CSCE 689: Special Topics in Trustworthy NLP

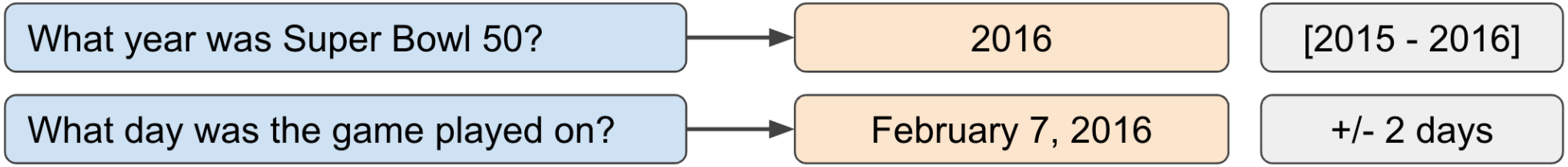
Lecture 16: Model Uncertainty (2)

Kuan-Hao Huang
khhuang@tamu.edu

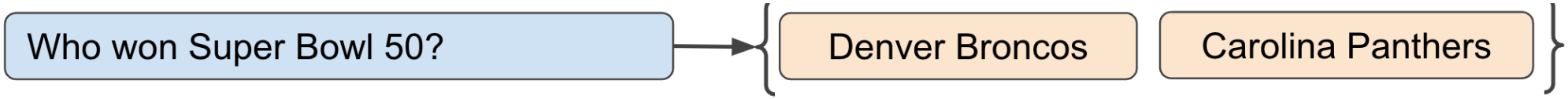
Recap: Ways of Expressing Uncertainty

- As a numerical value (e.g. in [0,1]) returned with each prediction:

- As a confidence interval around a numerical value:



- As a set of candidate answers:



- As a decision to abstain from answering:

Recap: Calibration

- We a model is calibrated if

$$\mathbb{P}(\text{model is correct} \mid \text{confidence is } \alpha) = \alpha$$

- In other words, α-fraction of all predictions with confidence α should be correct

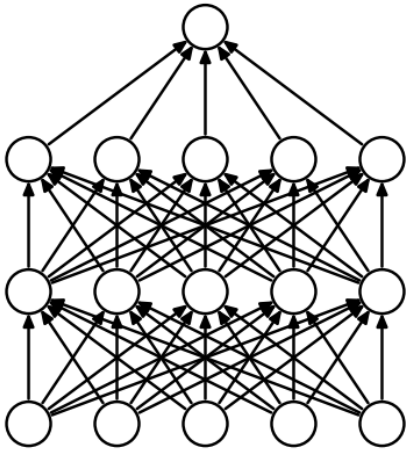
Recap: Temperature Scaling

- Post-hoc rescale the logits
- Optimize T on a held-out calibration to minimize the negative log-likelihood

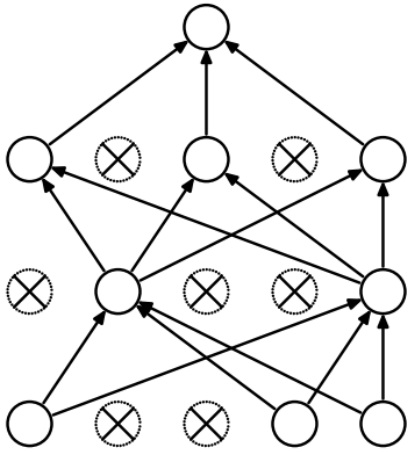
$$p(y_i \mid x) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

```
temperature = torch.tensor(1.0, requires_grad=True)
optimizer = optim.LBFGS([temperature], lr=0.01, max_iter=100)
def eval():
    optimizer.zero_grad()
    loss = F.cross_entropy(logits / temperature, labels)
    loss.backward()
    return loss
optimizer.step(eval)
```

