Session #27:
Calibrating Self-Supervised Models

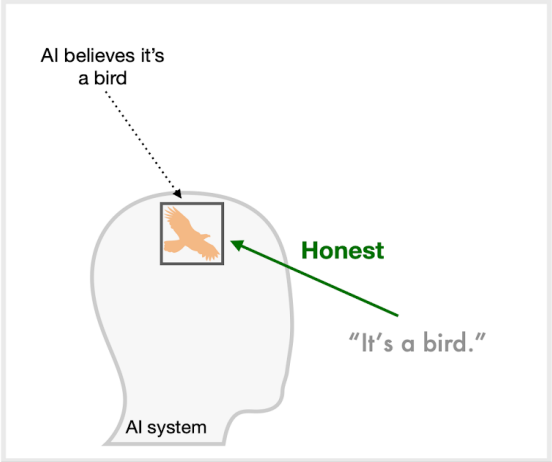
Tuesday, November 29
CSCI 601.771: Self-supervised Statistical Models

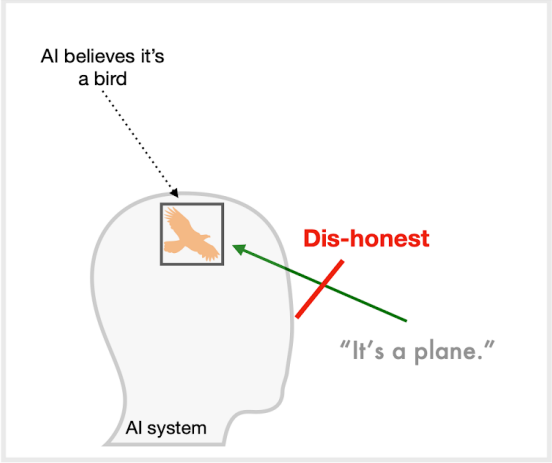
Motivation

- Desire AI systems that are **honest**
  - They know what they do
    and don't know with appropriate confidence

What is honest AI?

- If AI says S, then it believes S.
- Verify by checking if S matches belief.





: Iliana

Motivation

- Desire AI systems that are **honest**
    - They know what they do
      and don't know with appropriate confidence

- Experiments
    - Calibration
    - Self-evaluation
    - Self-knowledge

✍: Iliana

Calibration

- Calibrated when probability assigned to outcomes matches the frequency of the actual outcomes.

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

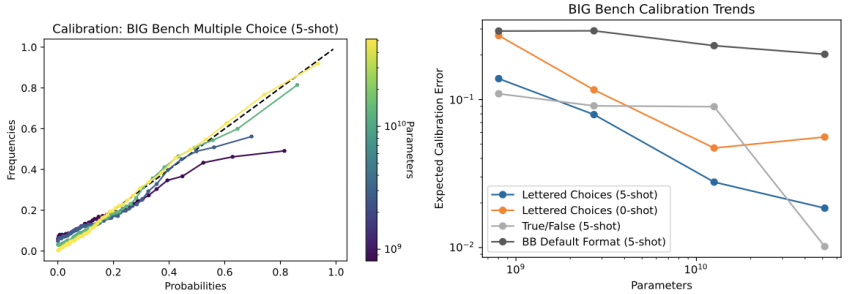


Figure 4 **(left)** We show calibration curves for various model sizes on all of the multiple choice tasks in BIG Bench, in the format described in section 2. We include a dashed line indicating perfect calibration. **(right)** Here we show trends in the expected calibration error on BIG Bench, for both multiple choice and a separate True/False format (see Section 3.2). We show the RMS calibration error in Figure 21 in the appendix.

