Calibration of prompting LLMs

Presented by: Howard Yen, Vishvak Murahari

10/3/2022

Agenda

1. Introduction
2. Calibrate Before Using
3. Surface Form Competition
4. Conclusion

In-Context Learning

Zero-shot

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

```
Translate English to French:     ← task description
cheese => .....................  ← prompt
```

One-shot

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

```
Translate English to French:     ← task description
sea otter => loutre de mer       ← example
cheese => .....................  ← prompt
```

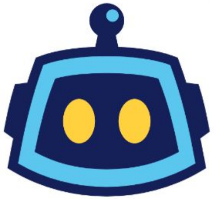
Few-shot

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

```
Translate English to French:     ← task description
sea otter => loutre de mer       ← examples
peppermint => menthe poivrée     ←
plush girafe => girafe peluche   ←
cheese => .....................  ← prompt
```

Brown et al., 2020

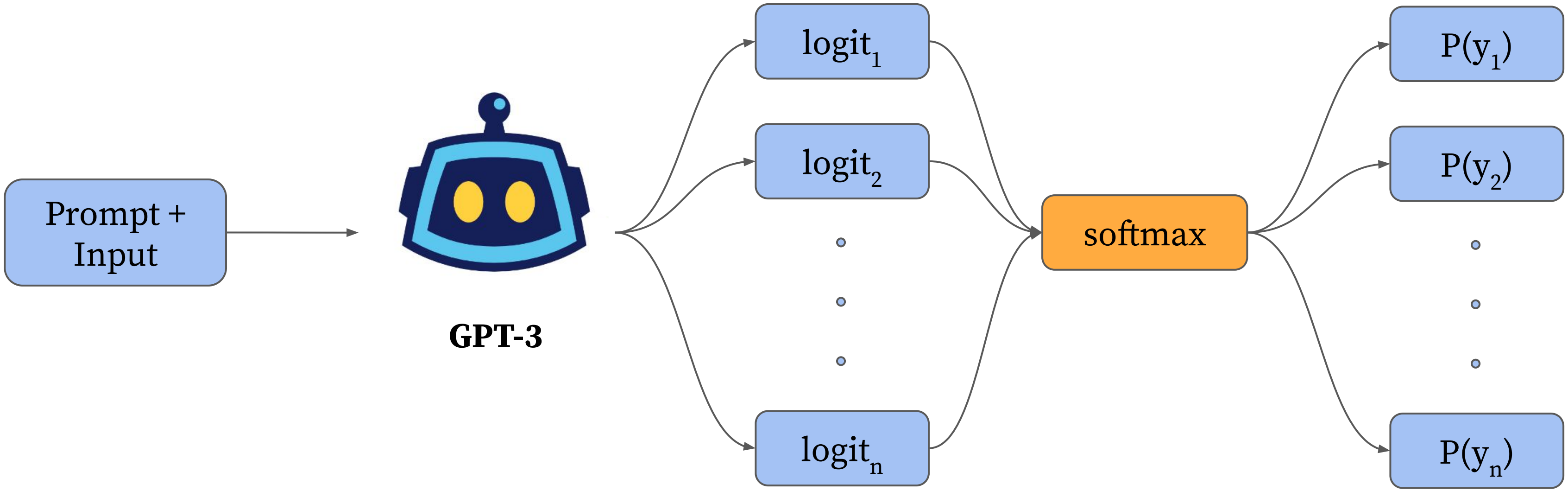
→



GPT-3

→ fromage

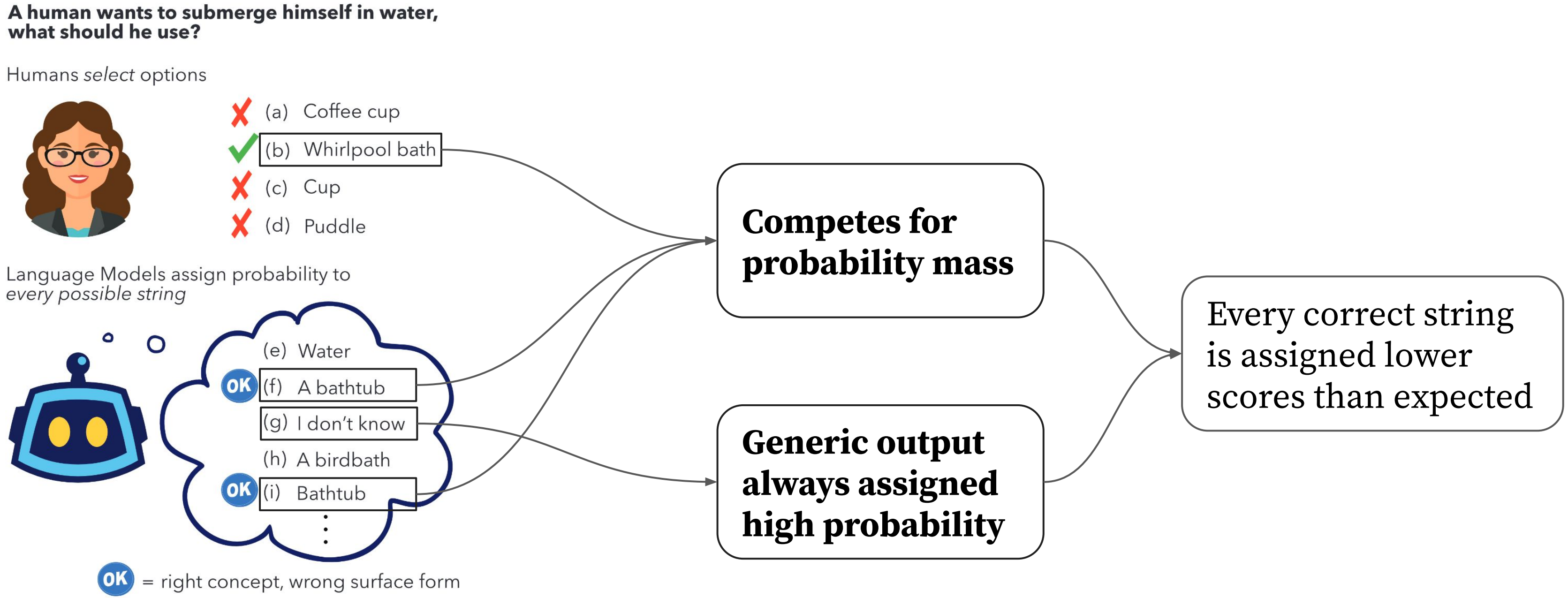
Language Modeling



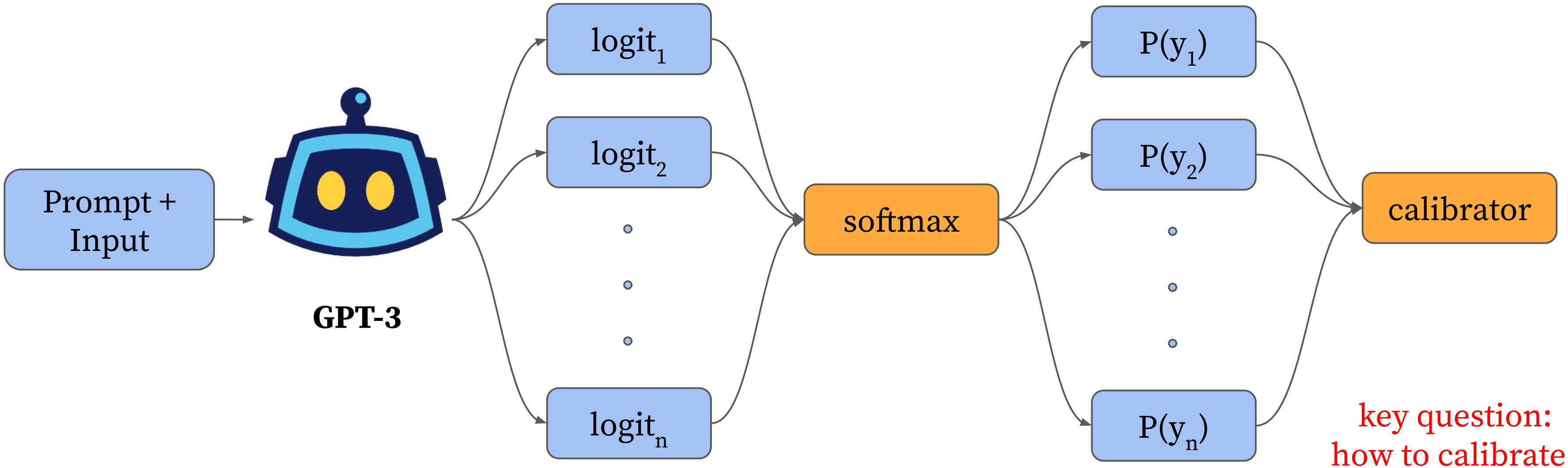
n = number of labels for close set classification tasks
n = number of words in the vocabulary for open set tasks

what are some possible flaws?

Surface Form Competition



Calibration



key question:
how to calibrate?

