北京大學
PEKING UNIVERSITY

Calibration, Self-Evaluation, Hallucination

On Calibration of Modern Neural Networks (2017)

北京大学 PEKING UNIVERSITY

Calibration

Predicting probability estimates representative of the true correctness likelihood

- Perfect calibration

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

- Bin

$$\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i), \qquad \text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i,$$

- Expected Calibration Error (ECE)

$$\mathop{\mathbb{E}}_{\hat{P}}\left[\left|\mathbb{P}\left(\hat{Y} = Y \mid \hat{P} = p\right) - p\right|\right] \qquad \text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|,$$

- Maximum Calibration Error (MCE)

$$\max_{p \in [0,1]} \left|\mathbb{P}\left(\hat{Y} = Y \mid \hat{P} = p\right) - p\right| \qquad \text{MCE} = \max_{m \in \{1,\ldots,M\}} |\text{acc}(B_m) - \text{conf}(B_m)|$$

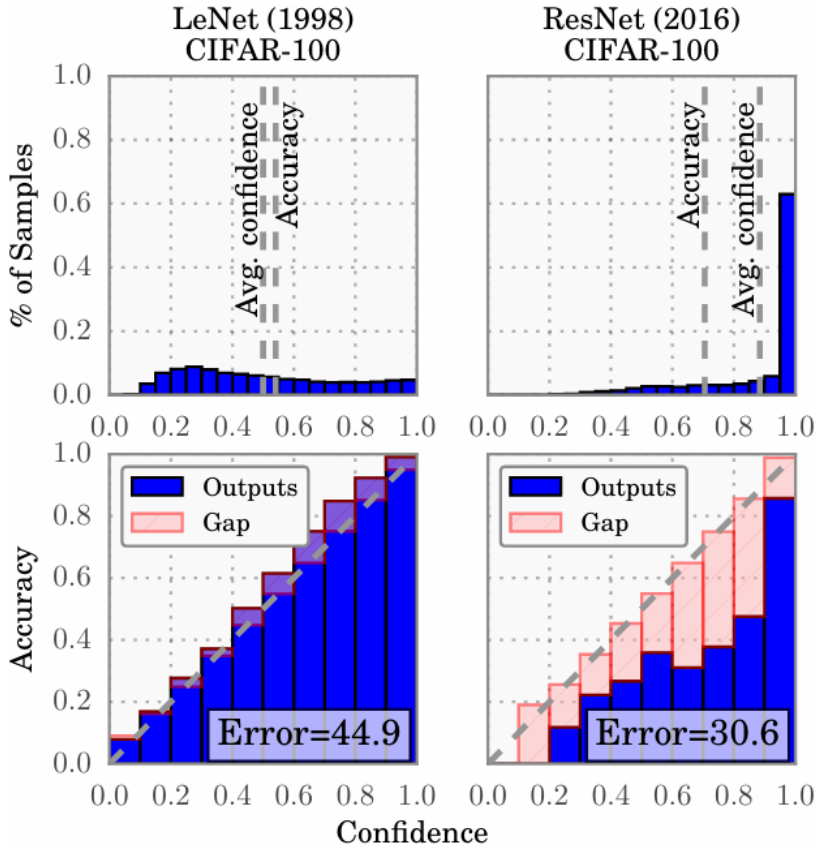

Figure 1. Confidence histograms (top) and reliability diagrams (bottom) for a 5-layer LeNet (left) and a 110-layer ResNet (right) on CIFAR-100. Refer to the text below for detailed illustration.

北京大學
PEKING UNIVERSITY

Language Models (Mostly) Know What They Know

Anthropic (2022)

Language Models (Mostly) Know What They Know (2022)

北京大学 PEKING UNIVERSITY

Calibration for LMs

- Multiple Choice Tasks

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

By default in BIG Bench [Srivastava et al., 2022] questions are posed in this way:

```
Question: Who was the first president of the United States?
 choice: Barack Obama
 choice: George Washington
 choice: Michael Jackson
Answer:
```

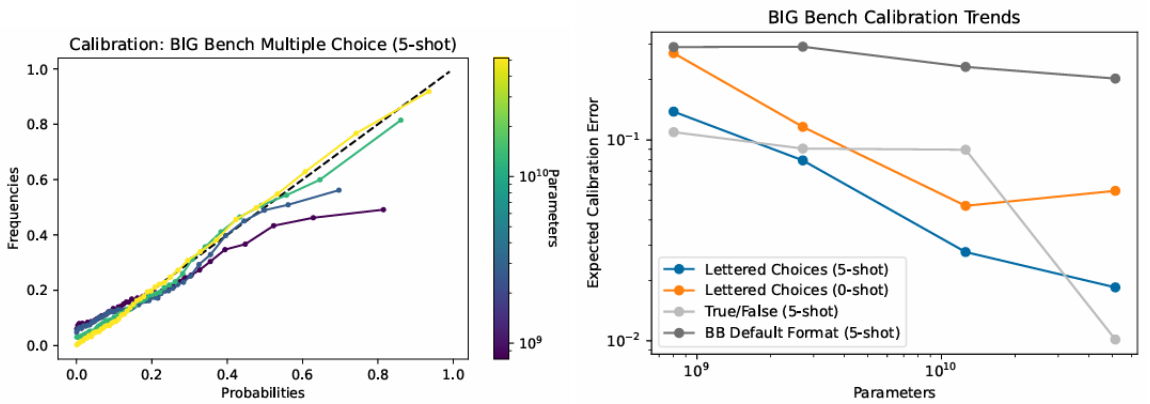

Figure 4 **(left)** We show calibration curves for various model sizes on all of the multiple choice tasks in BIG Bench, in the format described in section 2. We include a dashed line indicating perfect calibration. **(right)** Here we show trends in the expected calibration error on BIG Bench, for both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

