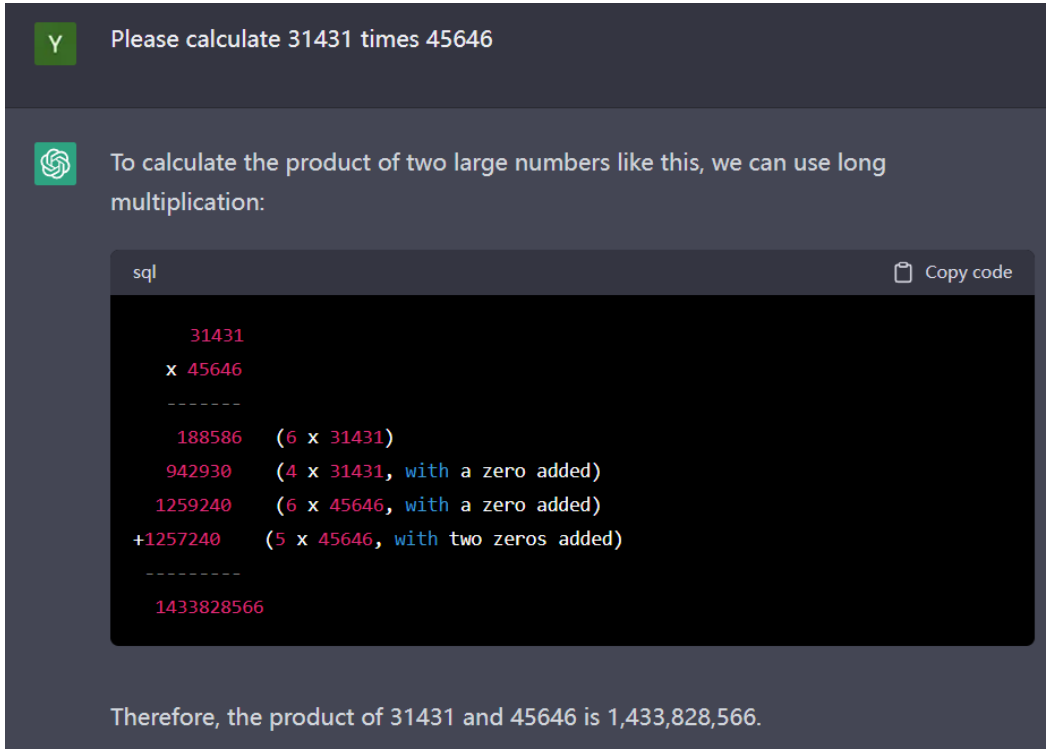
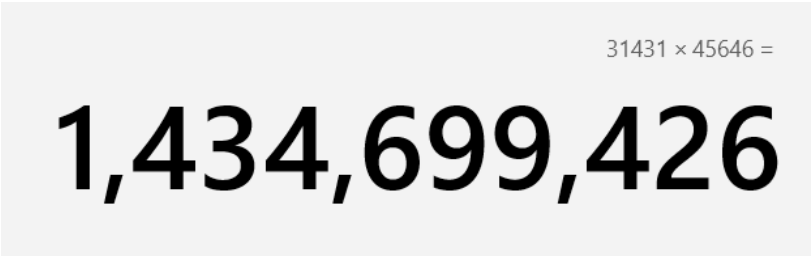
Language Models (Mostly) Know What They Know

Yusen Zhang yfz5488@psu.edu

Introduction

- Is AI system honest?
  - accurately and faithfully evaluate their level of confidence?





The answer given by ChatGPT is not correct, but it is very confident about its answer.

Larger Models are Calibrated on Diverse Multiple Choice Questions

- calibration:
  - A model makes calibrated predictions if the probability it assigns to outcomes coincides with the **frequency** with which these outcomes actually occur.
  - Perfect Calibration[1]:

$$\mathbb{P}\left(\hat{Y} = Y \mid \hat{P} = p\right) = p, \quad \forall p \in [0, 1]$$

Where Y hat is a class prediction and P hat its associated confidence, i.e. probability of correctness.

  - Expected Calibration Error (ECE)[1]:

$$\mathop{\mathbb{E}}_{\hat{P}}\left[\left|\mathbb{P}\left(\hat{Y} = Y \mid \hat{P} = p\right) - p\right|\right]$$

[1] Guo C, Pleiss G, Sun Y, et al. On calibration of modern neural networks[C]//International conference on machine learning. PMLR, 2017:

Larger Models are Calibrated on Diverse Multiple Choice Questions

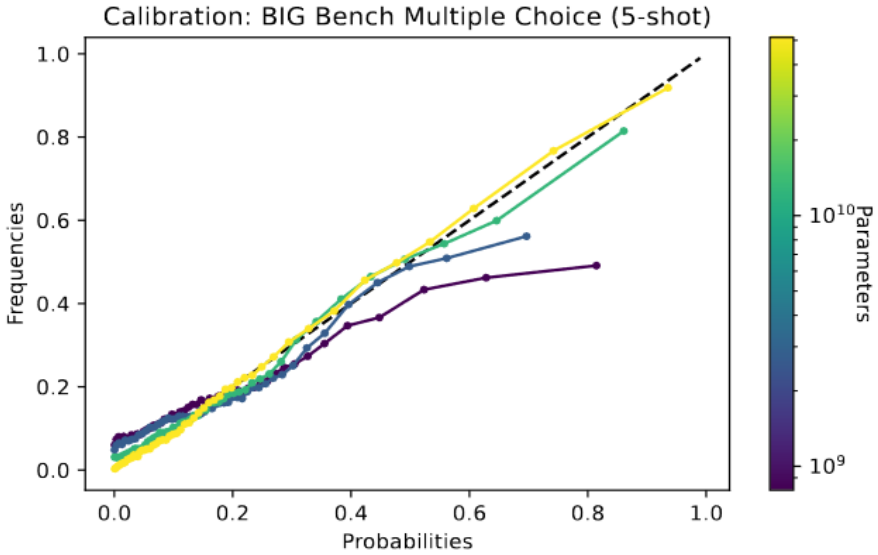
- calibration: Let the model answer the multiple choice questions!

