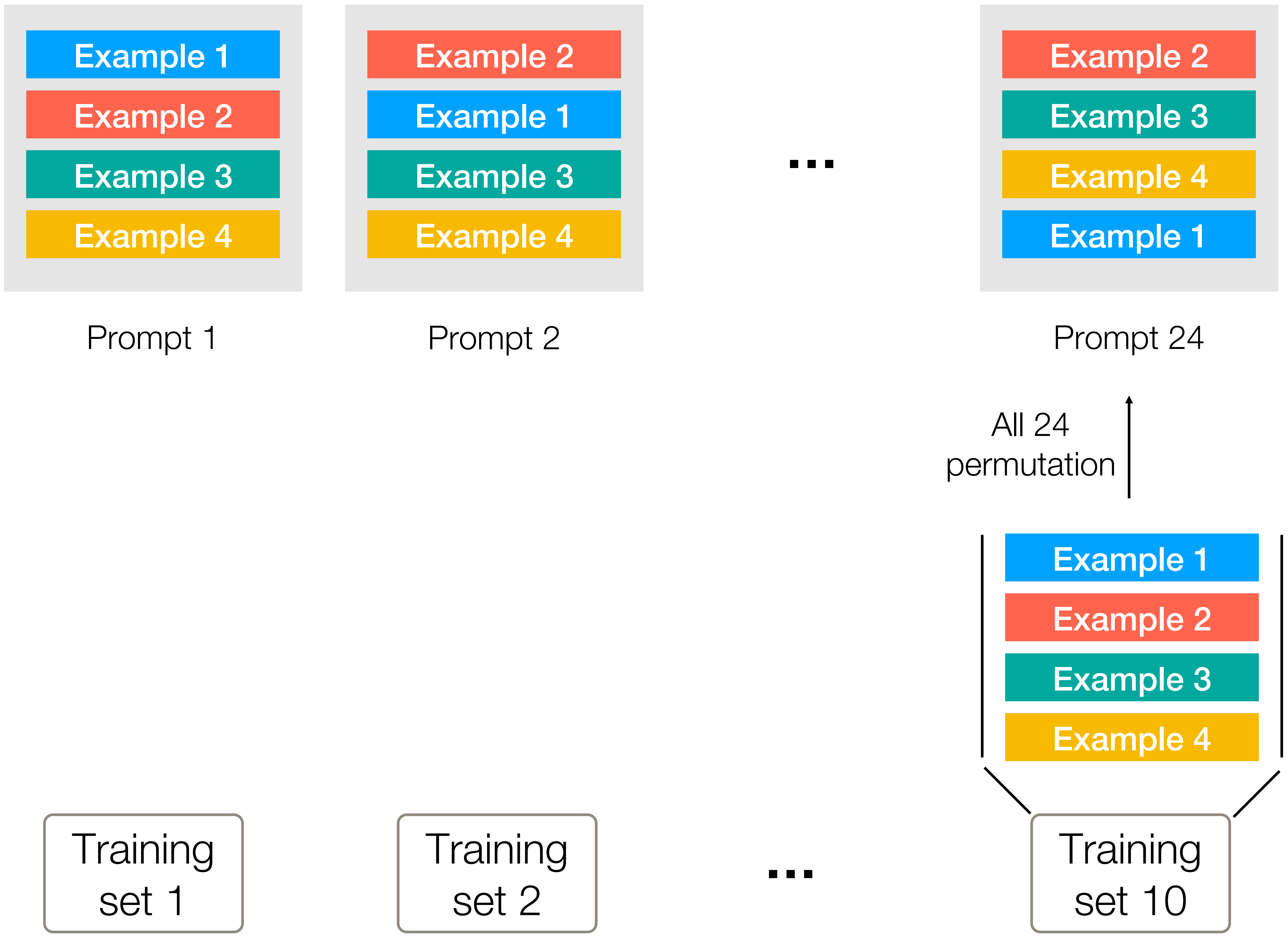
How important is the structure of the prompt for in-context learning?

Components of a prompt:

1 Prompt format

2 Training example selection

3 Training example permutation



Zhao et al., Calibrate before Use: Improving Few-Shot Performance of Language Models, ICML 2021

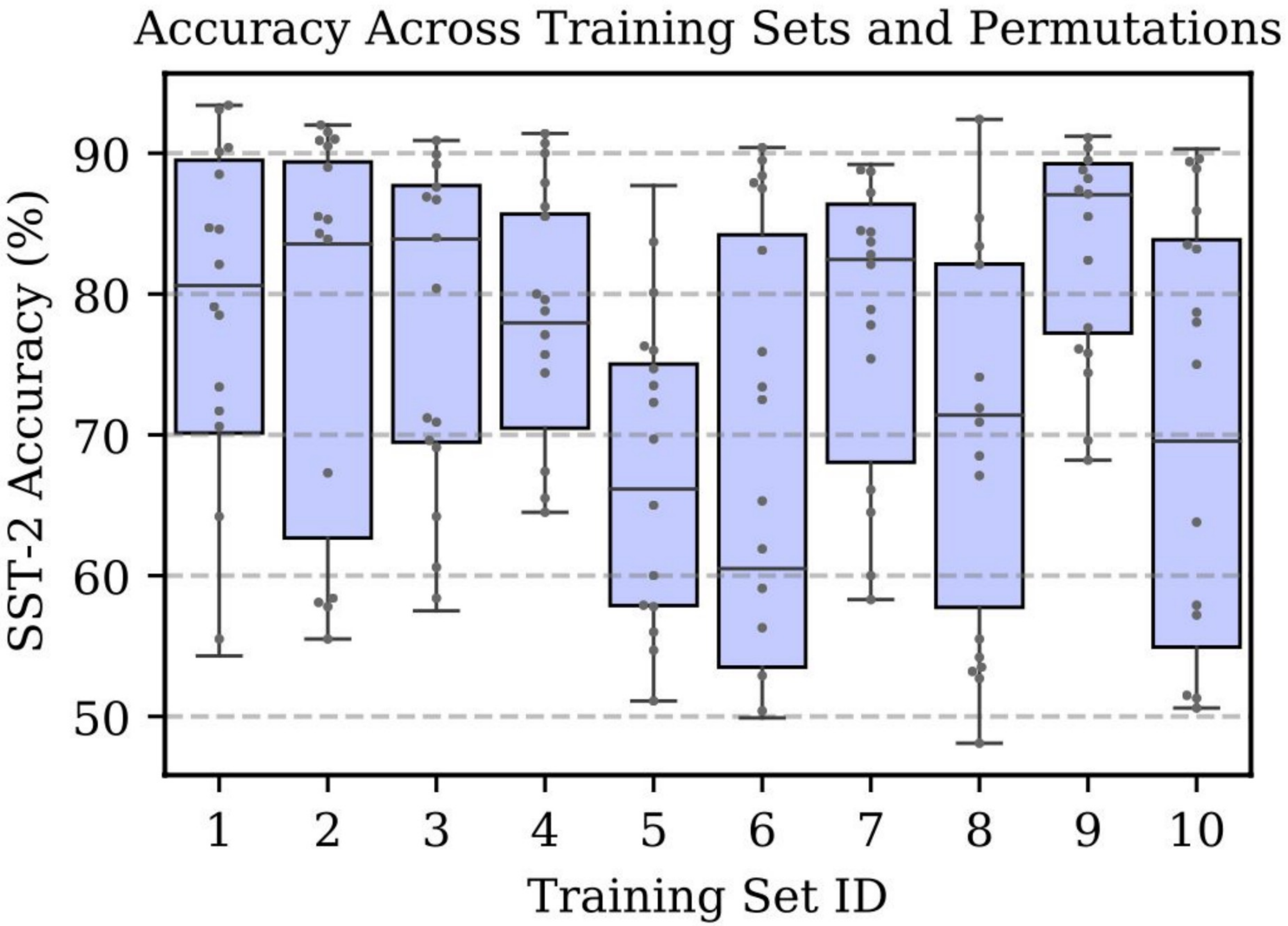
How important is the structure of the prompt for in-context learning?

Components of a prompt:

1 Prompt format

2 Training example selection

3 Training example permutation



Note

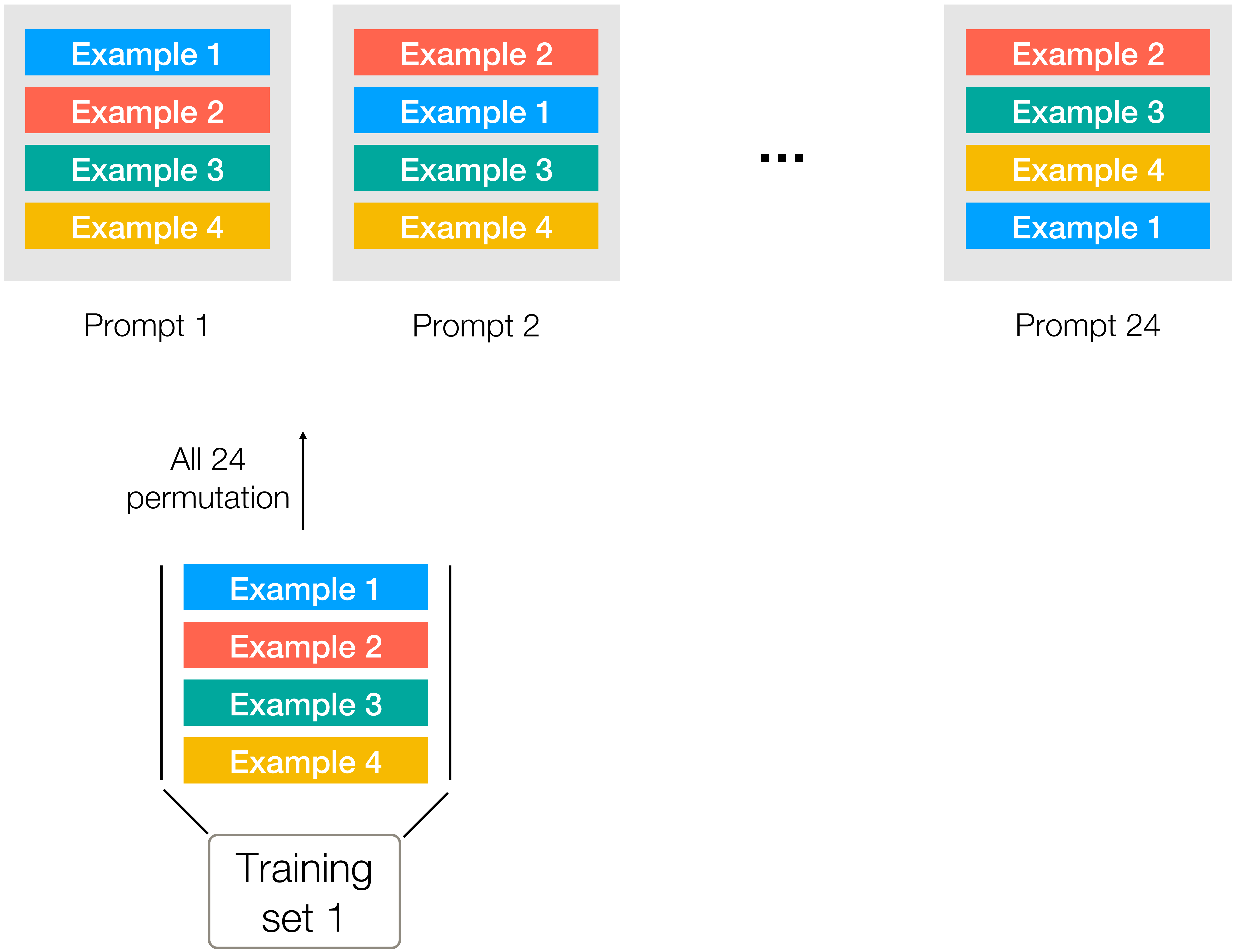
In-context learning is highly sensitive to example **selection**

Zhao et al., Calibrate before Use: Improving Few-Shot Performance of Language Models, ICML 2021

How important is the structure of the prompt for in-context learning?

Components of a prompt:

1. Prompt format
2. Training example selection
3. **Training example permutation**



Zhao et al., Calibrate before Use: Improving Few-Shot Performance of Language Models, ICML 2021

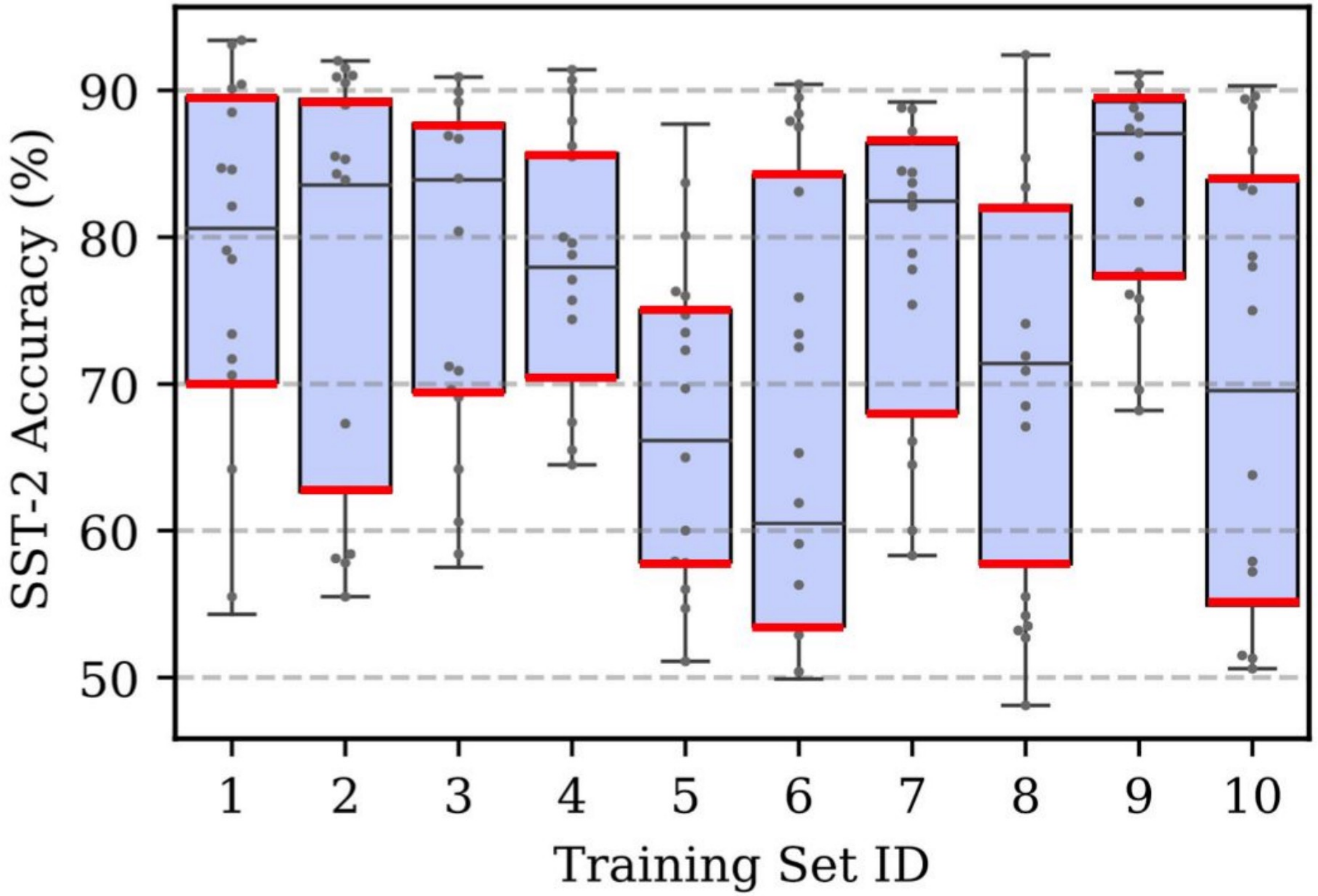
How important is the structure of the prompt for in-context learning?