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

Motivation

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:
```

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:

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:

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

Components of a prompt

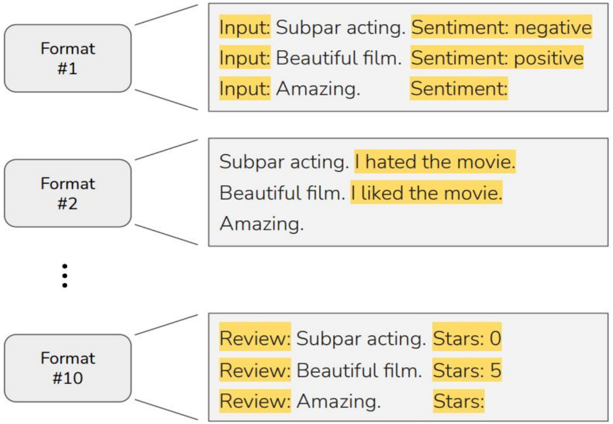
1. Prompt format
2. Training example selection
3. Training example permutation

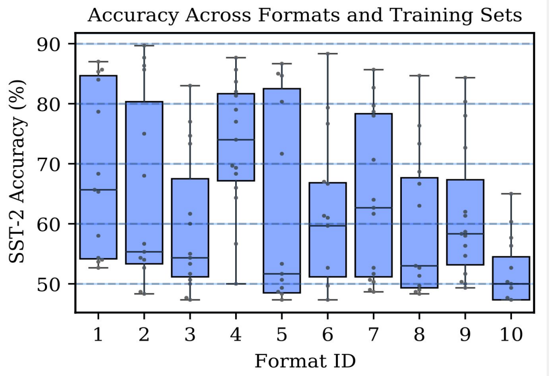
Let's try to ablate each component

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



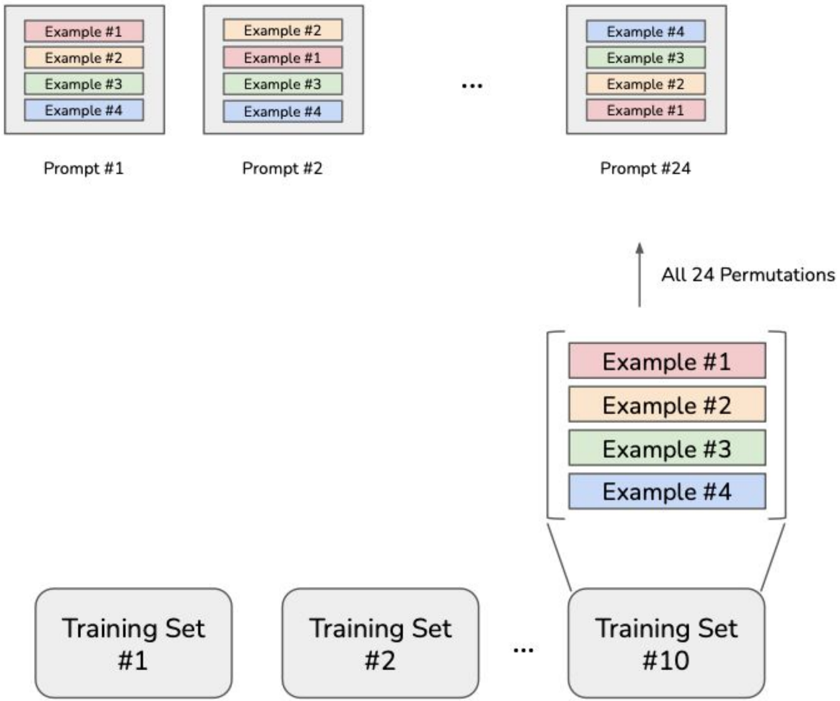


In-context learning is highly sensitive to prompt format

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

Components of a prompt

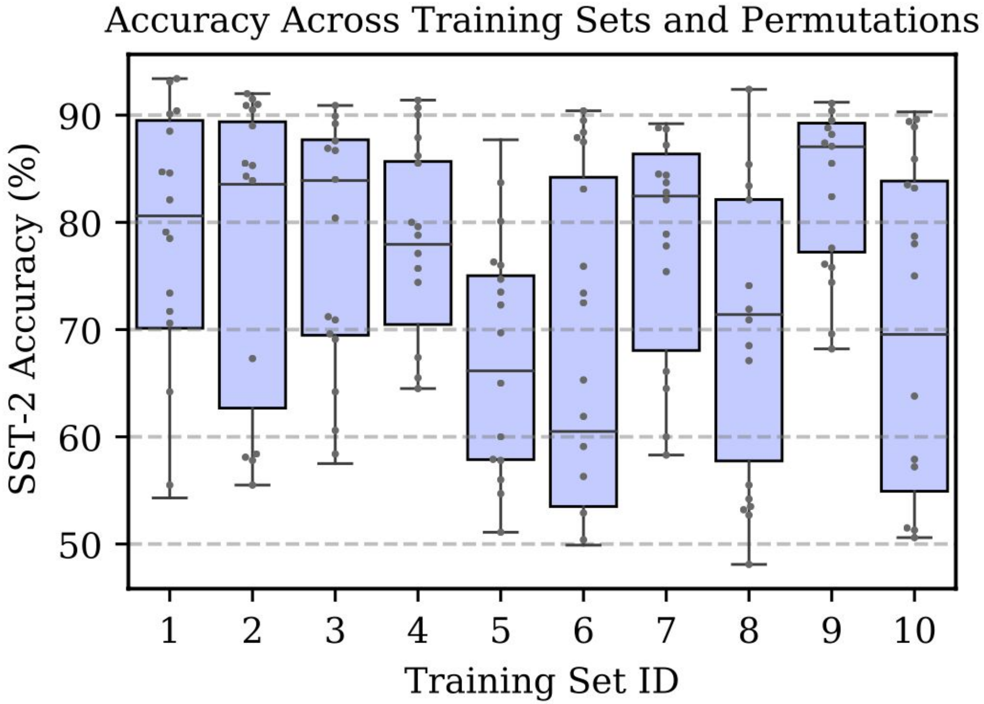
1. Prompt format
2. **Training example selection**
3. Training example permutation



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

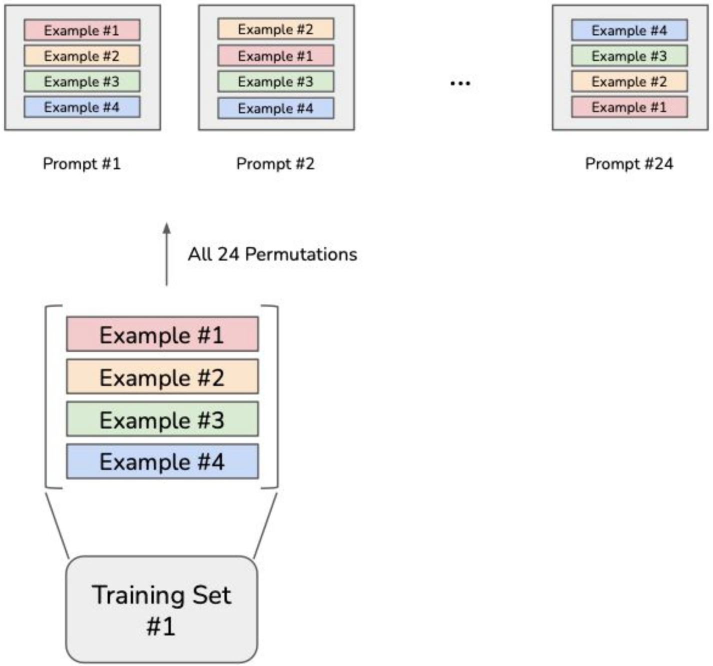


In-context learning is highly sensitive to example selection

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

Components of a prompt

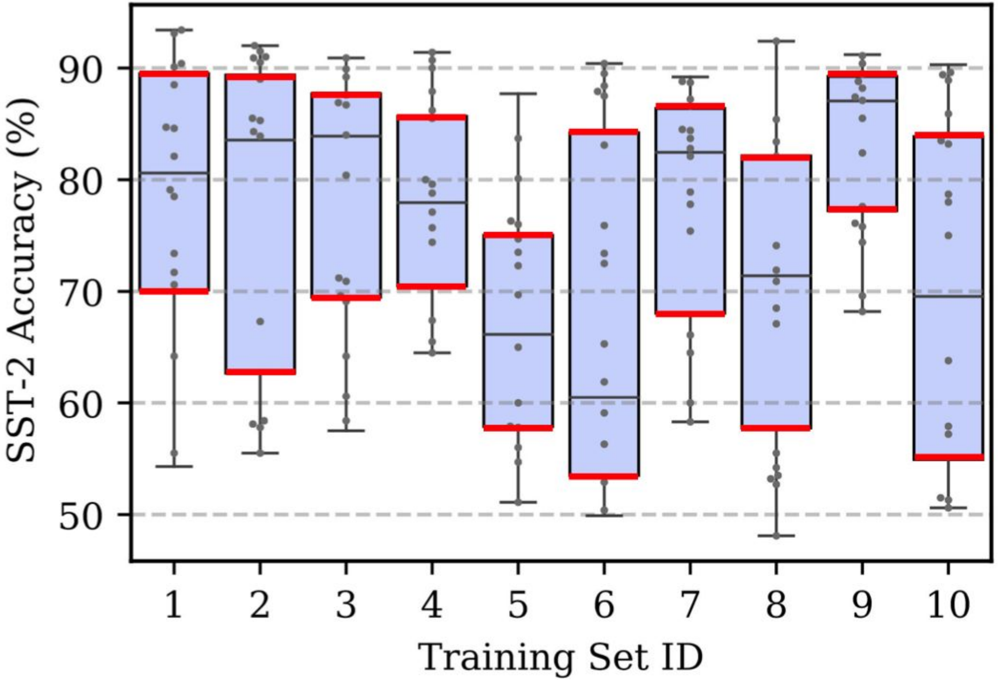
1. Prompt format
2. Training example selection
3. **Training example permutation**



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**



In-context learning is highly sensitive to example permutation

What causes this sensitivity?

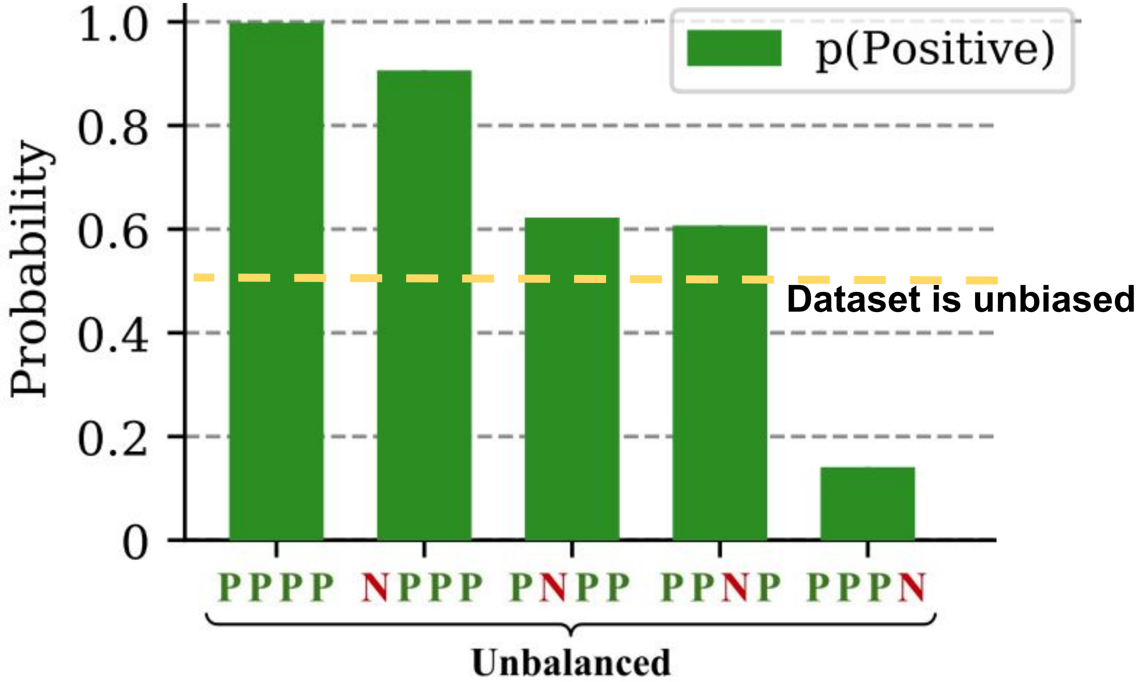
Three main reasons

1. Majority label bias
2. Common token bias
3. Recency bias

What causes this sensitivity?

Three main reasons

1. **Majority label bias**
2. Common token bias
3. Recency bias

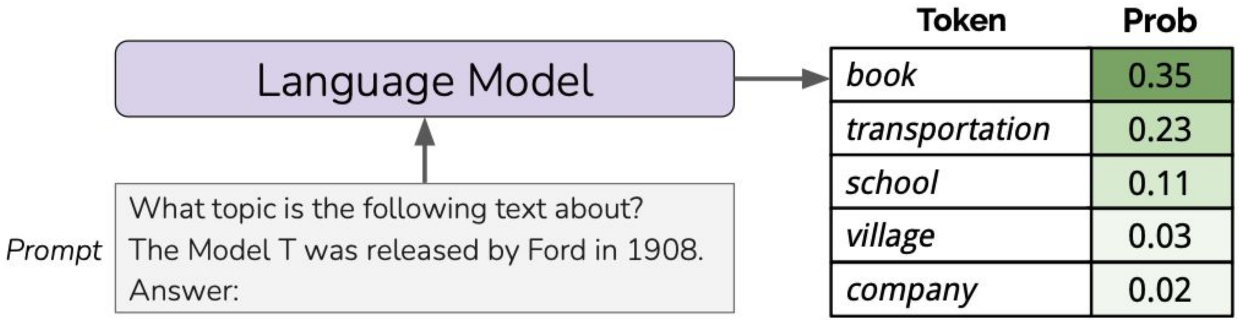


1. **Model prefers to predict positive when the majority labels is "P/Positive"**
2. **Surprising because the validation dataset is balanced!**

What causes this sensitivity?

Three main reasons

1. Majority label bias
2. **Common token bias**
3. Recency bias



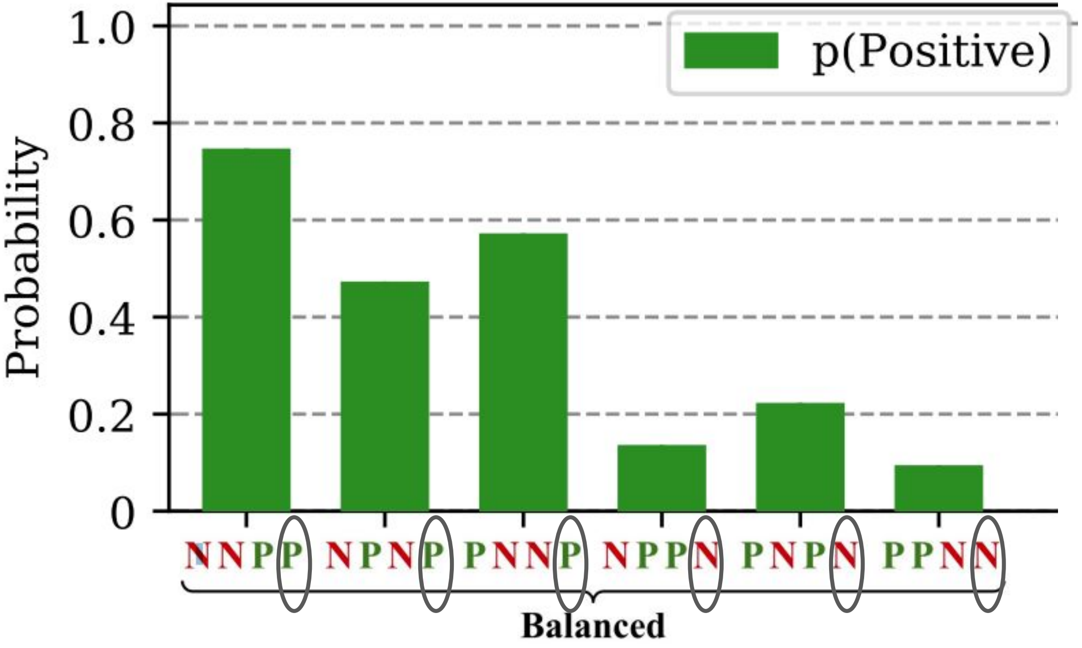
	Token	Web (%)	Label (%)	Prediction (%)
❌	*book*	0.026	9	29
✅	*transportation*	0.0000006	9	4

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

What causes this sensitivity?

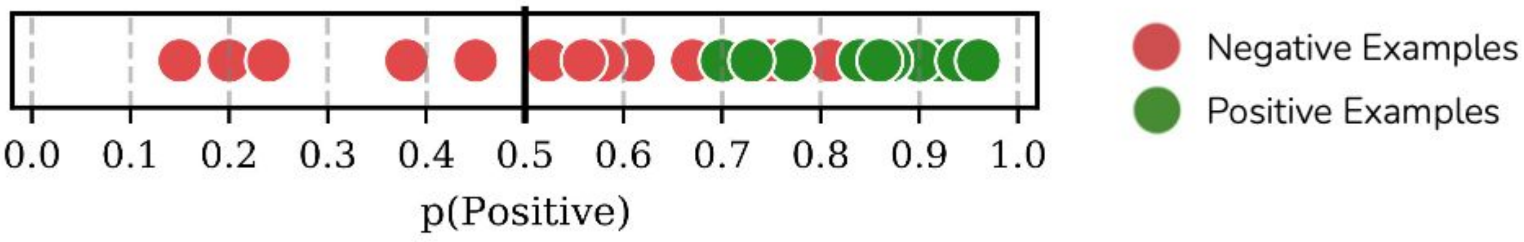
Three main reasons

1. Majority label bias
2. Common token bias
3. **Recency bias**



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

What is the impact of all these factors?



Visualizing predictions of 25 randomly sampled instances from SST2