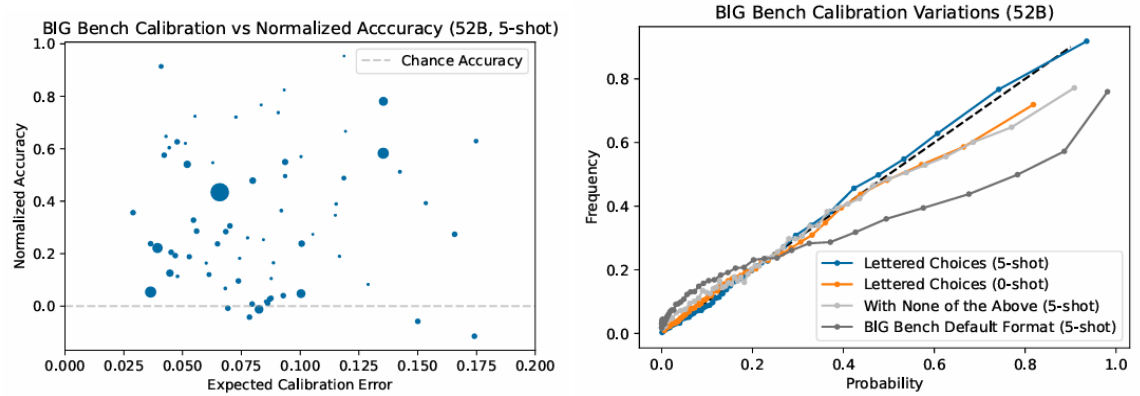

Figure 5 **(left)** We show expected calibration error versus normalized accuracy for all BIG Bench tasks; the number of problems in each task is represented by the marker sizes. We do not find any noticeable correlation between accuracy and calibration within BIG Bench. To normalize accuracies we linearly map chance accuracy to 0, keeping perfect accuracy at 1. **(right)** We compare calibration for several variations on BIG Bench evaluations: we vary between 0-shot and 5-shot, replace an answer option with "none of the above", and compare our format with letter-labeled choices to the default BIG Bench formatting.

Language Models (Mostly) Know What They Know (2022)

北京大学 PEKING UNIVERSITY

Calibration for LMs

- None of above

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

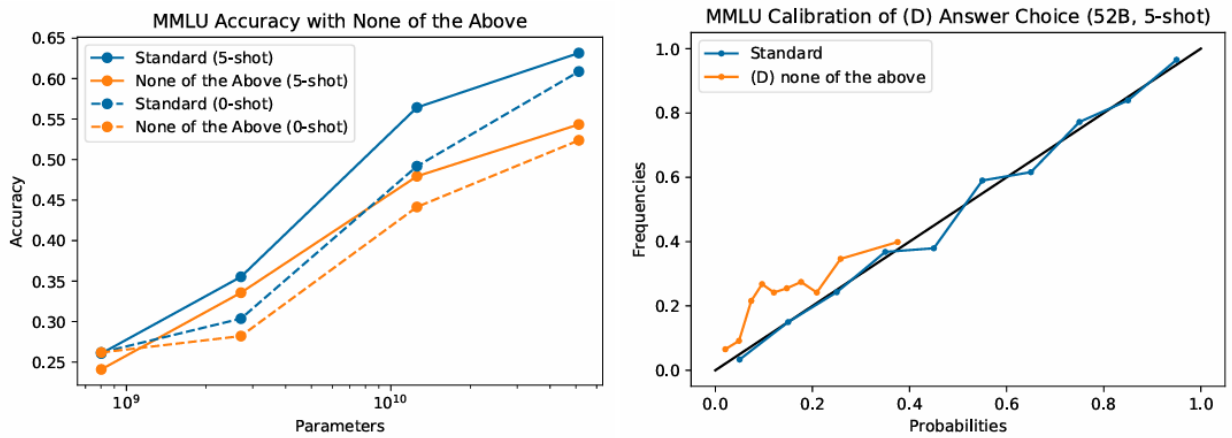

Figure 7 **(left)** We show accuracy on MMLU in the standard format, and after replacing option (D) with "none of the above". This replacement decreases accuracy very significantly. **(right)** We show calibration specifically for the (D) answer option, in the standard form of MMLU and with (D) as "none 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?