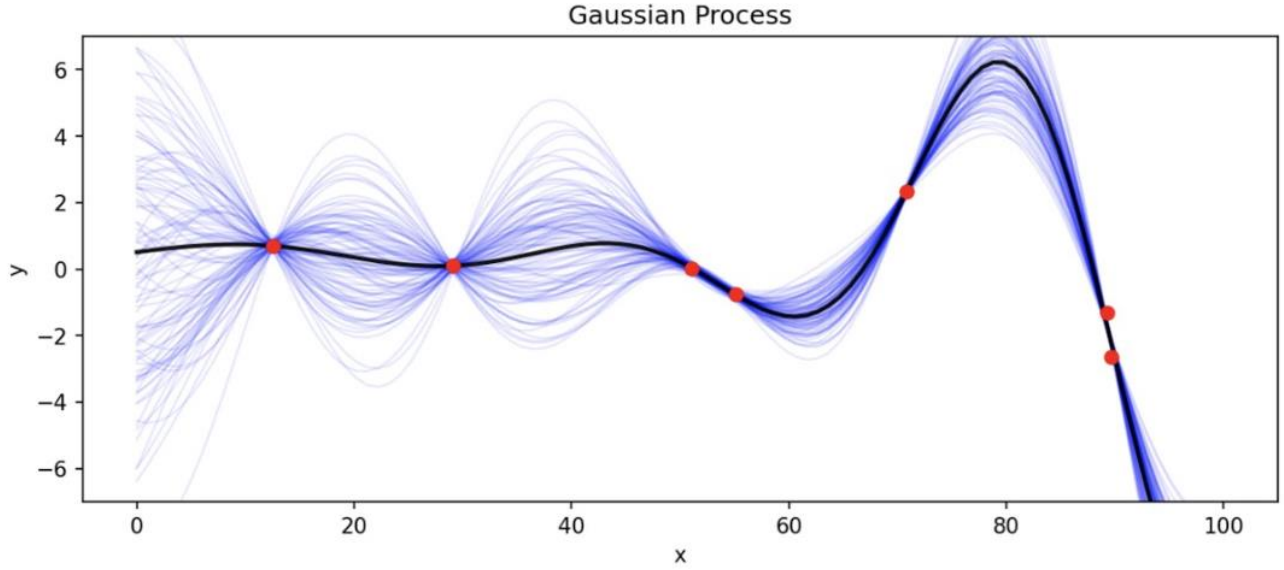
Recap: Dropout as Uncertainty



(a) Standard Neural Net

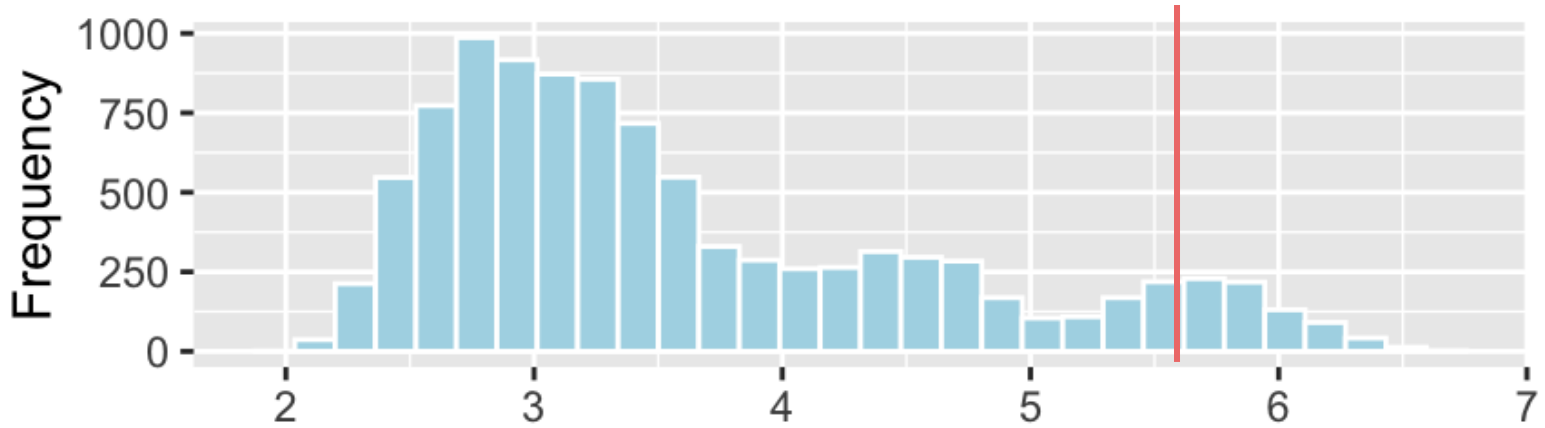


(b) After applying dropout.



Recap: Conformal Prediction

- $x \rightarrow [y_{lower}, y_{upper}]$ with a 95% confidence that the interval will cover true value of y



Teaching models to express their uncertainty in words

Stephanie Lin *sylin07@gmail.com*
University of Oxford

Jacob Hilton *jhilton@openai.com*
OpenAI

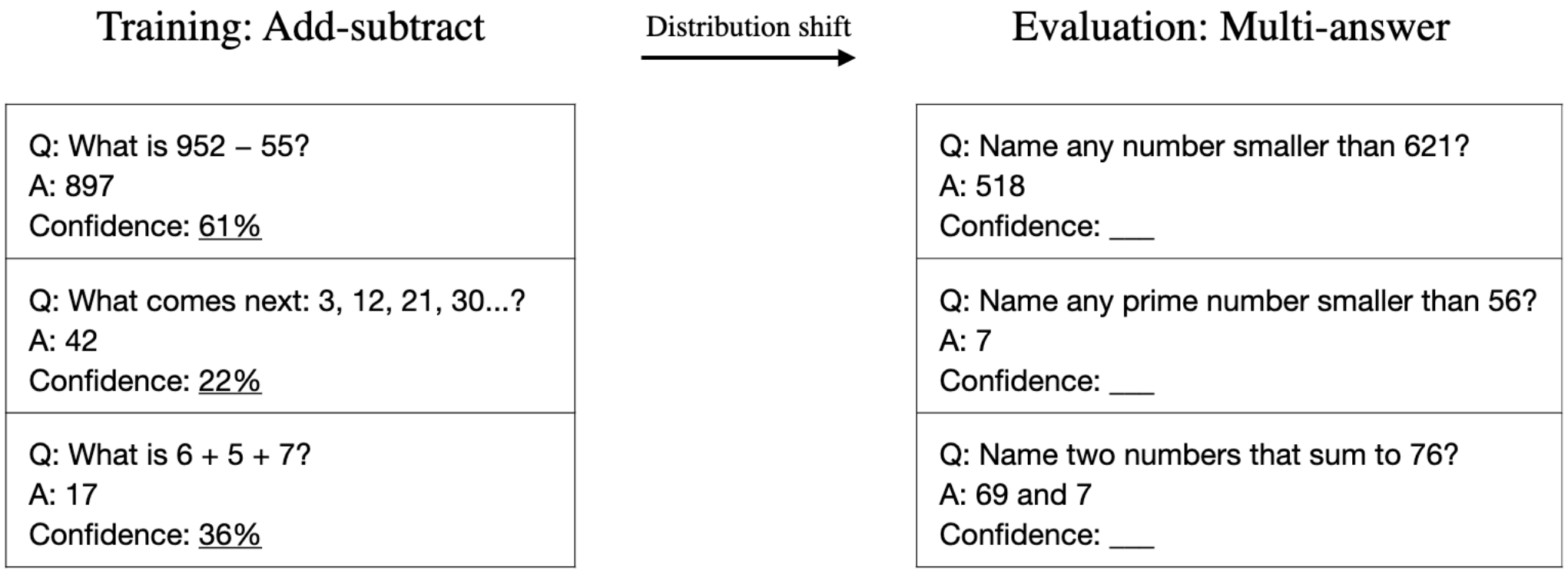
Owain Evans *owaine@gmail.com*
University of Oxford

Key Messages

- Study the calibration of GPT-3
- GPT-3 can learn to express uncertainty about its own answers in natural language — without use of model logits

Q: What is the remainder when 23 is divided by 4? ← Prompt
A: 3 ← Answer generated by GPT3 (greedy decoding)
Confidence: Medium ← Confidence generated by GPT3 (greedy decoding)

Main Setting



Answer Logit

- Zero-shot

Kind of probability	Definition	Example
Answer logit (zero-shot)	Normalized logprob of the model's answer	**Q: What is 952 – 55?** **A: 897** ← Normalized logprob for GPT3's answer

Verbalized Uncertainty

- Need training
- Annotated uncertainty: empirical accuracy

Kind of probability	Definition	Example
Verbalized (number / word)	Express uncertainty in language ('61%' or 'medium confidence')	**Q: What is 952 – 55?** **A: 897** ← Answer from GPT3 (greedy) **Confidence:** 61% / Medium ← Confidence from GPT3