<u>All the biases effectively shift the output distribution</u>

Pre-Lecture Question 1

Zhao et al., 2021 argue that the sensitivity of in-context learning results may be attributed to several biases in the prompts. What are the biases?

1. **Recency Bias:** Model is more likely to predict a label that occurs the most recently (towards the end of the prompt).
2. **Majority label bias:** Model predicts label that occurs more in the prompt.
3. **Common token bias:** Model tends to predict the token that is occurs more in the pretraining distribution.

Methodology

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:

Get model's prediction

positive	0.65
negative	0.35

Contextual calibration

Classification tasks: normalized scores of label words
Generation tasks: probabilities of the first token of the generation over the entire vocabulary

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

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

Calibrated probs — Original probs

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

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

W: diagonal matrix
b: bias set to zeros

Slide from http://ericswallace.com/calibrate

Contextual calibration -- technical details

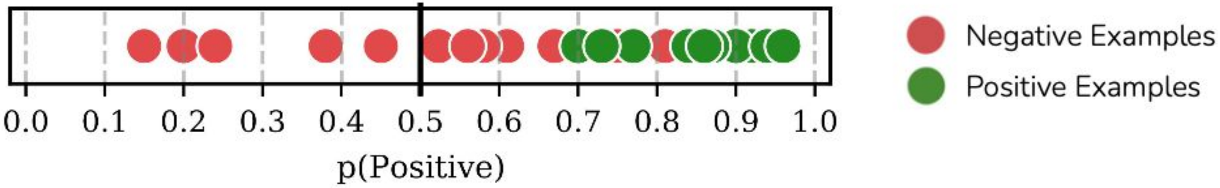
For generation tasks, why is only the first token calibrated?

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

Contextual calibration -- technical details

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

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



<u>All the biases effectively shift the output distribution</u>

Contextual calibration -- technical details

Why do they calibrate probabilities instead of calibrating logits?

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

Experimental Setup

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

– SST-2 (Socher et al., 2013)
– AGNews (Zhang et al., 2015)
– DBPedia (Zhang et al., 2015)
– TREC (Voorhees & Tice, 2000)
– RTE (Dagan et al., 2005)
– CB (de Marneffe et al., 2019)

Check out details of the other datasets in the appendix

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

Datasets: Fact Retrieval

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

LAMA (Petroni et al., 2019)

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

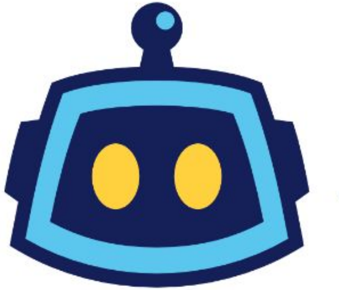
Datasets: Information Extraction