Knowing What You Know

- Does the model know if each answer option is correct?
  - Include a "none of the above" option

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

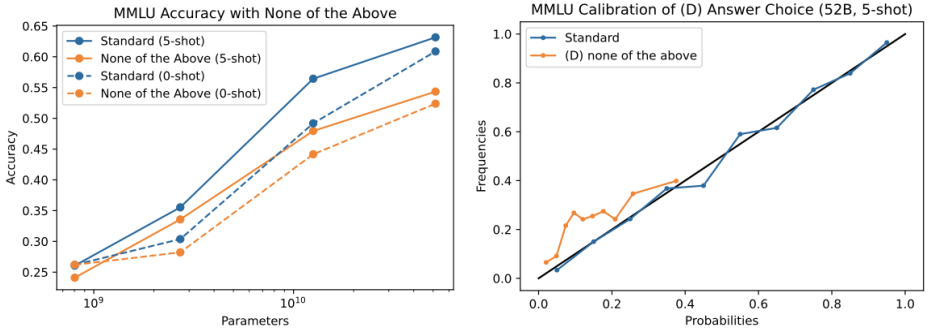


Figure 7 **(left)** We show accuracy on MMLU in the standard format, and after replacing option (D) with "none of the above". This replacement decreases accuracy very significantly. **(right)** We show calibration specifically for the (D) answer option, in the standard form of MMLU and with (D) as "none of the above". The latter makes calibration much worse, and in particular the model seems strongly biased against using this option, which also harms accuracy.

Knowing What You Know

- Switch to True/False
- Improves calibration for large models

```
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:
```

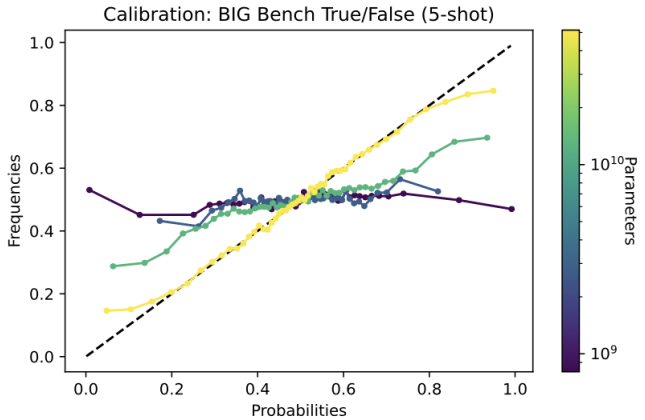


Figure 8 We show calibration curves for various model sizes on all of the multiple choice tasks in BIG Bench, reformulated as True/False questions on a mix of the correct answers, and randomly chosen incorrect answer options. The 52B model is very well-calibrated except near the tails, where it is overconfident.

Is Your Answer True or False?

- Can we ask language models about their own outputs?
  - Ask Question

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

  - Sample a model response

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

  - Ask model about sampled response

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

- Samples from smaller models are easier to categorize as correct/incorrect
- Zero-shot P(True) is poorly calibrated
  - ~ 50% for most samples

Is Your Answer True or False?

- Generate 5 sample answers from model and ask about validity of one.

```
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:
```

- Improves performance, but is still poor for zero-shot.

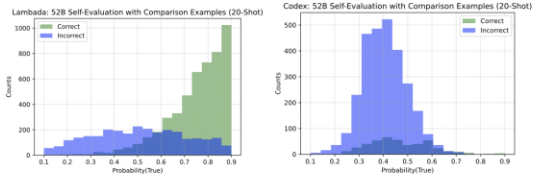


Figure 10 Models self-evaluate their own samples by producing a probability P(True) that the samples are in fact correct. Here we show histograms of P(True) for the correct and incorrect samples, in the evaluation paradigm where models also see five $T = 1$ samples for the same question, in order to improve their judgment. Here we show results only for Lambada and Codex, as these are fairly representative of short-answer and long-answer behavior; for full results see Figure 28 in the appendix.