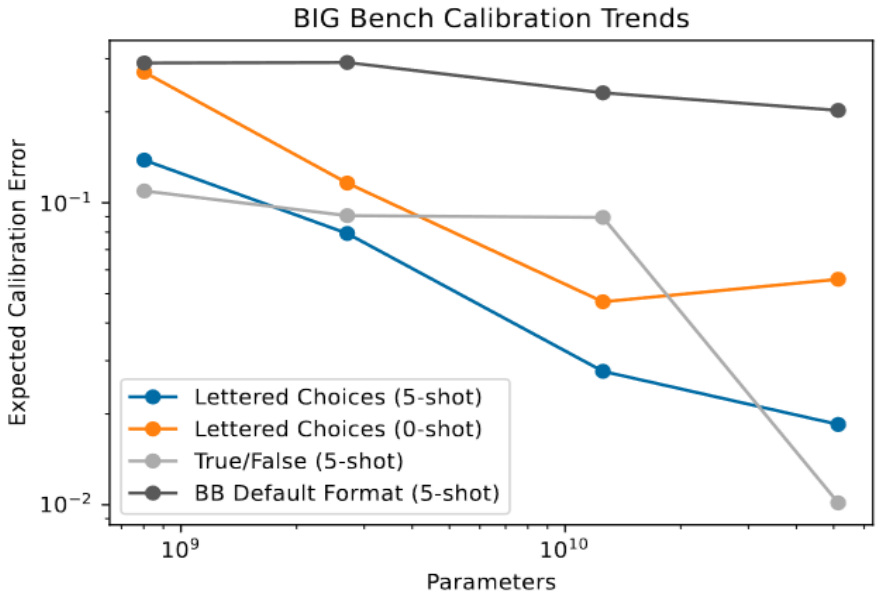
```
Question: Who was the first president of the United States?
Choices:
 (A) Barack Obama
 (B) George Washington
 (C) Michael Jackson
Answer:
```

And the model prediction should be (B).

Larger Models are Calibrated on Diverse Multiple Choice Questions

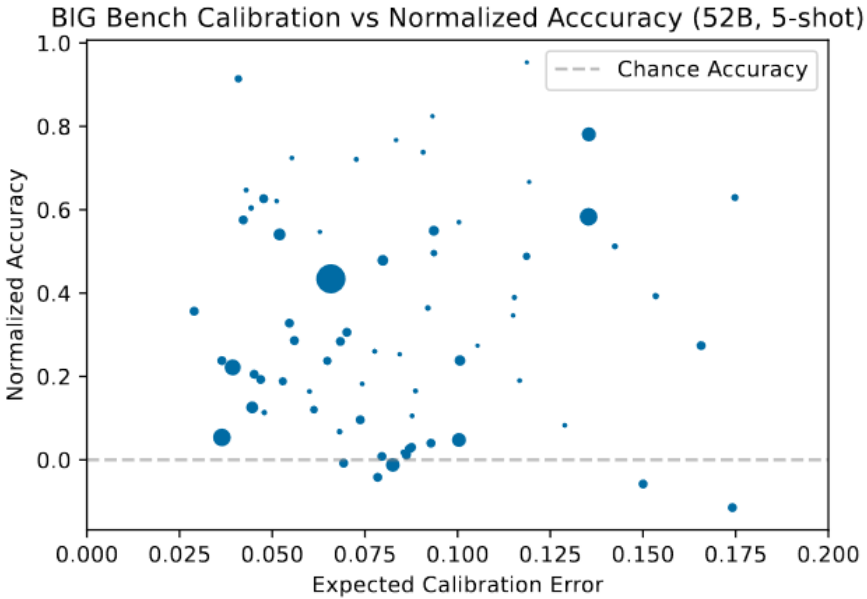
- calibration: Let the model answer the multiple choice questions!
  - Conclusion1: models are well-calibrated!

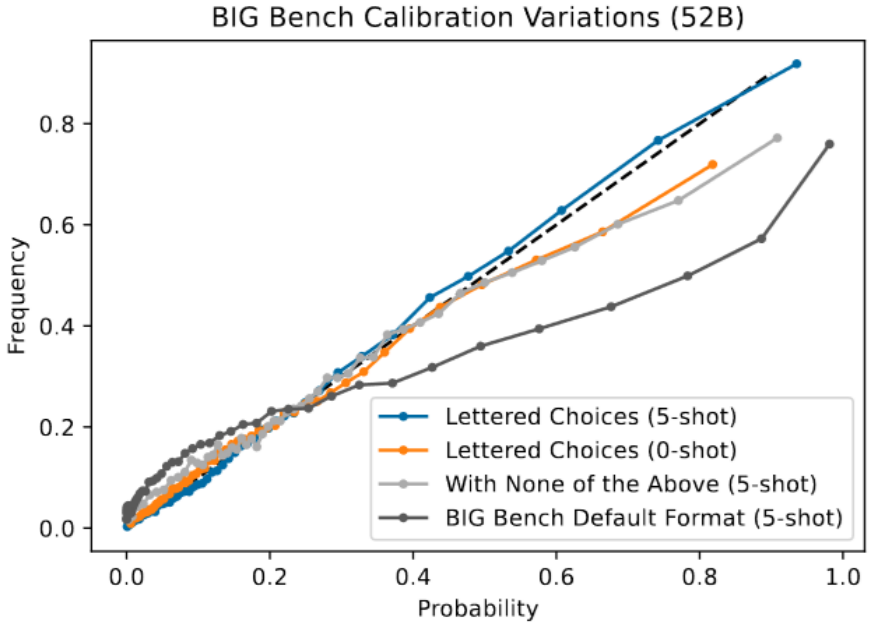




Larger Models are Calibrated on Diverse Multiple Choice Questions

- calibration: Let the model answer the multiple choice questions!
  - Conclusion2: no noticeable correlation between accuracy and calibration
  - Conclusion3: few-shot improves calibration, no use: None of the Above





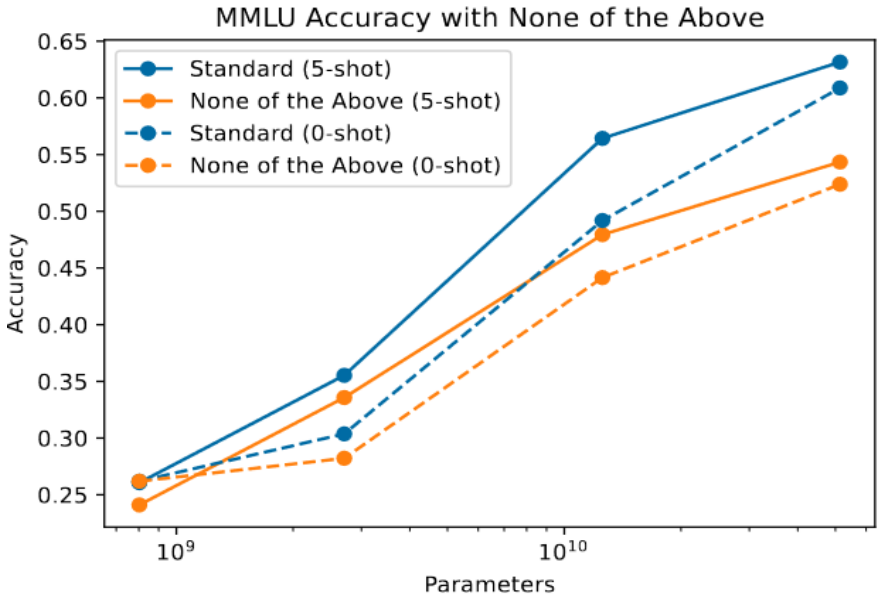
From Calibration to Knowing What You Know