Components of a prompt:

1 Prompt format

2 Training example selection

3 **Training example permutation**



Note

In-context learning is highly sensitive to example **permutation**

Zhao et al., Calibrate before Use: Improving Few-Shot Performance of Language Models, ICML 2021

What causes this sensitivity?

Three main reasons:

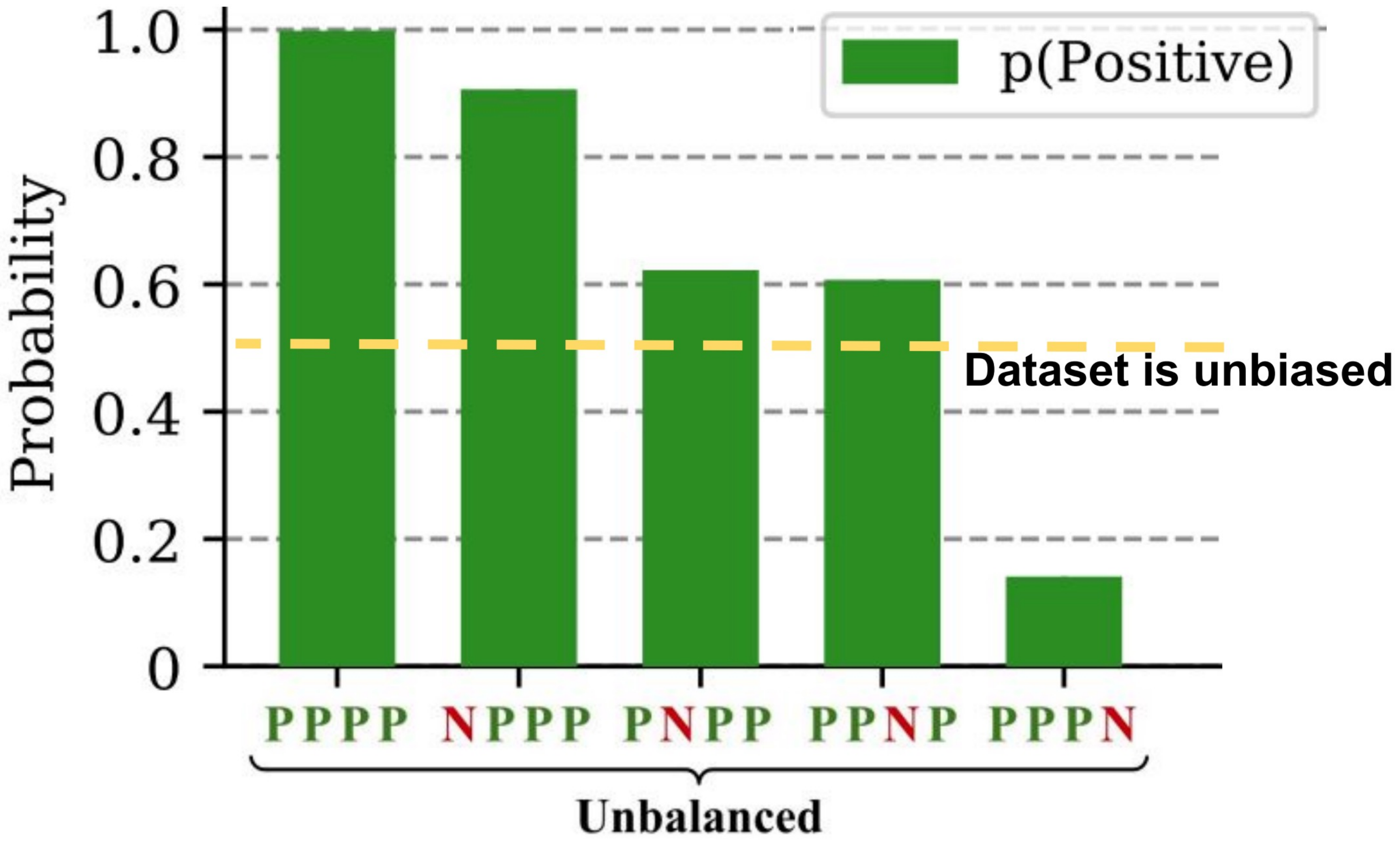
- Majority label bias
- Common token bias
- Recency bias

Zhao et al., Calibrate before Use: Improving Few-Shot Performance of Language Models, ICML 2021

What causes this sensitivity?

Three main reasons:

- **Majority label bias**
- Common token bias
- Recency bias



1. Model prefers to predict positive when the majority labels is "P/Positive"
2. Surprising because the validation dataset is balanced!

Zhao et al., Calibrate before Use: Improving Few-Shot Performance of Language Models, ICML 2021

What causes this sensitivity?

Three main reasons:

- Majority label bias
- **Common token bias**
- Recency bias

Language Model

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

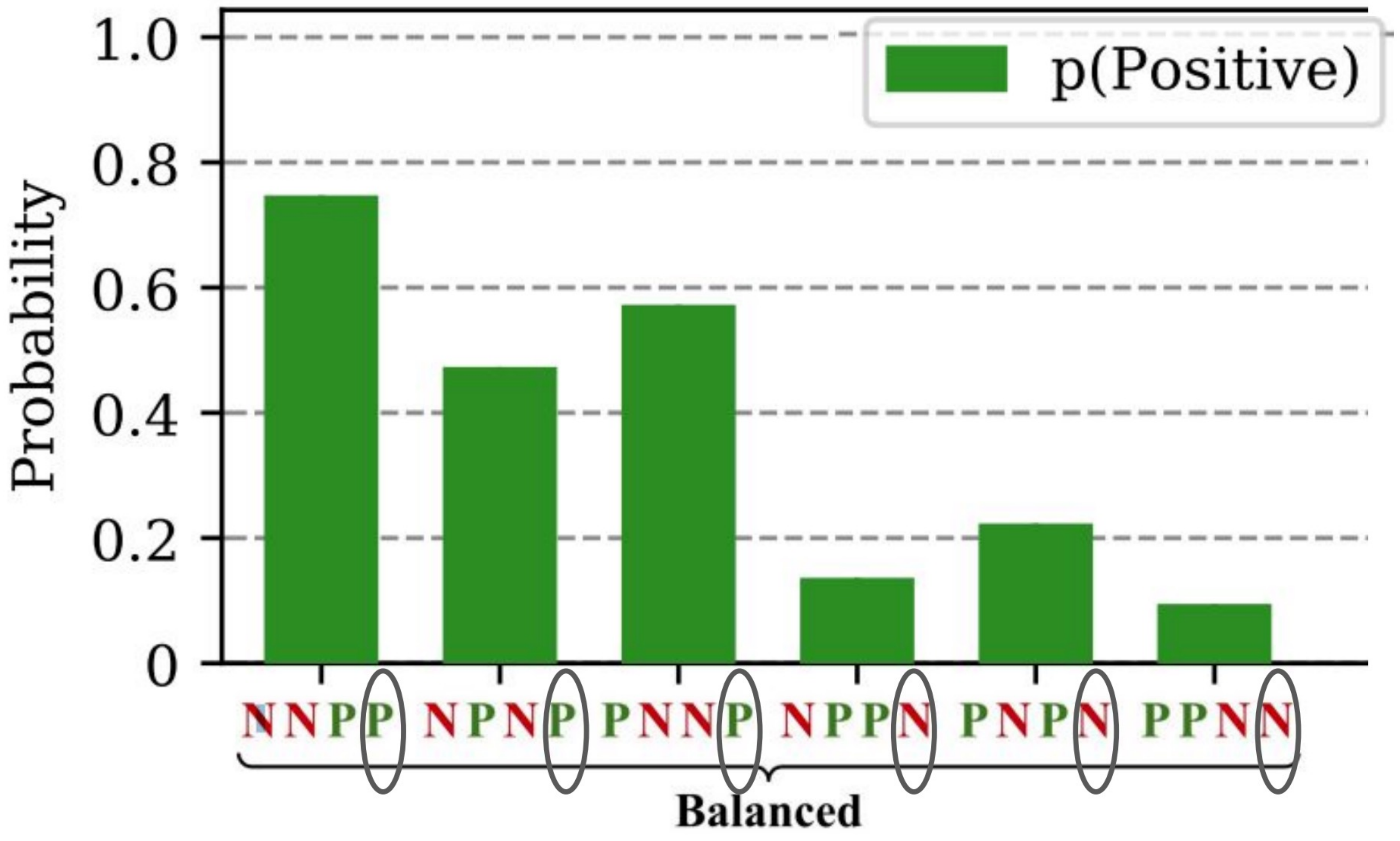
Model is biased towards predicting the **incorrect** frequent token "book" even when both "book" and "transportation" are equally likely labels in the dataset

Zhao et al., Calibrate before Use: Improving Few-Shot Performance of Language Models, ICML 2021

What causes this sensitivity?

Three main reasons:

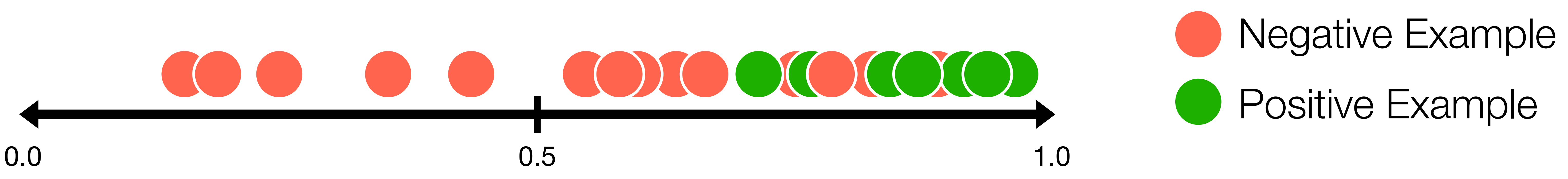
- Majority label bias
- Common token bias
- **Recency bias**



1. Model is heavily biased towards the most recent label
2. Again, dataset is balanced!

Zhao et al., Calibrate before Use: Improving Few-Shot Performance of Language Models, ICML 2021

What is the impact of all these factors?



Visualizing predictions of 25 randomly sampled instances from SST2

Zhao et al., Calibrate before Use: Improving Few-Shot Performance of Language Models, ICML 2021

How do we make in-context learning more robust?

Can we infer the shift in the output distribution caused by a given prompt?

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:**

Model

positive	0.65
negative	0.35

Note

Classification tasks: normalized scores of label words

Generation tasks: probabilities of the first token of the generation over the entire vocabulary

Step 2: Counter the bias

"Calibrate" predictions with affine transformation

$$\underset{\text{Calibrated probs}}{\hat{q}} = \textbf{softmax}(W\underset{\text{Original probs}}{\hat{p}} + b)$$

Fit W and b to cause uniform prediction for "N/A"

$$W = \begin{bmatrix} \frac{1}{0.65} & 0 \\ 0 & \frac{1}{0.35} \end{bmatrix} \qquad b = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Contextual calibration (technical details)

For generation tasks, why is only the first token calibrated?

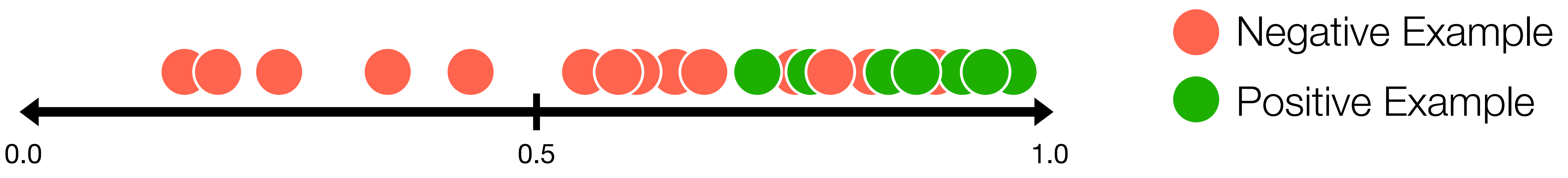
- Authors claim the first token has the most impact on future predictions
- Calibrating all generated tokens might be tricky as dimension of W is |V| x |V|

Zhao et al., Calibrate before Use: Improving Few-Shot Performance of Language Models, ICML 2021

Contextual calibration (technical details)

Why is W diagonal? Why can't we learn some fancy non-linear function?

- The biases effectively cause a simple shift in the output distribution, we don't need a fancy function
- Diagonal W is easy to invert, low computational overhead
- If we added a non-linearity, how would we learn W with a few samples?
  - Potentially gradient descent, but tricky with few samples



Zhao et al., Calibrate before Use: Improving Few-Shot Performance of Language Models, ICML 2021

Contextual calibration (technical details)

Why do they calibrate probabilities instead of calibrating logits?