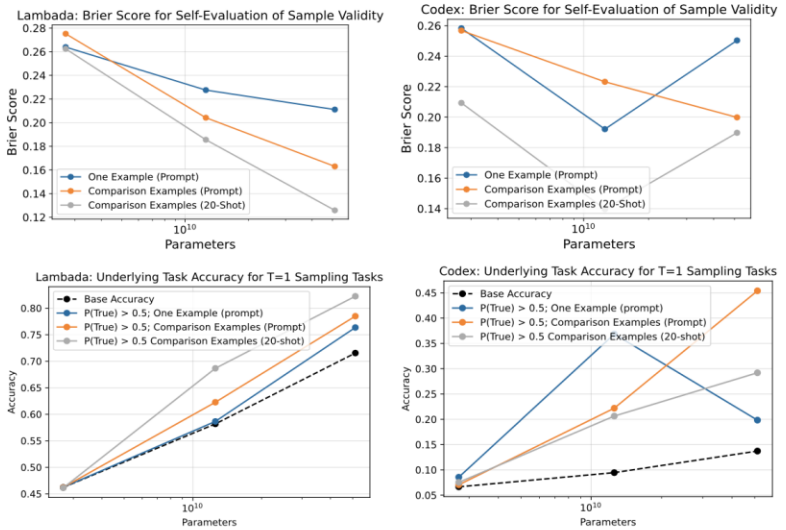


Figure 11 Here we show results only for Lambada and Codex, as these are fairly representative of short-answer and long-answer behavior; for full results see Figures 28, 29, 30, and 31 in the appendix. **Top**: Here we show the Brier scores for model self-evaluation with three methods: basic self-evaluation with a prompt, and self-evaluation with comparison samples, either with a fixed prompt or 20-shot. Note that the Brier score combines accuracy of the True/False determination with calibration, and 20-shot evaluation with comparison samples performs best in every case. Brier scores do not decrease with model size on evaluations like Codex because small model samples are almost always invalid, so that it's relatively trivial to achieve a small Brier score. **Bottom**: We show the base accuracy of our models on various sampling tasks, and then the accuracy among the responses where via self-evaluation we have P(True) > 0.5. For P(True) we evaluate with a single example and a prompt, and then both 20-shot and with a prompt with five comparison examples. Few-shot evaluation is important for obtaining good calibration.

✍: Iliana

Prediction of Knowledge

- Two approaches
  - Value Head – train P(IK) as logit from additional value "head" added to the model
  - Natural Language – train P(IK) by asking model what confidence they could answer a question
- Train model to predict whether they know the answer to a question
  - P(IK) - Probability of "I know the answer"

Figure 3 Examples of P(IK) scores from a 52B model. Token sequences that ask harder questions have lower P(IK) scores on the last token. To evaluate P(IK) on a specific full sequence, we simply take the P(IK) score at the last token. Note that we only train P(IK) on final tokens (and not on partial questions).

Prediction of Knowledge

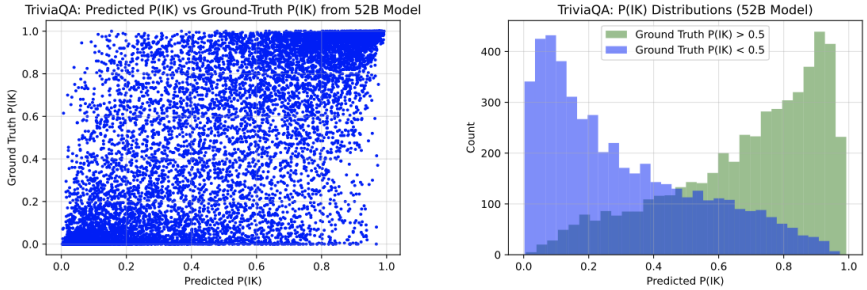


Figure 12 Testing a 52B classifier on a held-out set of TriviaQA questions. We see that the classifier predicts lower values of P(IK) for the questions it gets incorrect, and higher values of P(IK) for the questions it gets correct. We set the ground truth P(IK) as the fraction of samples at $T = 1$ that the model gets correct.