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

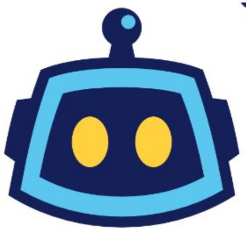
– ATIS (Hemphill et al., 2019)
– MIT Movies (Liu et al., 2012)

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

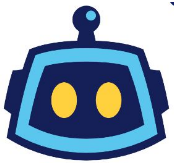
Model



GPT-3 - 175 billion

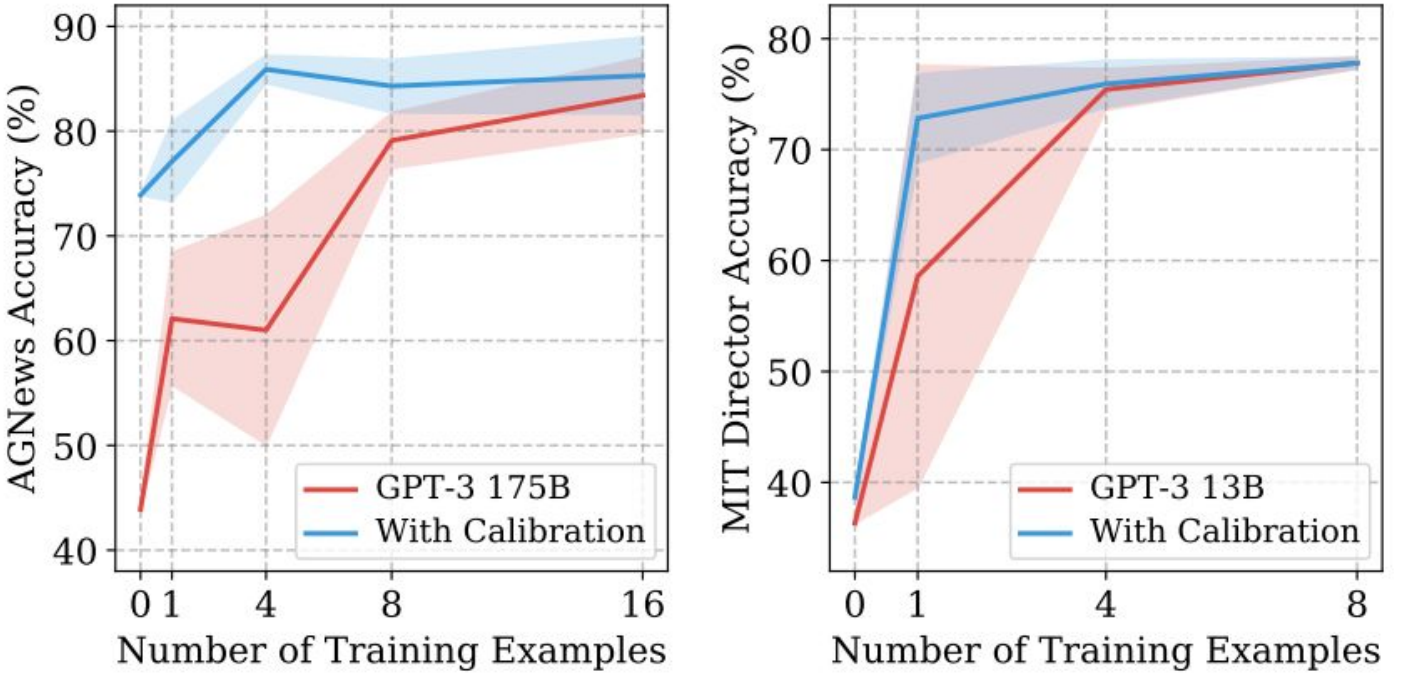


GPT-3 - 13 billion

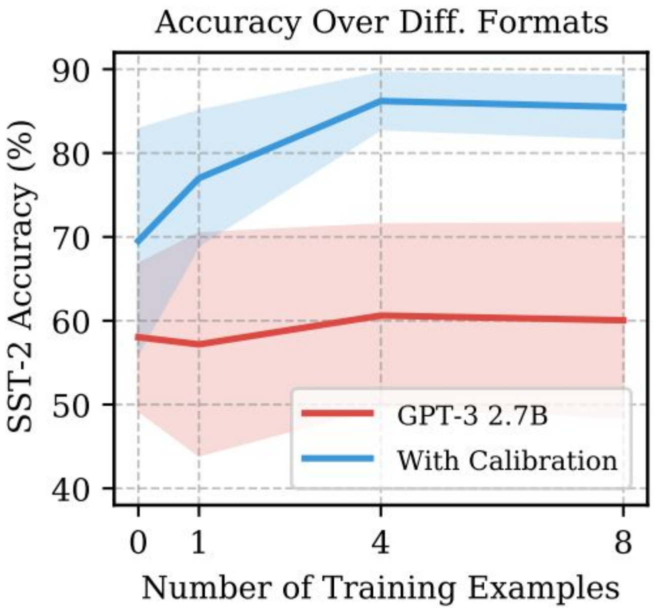


GPT-3 - 2.7 billion

Results



Reduces variance across training sets and permutations



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

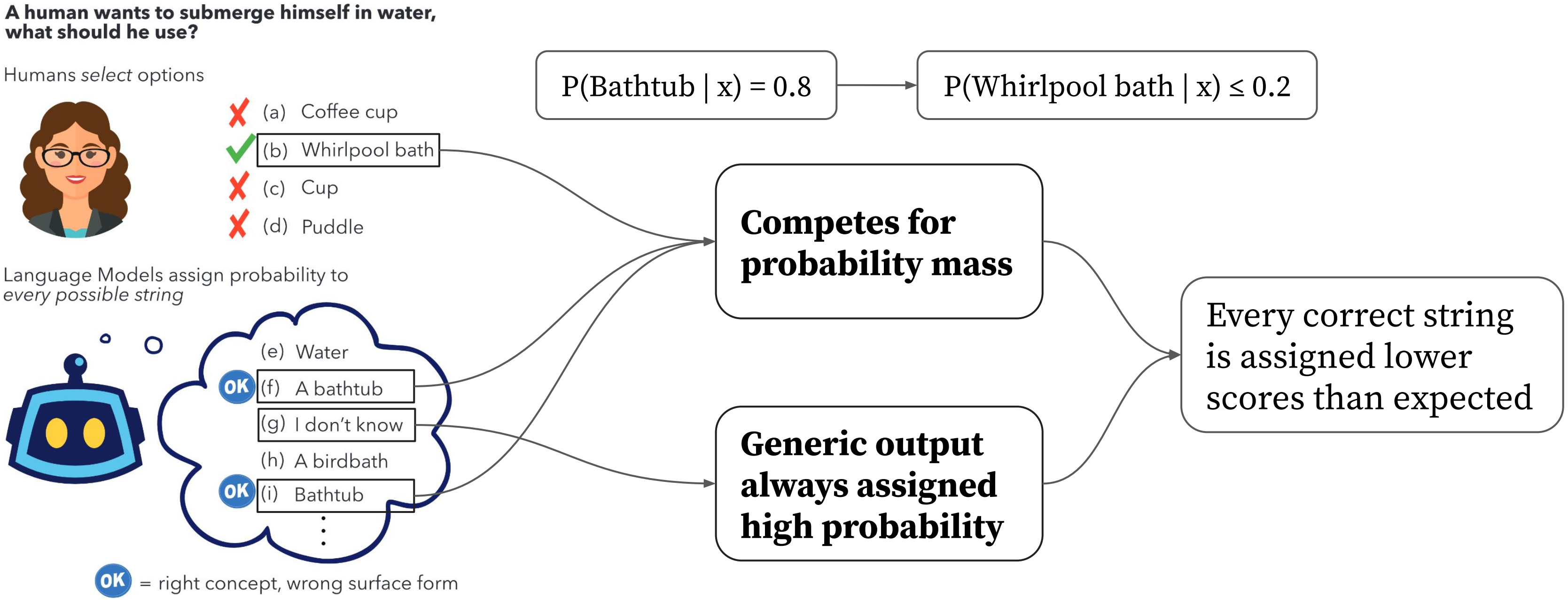
Surface Form Competition: Why the Highest Probability Answer Isn't Always Right

=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

Motivation

Surface Form Competition

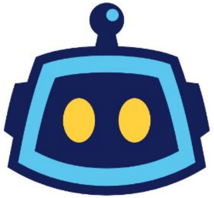


Choice of Plausible Alternatives (COPA) (Roemmele et al., 2011)

Premise ($\mathbf{x}$):
The bar closed because

Hypothesis 1 ($\mathbf{y}_1$):
it was crowded.

Hypothesis 2 ($\mathbf{y}_2$) :
it was 3am.



$P(y_1 \mid x) > P(y_2 \mid x)$

GPT-3

Methodology

Baselines

Template

Premise ($\mathbf{x}$):
The bar closed because

Domain Premise ($\mathbf{x}_{\text{domain}}$):
because

Hypothesis 1 ($\mathbf{y}_1$):
it was crowded.

Hypothesis 2 ($\mathbf{y}_2$) :
it was 3am.

Task: choose between Hypothesis y_1 and y_2 given Premise x

Baselines

Template

Premise ($\mathbf{x}$):
The bar closed because

Domain Premise ($\mathbf{x}_{\text{domain}}$):
because

Hypothesis 1 ($\mathbf{y}_1$):
it was crowded.

Hypothesis 2 ($\mathbf{y}_2$) :
it was 3am.

Scoring Functions

Probability (LM) $\quad \underset{i}{\text{argmax}}\ P(\mathbf{y}_i|\mathbf{x})$

Baselines

Template

Premise ($\mathbf{x}$):
The bar closed because

Domain Premise ($\mathbf{x}_{\text{domain}}$):
because

Hypothesis 1 ($\mathbf{y}_1$):
it was crowded.

Hypothesis 2 ($\mathbf{y}_2$) :
it was 3am.

Scoring Functions

Probability (LM)

$$\underset{i}{\operatorname{argmax}}\ P(\mathbf{y}_i|\mathbf{x})$$

Average Log-Likelihood (Avg)

$$\arg\max_i \frac{\sum_{j=1}^{\ell_i} P(y_i^j|\mathbf{x}, \mathbf{y}^{1\cdots j-1})}{\ell_i}$$

Baselines

Template

Premise ($\mathbf{x}$):
The bar closed because

Domain Premise ($\mathbf{x}_{\text{domain}}$):
because

Hypothesis 1 ($\mathbf{y}_1$):
it was crowded.

Hypothesis 2 ($\mathbf{y}_2$) :
it was 3am.

Scoring Functions

Probability (LM)

$$\underset{i}{\text{argmax}}\ P(\mathbf{y}_i|\mathbf{x})$$

Average Log-Likelihood (Avg)