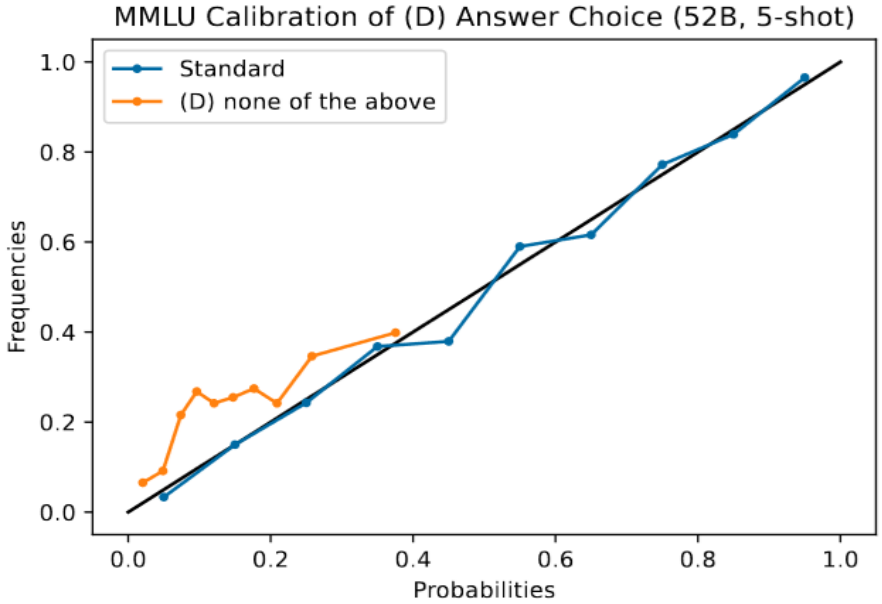
- Replacing an Option with ‘None of the Above’ Harms Performance and Calibration

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

From Calibration to Knowing What You Know

- Replacing an Option with 'None of the Above' Harms Performance and Calibration
  - Conclusion1: it harms the performance
  - Conclusion2: it makes calibration much worse





From Calibration to Knowing What You Know

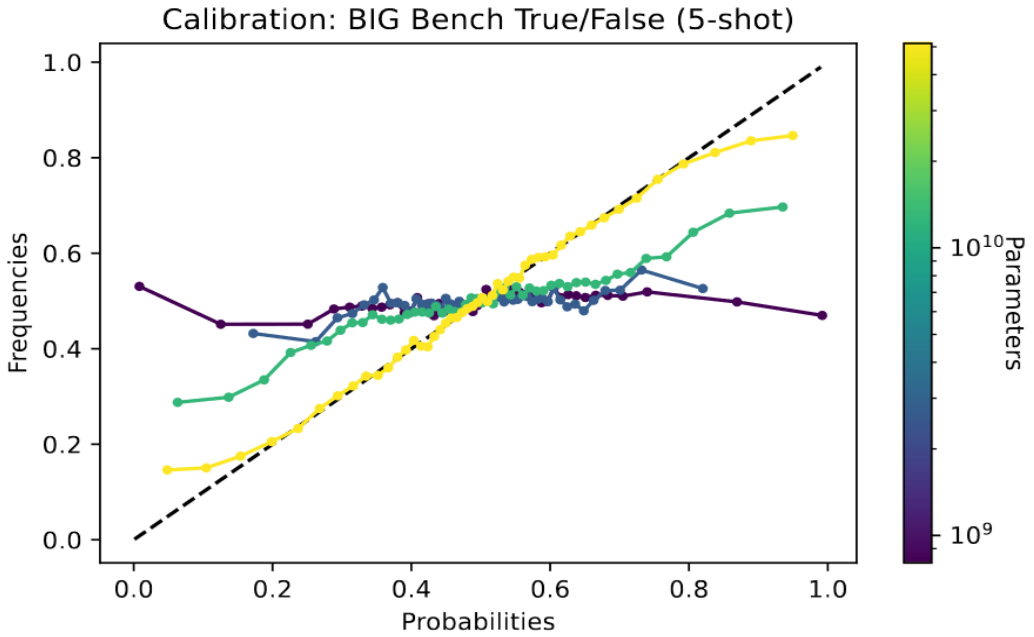
- Models are Well-Calibrated on True/False Tasks

```
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:
```

we expect either ‘ (A)’ or ‘ (B)’ as an answer

From Calibration to Knowing What You Know

- Models are Well-Calibrated on True/False Tasks
  - Conclusion1: The 52B model is very well-calibrated
  - Conclusion2: except near the tails, where it is overconfident.



From Calibration to Knowing What You Know

- RLHF Policy Miscalibration Can Be Remediated with a Temperature Tuning
  - Proximal Policy Optimization
  - Reinforment Learning from Human Feedbacks

$$L(s, a, \theta_k, \theta) = \min \left(\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)} A^{\pi_{\theta_k}}(s, a), \quad \text{clip} \left(\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)}, 1 - \epsilon, 1 + \epsilon \right) A^{\pi_{\theta_k}}(s, a) \right)$$

Advantage is positive: Suppose the advantage for that state-action pair is positive, in which case its contribution to the objective increases

Advantage is negative: Suppose the advantage for that state-action pair is negative, in which case its contribution to the objective reduces

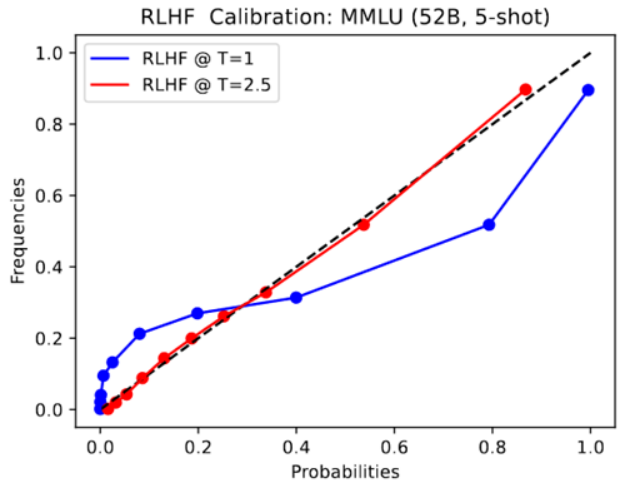
Ask the AI: Is your proposed answer True or False?