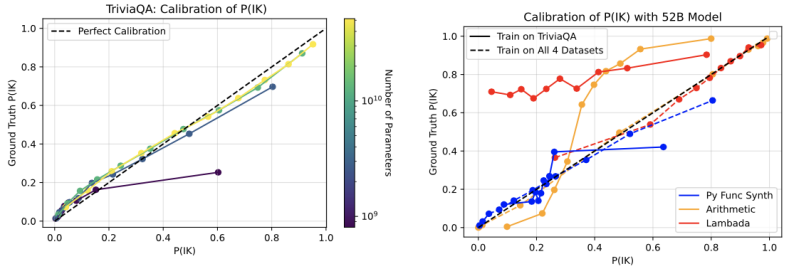


Figure 13 Left: Full calibration plot of P(IK) classifiers on TriviaQA over a range of model sizes. We see that the smallest models have higher calibration error than the biggest models. The larger classifiers are very well calibrated in-distribution. Right: We show calibration curves for P(IK) on three other sampling-based datasets, both in-distribution and out-of-distribution (trained only on TriviaQA). We see that OOD calibration of P(IK) is often quite poor, and for the most part models are underconfident.

Prediction of Knowledge

- For obscure questions, does including source material help P(IK)?
  - Compute P(IK) with background material

```
Here is some background information material: <BACKGROUND_MATERIAL>
Now, answer the following question.
Question: Which Lloyd Webber musical premiered in the US on 10th
    December 1993?

Answer:
```

  - Compute P(IK) without background material

Without Background Material:

```
Question: Which Lloyd Webber musical premiered in the US on 10th
    December 1993?

Answer:
```

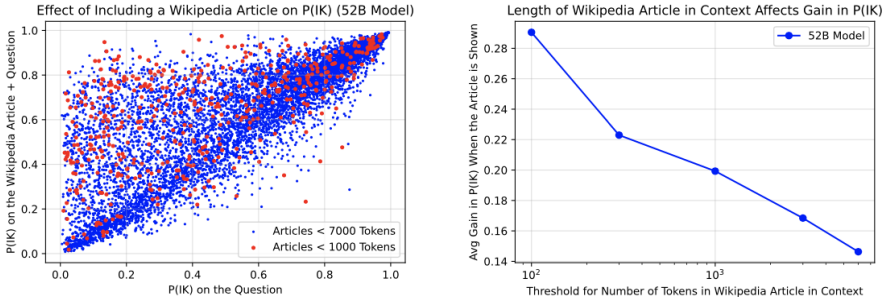


Figure 18 Effect of including Wikipedia article on P(IK) for TriviaQA Questions. We see that including a relevant Wikipedia article in the context boosts the average P(IK) on TriviaQA Questions. P(IK) increases more for shorter Wikipedia articles, from which it is presumably easier to identify the relevant facts.

Prediction of Knowledge

- For the hardest dataset GSM8k, can adding hints help performance?

```
Question: Students in class 3B are collecting school points for behavior
    . If they get enough points in total, they can go on a trip. In the
    class, there are Adam, Martha, Betty, and Tom. Adam has collected 50
     points. Betty was better than Adam and collected 30% more. Marta
    managed to collect 3 times more points than Tom, who has 30 points
    less than Betty. How many points is the class missing to go on the
    trip if the minimum threshold is 400 points?
Here is a hint: Betty has 30% more points than Adam, so it's 30/100 * 50
     = <<30/100*50=15>>15 points more.
Betty's total is therefore 50 + 15 = <<50+15=65>>65 points.
Tom has 30 points less than Betty, so he has 65 - 30 = <<65-30=35>>35
    points.
Marta has 3 times more points than Tom, so she has 3 * 35 =
    <<3*35=105>>105 points.
In total, all students collected 50 + 65 + 35 + 105 =
    <<50+65+35+105=255>>255 points.
So the class is missing 400 - 255 = <<400-255=145>>145 points to go on
    the trip.

Answer:
```

- Two types of hints
  - Vary amount of information in hint
  - Create good, incorrect, and distracting hints