$$\arg\max_i \frac{\sum_{j=1}^{\ell_i} P(y_i^j|\mathbf{x},\mathbf{y}^{1\cdots j-1})}{\ell_i}$$

Contextual Calibration (CC)

$$\arg\max_i \mathbf{w}P(\mathbf{y}_i|\mathbf{x}) + \mathbf{b}$$

Zhao et al., 2021

Baselines

Template

Premise ($\mathbf{x}$):
The bar closed because

Domain Premise ($\mathbf{x}_{\text{domain}}$):
because

Hypothesis 1 ($\mathbf{y}_1$):
it was crowded.

Hypothesis 2 ($\mathbf{y}_2$):
it was 3am.

Scoring Functions

Probability (LM)

$$\underset{i}{\text{argmax}}\ P(\mathbf{y}_i|\mathbf{x})$$

Average Log-Likelihood (Avg)

$$\arg\max_i \frac{\sum_{j=1}^{\ell_i} P(y_i^j|\mathbf{x},\mathbf{y}^{1\cdots j-1})}{\ell_i}$$

Contextual Calibration (CC)

$$\arg\max_i \mathbf{w}P(\mathbf{y}_i|\mathbf{x}) + \mathbf{b}$$

Zhao et al., 2021

Domain Conditional PMI (PMI_{DC})

$$\arg\max_i \frac{P(\mathbf{y}_i|\mathbf{x})}{P(\mathbf{y}_i|\mathbf{x}_{\text{domain}})}$$

note: this paper does not introduce any new modeling approaches, just a new scoring function

Pointwise Mutual Information (PMI)

Template

Premise ($\mathbf{x}$):
The bar closed because

Domain Premise ($\mathbf{x}_{\text{domain}}$):
because

Hypothesis 1 ($\mathbf{y}_1$):
it was crowded.

Hypothesis 2 ($\mathbf{y}_2$):
it was 3am.

$$\mathrm{PMI}(\mathbf{x}, \mathbf{y}) = \boxed{\log \frac{P(\mathbf{y}|\mathbf{x})}{P(\mathbf{y})}} = \boxed{\log \frac{P(\mathbf{x}|\mathbf{y})}{P(\mathbf{x})}}. \quad (2)$$

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

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

Domain Conditional Pointwise Mutual Information (PMI)

Template

Premise ($\mathbf{x}$):
The bar closed because

Domain Premise ($\mathbf{x}_{\text{domain}}$):
because

Hypothesis 1 ($\mathbf{y}_1$):
it was crowded.

Hypothesis 2 ($\mathbf{y}_2$):
it was 3am.

assumption: ending of the conditional premise x is a domain-relevant string x_{domain}

$$\text{PMI}(\mathbf{x}, \mathbf{y}) = \log \frac{P(\mathbf{y}|\mathbf{x})}{\boxed{P(\mathbf{y})}} = \log \frac{P(\mathbf{x}|\mathbf{y})}{P(\mathbf{x})}. \quad (2)$$

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

$$\text{PMI}_{\text{DC}}(\mathbf{x}, \mathbf{y}, \text{domain}) = \frac{P(\mathbf{y}|\mathbf{x}, \text{domain})}{P(\mathbf{y}|\text{domain})} = \frac{P(\mathbf{y}|\mathbf{x}, \text{domain})}{P(\mathbf{y}|\mathbf{x}_{\text{domain}})}$$

where domain is representative of the given task

Unconditional Baseline

$$\arg\max_i P(\mathbf{y}_i | \mathbf{x}_{\text{domain}}).$$