- OpenAI API only returns probabilities across the vocabulary
- Authors acknowledge that calibrating logits would have been more “natural”

Zhao et al., Calibrate before Use: Improving Few-Shot Performance of Language Models, ICML 2021

Datasets: Text Classification

Task	Prompt	Label Names
SST-2	Review: This movie is amazing! Sentiment: Positive	Positive, Negative
AGNews	Article: USATODAY.com - Retail sales bounced back a bit in July, and new claims for jobless benefits fell last week, the government said Thursday, indicating the economy is improving from a midsummer slump. Answer: Business	World, Sports, Business, Technology

Note

1. Label is just a single token
2. We calibrate probabilities of all the label words

Zhao et al., Calibrate before Use: Improving Few-Shot Performance of Language Models, ICML 2021

Datasets: Fact Retrieval

Task	Prompt
LAMA	Alexander Berntsson was born in Sweden
	Khalid Karami was born in

Note

1. Label is just a single token
2. We calibrate probabilities of all the words in the vocabulary

Zhao et al., Calibrate before Use: Improving Few-Shot Performance of Language Models, ICML 2021

Datasets: Information Extraction

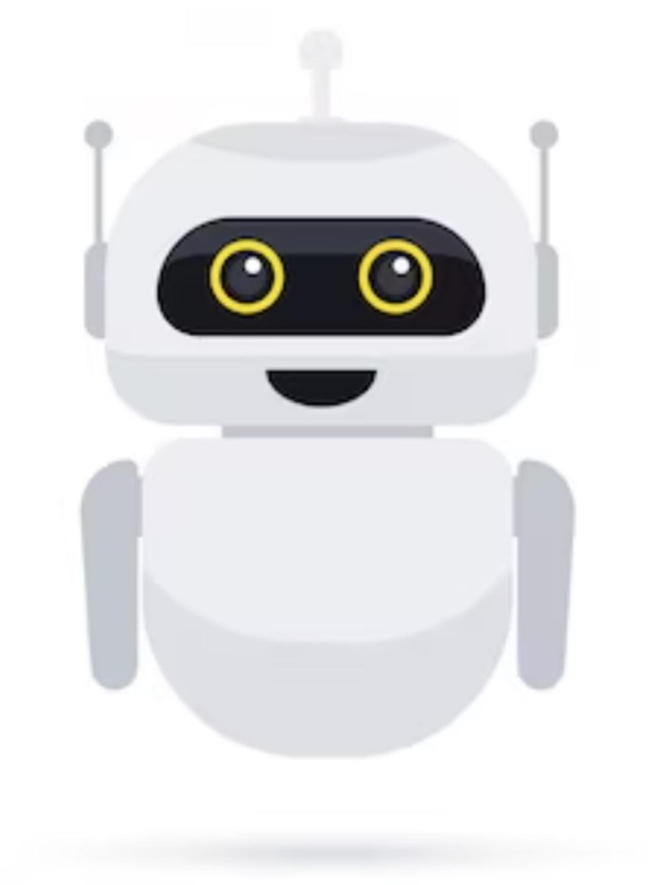
ATIS (Airline)	Sentence: what are the two american airlines flights that leave from dallas to san francisco in the evening Airline name: american airlines
MIT Movies (Genre)	Sentence: last to a famous series of animated movies about a big green ogre and his donkey and cat friends Genre: animated

Note

1. Label is multiple tokens
2. We calibrate probabilities of all the words in the vocabulary

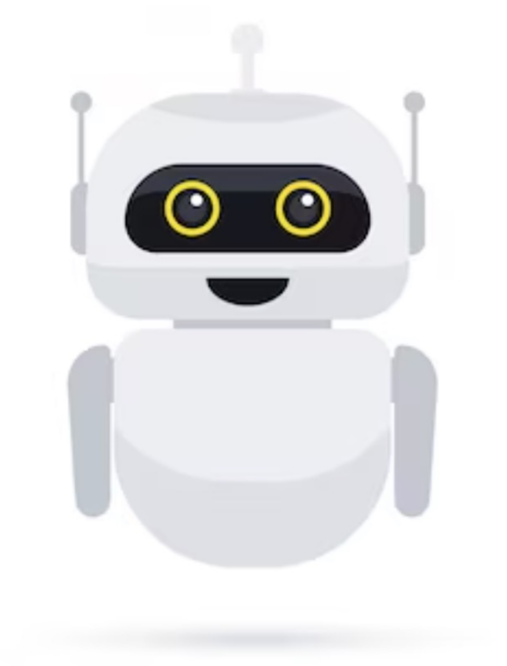
Zhao et al., Calibrate before Use: Improving Few-Shot Performance of Language Models, ICML 2021

Model



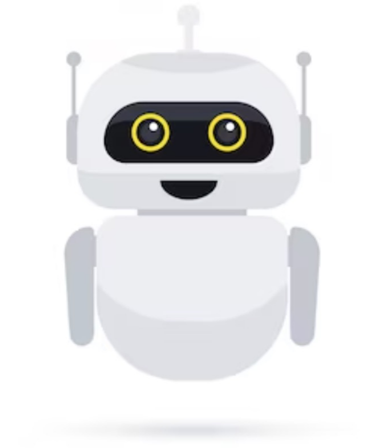
GPT-3

175 billion



GPT-3

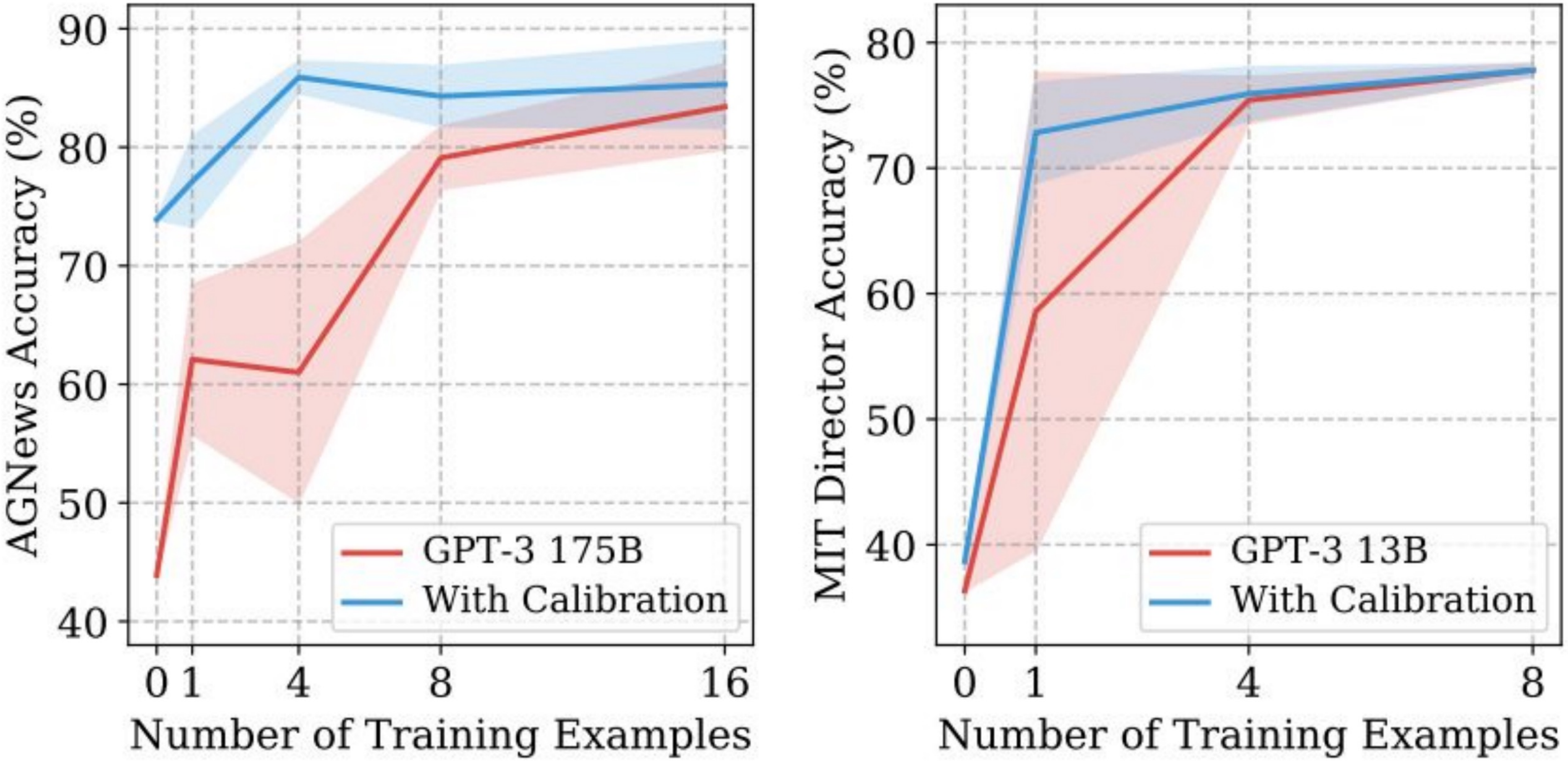
13 billion



GPT-3

2.7 billion

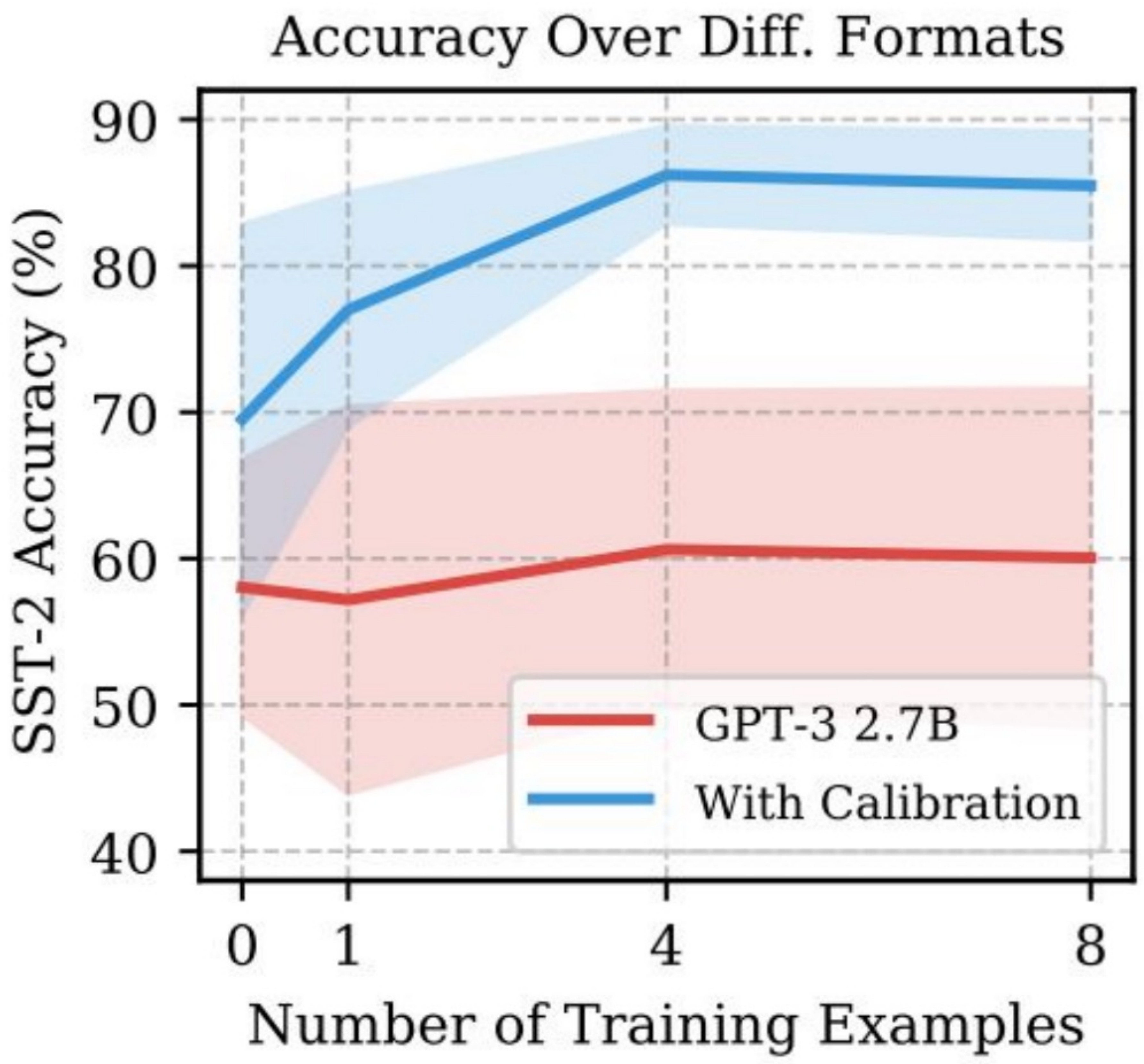
Results



Reduces variance across training sets and permutations

Zhao et al., Calibrate before Use: Improving Few-Shot Performance of Language Models, ICML 2021

Results



Format ID	Prompt	Label Names
1	Review: This movie is amazing! Answer: Positive Review: Horrific movie, don't see it. Answer:	Positive, Negative
2	Review: This movie is amazing! Answer: good Review: Horrific movie, don't see it. Answer:	good, bad
3	My review for last night's film: This movie is amazing! The critics agreed that this movie was good My review for last night's film: Horrific movie, don't see it. The critics agreed that this movie was	good, bad
4	Here is what our critics think for this month's films. One of our critics wrote "This movie is amazing!". Her sentiment towards the film was positive. One of our critics wrote "Horrific movie, don't see it". Her sentiment towards the film was	positive, negative

Reduces variance across 15 different prompt formats

Zhao et al., Calibrate before Use: Improving Few-Shot Performance of Language Models, ICML 2021

Surface Form Competition:
Why the Highest Probability Answer Isn't Always Right

=Ari Holtzman[1] =Peter West[1,2]
Vered Shwartz[1,2] Yejin Choi[1,2] Luke Zettlemoyer[1]
[1]Paul G. Allen School of Computer Science & Engineering, University of Washington
[2]Allen Institute for Artificial Intelligence
{ahai,pawest}@cs.washington.edu

EMNLP 2021

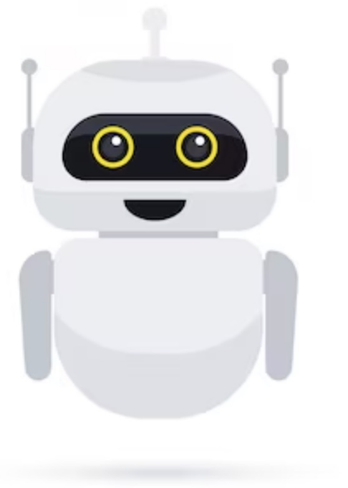
Surface Form Competition

A human wants to submerge himself in water, what should he use?