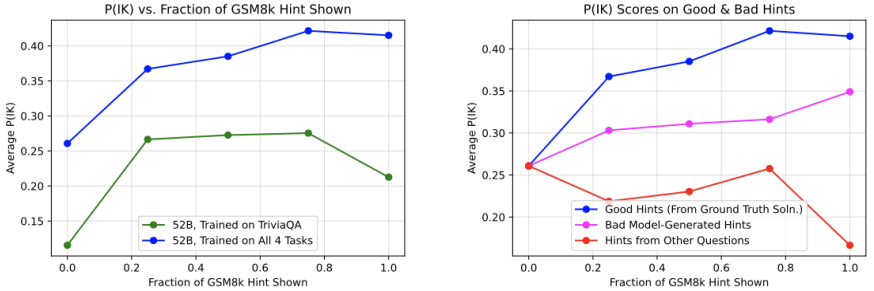


Figure 19 Effect of hints on P(IK) applied to GSM8k – all of these results represent generalization, as the models were not trained on GSM8k. Left: We see that showing more of the GSM8k hint results in higher P(IK). The effect is more consistent for the model trained on all 4 tasks (TriviaQA, LAMBADA, Arithmetic, and Python Function Synthesis), rather than the one trained on just TriviaQA. Right: We evaluate the model trained on all 4 other tasks on various hints. We see lower P(IK) scores for bad hints (though the models are partially fooled), and actual decreases in the P(IK) score when the hints are irrelevant because they come from other questions.

Prediction of Knowledge

- Is the P(IK) model truly capturing self-knowledge?
- Train two models on different data distributions
  - Four repetitions of high-quality dataset
  - Single copy of high-quality dataset mixed with lower-quality distribution of web data
- Each model has higher P(IK) on questions that model alone gets right compared to the other model's uniquely correct questions
- Second experiment – finetune both models on ground truth from each model.

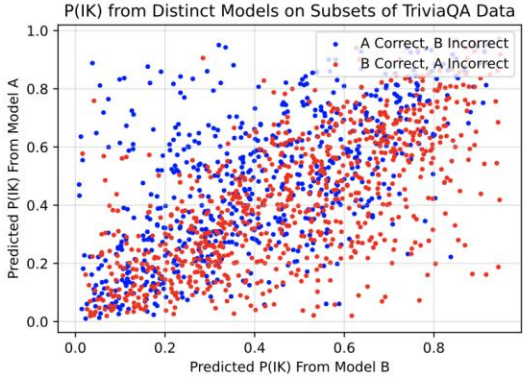


Figure 20 Scatterplot that disambiguates numbers in Table 2. We evaluate 2 distinct 12B models on TriviaQA, separating the data into questions each model gets correct and the other gets incorrect. The scatterplot depicts the P(IK) scores from each model on both of these data subsets.

	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

Table 3 'Cross-Experiments': We trained P(IK) classifiers on the ground-truth data from two models, starting from both models. Ideally, starting from pretrained model X should do better than starting from model Y when training P(IK) using data from model X. We see some signal that starting from model B does better than starting from model A when testing on data from model B. However, we see no difference between both initializations when testing on data from model A.

Empiricists

- https://colab.research.google.com/drive/1YIcelfDdNx6WxnK-T3wwinWtEI_HkVC9?usp=sharing

Strengths

- Quite exhaustive (models, datasets, experiments)
- Levels of study:
  - Simple Calibration tests
  - Asking models to evaluate their own predictions; calibration tests on top.
  - …
- Metacognition/Meta question: a deeper level

Weaknesses

- Did not attempt to address some findings like:
  - "None of the above"
  - Temperature adjustment fix
- Natural language P(IK) seems unmotivated and unnecessarily mentioned.
- Calibration vs Prompt Engineering.
- Focusing on history, not many up-to-date knowledge
  - Do models know whether they know who is the president of US now?