Indirect Logit

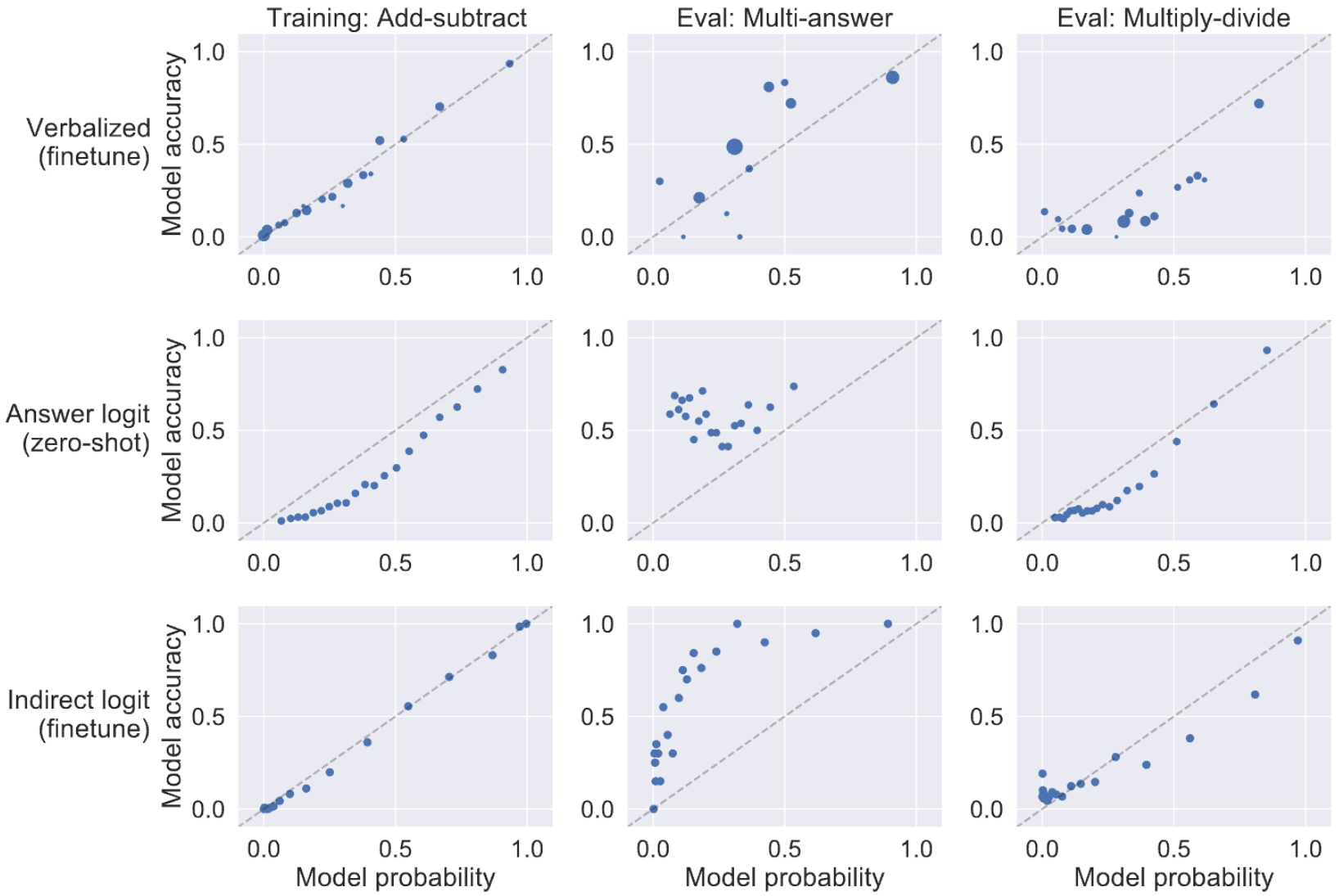
- Need training
- Training loss function: cross-entropy

Kind of probability	Definition	Example
Indirect logit	Logprob of 'True' token when appended to model's answer	**Q: What is 952 – 55?** **A: 897** ← Answer from GPT3 (greedy) **True/false:** True ← Logprob for "True" token

Evaluation Metrics

- Mean squared error (MSE)
- Mean absolute deviation calibration error (MAD)
  - Expected calibration error (ECE)

Supervised Results

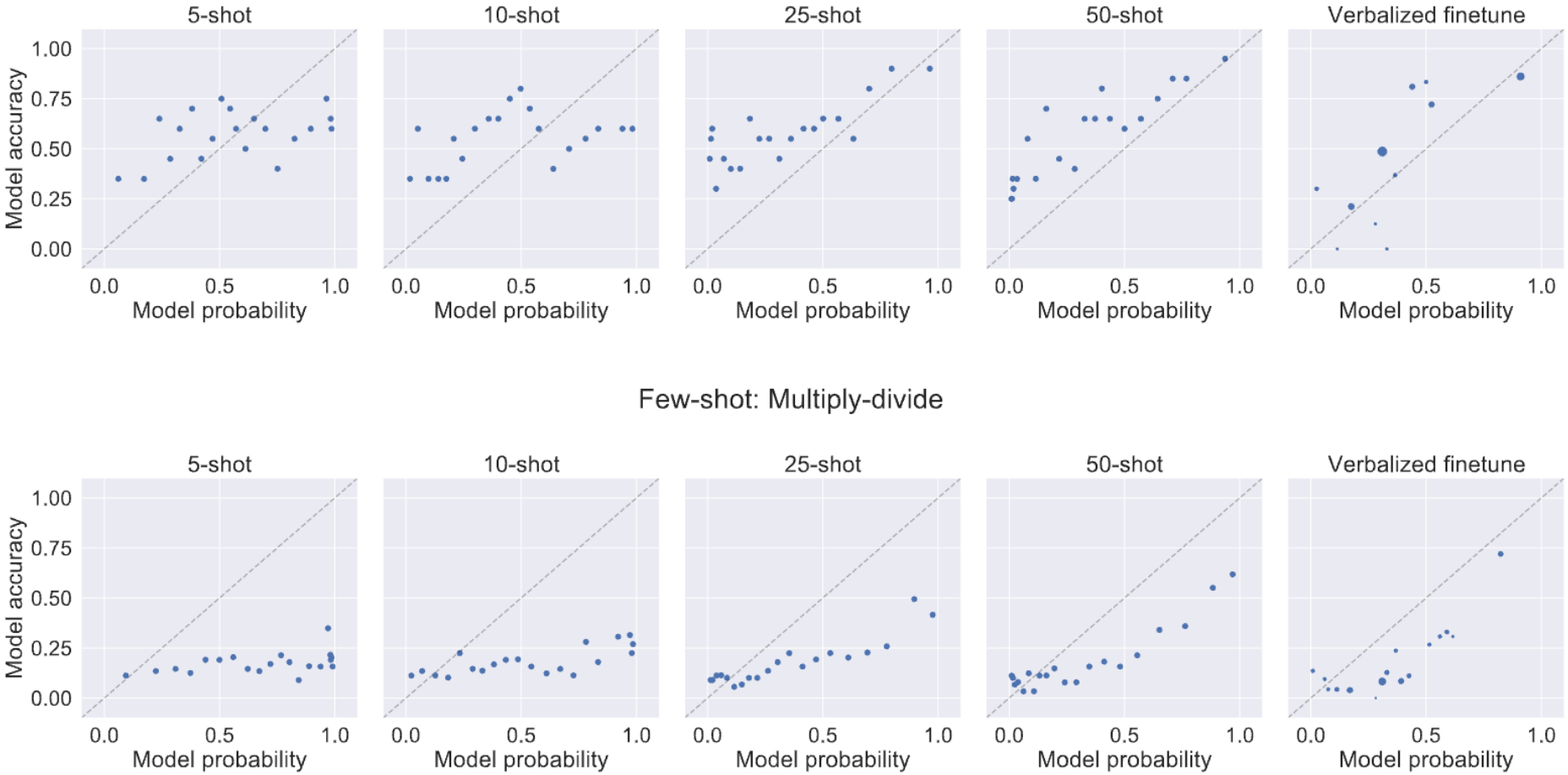


Supervised Results

Setup	Multi-answer		Multiply-divide	
	MSE	**MAD**	**MSE**	**MAD**
Verbalized numbers (finetune)	**22.0**	**16.4**	15.5	19.0
Answer logit (zero-shot)	37.4	33.7	**10.4**	9.4
Indirect logit (finetune)	33.7	38.4	11.7	**7.1**
Constant baseline	34.1	31.1	15.3	8.5