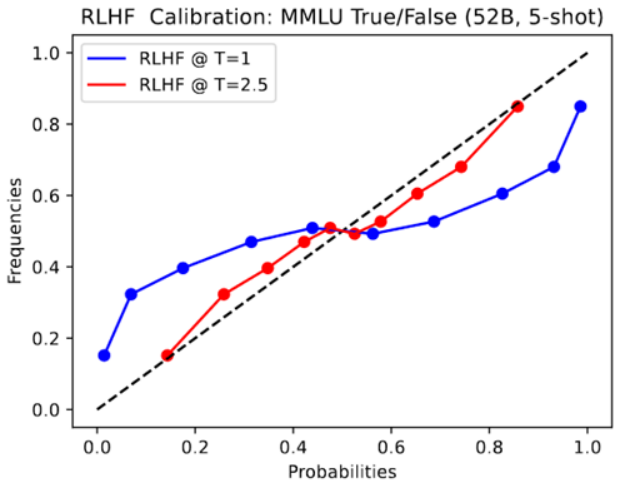
- RLHF Policy Miscalibration Can Be Remediated with a Temperature Tuning
  - Proximal Policy Optimization
  - Reinforment Learning from Human Feedbacks
  - Temperature: higher temperature encourage the model to generate more **new** content that is not in the source text.

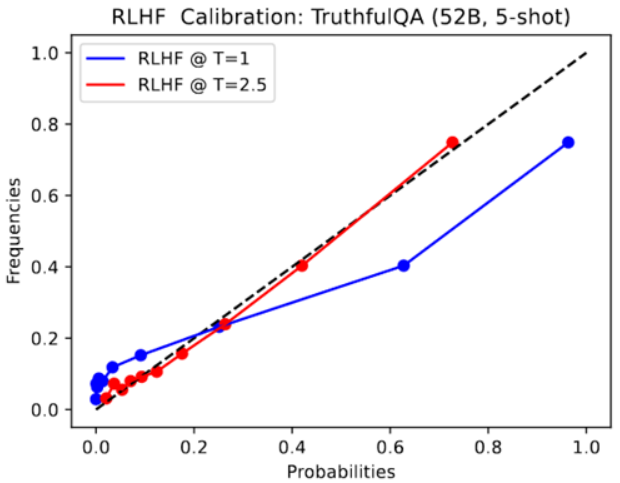
$$\frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

From Calibration to Knowing What You Know

- RLHF Policy Miscalibration Can Be Remediated with a Temperature Tuning
  - these policies naively appear very miscalibrated
  - a simple temperature adjustment helps a lot!







Ask the AI: Is your proposed answer True or False?

- Basic Self-Evaluation
  - First, ask the question:

```
Question: Who was the first president of the United States?
Answer:
```

  - Then, ask if it is correct:

```
Question: Who was the first president of the United States?
Proposed Answer: George Washington was the first president.

Is the proposed answer:
 (A) True
 (B) False
The proposed answer is:
```

Ask the AI: Is your proposed answer True or False?

- Showing Many T = 1 Samples Improves Self-Evaluation
  - improve performance further by showing the model other T = 1 samples, for comparison (Thomas, Johyn etc. 5 shots)

```
Question: Who was the third president of the United States?
Here are some brainstormed ideas: James Monroe
Thomas Jefferson
John Adams
Thomas Jefferson
George Washington
Possible Answer: James Monroe
Is the possible answer:
 (A) True
 (B) False
The possible answer is:
```

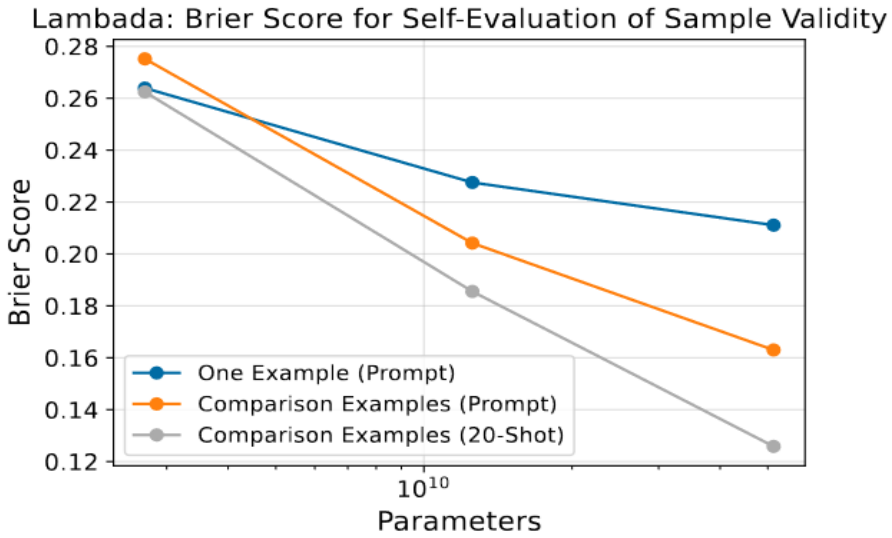
*Note: Brier Score for Multiple Classes:

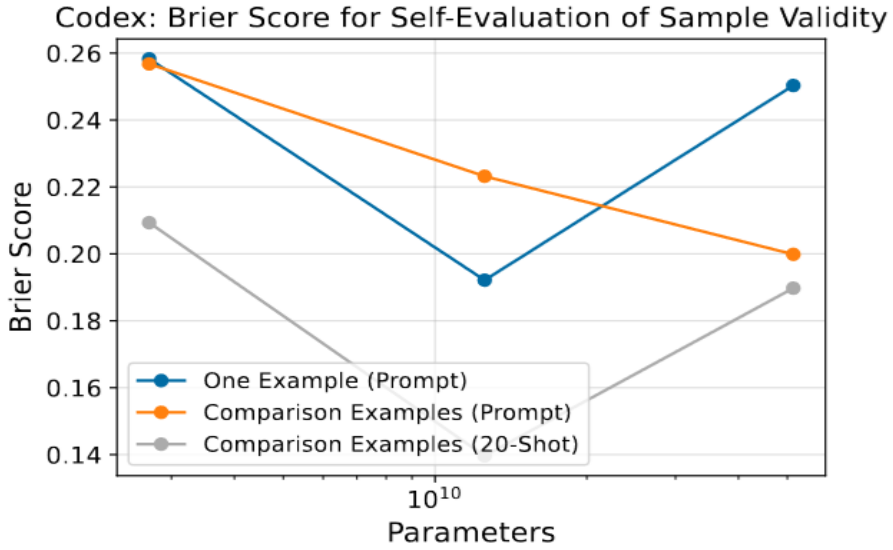
- f: prediction probability
- o: label (0 or 1)
- R: # classes
- N: # samples

$$BS = \frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{R}(f_{ti} - o_{ti})^2$$

Ask the AI: Is your proposed answer True or False?