- RLHF Post-training

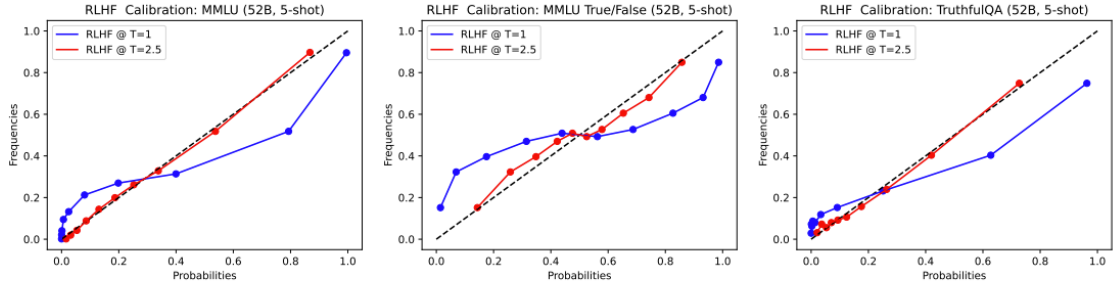

Figure 9 We show calibration curves for RLHF policies finetuned from our language models. Calibration of these models appears to be very poor, but simply adjusting the temperature of their probability distributions to $T = 2.5$ largely fixes calibration issues for three different evaluations.

Language Models (Mostly) Know What They Know (2022)

Calibration for LMs

- True / 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:
```

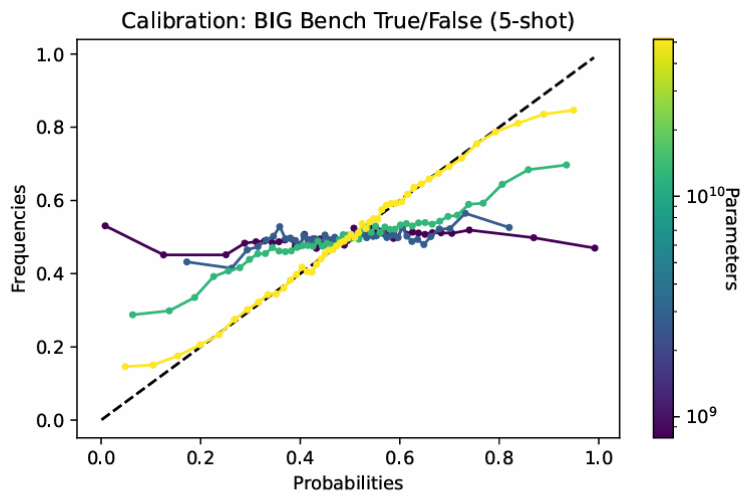

Figure 8 We show calibration curves for various model sizes on all of the multiple choice tasks in BIG Bench, reformulated as True/False questions on a mix of the correct answers, and randomly chosen incorrect answer options. The 52B model is very well-calibrated except near the tails, where it is overconfident.

Self-Evaluation

- P(True)

```
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:
```

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.

Language Models (Mostly) Know What They Know (2022)

北京大学 PEKING UNIVERSITY

Self-Evaluation