Humans select options

- (a) Coffee cup
- (b) Whirlpool bath
- (c) Cup
- (d) Puddle

Language Models assign probability to every possible string



- (e) Water
- (f) A bathtub
- (g) I don't know
- (h) A birdbath
- (i) Bathtub
- ⋮

= right concept, wrong surface form

$P(Bathtub|x) = 0.8$ → $P(Whirlpool\ bath|x) \leq 0.2$

Competes for probability mass

Generic output always assigned high probability

Every correct string is assigned lower scores than expected

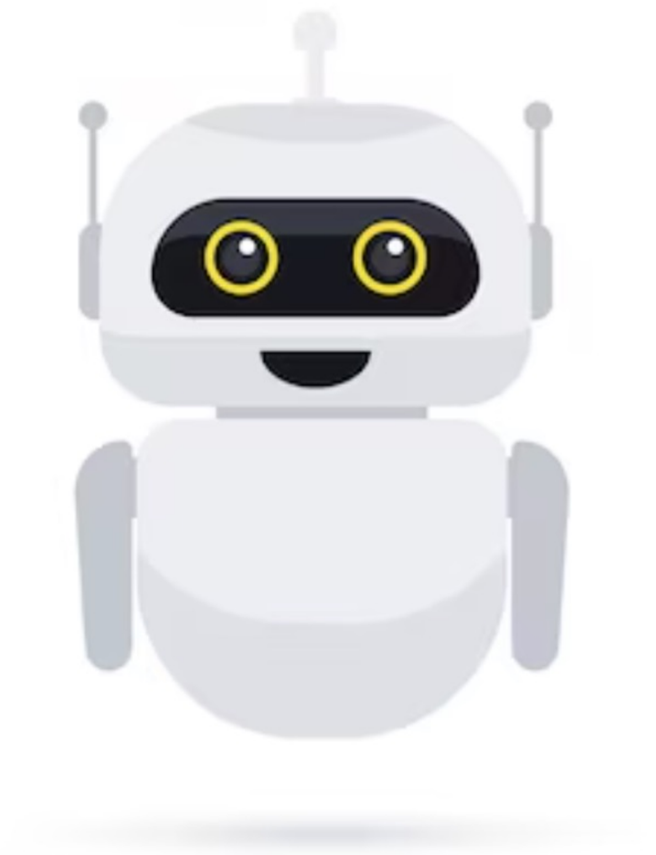
Holtzman et al., Surface Form Competition: Why the Highest Probability Answer Isn't Always Right, EMNLP 2021

Choice of Plausible Alternatives (COPA)

Premise (X): The bar closed because

Hypothesis 1 (y_1): it was crowded.

Hypothesis 2 (y_2): it was 3am.



$$P(y_1|X) > P(y_2|X)$$

GPT-3

Holtzman et al., Surface Form Competition: Why the Highest Probability Answer Isn't Always Right, EMNLP 2021

Baselines

Template:

Premise (X): The bar closed because

Domain Premise (X_{domain}): because

Hypothesis 1 (y_1): it was crowded.

Hypothesis 2 (y_2): it was 3am.

choose between
Hypothesis y_1 and y_2 given
Premise x

Holtzman et al., Surface Form Competition: Why the Highest Probability Answer Isn't Always Right, EMNLP 2021

Baselines

Template:

Premise (X): The bar closed because

Domain Premise (X_{domain}): because

Hypothesis 1 (y_1): it was crowded.

Hypothesis 2 (y_2): it was 3am.

Note

This paper does not introduce any new modeling approaches, just a new scoring function

Scoring Functions

Probability (LM)

$$\underset{i}{\mathrm{arg}max} P(y_i|x)$$

logit

Average Log-Likelihood (Ava)

$$\underset{i}{\mathrm{arg}max} \frac{\sum_{j=1}^{l_i} P(y_i^j|x, y^{1\ldots j-1})}{l_i}$$

Contextual Calibration (CC)

$$\underset{i}{\mathrm{argmax}}\, w_i P(y_i|x) + b$$

Zhao et al., 2021

Domain Conditional PMI (PMI_{DC})

$$\underset{i}{\mathrm{arg}max} \frac{P(y_i|x)}{P(y_i|x_{domain})}$$

Holtzman et al., Surface Form Competition: Why the Highest Probability Answer Isn’t Always Right, EMNLP 2021

Pointwise Mutual Information (PMI)

Template:

Premise (X): The bar closed because

Domain Premise (X_{domain}): because

Hypothesis 1 (y_1): it was crowded.

Hypothesis 2 (y_2): it was 3am.

$$PMI(x, y) = \boxed{\log \frac{P(y|x)}{P(y)}} = \boxed{\log \frac{P(x|y)}{P(x)}}$$

How much more likely does the hypothesis y becomes if we are given the premise x?

The probability of the premise x given the hypothesis y - “scoring by premise” (more on this later)

Holtzman et al., Surface Form Competition: Why the Highest Probability Answer Isn’t Always Right, EMNLP 2021

Domain Conditional Pointwise Mutual Information (PMI)

Template:

Premise (X): The bar closed because

Domain Premise (X_{domain}): because

Hypothesis 1 (y_1): it was crowded.

Hypothesis 2 (y_2): it was 3am.

$$PMI(x, y) = \log \frac{P(y|x)}{\boxed{P(y)}} = \log \frac{P(x|y)}{P(x)}$$

poorly calibrated because language models are not trained to produce unconditional generations

$$PMI_{DC}(x, y, domain) = \log \frac{P(y|x, domain)}{P(y|domain)} = \log \frac{P(y|x, domain)}{P(y|x_{domain})}$$

where domain is representative of the given task

Note

Assumption: ending of the conditional premise x is a domain-relevant string X_{domain}

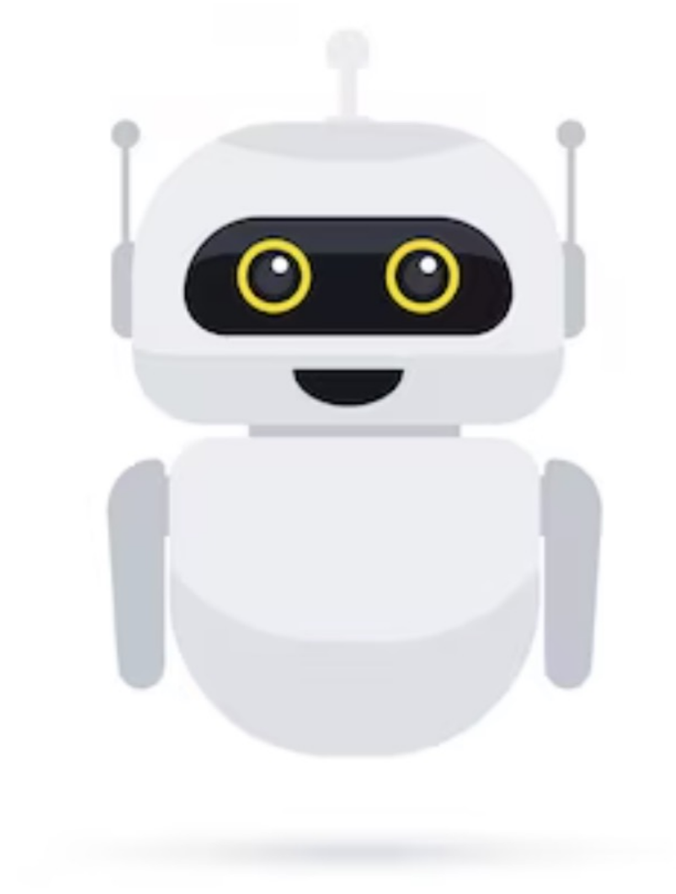
Dataset

$[Orginal\ Question]_P$
$[Domain\ premise]_{DP}$
$[Orginal\ answers]_{UH}$

Type	Dataset	Template
Continuation	COPA	[The man broke his toe]$_P$ [because]$_{DP}$ [he got a hole in his sock.]$_{UH}$
		[I tipped the bottle]$_P$ [so]$_{DP}$ [the liquid in the bottle froze.]$_{UH}$
	StoryCloze	[Jennifer has a big exam tomorrow. She got so stressed, she pulled an all-nighter. She went into class the next day, weary as can be. Her teacher stated that the test is postponed for next week.]$_P$ [The story continues:]$_{DP}$ [Jennifer felt bittersweet about it.]$_{UH}$
	HellaSwag	[A female chef in white uniform shows a stack of baking pans in a large kitchen presenting them. the pans]$_P$ [contain egg yolks and baking soda.]$_{UH}$
QA	RACE	[There is not enough oil in the world now. As time goes by, it becomes less and less, so what are we going to do when it runs out [...].]$_P$ question: [According to the passage, which of the following statements is true]$_P$ [?]$_{DP}$ answer: [There is more petroleum than we can use now.]$_{UH}$
	ARC	[What carries oxygen throughout the body?]$_P$ [the answer is:]$_{DP}$ [red blood cells.]$_{UH}$
	OBQA	[Which of these would let the most heat travel through?]$_P$ [the answer is:]$_{DP}$ [a steel spoon in a cafeteria.]$_{UH}$
	CQA	[Where can I stand on a river to see water falling without getting wet?]$_P$ [the answer is:]$_{DP}$ [bridge.]$_{UH}$
Boolean QA	BoolQ	title: [The Sharks have advanced to the Stanley Cup finals once, losing to the Pittsburgh Penguins in 2016 [...]]$_P$ question: [Have the San Jose Sharks won a Stanley Cup?]$_P$ [answer:]$_{DP}$ [No.]$_{UH}$
Entailment	RTE	[Time Warner is the world's largest media and Internet company.]$_P$ question: [Time Warner is the world's largest company.]$_P$ [true or false? answer:]$_{DP}$ [true.]$_{UH}$
	CB	question: Given that [What fun to hear Artemis laugh. She's such a serious child.]$_P$ Is [I didn't know she had a sense of humor.]$_P$ true, false, or neither? [the answer is:]$_{DP}$ [true.]$_{UH}$
Text Classification	SST-2	"[Illuminating if overly talky documentary]$_P$" [[The quote] has a tone that is]$_{DP}$ [positive.]$_{UH}$
	SST-5	"[Illuminating if overly talky documentary]$_P$" [[The quote] has a tone that is]$_{DP}$ [neutral.]$_{UH}$
	AG's News	title: [Economic growth in Japan slows down as the country experiences a drop in domestic and corporate [...]]$_P$ summary: [Expansion slows in Japan]$_P$ [topic:]$_{DP}$ [Sports.]$_{UH}$
	TREC	[Who developed the vaccination against polio?]$_P$ [The answer to this question will be]$_{DP}$ [a person.]$_{UH}$

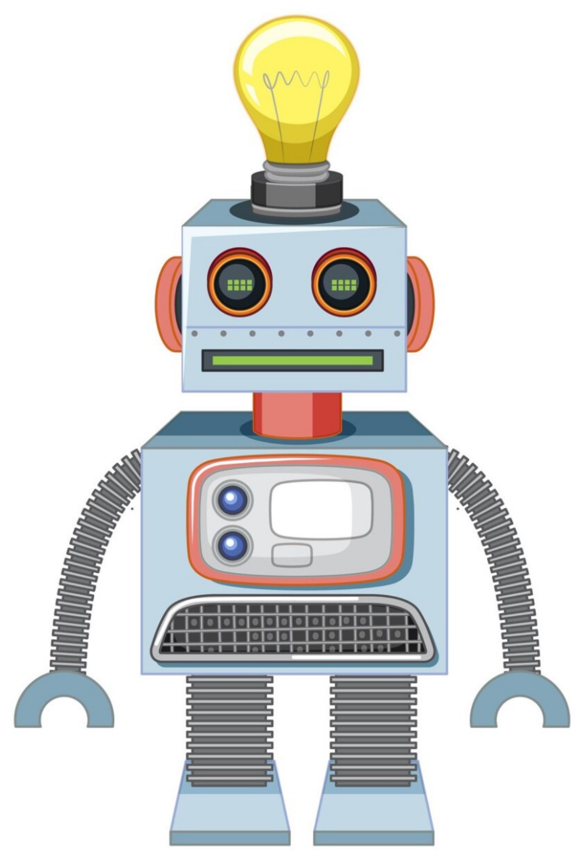
Holtzman et al., Surface Form Competition: Why the Highest Probability Answer Isn't Always Right, EMNLP 2021

Model



GPT-3

Zero-shot



GPT-2

Reported but won't be the focus of the results

Zero-shot Multiple Choice Accuracy

Holtzman et al., 2021

Params.	2.7B					6.7B				13B				175B				
	Unc	LM	Avg	PMI_{DC}	CC	Unc	LM	Avg	PMI_{DC}	Unc	LM	Avg	PMI_{DC}	Unc	LM	Avg	PMI_{DC}	CC
COPA	54.8	68.4	68.4	**74.4**	-	56.4	75.8	73.6	**77.0**	56.6	79.2	77.8	**84.2**	56.0	85.2	82.8	**89.2**	-
SC	50.9	66.0	68.3	**73.1**	-	51.4	70.2	73.3	**76.8**	52.0	74.1	77.8	**79.9**	51.9	79.3	83.1	**84.0**	-
HS	31.1	34.5	**41.4**	34.2	-	34.7	40.8	**53.5**	40.0	38.8	48.8	**66.2**	45.8	43.5	57.6	**77.2**	53.5	-
R-M	22.4	37.8	42.4	**42.6**	-	21.2	43.3	45.9	**48.5**	22.9	49.6	50.6	**51.3**	22.5	55.7	**56.4**	55.7	-
R-H	21.4	30.3	32.7	**36.0**	-	22.0	34.8	36.8	**39.8**	22.9	38.2	39.2	**42.1**	22.2	42.4	43.3	**43.7**	-
ARC-E	31.6	**50.4**	44.7	44.7	-	33.5	**58.2**	52.3	51.5	33.8	**66.2**	59.7	57.7	36.2	**73.5**	67.0	63.3	-
ARC-C	21.1	21.6	25.5	**30.5**	-	21.8	26.8	29.8	**33.0**	22.3	32.1	34.3	**38.5**	22.6	40.2	43.2	**45.5**	-
OBQA	10.0	17.2	27.2	**42.8**	-	11.4	22.4	35.4	**48.0**	10.4	28.2	41.2	**50.4**	10.6	33.2	43.8	**58.0**	-
CQA	15.9	33.2	36.0	**44.7**	-	17.4	40.0	42.9	**50.3**	16.4	48.8	47.9	**58.5**	16.3	61.0	57.4	**66.7**	-
BQ	**62.2**	58.5	58.5	53.5	-	37.8	**61.0**	**61.0**	**61.0**	**62.2**	61.1	61.1	60.3	37.8	62.5	62.5	**64.0**	-
RTE	47.3	48.7	48.7	**51.6**	49.5	52.7	**55.2**	**55.2**	48.7	52.7	52.7	52.7	**54.9**	47.3	56.0	56.0	**64.3**	57.8
CB	08.9	51.8	51.8	**57.1**	50.0	08.9	33.9	33.9	**39.3**	08.9	**51.8**	**51.8**	50.0	08.9	48.2	48.2	**50.0**	48.2
SST-2	49.9	53.7	53.76	**72.3**	71.4	49.9	54.5	54.5	**80.0**	49.9	69.0	69.0	**81.0**	49.9	63.6	63.6	71.4	**75.8**
SST-5	18.1	20.0	20.4	**23.5**	-	18.1	27.8	22.7	**32.0**	18.1	18.6	**29.6**	**19.1**	17.6	27.0	27.3	**29.6**	-
AGN	25.0	69.0	69.0	**67.9**	63.2	25.0	**64.2**	**64.2**	57.4	25.0	69.8	69.8	**70.3**	25.0	**75.4**	**75.4**	74.7	73.9
TREC	13.0	29.4	19.2	**57.2**	38.8	22.6	30.2	22.8	**61.6**	22.6	**34.0**	21.4	32.4	22.6	47.2	25.4	**58.4**	57.4

$\arg\max_i P(y_i|x_{domain})$

ignore the premise completely!