ignore the premise
completely!

Experimental Setup

[original question]$_P$
[domain premise]$_{DP}$
[original answers]$_{UH}$

Datasets

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.]$_{UH}$

Continuation: requires the model to select a continuation to previous text

- Choice of Plausible Alternatives (Roemmele et al., 2011)
- StoryCloze (Mostafazadeh et al., 2017)
- HellaSwag (Zellers et al., 2019)

Datasets

$[\text{original question}]_P$
$[\text{domain premise}]_{DP}$
$[\text{original answers}]_{UH}$

Type	Dataset	Template
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.]$_{UH}$

Question Answering (QA):

- RACE (Lai et al., 2017)
- ARC (Clark et al., 2018)
- Open Book Question Answering (Mihaylov et al., 2018)
- CommonsenseQA (Talmor et al., 2019)

[original question]$_P$
[domain premise]$_{DP}$
[original answers]$_{UH}$

Datasets

Type	Dataset	Template
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.]$_{UH}$

Boolean Question Answering:

- BoolQ (Clark et al., 2019)

[original question]$_P$
[domain premise]$_{DP}$
[original answers]$_{UH}$

Datasets

Type	Dataset	Template
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.]$_{UH}$

Entailment: if a premise sentence entails a hypothesis sentence

- [Recognizing Textual Entailment (Dagan et al., 2015)]
- [Commitment Bank (De Marneffe et al. 2019)]

$[\text{original question}]_P$
$[\text{domain premise}]_{DP}$
$[\text{original answers}]_{UH}$

Datasets

Type	Dataset	Template
Text Classification	SST-2	"$[\text{Illuminating if overly talky documentary}]_P$" $[\text{[The quote] has a tone that is}]_{DP}$ $[\text{positive.}]_{UH}$
	SST-5	"$[\text{Illuminating if overly talky documentary}]_P$" $[\text{[The quote] has a tone that is}]_{DP}$ $[\text{neutral.}]_{UH}$
	AG's News	title: $[\text{Economic growth in Japan slows down as the country experiences a drop in domestic and corporate [...]}]_P$ summary: $[\text{Expansion slows in Japan}]_P$ $[\text{topic:}]_{DP}$ $[\text{Sports.}]_{UH}$
	TREC	$[\text{Who developed the vaccination against polio?}]_P$ $[\text{The answer to this question will be}]_{DP}$ $[\text{a person.}]_{UH}$

Text Classification: if a premise sentence entails a hypothesis sentence

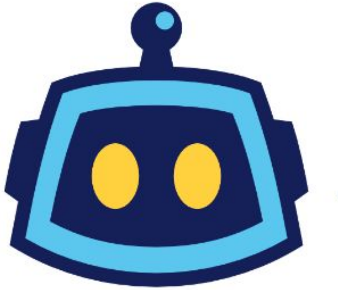
- Stanford Sentence Treebank (Socher et al., 2013)
- AG's News (Zhang et al., 2015)
- TREC (Li and Roth, 2002)

Datasets

[original question]$_{P}$
[domain premise]$_{DP}$