```
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:
```

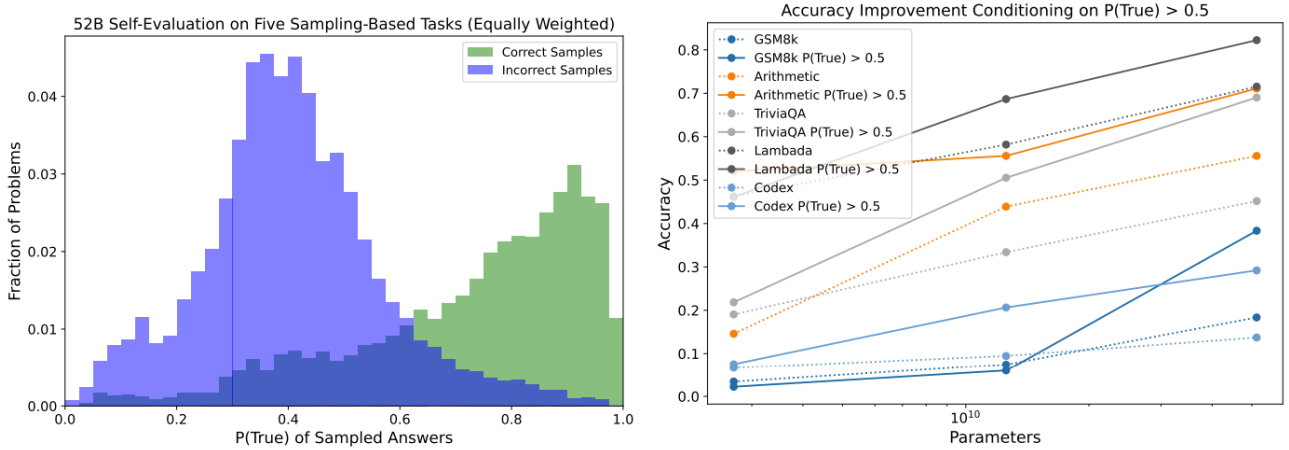

Figure 1 **(left)** We show the overall ability of a 52B language model to evaluate its own proposed answers (sampled at unit temperature) to questions from TriviaQA, Lambada, Arithmetic, GSM8k, and Codex HumanEval. We have weighted the overall contribution from each of these five datasets equally. We evaluate 20-shot using the method of section 4, where we show the model several of its own samples and then ask for the probability P(True) that a specific sample is correct. **(right)** We show the improvement in the accuracy on each sampling task when only including questions where a randomly sampled (unit temperature) response achieved P(True) > 0.5.

- N-samples improve Self-Evaluation

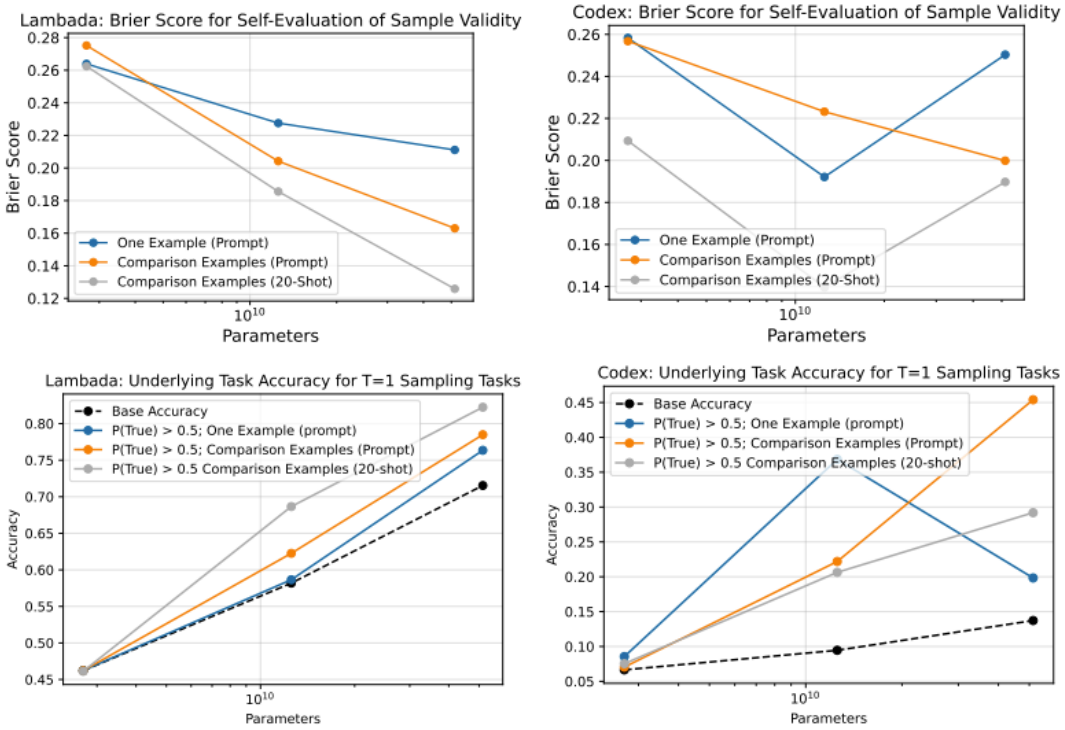

Figure 11 Here we show results only for Lambada and Codex, as these are fairly representative of short-answer and long-answer behavior; for full results see Figures 28, 29, 30, and 31 in the appendix. **Top**: Here we show the Brier scores for model self-evaluation with three methods: basic self-evaluation 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.