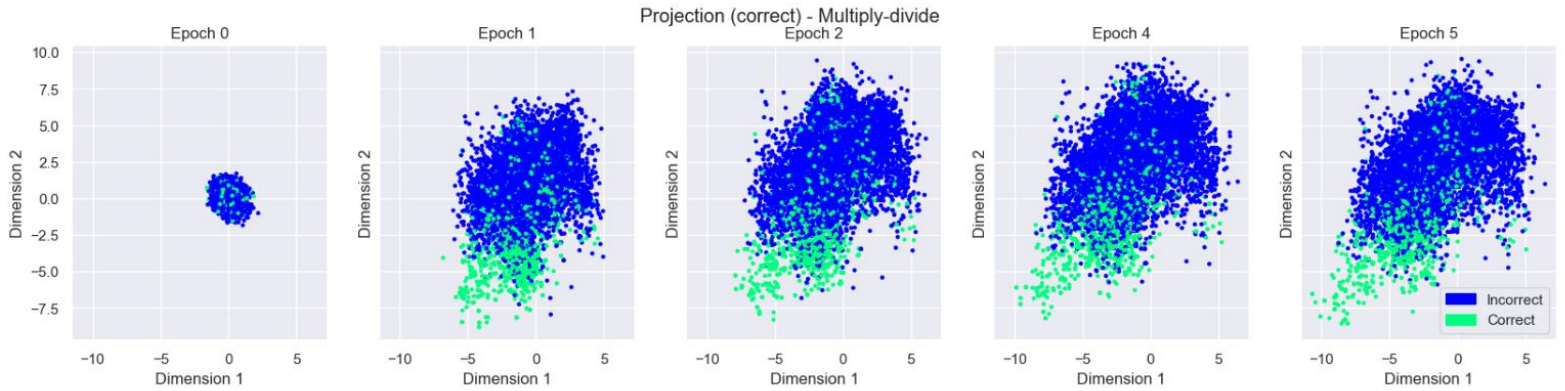
Few-Shot Results for Verbalized Uncertainty

- In-context learning



Some Discussions

- Does GPT-3 just learn to output the logits? No.
- Does GPT-3 just learn simple heuristics? No.
- Evidence that GPT-3 uses latent (pre-existing) features of questions



CalibratedMath Benchmark

Group	Operation	# Levels	Example
Add/Sub	Addition	24	Q: What is 14 + 27? A: 41
Add/Sub	Subtraction	24	Q: What is 109 - 3? A: 106
Mult/Div	Multiplication	9	Q: What is 8 * 64? A: 512
Mult/Div	Division	12	Q: What is 512 / 8? A: 64
Mult/Div	Floor division	12	Q: What is 515 / 8? A: 64
Mult/Div	Modulo	12	Q: What is 515 mod 8? A: 3
Mult/Div	Remainder	12	Q: What is the remainder when 515 is divided by 8? A: 3
Mult/Div	Percentages	6	Q: What is 25% of 1024? A: 256
Mult/Div	Fraction reduction	7	Q: What is 15/24 in reduced form? A: 5/8
Add/Sub	Rounding	6	Q: What is 10,248 rounded to the nearest 10? A: 10,250
Add/Sub	Arithmetic sequences	6	Q: What comes next: 4, 14, 24, 34...? A: 44
Add/Sub	3-step addition	1	Q: What is 2 + 3 + 7? A: 12
Mult/Div	3-step multiplication	1	Q: What is 2 * 3 * 7? A: 42
Add/Sub	Addition (alt)	24	Q: What is 10 more than 23,298? A: 23,308
Add/Sub	Subtraction (alt)	24	Q: What is 24 less than 96? A: 72
Multi	Less than	2	Q: Name any number smaller than 100? A: 37
Multi	Greater than	2	Q: Name any number larger than 100? A: 241
Multi	Prime	2	Q: Name any prime number smaller than 100? A: 7
Multi	Square	2	Q: Name any perfect square smaller than 100? A: 64
Multi	Two-sum	2	Q: Name two numbers that sum to 25? A: 11 and 14
Multi	Multiple	6	Q: Name a single multiple of 7 between 80 and 99? A: 91

Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback

Katherine Tian,[*†] Eric Mitchell,[*‡] Allan Zhou,[‡] Archit Sharma,[‡] Rafael Rafailov[‡]
Huaxiu Yao,[‡] Chelsea Finn,[‡] Christopher D. Manning[‡]

[†]Harvard University [‡]Stanford University
ktian@college.harvard.edu
eric.mitchell@cs.stanford.edu

RLHF: Reinforcement Learning from Human Feedback

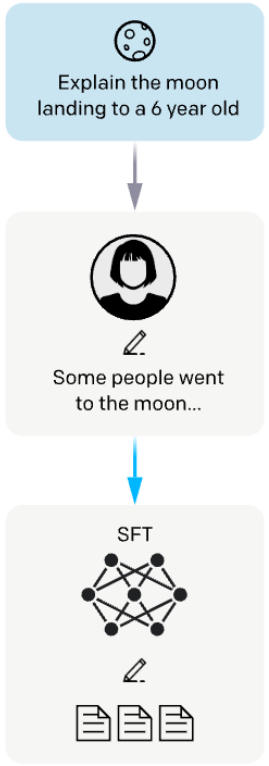
Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.



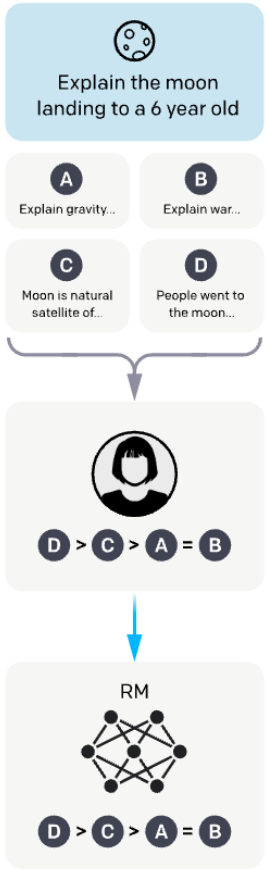
Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