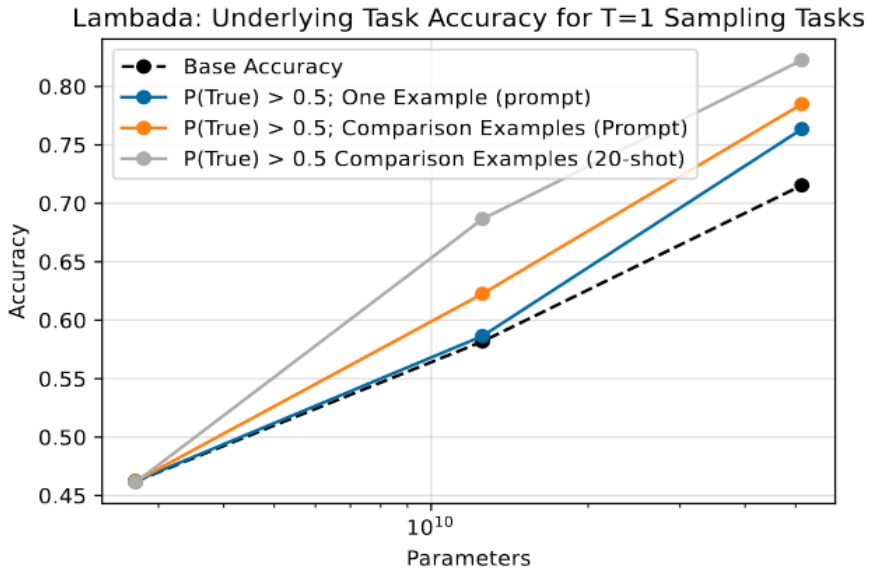
- Showing Many T = 1 Samples Improves Self-Evaluation
  - Conclusion2: Overall, if given a few examples from a given distribution, models can generate samples and then self-evaluate them to productively differentiate correct and incorrect samples, with reasonably calibrated confidence.

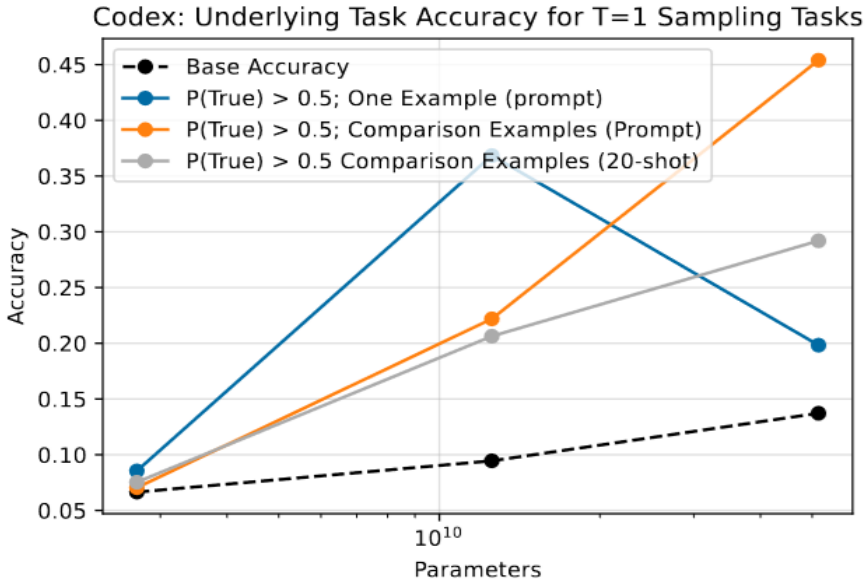




Ask the AI: Is your proposed answer True or False?

- Showing Many T = 1 Samples Improves Self-Evaluation
  - Conclusion2: conditional accuracies are substantially higher than the overall accuracy