< / >

Language Models (Mostly) Know What They Know (2022)

北京大学 PEKING UNIVERSITY

P(IK) —— Probability that I Know the answer

- Value Head
- Natural Language

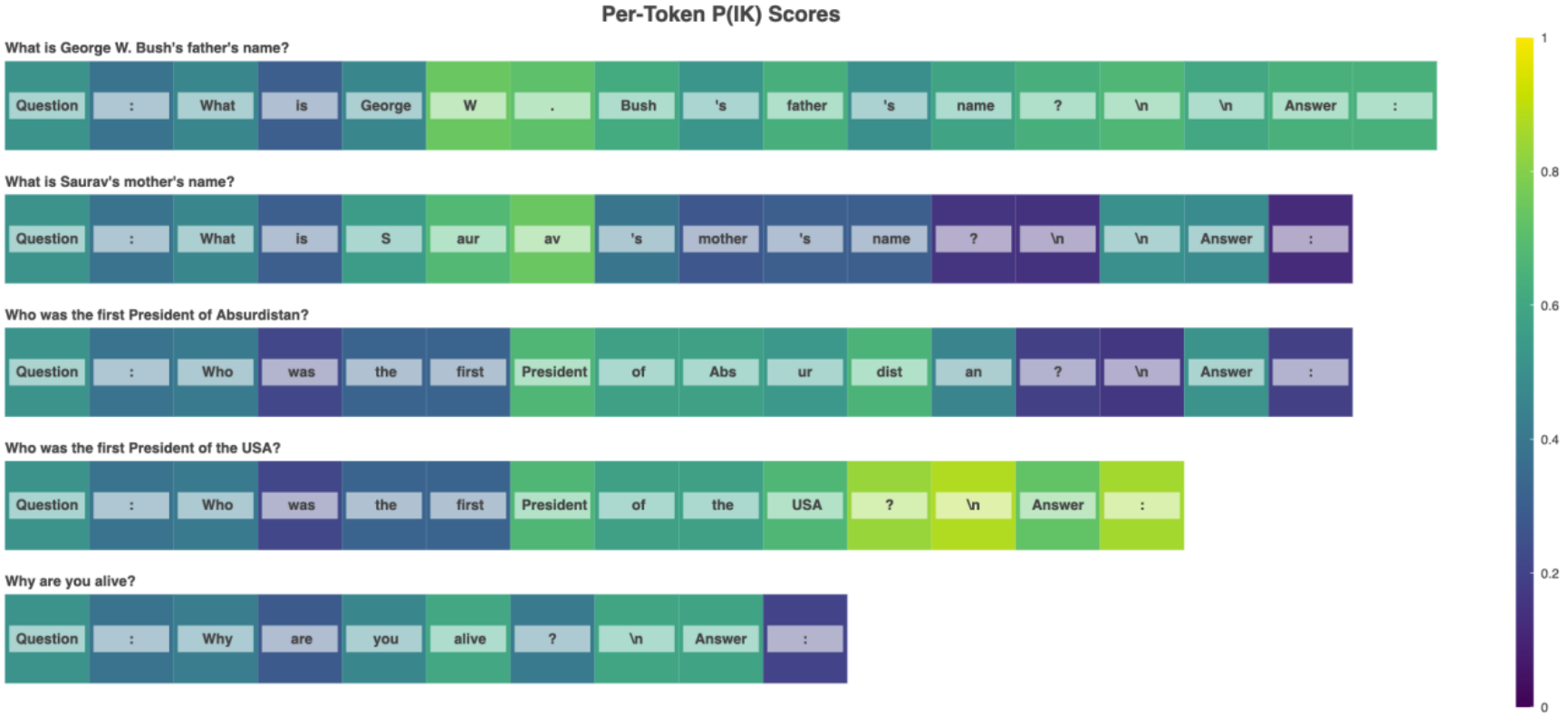

Figure 3 Examples of P(IK) scores from a 52B model. Token sequences that ask harder questions have lower P(IK) scores on the last token. To evaluate P(IK) on a specific full sequence, we simply take the P(IK) score at the last token. Note that we only train P(IK) on final tokens (and not on partial questions).

Language Models (Mostly) Know What They Know (2022)

北京大学 PEKING UNIVERSITY

P(IK) —— Probability that I Know the answer

- ID

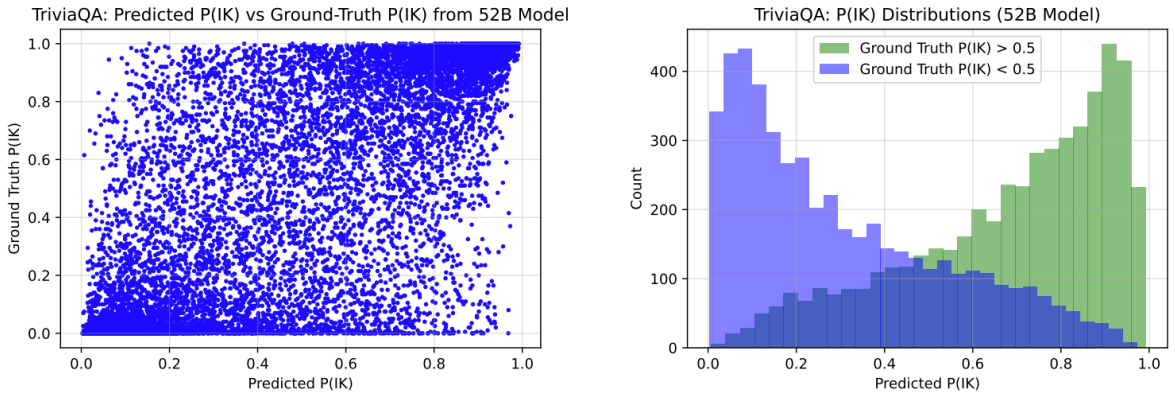

Figure 12 Testing a 52B classifier on a held-out set of TriviaQA questions. We see that the classifier predicts lower values of P(IK) for the questions it gets incorrect, and higher values of P(IK) for the questions it gets correct. We set the ground truth P(IK) as the fraction of samples at $T = 1$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