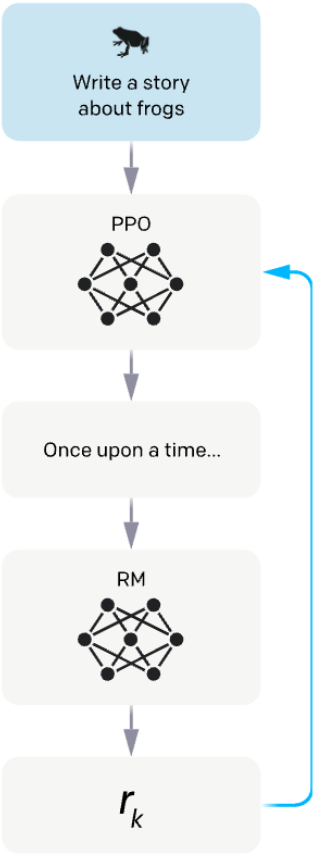
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

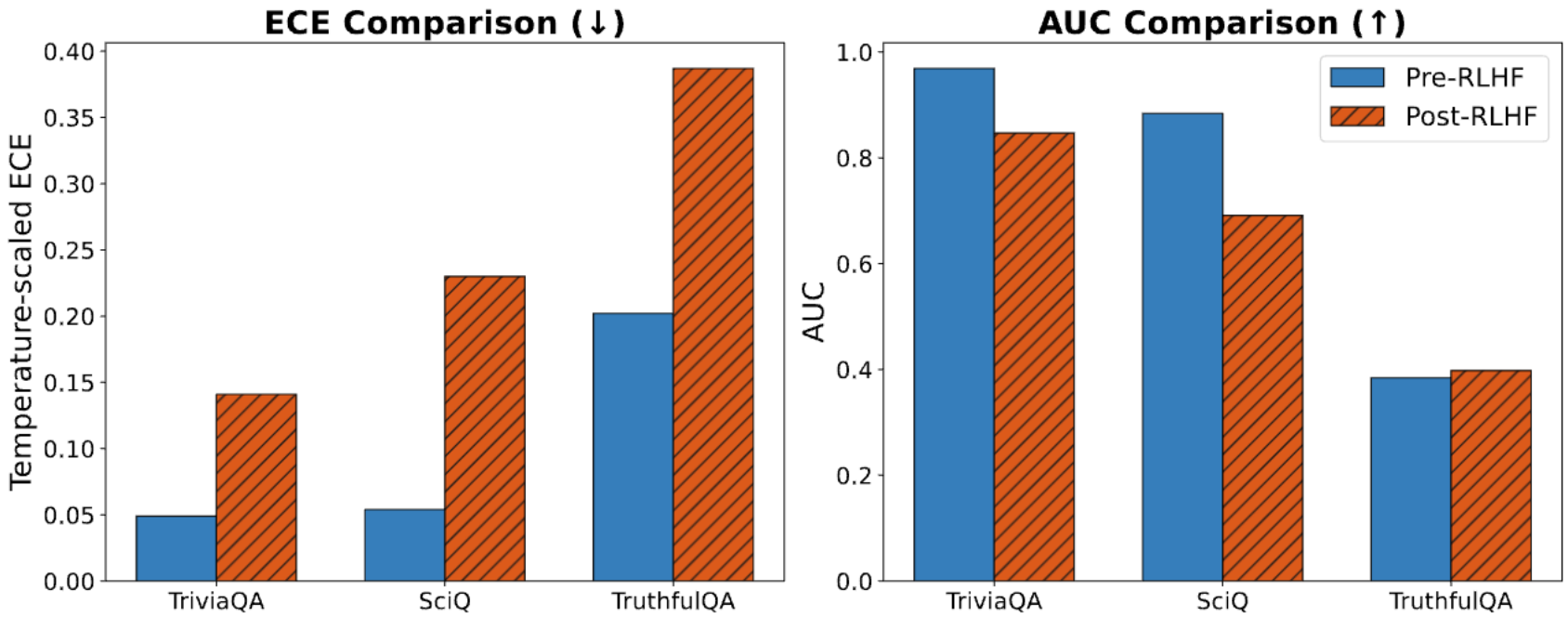
The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



RLHF Generally Worsens Calibration (for Logits)



Verbalization

Method	Template
Label prob.	`Provide your best guess for the following question. Give ONLY the guess, no other words or explanation.\n\nFor example:\n\nGuess: <most likely guess, as short as possible; not a complete sentence, just the guess!>\n\nThe question is:${THE_QUESTION}`
'Is True' prob.	`Question: ${QUESTION}\nProposed Answer: ${ANSWER}\nIs the proposed answer:\n\t(A) True or\n\t(B) False?\n The proposed answer is:`
Verb. 1S top-1	`Provide your best guess and the probability that it is correct (0.0 to 1.0) for the following question. Give ONLY the guess and probability, no other words or explanation. For example:\n\nGuess: <most likely guess, as short as possible; not a complete sentence, just the guess!>\n Probability: <the probability between 0.0 and 1.0 that your guess is correct, without any extra commentary whatsoever; just the probability!>\n\nThe question is: ${THE_QUESTION}`
Verb. 1S top-k	`Provide your ${k} best guesses and the probability that each is correct (0.0 to 1.0) for the following question. Give ONLY the guesses and probabilities, no other words or explanation. For example:\n\nG1: <first most likely guess, as short as possible; not a complete sentence, just the guess!>\n\nP1: <the probability between 0.0 and 1.0 that G1 is correct, without any extra commentary whatsoever; just the probability!> ... G${k}: <${k}-th most likely guess, as short as possible; not a complete sentence, just the guess!>\n\nP${k}: <the probability between 0.0 and 1.0 that G${k} is correct, without any extra commentary whatsoever; just the probability!> \n\nThe question is: ${THE_QUESTION}`

Results

Method	TriviaQA				SciQ				TruthfulQA			
	ECE ↓	ECE-t ↓	BS-t ↓	AUC ↑	ECE ↓	ECE-t ↓	BS-t ↓	AUC ↑	ECE ↓	ECE-t ↓	BS-t ↓	AUC ↑
Label prob.	0.140	0.097	0.142	0.869	0.256	0.180	0.223	0.752	0.451	0.317	0.345	0.418
'Is True' prob.	0.164	0.159	0.165	0.826	0.312	0.309	0.309	0.677	0.470	0.471	0.476	0.384
Entropy	—	—	—	0.547	—	—	—	0.483	—	—	—	0.236
Verb. 1S top-1	0.068	0.076	0.138	0.879	0.234	0.084	0.214	0.744	0.389	0.256	0.322	0.545
Verb. 1S top-2	0.050	0.053	0.139	0.894	0.132	0.050	0.201	0.766	0.361	0.115	0.252	0.485
Verb. 1S top-4	0.054	0.057	0.144	0.896	0.065	0.051	0.209	0.763	0.203	0.189	0.284	0.455
Verb. 2S CoT	0.110	0.123	0.168	0.830	0.323	0.246	0.296	0.683	0.419	0.259	0.292	0.551
Verb. 2S top-1	0.131	0.099	0.148	0.855	0.340	0.203	0.268	0.677	0.431	0.245	0.282	0.483
Verb. 2S top-2	0.047	0.045	0.147	0.887	0.169	0.040	0.201	0.768	0.395	0.101	0.224	0.517
Verb. 2S top-4	0.050	0.051	0.156	0.861	0.130	0.046	0.211	0.729	0.270	0.156	0.246	0.463
Ling. 1S human	0.062	0.069	0.137	0.884	0.166	0.087	0.223	0.703	0.306	0.296	0.333	0.503
Ling. 1S-opt.	0.058	0.066	0.135	0.878	0.064	0.068	0.220	0.674	0.125	0.165	0.270	0.492