[original 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.]$_{UH}$

Model



GPT-3

Zero-shot

GPT-2

Reported but won't be the focus of the results

Results

Zero-shot Multiple Choice Accuracy

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

Consistently beat or tie other methods across model sizes and datasets

Summarized Results

Percent of Ties or Wins by Method

	Method	Unc	LM	Avg	PMI_{DC}	CC
GPT-2	125M	12.50	6.25	12.50	**68.75**	-
	350M	6.25	18.75	12.50	**68.75**	-
	760M	6.25	6.25	12.50	**75.00**	-
	1.6B	6.25	12.50	12.50	**80.00**	20.00
GPT-3	2.7B	6.25	6.25	6.25	**86.66**	0.00
	6.7B	6.25	25.00	25.00	**75.00**	-
	13B	6.25	18.75	18.75	**68.75**	-
	175B	6.25	12.50	18.75	**62.50**	6.25

consistently beat or tie for best results when compared to other methods

Zero-shot Multiple Choice Accuracy

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
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	-
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**	-
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**	-

HellaSwag (HS): generated by GPT-2

ARC Easy (ARC-E): too simple, harder questions are in ARC Challenge (ARC-C)

BoolQ (BQ): complex questions requiring high level reasoning → random guess

Prompt Robustness

Prompt Robustness on SST-2

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

maintain the highest mean using 15 different templates for SST-2 in Zhao et al. (2021)

but still high variance

Ablations

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

Removing Surface Form Competition

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

Premise ($\mathbf{x}$): The bar closed *because*

Domain Premise ($\mathbf{x}_{\text{domain}}$): *because*

Hypothesis 1 ($\mathbf{y}_1$): it was crowded.

Hypothesis 2 ($\mathbf{y}_2$): it was 3 AM.

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

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

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

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

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

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

Removing Surface Form Competition

Premise ($\mathbf{x}$): The bar closed *because*

Domain Premise ($\mathbf{x}_{\text{domain}}$): *because*

Hypothesis 1 ($\mathbf{y}_1$): it was crowded.

Hypothesis 2 ($\mathbf{y}_2$): it was 3 AM.

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

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