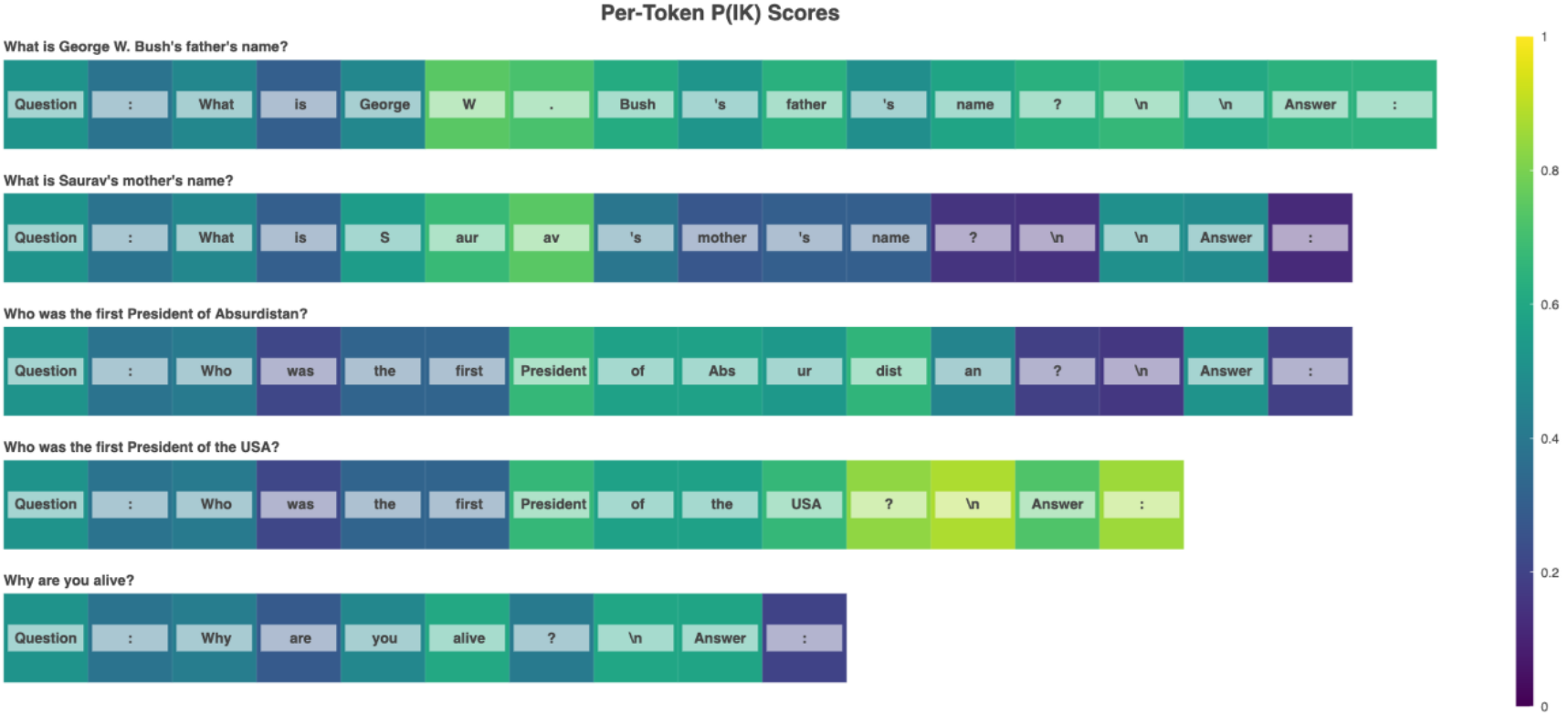


Training Models to Predict Whether They Can Answer Questions Correctly

- Showing Many T = 1 Samples Improves Self-Evaluation
  - **Value Head**: We train P(IK) as the logit from an additional value ‘head’ added to the model (independent of the logits for language modeling). An advantage of this approach is that we can easily probe P(IK) at general token positions.
  - **Natural Language**: We train P(IK) by asking the model to literally address "With what confidence could you answer this question?", and output an answer like 0%, 10%, 20%, · · · 100%.

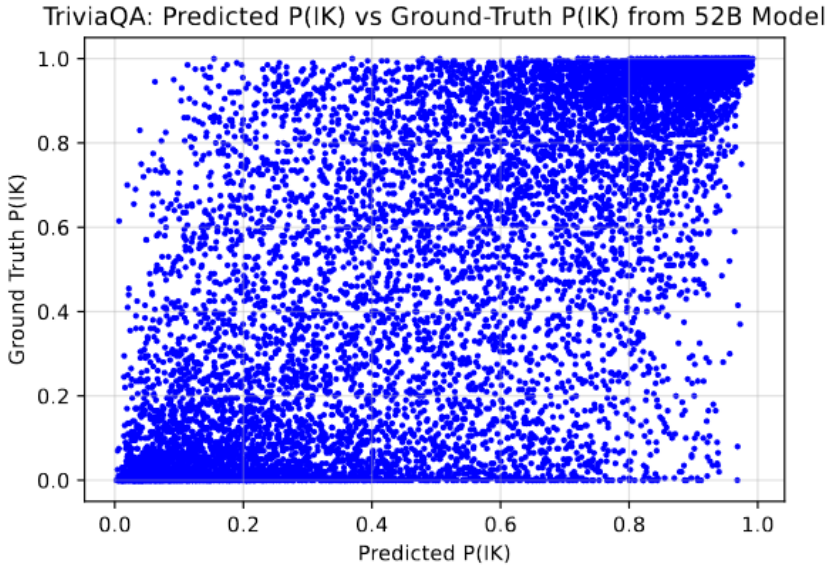
Training Models to Predict Whether They Can Answer Questions Correctly

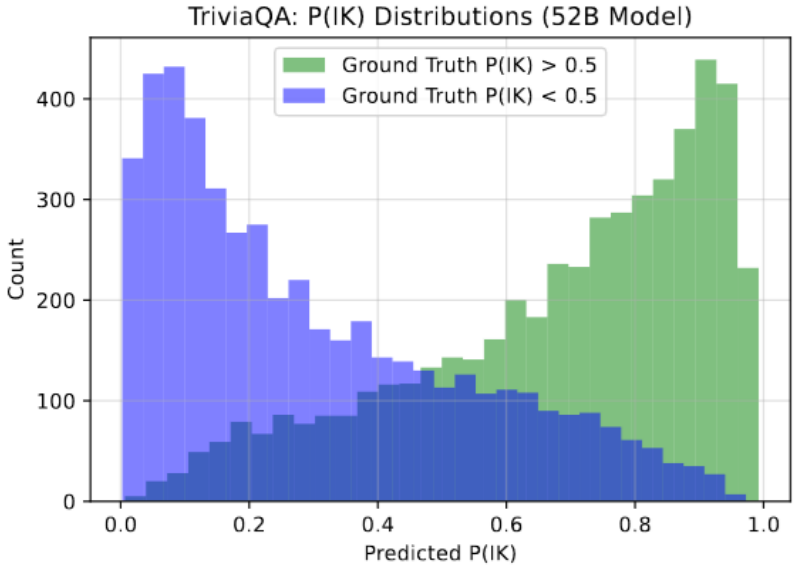
- Example: harder quesitons, lower P(IK) at the last token



Training Models to Predict Whether They Can Answer Questions Correctly

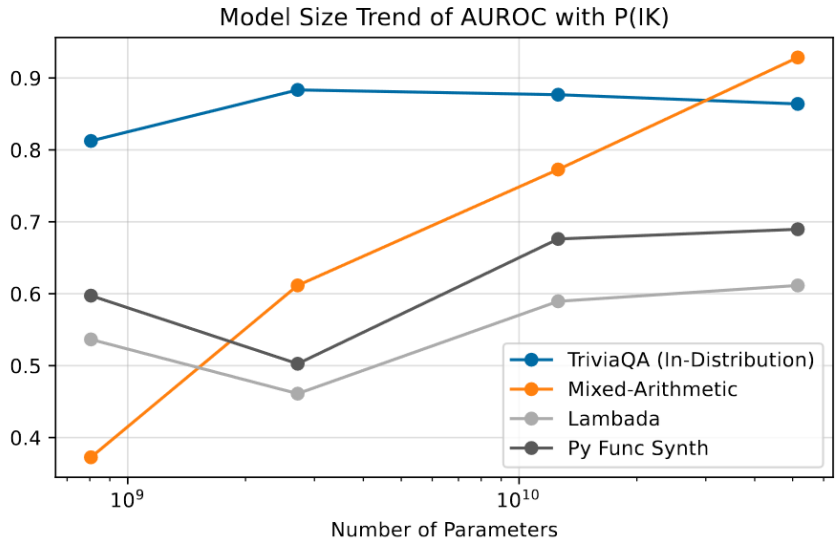
- Evaluating P(IK) Training and Model Size Trends
  - P(IK) is very well calibrated on TriviaQA