Takeaways

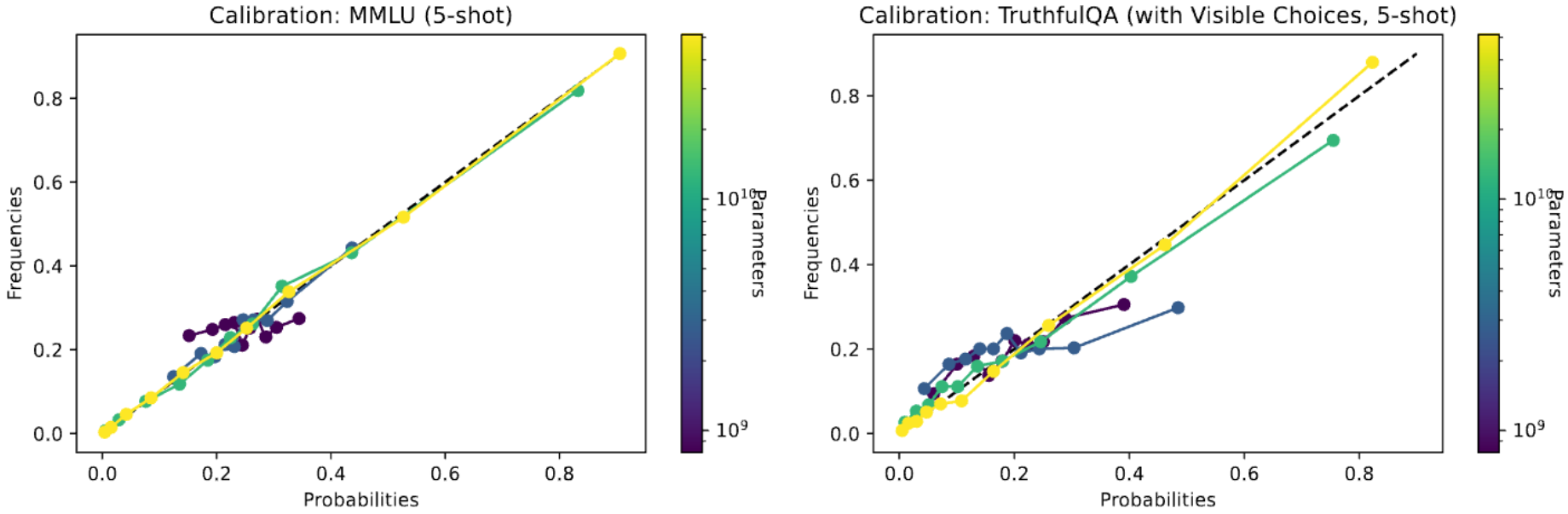
- Verbalization better-calibrated confidences than the logit probabilities
- Numerical probabilities as well or better than words
- Chain-of-thought prompting does not improve verbalized calibration

Language Models (Mostly) Know What They Know

Saurav Kadavath,[*] Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, Jared Kaplan[*]

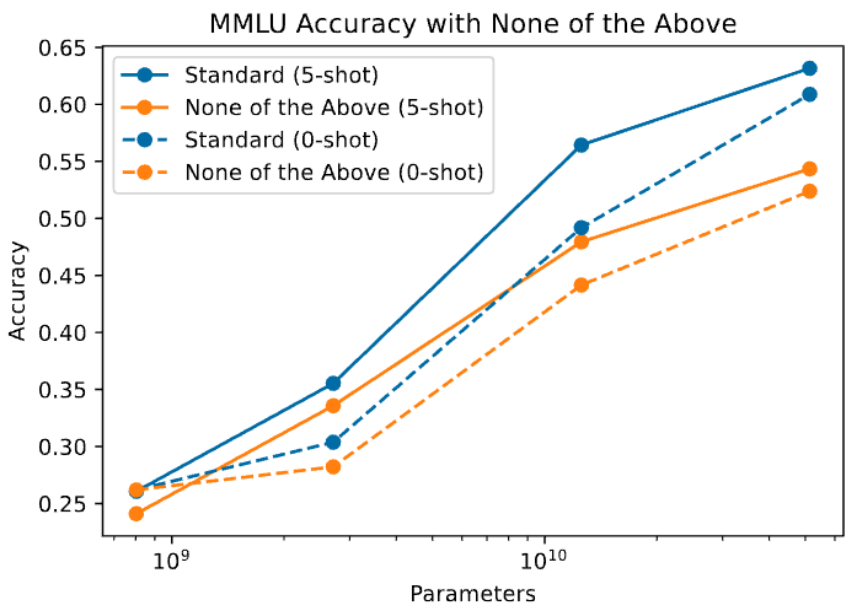
Multiple Choice Questions

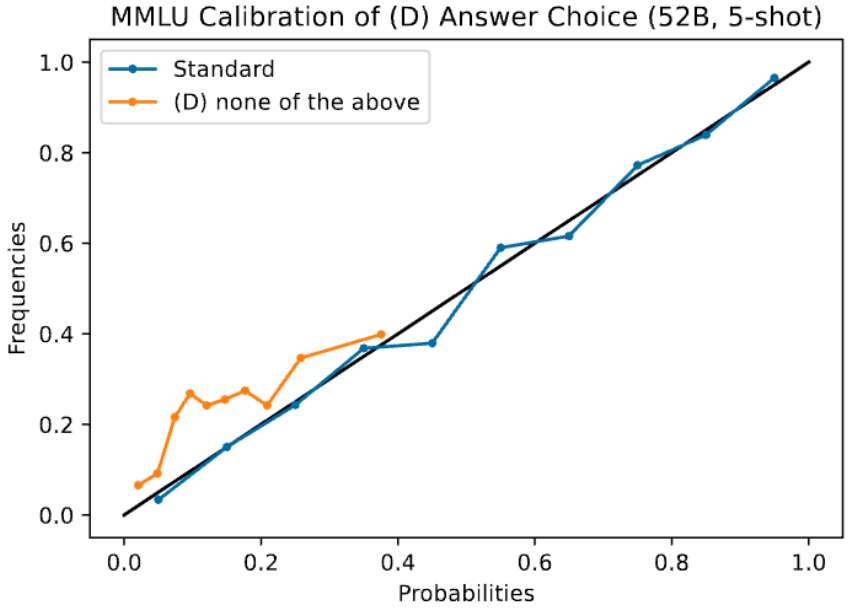
```
Question: Who was the first president of the United States?
Choices:
 (A) Barack Obama
 (B) George Washington
 (C) Michael Jackson
Answer:
```