Consistently beat or tie other methods across model sizes and datasets

Prompt Robustness

Prompt Robustness on SST-2

	Method	Unc	LM	PMI$_{DC}$
GPT-2	125M	49.9$_{0}$	56.8$_{7.3}$	**58.8**$_{7.6}$
	350M	49.9$_{0}$	58.0$_{11.3}$	**60.3**$_{11.4}$
	760M	49.9$_{0}$	57.0$_{9.2}$	**67.7**$_{13.4}$
	1.6B	49.9$_{0}$	57.3$_{8.2}$	**69.8**$_{13.3}$
GPT-3	2.7B	49.9$_{0}$	56.1$_{9.0}$	**66.2**$_{15.7}$
	6.7B	49.9$_{0}$	59.5$_{10.7}$	**67.9**$_{13.6}$
	13B	49.9$_{0}$	63.0$_{14.9}$	**71.7**$_{16.1}$
	175B	49.9$_{0}$	72.5$_{15.7}$	**74.8**$_{14.0}$

maintain the highest mean using
15 different templates for SST-2

but still high variance

4-shot Inference Results

	SST-2			CQA			
Method	Unc	LM	PMI$_{DC}$	Unc	LM	Avg	PMI$_{DC}$
125M	49.9$_{0}$	63.6$_{7.4}$	**71.7**$_{5.1}$	15.5$_{0}$	29.9$_{1.6}$	32.7$_{1.4}$	**38.3**$_{1.7}$
350M	49.9$_{0}$	76.3$_{13.8}$	**76.4**$_{8.1}$	16.5$_{0}$	37.6$_{2.3}$	40.4$_{2.3}$	**45.7**$_{2.4}$
760M	49.9$_{0}$	85.9$_{7.2}$	**87.1**$_{3.0}$	16.1$_{0}$	41.5$_{2.6}$	42.4$_{2.5}$	**47.0**$_{1.5}$
1.6B	49.9$_{0}$	85.4$_{1.7}$	**89.4**$_{4.0}$	16.0$_{0}$	46.2$_{1.5}$	47.7$_{1.9}$	**52.3**$_{2.1}$
2.7B	49.9$_{0}$	**88.1**$_{4.9}$	87.7$_{5.5}$	16.6$_{0}$	43.0$_{1.7}$	45.6$_{1.9}$	**50.4**$_{1.1}$
6.7B	49.9$_{0}$	**92.9**$_{2.1}$	79.8$_{6.9}$	16.9$_{0}$	52.3$_{1.4}$	53.4$_{1.0}$	**56.5**$_{1.6}$
13B	49.9$_{0}$	85.4$_{9.0}$	**86.9**$_{7.5}$	16.7$_{0}$	58.4$_{2.0}$	59.3$_{1.5}$	**63.4**$_{1.4}$
175B	49.9$_{0}$	89.9$_{5.5}$	**95.5**$_{0.7}$	16.5$_{0}$	69.1$_{1.9}$	69.4$_{0.8}$	**72.0**$_{0.9}$

Holtzman et al., Surface Form Competition: Why the Highest Probability Answer Isn't Always Right, EMNLP 2021

Removing Surface Form Competition

COPA

because

so

The bar closed because it was 3 AM
I tipped the bottle so the liquid in the bottle poured out

Holtzman et al., Surface Form Competition: Why the Highest Probability Answer Isn't Always Right, EMNLP 2021

Removing Surface Form Competition

COPA		"Flipped"
because	$\longrightarrow$	so
so	$\longrightarrow$	because

Premise (x): The bar closed **because**

Domain Premise (x_{domain}): **because**

Hypothesis 1 (y_1): it was crowded.

Hypothesis 2 (y_2): it was 3 AM.

Premise 1 ($\hat{x}_1$): It was crowded **so**

Premise 2 ($\hat{x}_2$): It was 3 AM **so**

Hypothesis (y): the bar closed.

Holtzman et al., Surface Form Competition: Why the Highest Probability Answer Isn't Always Right, EMNLP 2021

Removing Surface Form Competition

50.0 because the outputs are now the same for the two different inputs

	COPA				COPA Flipped			
Method	Unc	LM	Avg	PMI_{DC}	Unc	LM	Avg	PMI_{DC}
125M	56.4	61.0	63.2	62.8	50.0	63.2	63.2	63.2
350M	55.8	67.0	66.0	70.0	50.0	66.4	66.4	66.4
760M	55.6	69.8	67.6	69.4	50.0	70.8	70.8	70.8
1.6B	56.0	69.0	68.4	71.6	50.0	73.0	73.0	73.0
2.7B	54.8	68.4	68.4	74.4	50.0	68.4	68.4	68.4
6.7B	56.4	75.8	73.6	77.0	50.0	76.8	76.8	76.8
13B	56.6	79.2	77.8	84.2	50.0	79.0	79.0	79.0
175B	56.0	85.2	82.8	89.2	50.0	83.6	83.6	83.6

LM, Avg, and PMI_{DC} are the same without surface form competition

better on COPA than COPA Flipped since "because" and "so" are not perfectly invertible and the original phrases sound more natural

Holtzman et al., Surface Form Competition: Why the Highest Probability Answer Isn't Always Right, EMNLP 2021

Removing Surface Form Competition

Premise (x): The bar closed **because**

Domain Premise (x_{domain}): because

Hypothesis 1 (y_1): it was crowded.

Hypothesis 2 (y_2): it was 3 AM.

Premise 1 ($\hat{x}_1$): It was crowded **so**

Premise 2 ($\hat{x}_2$): It was 3 AM **so**

Hypothesis ($\hat{y}$): the bar closed.

Hypothesis 2' (y_2'): it was 3:30AM.

Premise 2' ($\hat{x}_2'$): It was 3:30AM so

$$P(y_1|x) > p(y_2'|x)$$

$$P(\hat{y}|\hat{x}_2') > P(\hat{y}|\hat{x}_1')$$

$$\frac{P(y_2'|x)}{P(y_2'|x_{domain})} > \frac{P(y_1|x)}{P(y_1|x_{domain})}$$

$$\log P(y_2|x) \approx -16$$

$$\log P(y_2'|x) \approx -20$$

both probabilities low due to surface form competition!

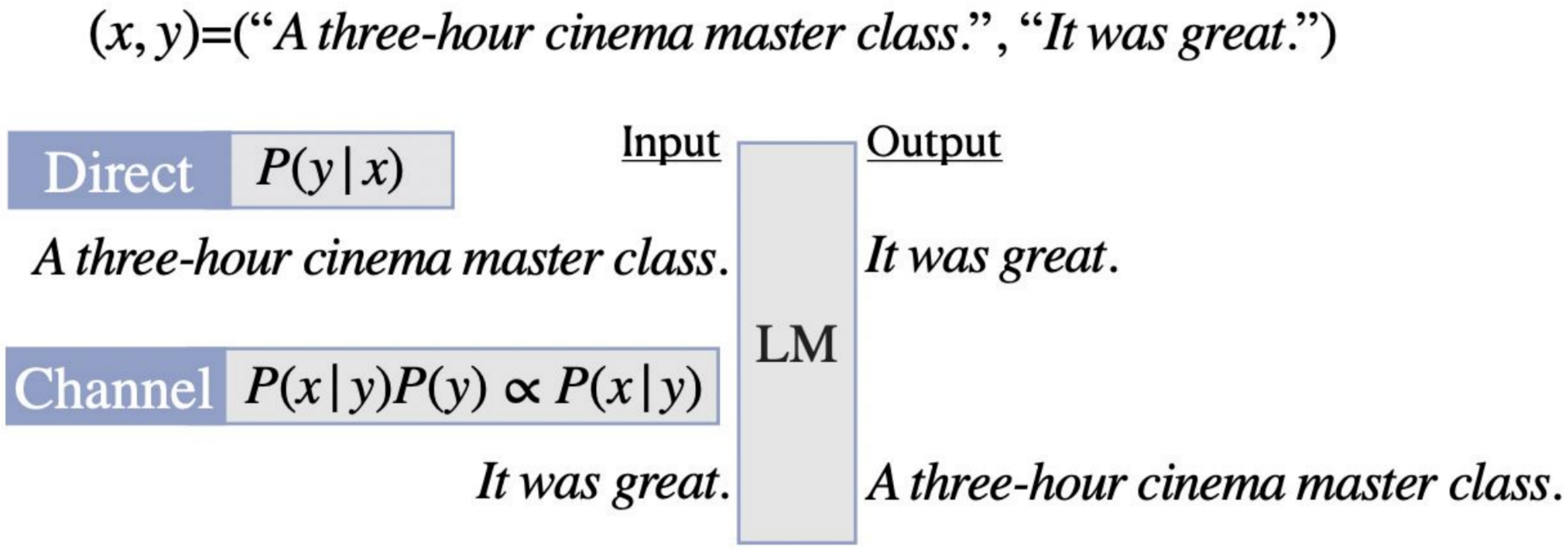
$$\log P(\hat{y}|\hat{x}_2) \approx -12$$

$$\log P(\hat{y}|\hat{x}_2') \approx -12$$

no competition → similarly high probabilities

Holtzman et al., Surface Form Competition: Why the Highest Probability Answer Isn't Always Right, EMNLP 2021

Noisy Channel (Min et al., 2022)



Note

another alternative to calibrate the probability of final output

So far …

$$s^{(i)} = \text{Template}(x^{(i)}, y^{(i)})$$

$$C = \text{Concat}\left(s^{(i)}, \ldots, s^{(k)}\right)$$

$$p(y|x, C)$$

Contextual Calibration (CC)

$$\underset{i}{\text{argmax}} \frac{P(y_i|x, C)}{p(y_i|[N/A], C)}$$

effective for single token outputs but not suited for multi-token generation.

Domain Conditional PMI (PMI_{DC})

$$\underset{i}{\text{arg}max} \frac{P(y_i|x, C)}{P(y_i|x_{domain}, C)}$$

removes surface form competition and generic output bias. However, domain specific string is subjective and difficult to choose the best one to use.

both papers focuses on novel ways to calculate the probabilities for language modeling

↓

improve performance with minimal changes

Mitigating Label Biases for In-context Learning

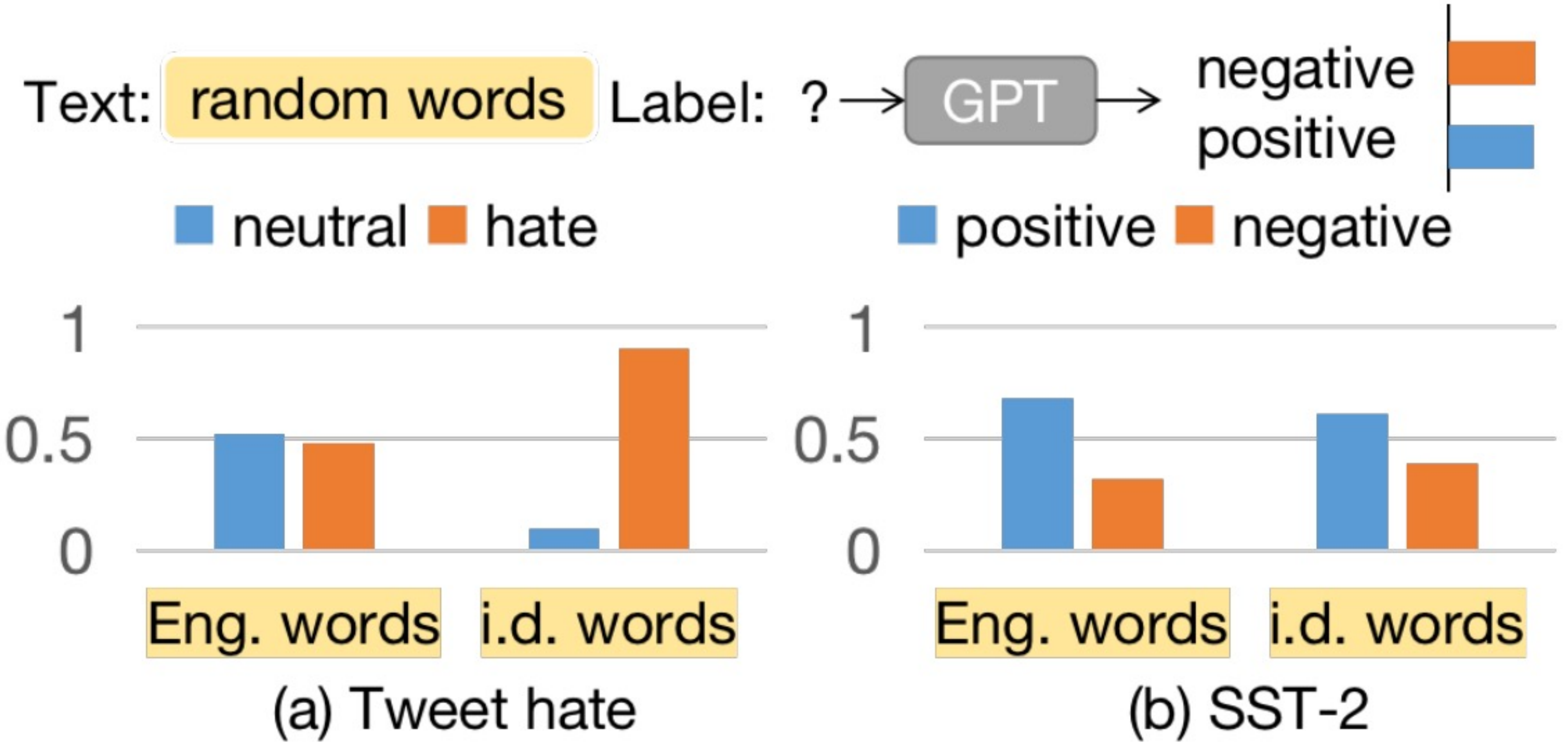
Yu Fei[†1], Yifan Hou[*2], Zeming Chen[*3], Antoine Bosselut[3]
[1]UC Irvine, [2]ETH Zurich, [3]NLP Lab, IC, EPFL, Switzerland
yu.fei@uci.edu, yifan.hou@inf.ethz.ch,
{zeming.chen, antoine.bosselut}@epfl.ch