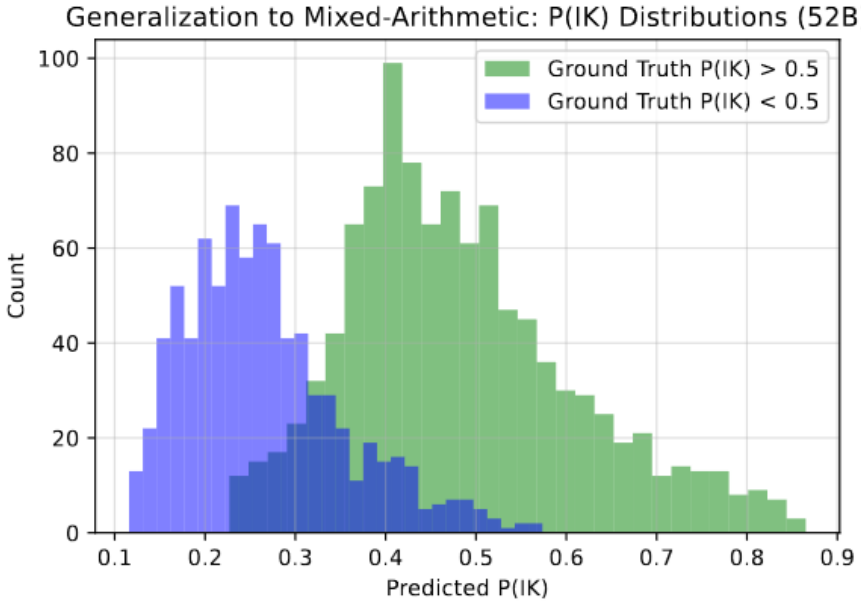
Training Models to Predict Whether They Can Answer Questions Correctly

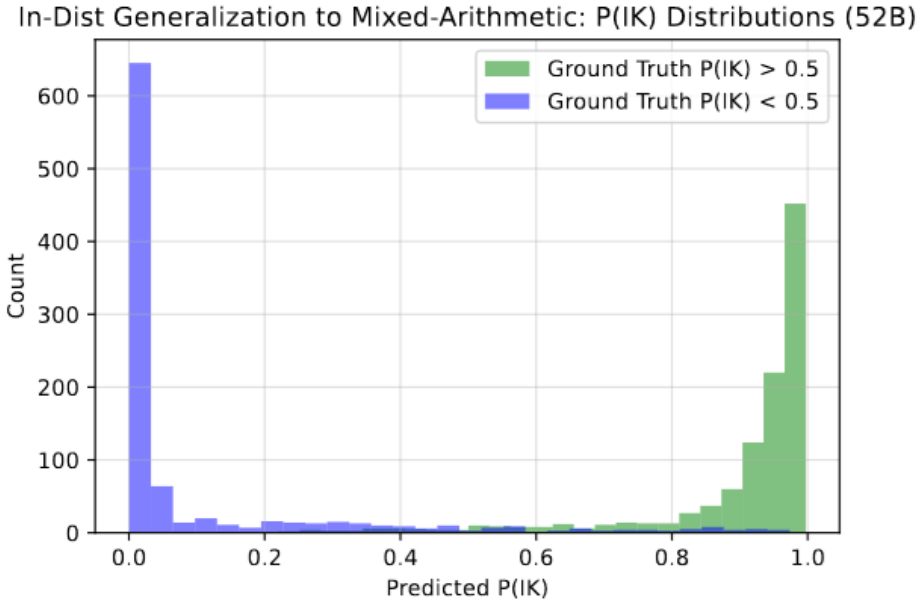
- Out of Distribution Generalization of P(IK)
  - the generalization of P(IK) when training only on TriviaQA and then evaluating on other datasets (Lmababa etc.)
  - Conclusion1: increasing AUROC as model size increases for all three out-of-distribution evals



Training Models to Predict Whether They Can Answer Questions Correctly

- Out of Distribution Generalization of P(IK)
  - Conclusion2: nontrivial generalization from TriviaQA to the other tasks, but training on the other tasks improves performance greatly





Other Discussions

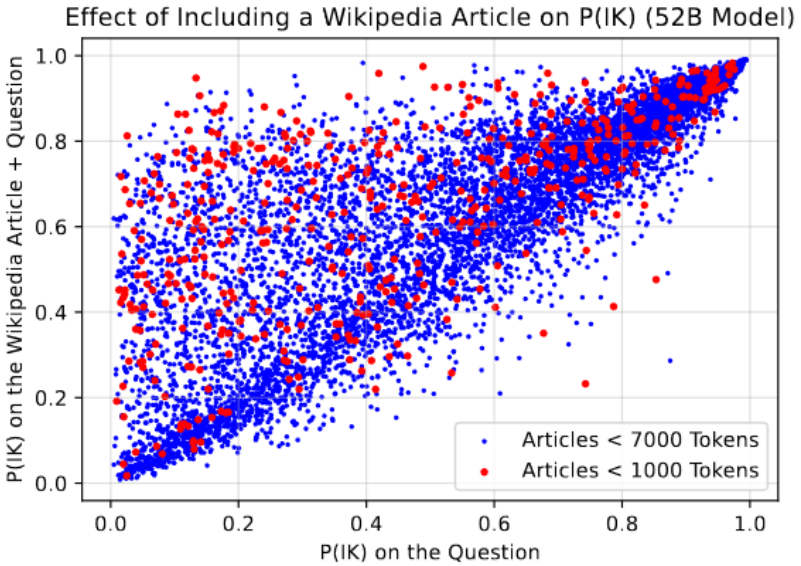
- P(IK) Generalizes to Account for Source Materials
  - then P(IK) appropriately predicts a low value, specifically 18% for a 52B model. However, if we prepend a Wikipedia article, then the P(IK) score rises to 78%.