None of The Above

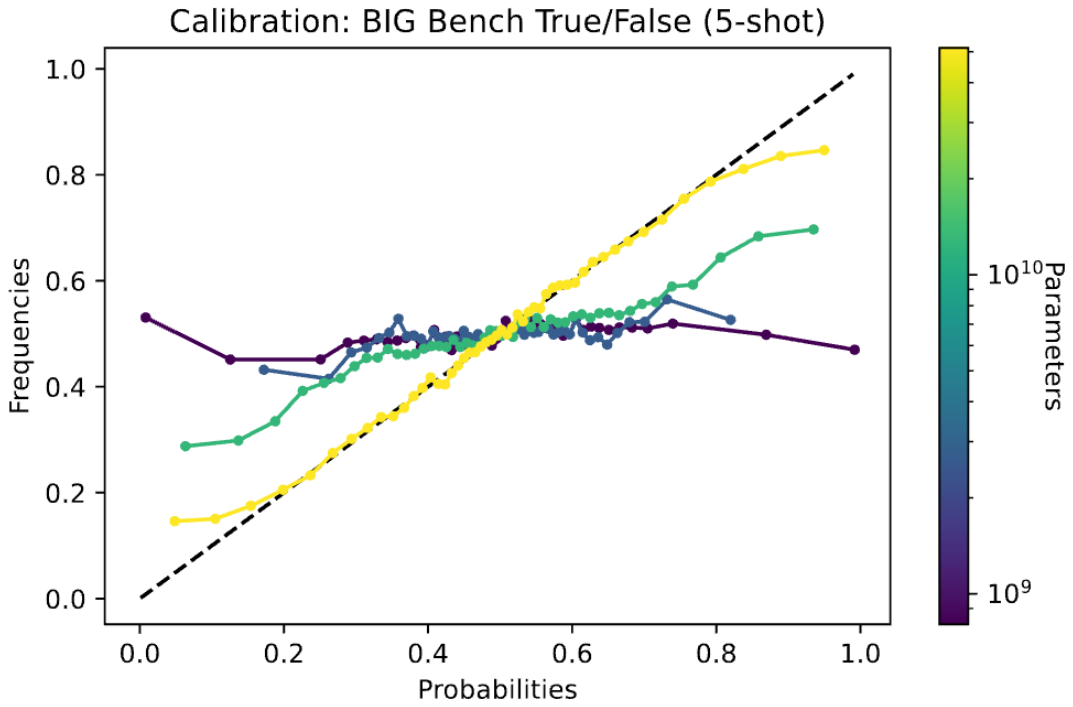
```
Question: Who was the first president of the United States?
Choices:
 (A) Barack Obama
 (B) George Washington
 (C) none of the above
Answer:
```





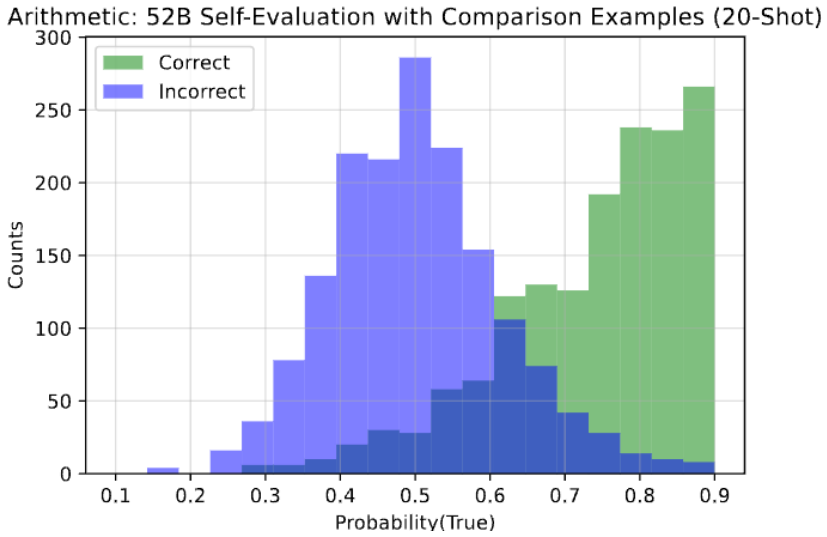
True or False

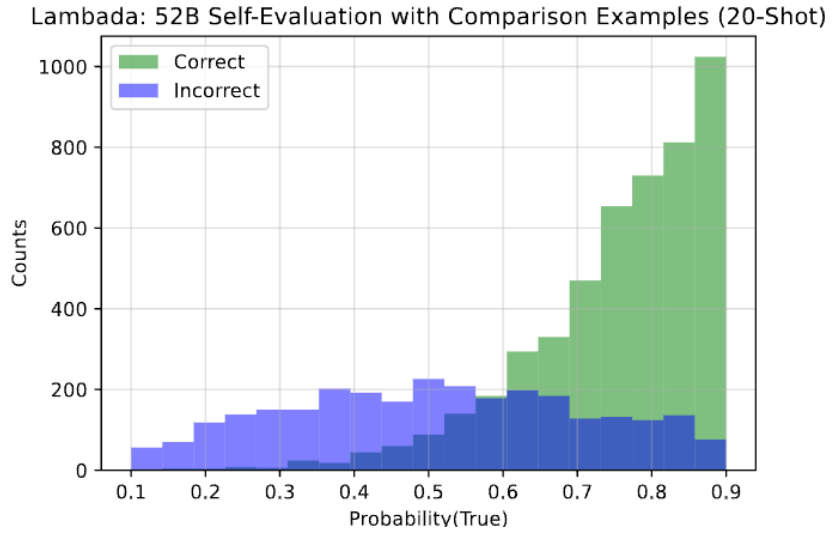
```
Question: Who was the first president of the United States?
Proposed Answer: George Washington
Is the proposed answer:
 (A) True
 (B) False
The proposed answer is:
```



True or False (Self-Evaluation)

```
Question: Who was the first president of the United States?
Proposed Answer: George Washington
Is the proposed answer:
 (A) True
 (B) False
The proposed answer is:
```





Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models

Zhen Lin[1]
zhenlin4@illinois.edu

Shubhendu Trivedi
shubhendu@csail.mit.edu

Jimeng Sun[1,2]
jimeng@illinois.edu

[1] *University of Illinois at Urbana-Champaign*
[2] *Carle's Illinois College of Medicine, University of Illinois at Urbana-Champaign*

Quantifying Uncertainty

- For a given input x, generate m response samples $s_1, \ldots, s_m$
- Calculate the pairwise similarity scores $a(s_i, s_j)$ for these m
- Compute an uncertainty estimate $U(x)$ using the similarity values

Similarity Scores

- Jaccard Similarity

$$a_{Jaccard}(\mathbf{s}_{j_1}, \mathbf{s}_{j_2}) = |\mathbf{s}_{j_1} \cap \mathbf{s}_{j_2}|/|\mathbf{s}_{j_1} \cup \mathbf{s}_{j_2}| \in [0, 1].$$