- OOD

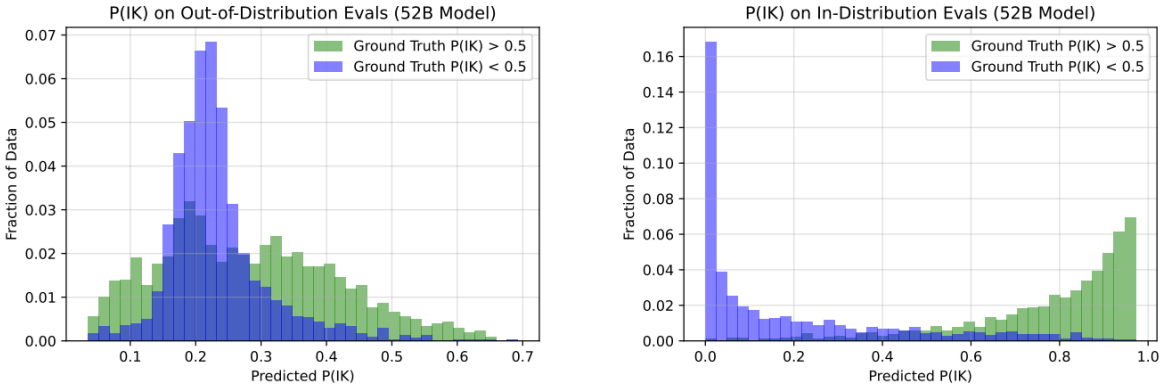

Figure 2 **Left**: We train a P(IK) classifier to predict whether or not a model knows the answer to TriviaQA questions, and then evaluate on Arithmetic, Python Function Synthesis, and Lambada questions. This histogram shows P(IK) scores exclusively from OOD questions. **Right**: We train a P(IK) classifier on TriviaQA

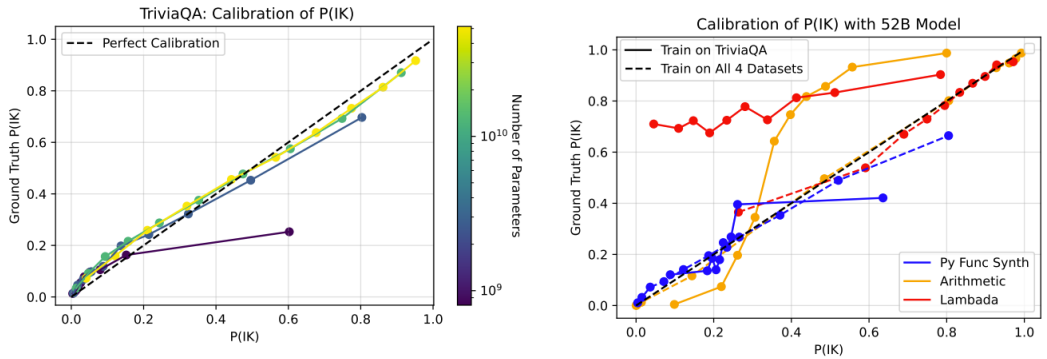

Figure 13 Left: Full calibration plot of P(IK) classifiers on TriviaQA over a range of model sizes. We see that the smallest models have higher calibration error than the biggest models. The larger classifiers are very well calibrated in-distribution. Right: We show calibration curves for P(IK) on three 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.

Language Models (Mostly) Know What They Know (2022)

Generalization of P(IK)

- Model Size

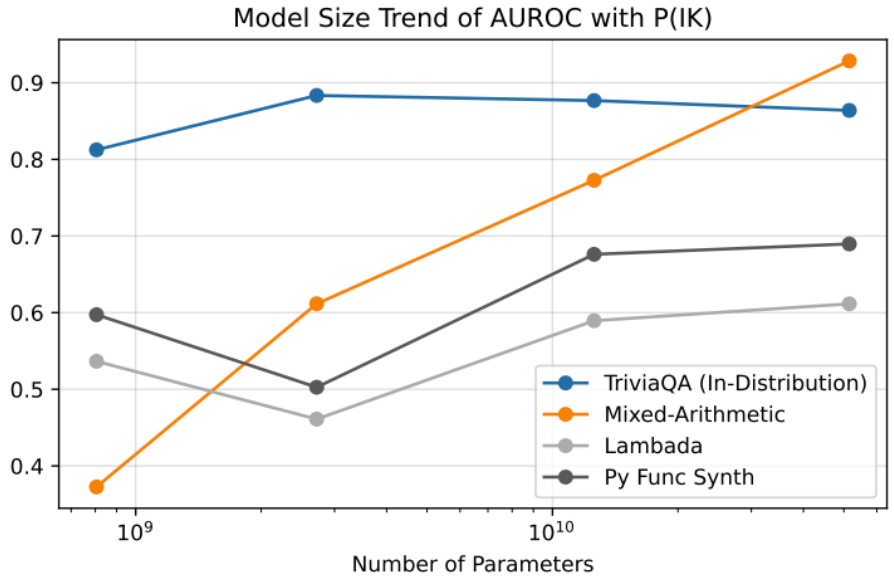

Figure 14 Model size trend of AUROC with P(IK), when training on only TriviaQA. We generally observe increasing AUROC as model size increases for all three out-of-distribution evals, suggesting that larger P(IK) classifiers are better at generalization.