Label Biases in ICL

- Vanilla-label bias
- Context-label bias
- Domain-label bias

Yu Fei et al., Mitigating label biases for in-context learning, ACL 2023

Domain label bias



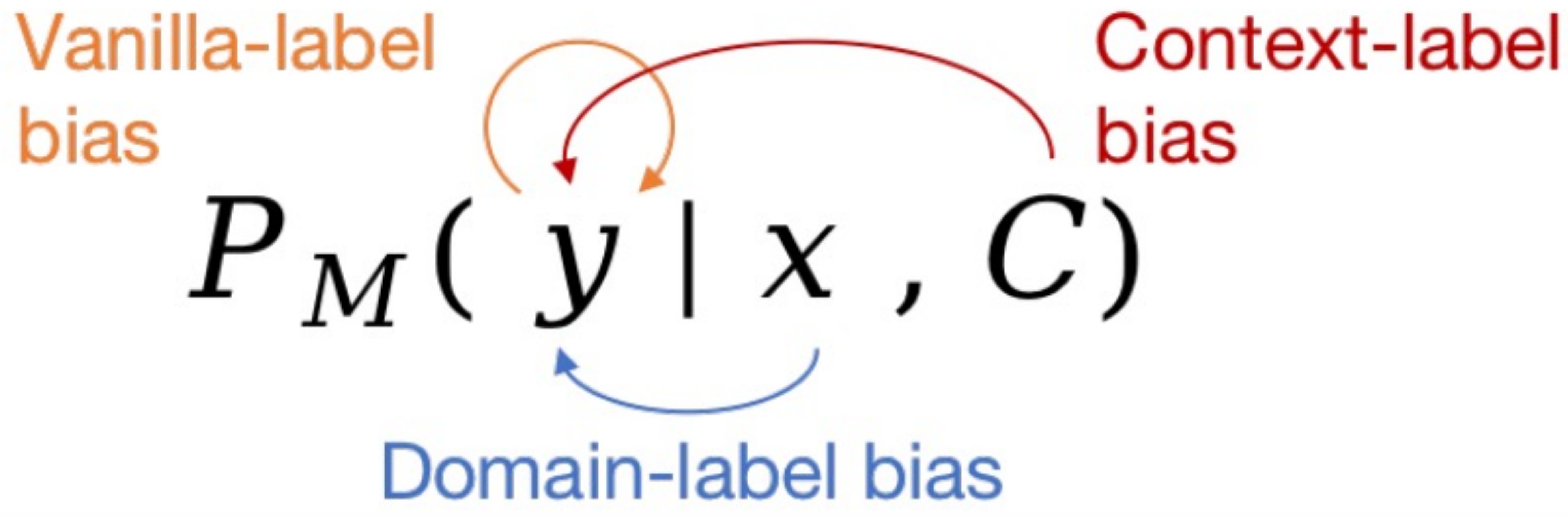
$$bias = \frac{1}{2}\sum_{y\in\mathcal{L}} \left| p(y|x_{Eng.}) - p(y|x_{i.d.}) \right|$$

Yu Fei et al., Mitigating label biases for in-context learning, ACL 2023

Domain-Context Calibration

$$\bar{p}(y|C) = \frac{1}{T}\sum_{t=1}^{T} p(y|[\text{Random i.d. text}]_t, C)$$

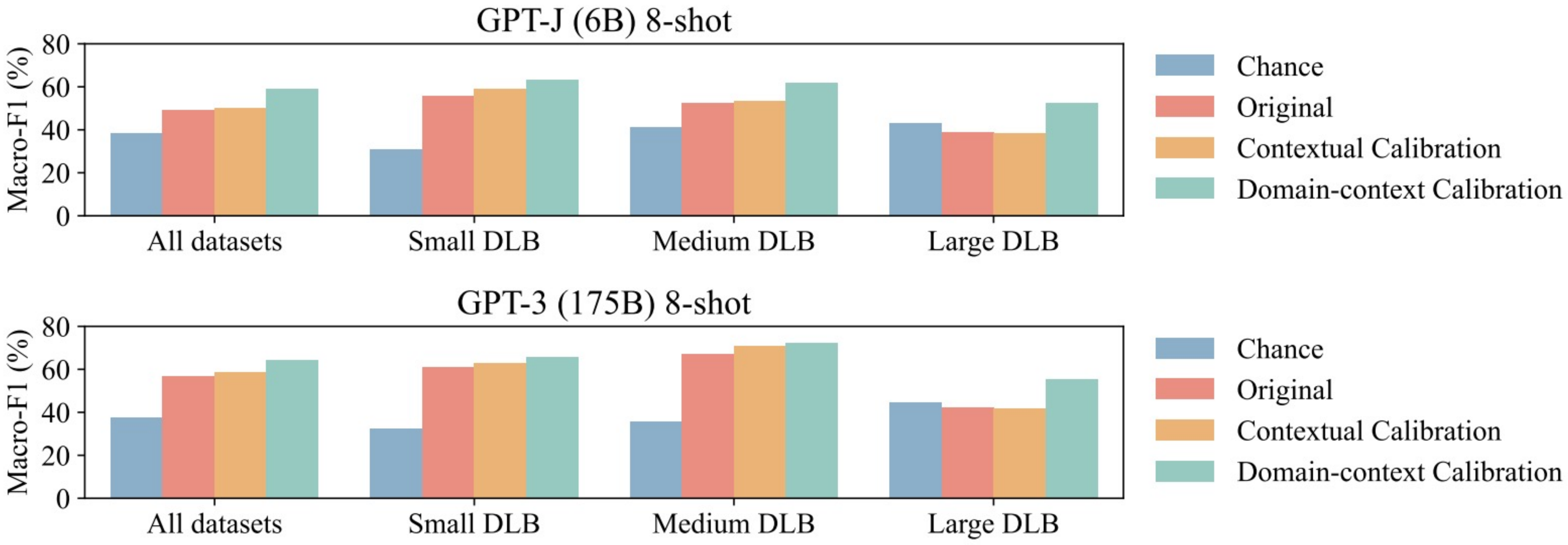
$$\hat{y}_i = \underset{y\in\mathcal{L}}{\operatorname{argmax}} \frac{p(y|x_i, C)}{\bar{p}(y|C)}$$



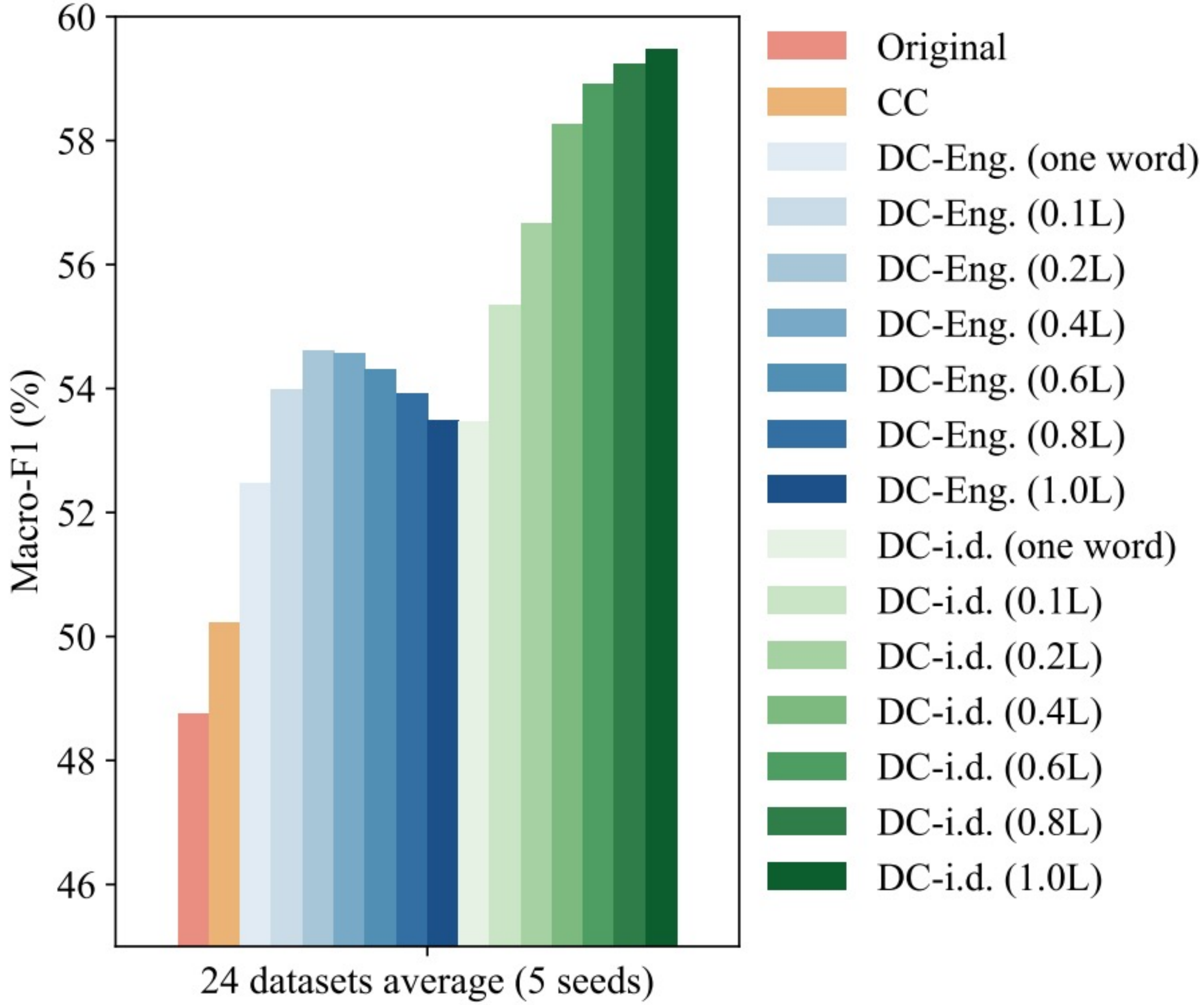
	Vanilla-lab.	Context-lab.	Domain-lab.
CC	✓	✓	✗
DC	✓	✓	✓

Yu Fei et al., Mitigating label biases for in-context learning, ACL 2023

Domain-Context Calibration



Yu Fei et al., Mitigating label biases for in-context learning, ACL 2023



Yu Fei et al., Mitigating label biases for in-context learning, ACL 2023

PROTOTYPICAL CALIBRATION FOR FEW-SHOT LEARNING OF LANGUAGE MODELS

Zhixiong Han,* Yaru Hao, Li Dong, Yutao Sun,* Furu Wei
Microsoft Research
{zhixhan8,sunyutao20001121}@gmail.com,
{yaruhao,lidong1,fuwei}@microsoft.com

Prototypical Calibration for Few-shot Learning

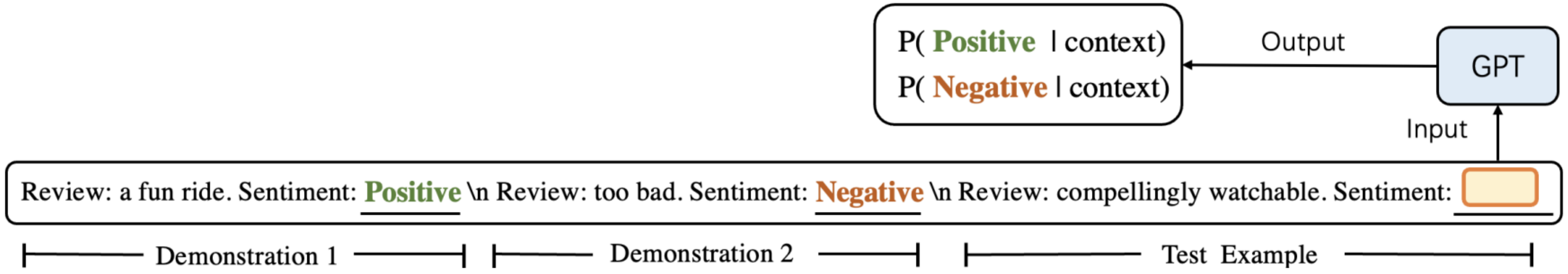
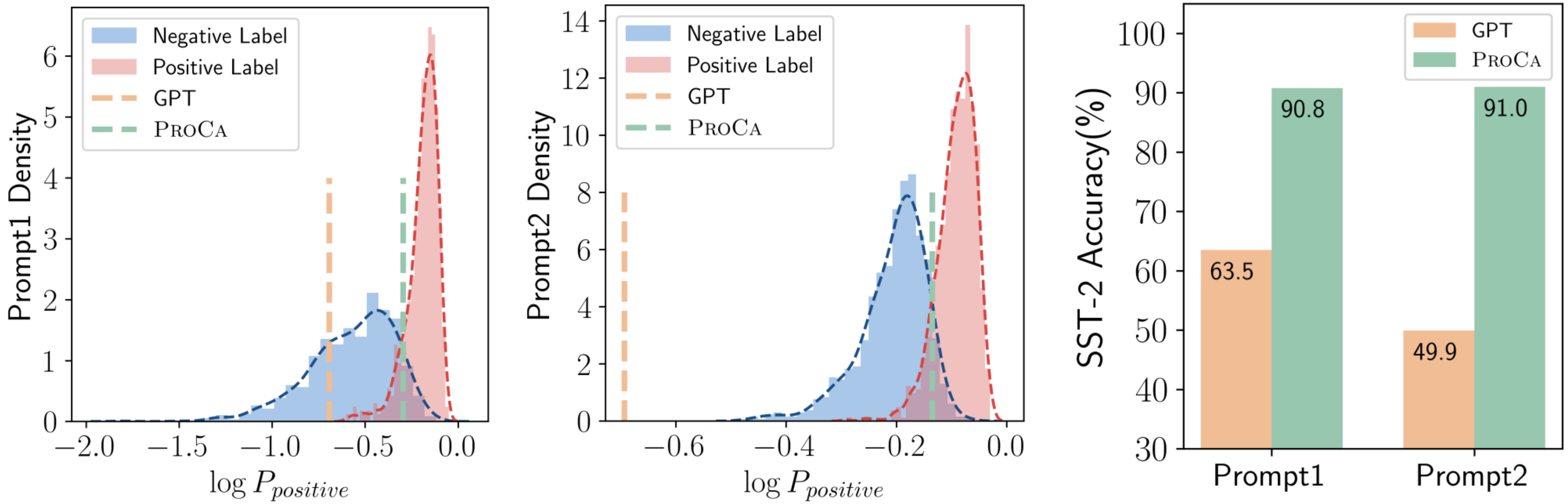
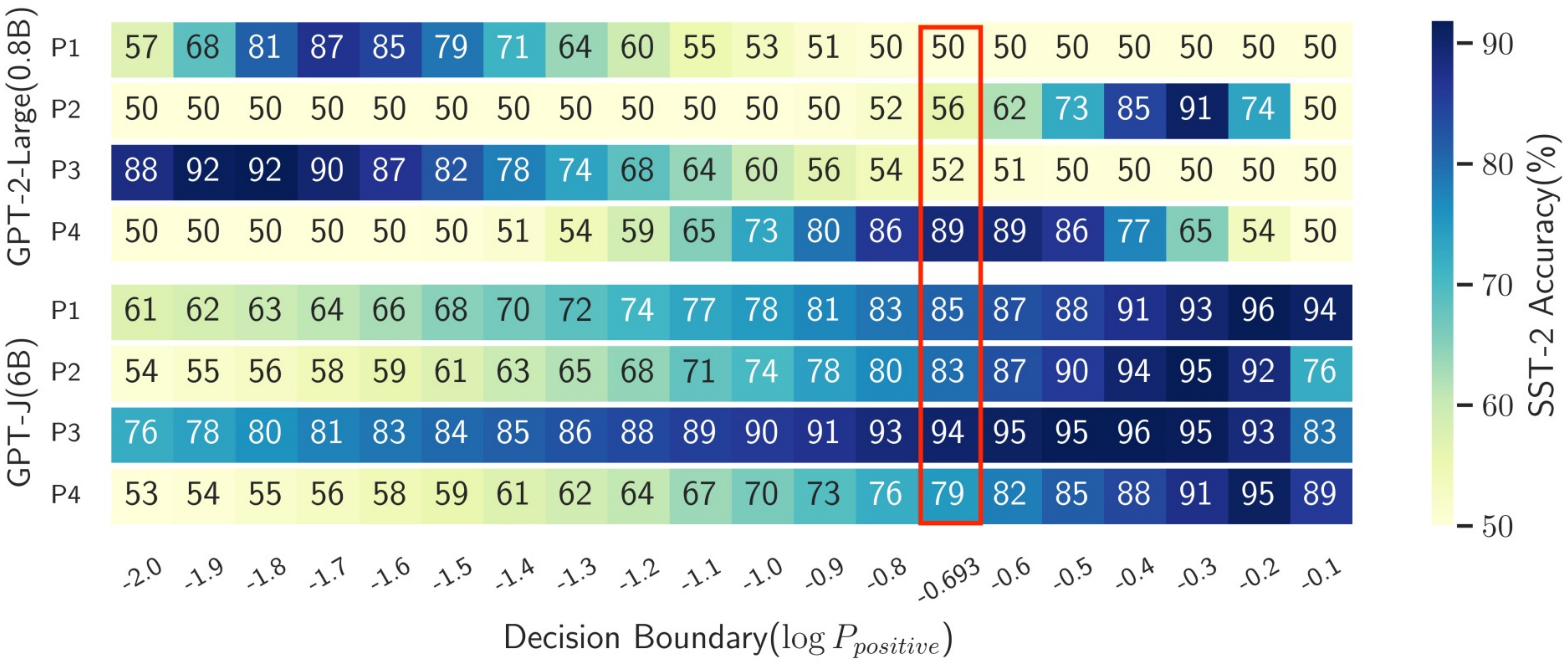