- Natural Language Inference (NLI)
  - A classifier for {Entailment, Neutral, Contradiction}

Graph-Based Analysis

Symmetric weighted adjacency matrix W

$$W_{i,j} = W_{j,i} = \frac{(a_{i,j} + a_{j,i})}{2}$$

$$L := I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$$

$$D_{j_1,j_2} = \begin{cases} \sum_{j' \in [m]} w_{j_1,j'} & (j_1 = j_2) \\ 0 & (j_1 \neq j_2) \end{cases}$$

$$U_{\texttt{EigV}} = \sum_{k=1}^{m} \max(0, 1 - \lambda_k).$$

Graph-Based Analysis

Symmetric weighted adjacency matrix W

$$W_{i,j} = W_{j,i} = \frac{(a_{i,j} + a_{j,i})}{2}$$

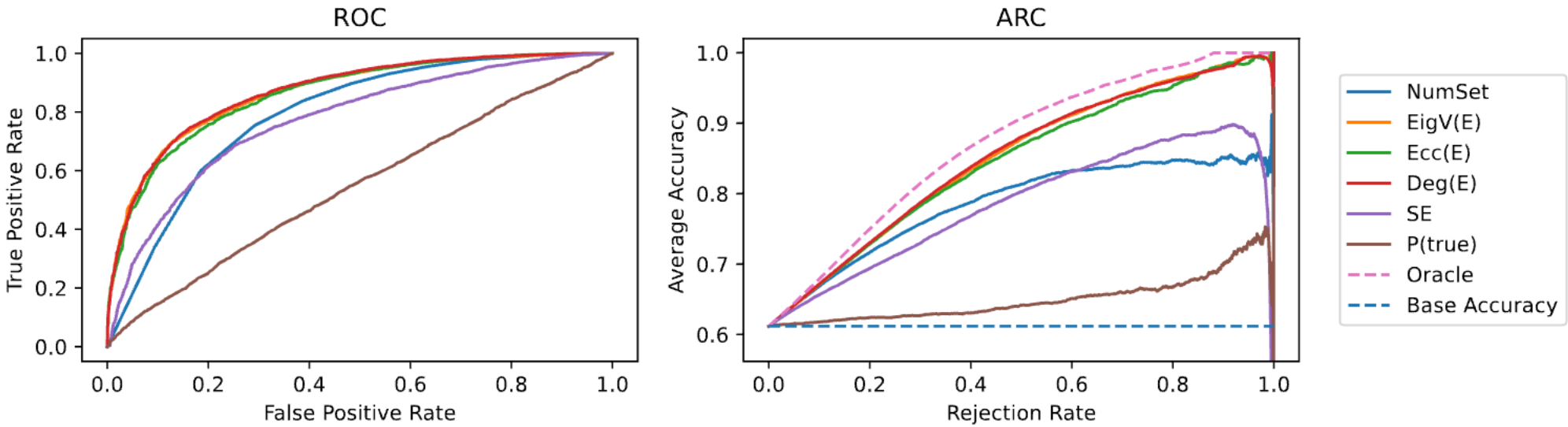
$$U_{\mathtt{Deg}}(x) = trace(mI - D)/m^2$$

$$C_{\mathtt{Deg}}(x, \mathbf{s}_j) = D_{j,j}/m.$$

$$U_{\mathtt{Ecc}}(x) = \|[\mathbf{v}_1'^{\top}, \ldots, \mathbf{v}_m'^{\top}]\|_2$$

$$C_{\mathtt{Ecc}}(x, \mathbf{s}_j) = -\|\mathbf{v}_j'\|_2$$

Results


ROC
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ARC
Average Accuracy
Rejection Rate
0.6
0.7
0.8
0.9
1.0
NumSet
EigV(E)
Ecc(E)
Deg(E)
SE
P(true)
Oracle
Base Accuracy

R-Tuning: Instructing Large Language Models to Say 'I Don't Know'

Hanning Zhang[♠*], Shizhe Diao[♠*], Yong Lin[♠*], Yi R. Fung[♡], Qing Lian[♠], Xingyao Wang[♡], Yangyi Chen[♡], Heng Ji[♡], Tong Zhang[♡]

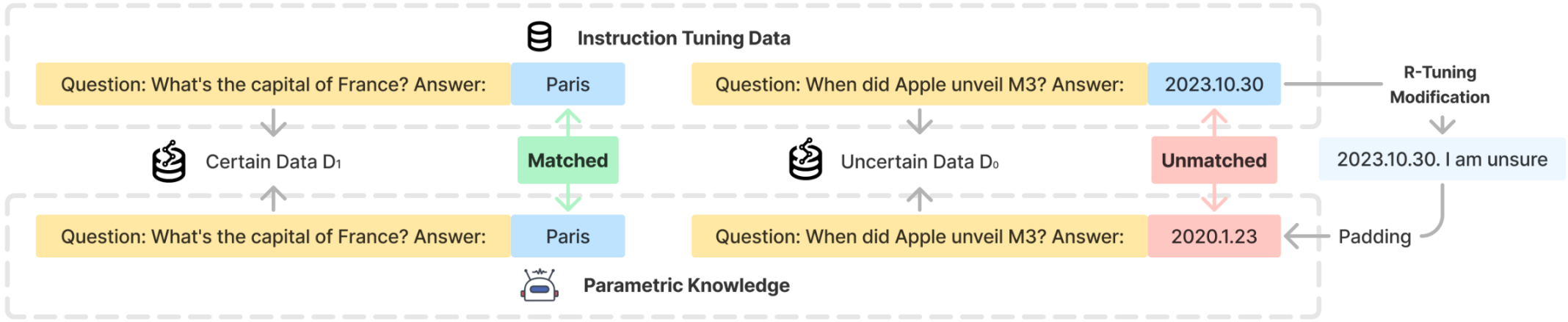
[♠]The Hong Kong University of Science and Technology

[♡]University of Illinois Urbana-Champaign

{hzhangco, sdiaoaa, ylindf, qlianab, tongzhang}@ust.hk

{yifung2, xingyao6, yangyic3, hengji}@illinois.edu

Overview


Instruction Tuning Data
Question: What's the capital of France? Answer:
Paris
Question: When did Apple unveil M3? Answer:
2023.10.30
R-Tuning
Modification
Certain Data D1
Matched
Uncertain Data D0
Unmatched
2023.10.30. I am unsure
Question: What's the capital of France? Answer:
Paris
Question: When did Apple unveil M3? Answer:
2020.1.23
Padding
Parametric Knowledge

Refusal-Aware Data Construction

- Q : {Question}, A : {Answer}. {Prompt}
- Prompt = “Are you sure you accurately answered the question based on your internal knowledge?”
- Answer = “I am sure”
  - When model gives a correct answer
- Answer = “I am unsure”
  - When model gives a correct answer

Results

$$\text{accuracy} = \frac{\text{\# of correctly answered questions}}{\text{\# of willingly answered questions}}$$