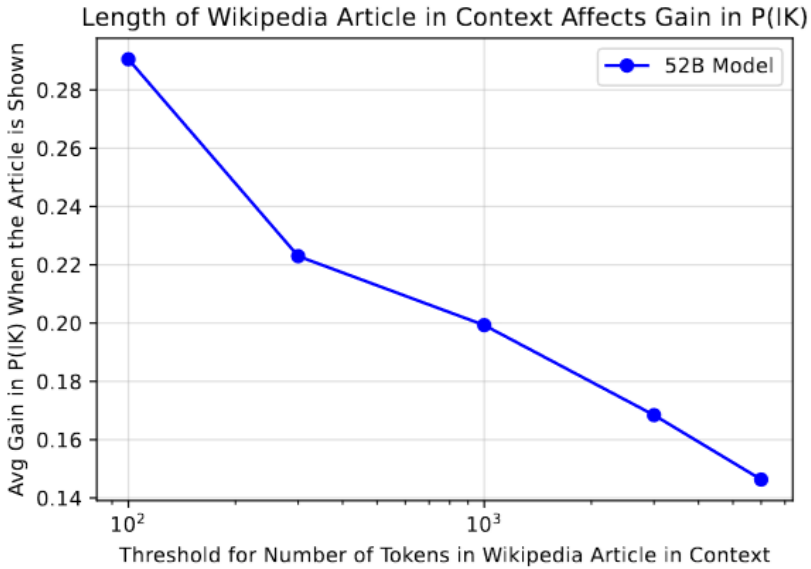
```
Wikipedia: The Idaho Rodeo Hall of Fame was established as a 501 (c)(3) non-
    profit organization on May 6, 2013. Lonnie and Charmy LeaVell are the
    founders of the organization. The actual charitable nonprofit status was
    received from the IRS on February 19, 2014. The IRHF hosts a reunion and
    induction ceremony annually every October. The Idaho Hall of Fame preserves
    and promotes the Western lifestyle and its heritage. The hall exists to
    dedicate the men and women in rodeo who contribute to ranching and farming
    through their sport. It also extends its reach to continue these western
    ways to the youth in the communities to ensure that these traditions
    continue for many generations. In 2015, the hall was awarded the Historic
    Preservation Recognition Award by National Society of the Daughters of the
    American Revolution.

What state's rodeo hall of fame was established in 2013?
```

Other Discussions

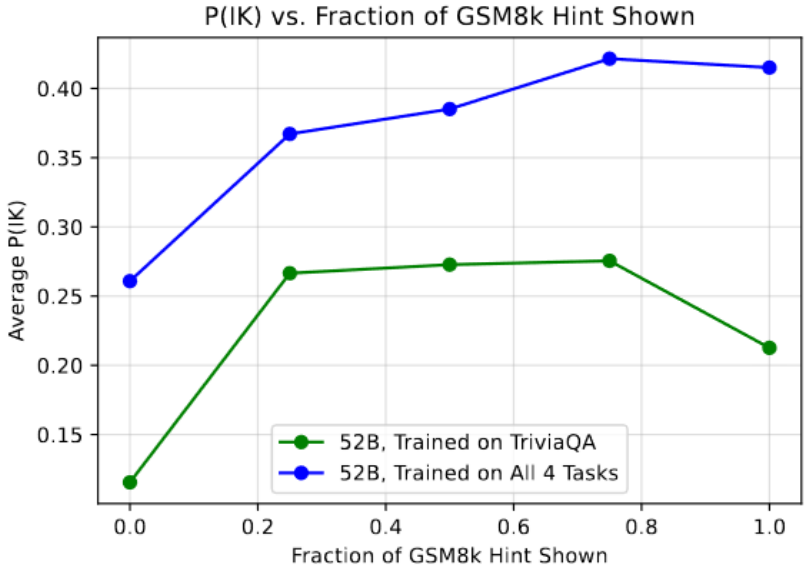
- P(IK) Generalizes to Account for Source Materials
  - then P(IK) appropriately predicts a low value, specifically 18% for a 52B model. However, if we prepend a Wikipedia article, then the P(IK) score rises to 78%.

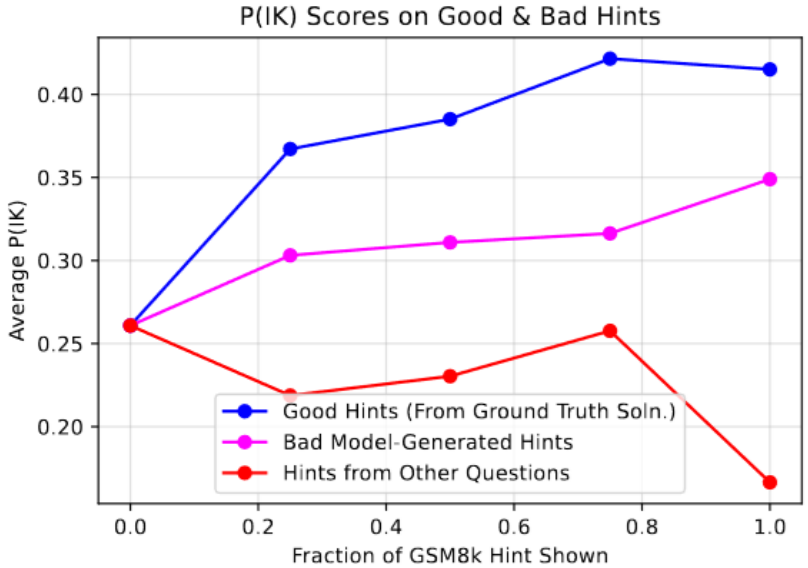




Other Discussions

- P(IK) Generalizes to Account for Hints Towards GSM8k Solutions
  - Conclusion1: showing more of the hint generally leads to higher P(IK),
  - Conclusion2: good hints that lead to the correct answer result in higher P(IK) scores than bad hints,





Other Discussions

- Comparing Models Trained with Distinct Pretraining Distributions
  - Is P(IK) really showing self-knowledge or just the difficulty of the task?
  - 2 same models: A was trained with four repetitions of a high quality dataset, while B uses a single copy lowerquality distribution of webdata.
  - Table shows: there is some signal that P(IK) is encoding model-specific information (because the samples are the same)
  - However, the difference is not obvious 6%

	Questions that only A gets right	Questions only B gets right
A's average P(IK)	0.463	0.408
B's average P(IK)	0.409	0.477

Other Discussions

- Comparing Models Trained with Distinct Pretraining Distributions
  - training a P(IK) classifier to predict whether or not model A knows the answer to a question would work better when starting from model A itself as the initial checkpoint, as compared to if we start from model B
  - starting from pretrained model X should do better than starting from model Y when training P(IK) using data from model X

	Test on Ground-Truth from Model A AUROC / Brier Score	Test on Ground-Truth from Model B AUROC / Brier Score
Starting from Model A	0.8633 / 0.1491	0.8460 / 0.1582
Starting from Model B	0.8631 / 0.1497	0.8717 / 0.1443

Conclusions and Comments

- Quesiton: Language Models Know What They Know ?
- Calibration: Multiple Choice, None of the Above, and True/False
- Self-Evaluation of Model-Generated Samples, without Finetuning
- Finetuning to Identify the Questions Models Can Correctly Answer

This paper investigate an important issue. ChatGPT sometimes cannot confess it limitations. What if we ask the P(IK) to ChatGPT?

I think the analysis is thorough, I can see their motivation and storyline, however, needs more proof reading.