Figure 1: Example of few-shot learning with GPT.



Han et al., Prototypical Calibration for Few-shot Learning, ICLR 2023

Decision boundary greatly influences the few-shot performance



Han et al., Prototypical Calibration for Few-shot Learning, ICLR 2023

Prototypical Calibration for Few-shot Learning

- Performant decision boundaries are inconsistent across language models and prompts.

- PC adaptively learn a decision boundary for few-shot classification:
  - It estimates N prototypical clusters for the model output p for N classes

$$P_{\text{GMM}}(X) = \sum_{n=1}^{N} \alpha_n P_{\text{G}}(X|\boldsymbol{\mu_n}, \boldsymbol{\Sigma_n}),$$

  - Then, assign labels to clusters according to labels of few-shot examples

- Inference time:

$$\tilde{n} = \underset{n=1,\cdots,N}{\arg\max}\ P_{\text{G}}(x|\mu_n^*, \Sigma_n^*).$$

Han et al., Prototypical Calibration for Few-shot Learning, ICLR 2023

Shot	Method	SST-2	SST-5	MR	Subj	AP	AGNews	DBpedia	RTE	TREC	Avg
					GPT-2-XL 1.5B						
	GPT	58.7$_{0.0}$	28.4$_{0.0}$	58.9$_{0.0}$	57.6$_{0.0}$	**51.8**$_{0.0}$	41.6$_{0.0}$	60.3$_{0.0}$	50.0$_{0.0}$	28.6$_{0.0}$	48.4
0-shot	ConCa	69.3$_{0.0}$	22.6$_{0.0}$	66.9$_{0.0}$	72.9$_{0.0}$	49.8$_{0.0}$	**67.7**$_{0.0}$	54.3$_{0.0}$	**50.4**$_{0.0}$	**42.8**$_{0.0}$	55.2
	ProCa	**84.8**$_{0.2}$	**45.0**$_{1.3}$	**82.0**$_{0.2}$	**73.3**$_{0.1}$	49.8$_{0.3}$	64.6$_{1.4}$	**73.6**$_{3.0}$	49.2$_{0.7}$	42.0$_{2.7}$	**62.7**
	GPT	59.8$_{14.0}$	26.2$_{8.5}$	51.3$_{0.6}$	54.5$_{8.6}$	51.0$_{0.1}$	37.4$_{6.7}$	51.3$_{12.7}$	**53.8**$_{1.0}$	29.1$_{6.5}$	46.0
1-shot	ConCa	76.4$_{2.2}$	30.2$_{5.7}$	69.4$_{5.0}$	62.0$_{7.0}$	60.3$_{4.0}$	65.0$_{3.8}$	70.9$_{7.4}$	53.1$_{0.9}$	40.5$_{3.3}$	58.6
	ProCa	**89.4**$_{2.4}$	**42.5**$_{2.9}$	**84.3**$_{1.0}$	**71.8**$_{5.7}$	**69.8**$_{8.2}$	**69.8**$_{4.3}$	**79.9**$_{3.8}$	49.5$_{1.9}$	**43.6**$_{5.0}$	**66.7**
	GPT	66.3$_{13.7}$	31.3$_{7.4}$	56.5$_{5.9}$	53.4$_{4.9}$	50.9$_{0.1}$	40.9$_{13.0}$	61.3$_{7.6}$	52.0$_{3.5}$	23.8$_{5.7}$	48.5
4-shot	ConCa	79.9$_{10.2}$	33.5$_{3.5}$	67.7$_{8.9}$	68.0$_{8.7}$	75.6$_{5.9}$	59.9$_{6.3}$	74.9$_{5.0}$	**52.9**$_{0.7}$	41.1$_{4.3}$	61.5
	ProCa	**90.4**$_{0.6}$	**39.6**$_{4.5}$	**78.1**$_{11.8}$	**74.8**$_{10.2}$	**80.1**$_{7.1}$	**67.4**$_{13.5}$	**87.2**$_{4.9}$	52.2$_{1.5}$	**46.0**$_{2.5}$	**68.4**
	GPT	57.0$_{9.0}$	30.5$_{7.9}$	65.2$_{12.7}$	57.9$_{11.2}$	50.9$_{0.0}$	42.9$_{4.2}$	67.9$_{7.1}$	53.0$_{2.1}$	37.2$_{4.9}$	51.4
8-shot	ConCa	73.9$_{11.6}$	28.7$_{3.4}$	74.1$_{8.4}$	68.3$_{8.3}$	71.1$_{7.4}$	55.9$_{14.0}$	75.0$_{4.2}$	**53.1**$_{0.2}$	45.8$_{1.7}$	60.7
	ProCa	**88.0**$_{1.3}$	**36.5**$_{4.4}$	**80.8**$_{6.4}$	**80.2**$_{3.3}$	**79.3**$_{7.8}$	**75.5**$_{3.2}$	**89.4**$_{0.7}$	51.3$_{2.0}$	**46.0**$_{2.5}$	**69.7**
					GPT-J 6B						
	GPT	66.6$_{0.0}$	26.6$_{0.0}$	65.9$_{0.0}$	67.9$_{0.0}$	54.2$_{0.0}$	33.7$_{0.0}$	21.8$_{0.0}$	55.2$_{0.0}$	23.4$_{0.0}$	46.1
0-shot	ConCa	57.7$_{0.0}$	35.4$_{0.0}$	57.1$_{0.0}$	59.9$_{0.0}$	63.1$_{0.0}$	**60.1**$_{0.0}$	49.9$_{0.0}$	55.6$_{0.0}$	42.2$_{0.0}$	53.4
	ProCa	**74.2**$_{0.2}$	**42.1**$_{0.8}$	**73.1**$_{0.4}$	**69.5**$_{0.2}$	**63.3**$_{0.2}$	55.1$_{0.4}$	**66.1**$_{1.5}$	**57.0**$_{1.0}$	**53.4**$_{6.1}$	**61.5**
	GPT	67.7$_{7.3}$	31.7$_{4.9}$	68.1$_{4.1}$	65.0$_{10.9}$	92.9$_{2.7}$	65.6$_{14.6}$	65.6$_{14.8}$	52.6$_{4.6}$	41.8$_{9.0}$	61.2
1-shot	ConCa	89.3$_{2.2}$	46.5$_{3.4}$	**88.5**$_{1.1}$	58.8$_{3.0}$	93.5$_{1.3}$	75.5$_{5.7}$	79.9$_{3.3}$	53.1$_{0.8}$	**64.7**$_{5.3}$	72.2
	ProCa	**90.8**$_{1.7}$	**47.6**$_{2.5}$	87.9$_{1.5}$	**77.9**$_{4.8}$	**95.1**$_{0.5}$	**79.8**$_{5.4}$	**90.0**$_{2.2}$	**56.7**$_{3.1}$	55.3$_{6.4}$	**75.7**
	GPT	88.6$_{4.3}$	44.7$_{3.3}$	84.4$_{8.2}$	58.2$_{6.3}$	89.4$_{10.0}$	72.1$_{6.5}$	80.5$_{13.2}$	55.6$_{6.7}$	38.1$_{5.4}$	68.0
4-shot	ConCa	92.9$_{3.7}$	**47.7**$_{4.4}$	87.8$_{1.8}$	66.5$_{11.7}$	93.4$_{1.0}$	76.4$_{4.0}$	88.6$_{3.0}$	54.7$_{1.5}$	48.5$_{4.9}$	72.9
	ProCa	**95.0**$_{0.4}$	46.2$_{4.6}$	**89.4**$_{1.9}$	**79.4**$_{5.8}$	**95.8**$_{0.8}$	**79.9**$_{6.6}$	**91.9**$_{2.6}$	**61.2**$_{2.7}$	**57.1**$_{5.3}$	**77.3**
	GPT	91.1$_{6.2}$	44.9$_{2.9}$	89.5$_{2.3}$	82.1$_{3.9}$	95.2$_{1.7}$	76.9$_{9.7}$	87.7$_{3.1}$	61.0$_{3.9}$	44.4$_{5.6}$	74.8
8-shot	ConCa	93.4$_{1.8}$	46.6$_{4.4}$	90.1$_{0.5}$	80.5$_{5.8}$	**96.2**$_{0.3}$	79.9$_{6.4}$	90.8$_{2.0}$	59.6$_{4.8}$	53.5$_{7.9}$	76.7
	ProCa	**94.4**$_{1.0}$	**47.4**$_{4.4}$	**90.7**$_{0.7}$	**83.6**$_{4.2}$	96.1$_{0.5}$	**84.2**$_{1.8}$	**95.1**$_{0.5}$	**61.7**$_{7.2}$	**61.0**$_{7.6}$	**79.4**
					Bloom 176B						
	Bloom	73.4$_{0.0}$	26.0$_{0.0}$	71.0$_{0.0}$	53.3$_{0.0}$	60.1$_{0.0}$	27.1$_{0.0}$	48.5$_{0.0}$	62.5$_{0.0}$	**59.0**$_{0.0}$	53.4
0-shot	ConCa	73.9$_{0.0}$	25.3$_{0.0}$	71.8$_{0.0}$	49.0$_{0.0}$	51.1$_{0.0}$	38.2$_{0.0}$	61.0$_{0.0}$	53.8$_{0.0}$	41.0$_{0.0}$	51.7
	ProCa	**76.4**$_{0.1}$	**31.8**$_{0.2}$	**73.4**$_{0.4}$	**61.3**$_{0.3}$	**80.4**$_{0.8}$	**60.1**$_{3.5}$	**75.8**$_{0.1}$	**62.6**$_{0.2}$	52.9$_{0.5}$	**63.9**
	Bloom	91.7$_{2.6}$	31.1$_{7.5}$	84.6$_{2.3}$	60.4$_{8.5}$	**96.1**$_{0.1}$	67.6$_{0.9}$	81.8$_{2.0}$	61.2$_{3.4}$	55.1$_{7.1}$	70.0
1-shot	ConCa	91.8$_{1.6}$	38.9$_{4.3}$	86.8$_{1.6}$	51.2$_{2.5}$	**96.1**$_{0.4}$	78.4$_{0.5}$	80.4$_{1.9}$	54.0$_{5.6}$	**69.3**$_{1.3}$	71.9
	ProCa	**93.6**$_{0.6}$	**47.5**$_{2.8}$	**88.0**$_{0.8}$	**72.0**$_{1.8}$	95.7$_{0.4}$	**81.6**$_{0.7}$	**83.7**$_{1.8}$	**65.7**$_{0.4}$	67.5$_{2.5}$	**77.3**
	Bloom	**96.3**$_{0.1}$	46.7$_{0.8}$	87.3$_{5.3}$	72.2$_{6.4}$	94.2$_{2.5}$	68.8$_{3.2}$	86.2$_{1.4}$	64.1$_{2.4}$	29.1$_{0.9}$	71.7
4-shot	ConCa	96.0$_{0.1}$	46.9$_{2.9}$	89.7$_{1.1}$	70.4$_{7.7}$	94.2$_{1.9}$	78.0$_{0.1}$	86.6$_{2.4}$	56.3$_{0.7}$	**64.8**$_{7.6}$	75.9
	ProCa	95.7$_{0.2}$	**50.2**$_{2.6}$	**91.2**$_{0.1}$	**78.5**$_{0.5}$	**95.8**$_{0.5}$	**82.7**$_{1.2}$	**87.0**$_{1.3}$	**68.6**$_{0.4}$	56.8$_{4.8}$	**78.5**
	Bloom	94.6$_{2.0}$	43.2$_{3.5}$	90.9$_{0.8}$	78.6$_{2.2}$	**96.0**$_{0.9}$	75.4$_{1.9}$	88.4$_{2.1}$	65.9$_{2.4}$	48.9$_{6.7}$	75.8
8-shot	ConCa	**96.1**$_{0.2}$	42.2$_{5.5}$	91.0$_{0.9}$	75.8$_{1.7}$	95.9$_{0.4}$	81.9$_{2.0}$	**89.5**$_{2.6}$	59.0$_{0.5}$	**73.9**$_{1.1}$	78.4
	ProCa	95.3$_{1.3}$	**53.1**$_{1.6}$	**92.0**$_{0.6}$	**80.6**$_{1.9}$	95.6$_{0.8}$	**82.1**$_{2.0}$	85.1$_{3.7}$	**69.5**$_{2.7}$	68.6$_{7.8}$	**80.2**