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

Hypothesis 2'($\mathbf{y}_2'$): it was 3:30AM.

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

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

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

$$\frac{P(\mathbf{y}_2'|\mathbf{x})}{P(\mathbf{y}_2'|\mathbf{x}_{\text{domain}})} > \frac{P(\mathbf{y}_1|\mathbf{x})}{P(\mathbf{y}_1|\mathbf{x}_{\text{domain}})}$$

$$\log P(\mathbf{y}_2|\mathbf{x}) \approx -16$$
$$\log P(\mathbf{y}_2'|\mathbf{x}) \approx -20$$

both probabilities low due to surface form competition!

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

no competition $\longrightarrow$ similarly high probabilities

Conclusion

Summary

Contextual Calibration (CC)

$$\arg\max_i \mathbf{w}P(\mathbf{y}_i|\mathbf{x}) + \mathbf{b}$$

Domain Conditional PMI (PMI_{DC})

$$\arg\max_i \frac{P(\mathbf{y}_i|\mathbf{x})}{P(\mathbf{y}_i|\mathbf{x}_{\text{domain}})}$$

both papers focuses on novel ways to calculate the probabilities for language modeling $\longrightarrow$ improve performance with minimal changes

Pre-Lecture Question 2

Explain the two calibration methods proposed in these two papers (Zhao et al., 2021) and (Hoffman et al., 2021). Can you think of any pros and cons comparison the two methods?

Zhao et al. (2021) proposes first calculating the output probabilities using content-free strings, and then use a linear classifier to calibrate the final probability. This method is computationally efficient and effective for single token outputs but not suited for multi-token generation.

Hoffman et al. (2021) proposes using a domain specific string to calculate the PMI of output strings. This calculates how much more likely an output becomes given the input, which removes surface form competition and generic output bias. However, domain specific string is subjective and difficult to choose the best one to use.

Related Work: Noisy Channel (Min et al., 2022)

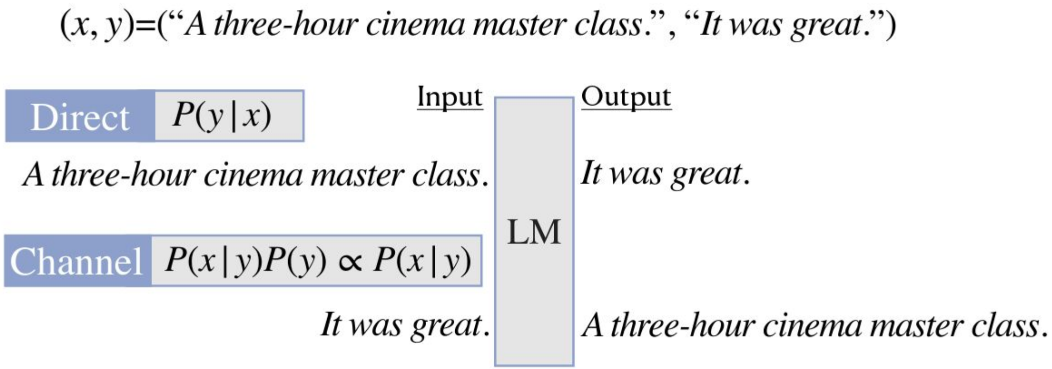


Figure 1: An illustration of the direct model and the channel model for language model prompting in the sentiment analysis task.

another alternative to calibrate the probability of final output

Pre-Lecture Question 3

Can you brainstorm other calibration ideas for improving prompting performance (and reducing the variance)? This can be either zero-shot or few-shot in-context learning.

Thank you!

Questions?