Pros and cons

- Diagram which support its idea and make comparison.
- The article did not mentioned how they select the hyperparameter of the training model, and they just show the results of the mathematical , but no proof.
- There are so many definition variables which might make the article more complicated
- We suppose there only one answer is correct
  - Models are Well-Calibrated on True/False Tasks

Visionary

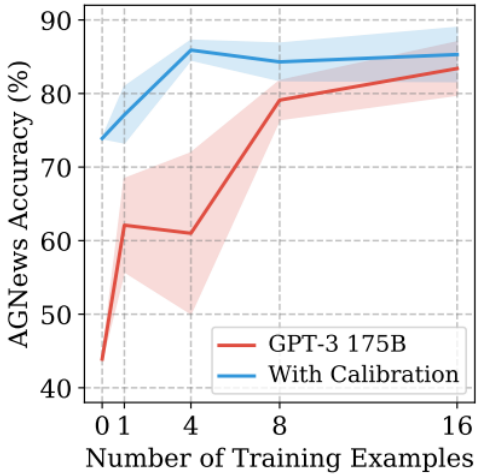
- Calibration
  - Bucket
  - Abstention
- Acc: accuracy
- Conf: model confidence (e.g. P(IK) or P(True))

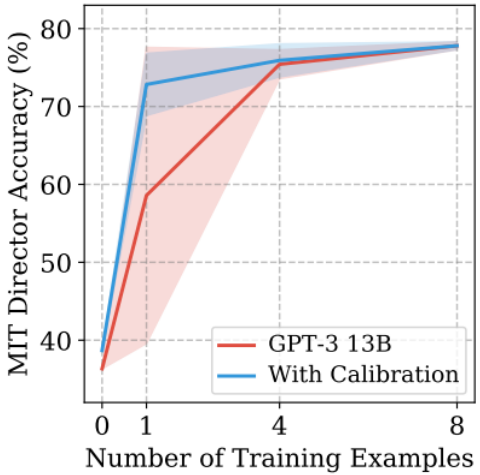
$$P(\hat{Y} = Y | P_N(\hat{Y}|X) = p) = p, \forall p \in [0, 1].$$

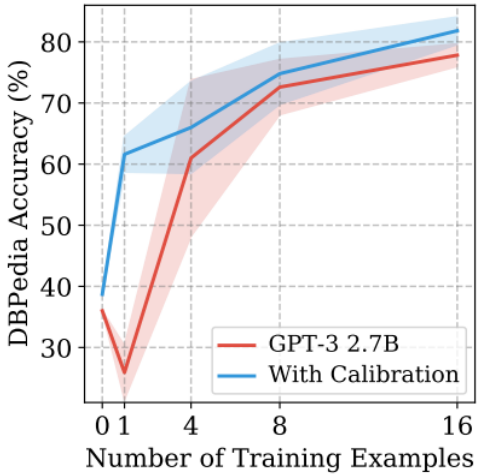
$$\sum_{m=1}^{M} \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|,$$

Visionary

- Calibration methods
  - **Calibrate P(True), "calibrate before use"**







Visionary

- Calibration methods
  - Calibrate P(True)
  - **P(IK) or P(True) is baseline, better calibration methods?**
    - **Sensitivity**
    - **Mutual Information**
    - **Flatness**

Beyond Calibration: Uncertainty Quantification/Conformal Prediction

- Guarantee to cover the true value with high probability
  - P(prediction in a **set**) > some probability
  - For a **desired probability,** what is the set in which the answer is guaranteed to exist?
- More detailed understanding of model probability space

{ Chihuahua, toy terrier, Italian greyhound, Boston bull, miniature pinscher }

{ English springer, Welsh springer spaniel, collie, boxer, Saint Bernard, Leonberg }

{ face powder, hamper, lotion, packet, shopping basket }

https://towardsdatascience.com/conformal-prediction-4775e78b47b6

Is Model Soup Well Calibrated?

- Model Soup:
  - Averaging weights of multiple models with different parameters
- AdaMix:
  - Mixture of Adapters
- Extension:
  - How well are these models calibrated?
  - What about Model soup of LLMs?