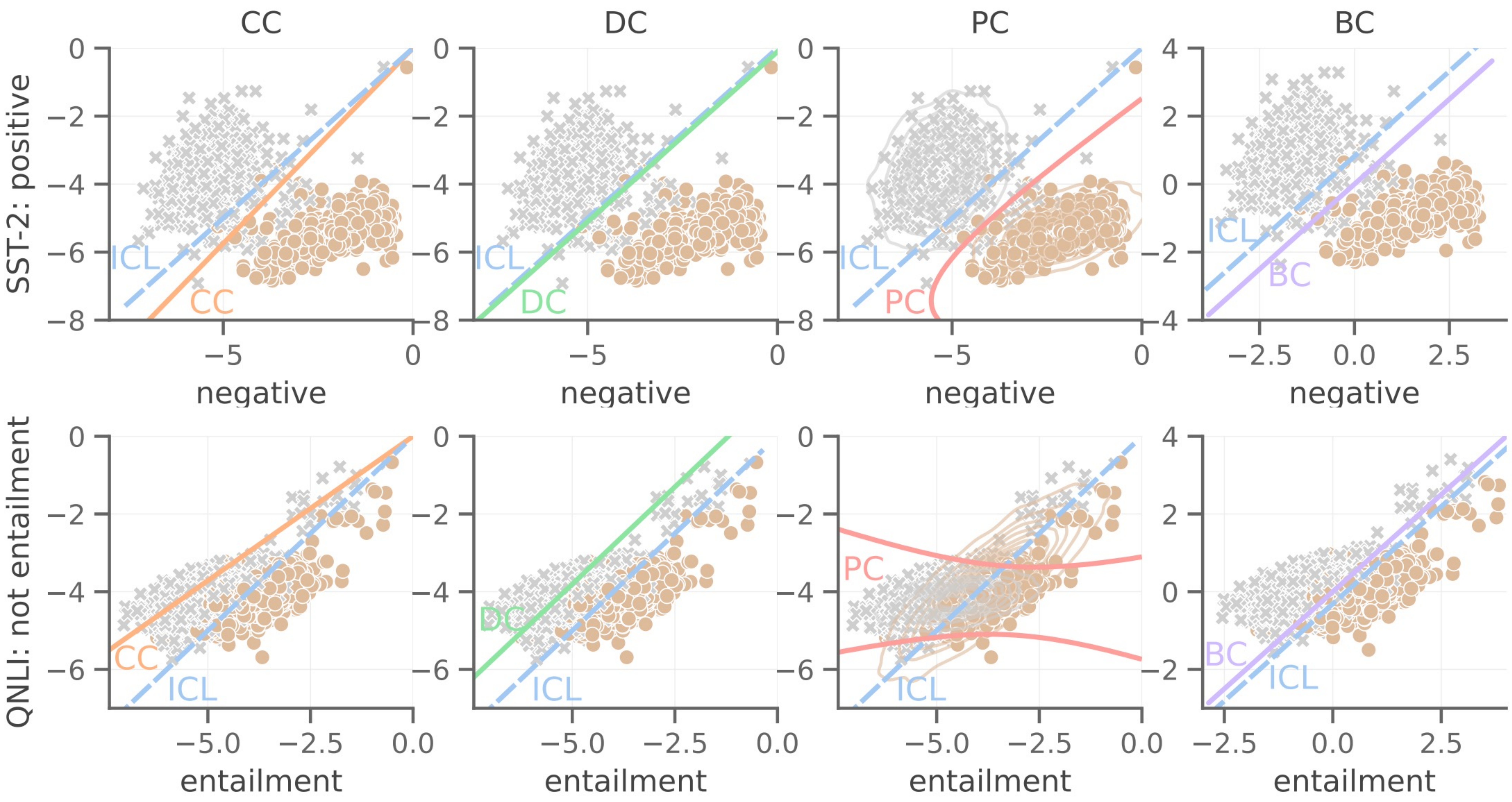
Han et al., Prototypical Calibration for Few-shot Learning, ICLR 2023

Batch Calibration: Rethinking Calibration for In-Context Learning and Prompt Engineering

Han Zhou[1,2,*] **Xingchen Wan**[1] **Lev Proleev**[1] **Diana Mincu**[1] **Jilin Chen**[1]

Katherine Heller[1] **Subhrajit Roy**[1]

[1] Google Research [2] University of Cambridge



Zhou et al., Batch calibration: Rethinking calibration for in-context learning and prompt engineering

Questions

- What is the disadvantage of non-linear decision boundaries?
  - non-linear decision boundaries learned by PC tend to be susceptible to overfitting and may suffer from instability in EM-GMM

- Is content-free input a good estimator of the contextual prior?
  - relying on content-free tokens for calibration is not always optimal and may even introduce additional bias, depending on the task type.

Zhou et al., Batch calibration: Rethinking calibration for in-context learning and prompt engineering

Batch Calibration

- Batch Calibration (BC), a zero-shot and self-adaptive (inference-only) calibration
  - only involves unlabeled test samples
- BC accurately models the bias from the prompt context (i.e. contextual bias) by marginalizing the LLM scores in the batched input.
- extends BC to the black-box few-shot learning (BCL)
  - introducing a single learnable parameter into BC, which enables it to adapt and learn the contextual bias from the available data resources.

Zhou et al., Batch calibration: Rethinking calibration for in-context learning and prompt engineering

Batch Calibration

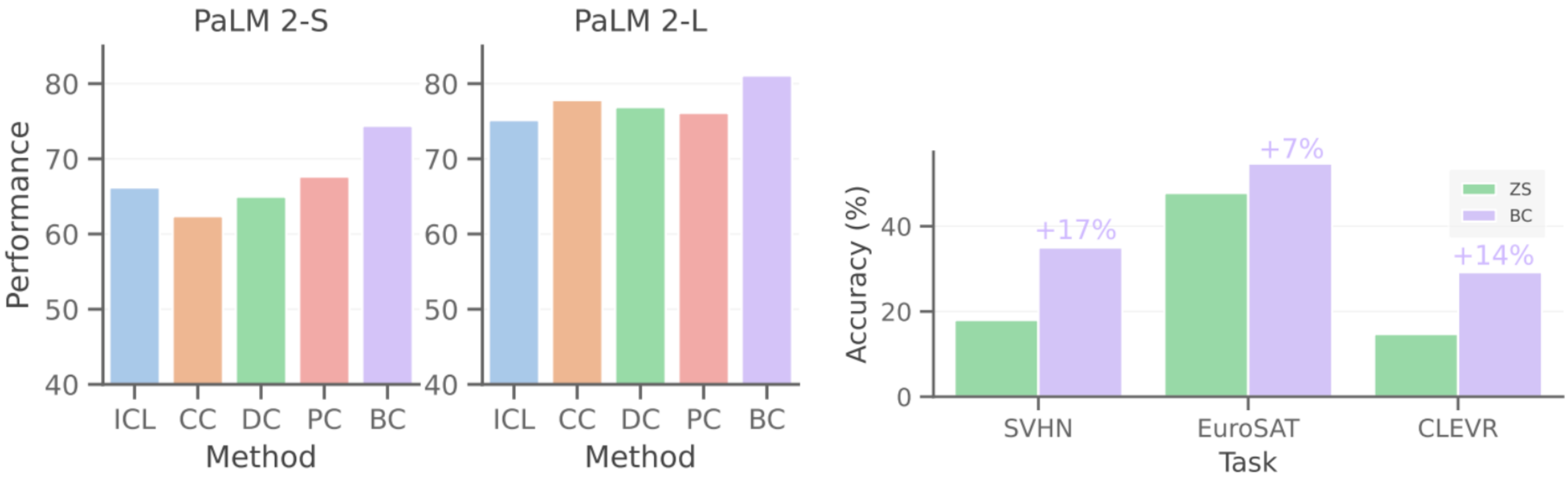
- Uses linear decision boundary for its robustness

- Instead of relying on content-free tokens, estimates the contextual bias for each class from a batch with M samples:

$$\mathbf{p}(y = y_j | C) = \mathop{\mathbb{E}}_{x \sim P(x)} \left[\mathbf{p}(y = y_j | x, C) \right] \approx \frac{1}{M} \sum_{i=1}^{M} \mathbf{p}(y = y_j | x^{(i)}, C) \, \forall \, y_j \in \mathcal{Y}.$$

Zhou et al., Batch calibration: Rethinking calibration for in-context learning and prompt engineering

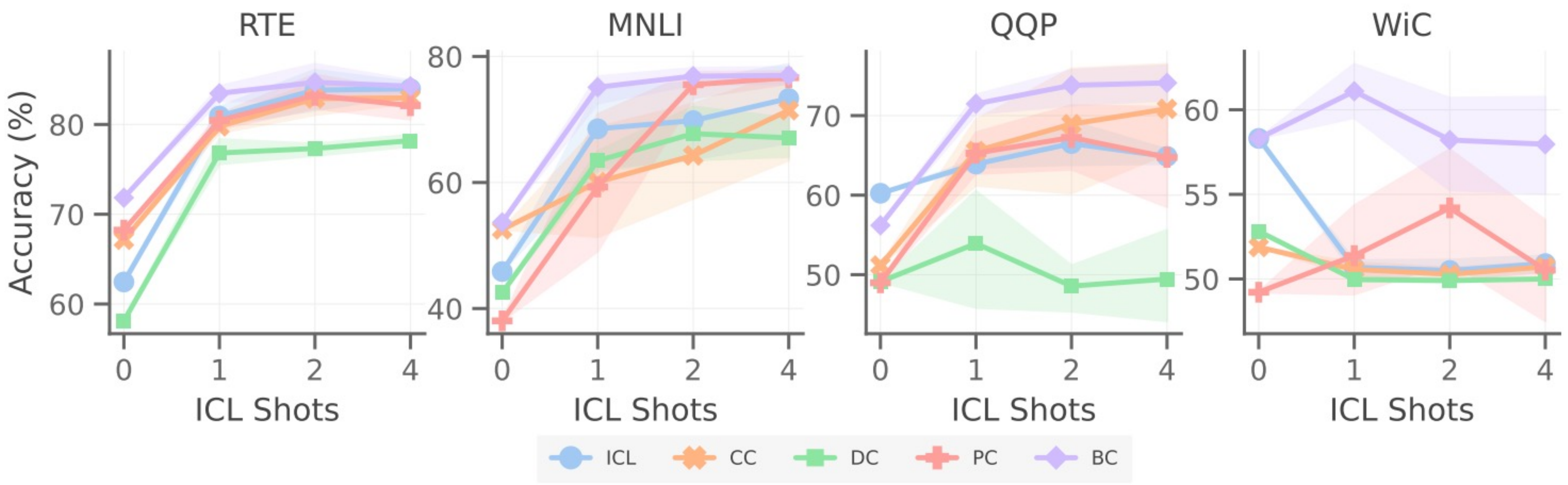
Results



Batch Calibration (BC) achieves the best performance on 1-shot ICL over CC, DC, and PC on an average of 13 NLP tasks on PaLM 2 and outperforms the zero-shot CLIP on image tasks.

Zhou et al., Batch calibration: Rethinking calibration for in-context learning and prompt engineering

Results on PaLM 2-S



Zhou et al., Batch calibration: Rethinking calibration for in-context learning and prompt engineering

Unified framework

Method	Token	#Forward	Comp. Cost	Cali. Form	Learning Term	Decision Boundary $h(\mathbf{p})$	Multi-Sentence	Multi-Class
CC	N/A	$1+1$	Inverse	$\mathbf{Wp}+\mathbf{b}$	$\mathbf{W}=\text{diag}(\hat{\mathbf{p}})^{-1}, \mathbf{b}=\mathbf{0}$	$p_0=\alpha p_1$	✗	✓
DC	Random	$20+1$	Add	$\mathbf{Wp}+\mathbf{b}$	$\mathbf{W}=\mathbf{I}, \mathbf{b}=-\frac{1}{T}\sum_t \mathbf{p}(y\|\text{text}_j, C)$	$p_0=p_1+\alpha$	✗	✓
PC	-	1	EM-GMM	-	$\sum_j \alpha_j P_G(\mathbf{p}\|\boldsymbol{\mu}_\mathbf{j}, \boldsymbol{\Sigma}_\mathbf{j})$	$P_\text{G}(\mathbf{p}\|\mu_0,\Sigma_0)=P_\text{G}(\mathbf{p}\|\mu_1,\Sigma_1)$	✓	✗
BC (Ours)	-	1	Add	$\mathbf{Wp}+\mathbf{b}$	$\mathbf{W}=\mathbf{I}, \mathbf{b}=-\mathbb{E}_x\left[\mathbf{p}(y\|x,C)\right]$	$p_0=p_1+\alpha$	✓	✓

- CC: $\hat{p} = p(y|[N/A], C)$
- DC: $\hat{p}(y|C) = \frac{1}{T}\sum_{t=1}^{T} p(y|[\text{RANDOM TEXT}]_t, C)$
- PC: $\tilde{n} = \arg\max_{n=1,\cdots,N} P_\text{G}(x|\mu_n^*, \Sigma_n^*).$
- BC: $\hat{p}(y|C) = \mathbb{E}_x[p(y|x,C)] \approx \frac{1}{M}\sum_{i=1}^{M} p(y|x^{(i)}, C)$

Zhou et al., Batch calibration: Rethinking calibration for in-context learning and prompt engineering

Conclusion

- Contextual Calibration (CC): calibrates the LLM given content-free tokens ("N/A")
- PMI-DC: calibrates the LLM given domain tokens (e.g., "?", "because")
- Domain-context Calibration (DC): calibrates the LLM given random i.d. tokens
- Prototypical Calibration (PC): learning a robust non-linear decision boundary using unlabeled samples
- Batch Calibration (BC): estimates the contextual bias for each class from a batch of unlabeled samples

Questions