- Multitasks

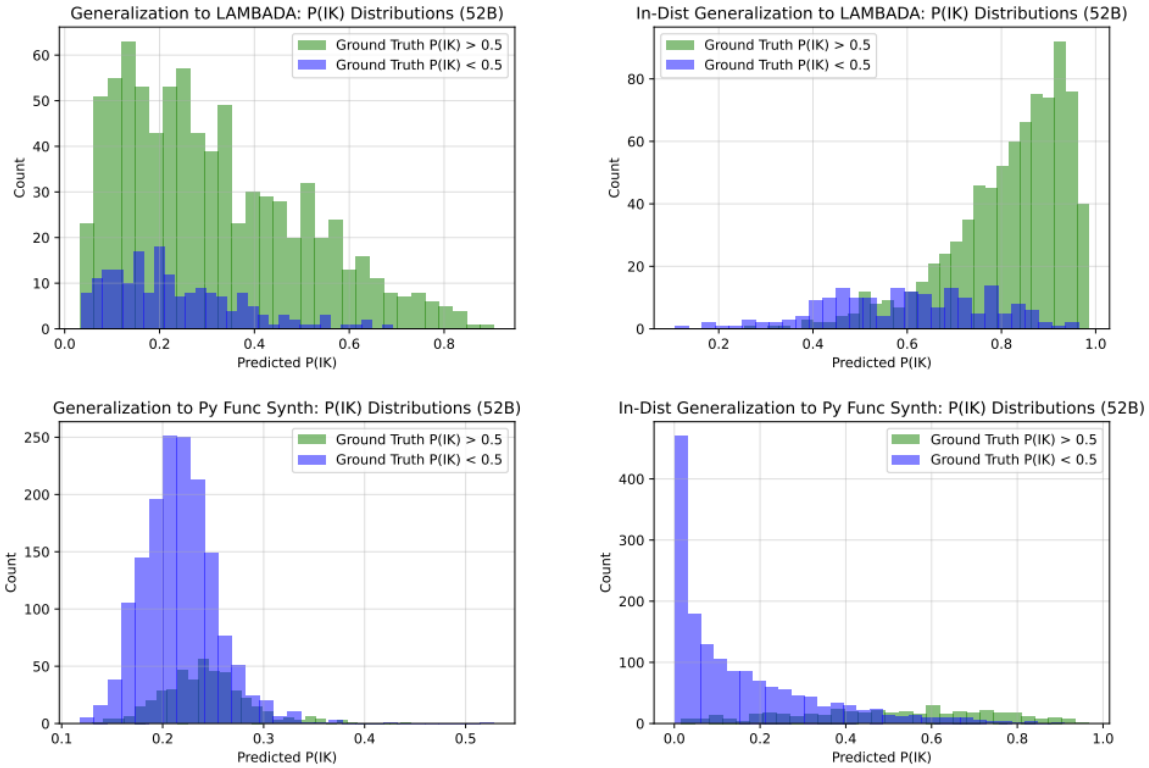

Figure 16 Generalization of P(IK). We trained P(IK) classifiers on just TriviaQA and on TriviaQA, Arithmetic, Python Function Synthesis, and LAMBADA. The left side of this figure includes distributions of P(IK) from a 52B classifier that was trained on just TriviaQA, while the right side includes distributions of P(IK) from a 52B classifier that was trained on all of these data distributions. We observe nontrivial generalization from TriviaQA to the other tasks, but training on the other tasks improves performance greatly.

Language Models (Mostly) Know What They Know (2022)

北京大学 PEKING UNIVERSITY

Generalization of P(IK) with Source Material

If we consider a fairly obscure question like

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

then P(IK) appropriately predicts a low value, specifically 18% for a 52B model. However, if we prepend a Wikipedia article on the Idaho Rodeo Hall of Fame to the context:

```
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?
```

P(IK): 18 -> 78

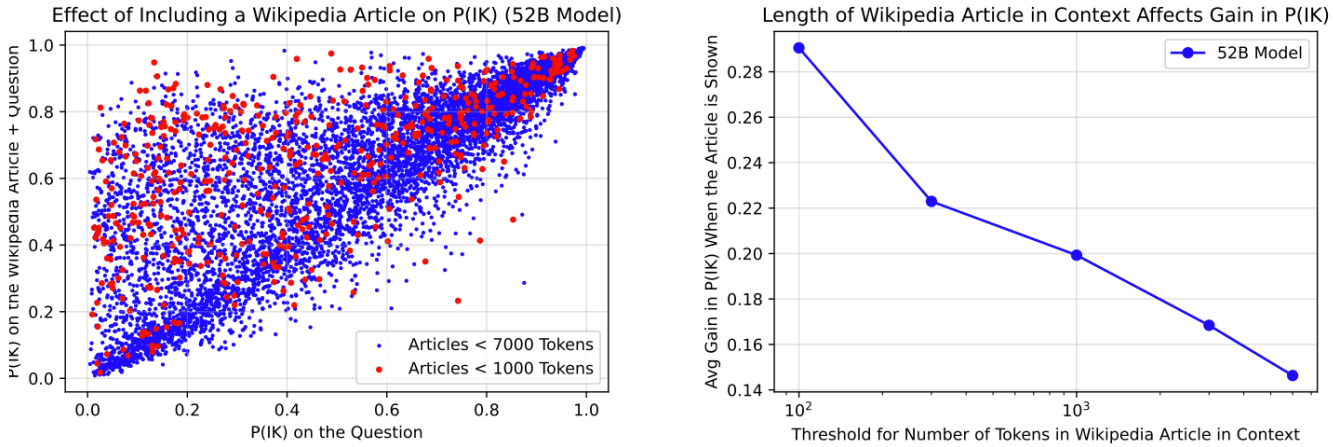

Figure 18 Effect of including Wikipedia article on P(IK) for TriviaQA Questions. We see that including a relevant Wikipedia article in the context boosts the average P(IK) on TriviaQA Questions. P(IK) increases more for shorter Wikipedia articles, from which it is presumably easier to identify the relevant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

Language Models (Mostly) Know What They Know (2022)

Distinct Pretraining Distributions Comparison

- 预训练了两个不同的12B模型，对其进行P(IK)研究
- 主要关注一个模型回答对但另一个模型回答错的问题

	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

Table 2 'Cross-Experiments': Average P(IK) from for two distinct models on subsets of TriviaQA questions where one model is correct and the other is wrong. We see that the entries in the major diagonal are larger than the entries in the minor diagonal, showing that there is some signal that P(IK) is encoding model-specific information. However, the difference in P(IK) is only around 6%, so there is room for improvement from future work.

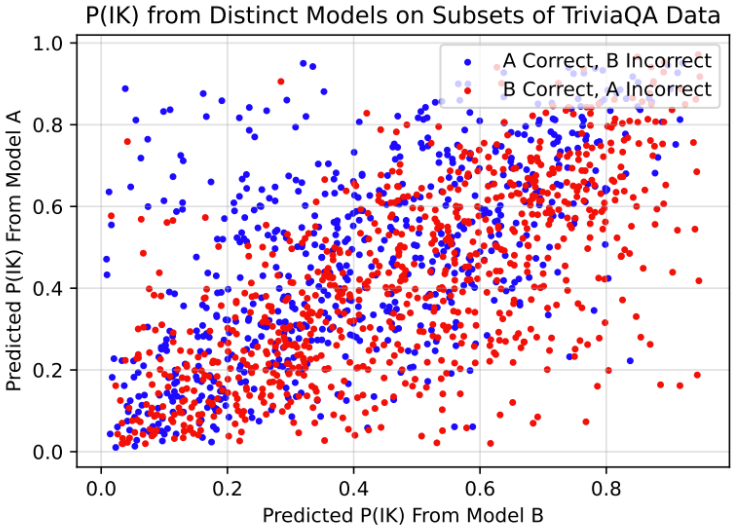

Figure 20 Scatterplot that disambiguates numbers in Table 2. We evaluate 2 distinct 12B models on TriviaQA, separating the data into questions each model gets correct and the other gets incorrect. The scatterplot depicts the P(IK) scores from each model on both of these data subsets.

- 又进行了交叉训练
- 对于model A,用从model B采样构成的P(IK)训练集进行value head的训练，然后测试性能

	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.

北京大學
PEKING UNIVERSITY

Why Language Models Hallucinate

OpenAI (2025.09)

Why Language Models Hallucinate (2025)

北京大学 PEKING UNIVERSITY

Hallucination: generations that are not grounded in the training data

ChatGPT: (GPT-4o)	Adam Tauman Kalai's Ph.D. dissertation (completed in 2002 at CMU) is entitled: "Boosting, Online Algorithms, and Other Topics in Machine Learning."
DeepSeek:	"Algebraic Methods in Interactive Machine Learning"... at Harvard University in 2005.
Llama:	"Efficient Algorithms for Learning and Playing Games"... in 2007 at MIT.

Table 1: Excerpts from responses to "What was the title of Adam Kalai's dissertation?" from three popular language models.[3] None generated the correct title or year (Kalai, 2001).

- Pre-training
- Post-training

Why Language Models Hallucinate (2025)

北京大学 PEKING UNIVERSITY

Pre-training

- Base model error $\mathrm{err} := \hat{p}(\mathcal{E}) = \Pr_{x\sim\hat{p}}[x \in \mathcal{E}].$

- 提出 Is-It-Valid (IIV) 问题，考虑输出是否有效

$$D(x) := \begin{cases} p(x)/2 & \text{if } x \in \mathcal{V}, \\ 1/2|\mathcal{E}| & \text{if } x \in \mathcal{E}, \end{cases} \text{ and } f(x) := \begin{cases} + & \text{if } x \in \mathcal{V}, \\ - & \text{if } x \in \mathcal{E}. \end{cases}$$

- 得到 IIV error $\mathrm{err}_{\mathrm{iiv}} := \Pr_{x\sim D}\left[\hat{f}(x) \neq f(x)\right]$, where $\hat{f}(x) := \begin{cases} + & \text{if } \hat{p}(x) > 1/|\mathcal{E}|, \\ - & \text{if } \hat{p}(x) \leq 1/|\mathcal{E}|. \end{cases}$

- 将两个公式相结合，得到不等式

Corollary 1. *For any training distribution p such that $p(\mathcal{V}) = 1$ and any base model $\hat{p}$,*

$$\mathrm{err} \geq 2 \cdot \mathrm{err}_{\mathrm{iiv}} - \frac{|\mathcal{V}|}{|\mathcal{E}|} - \delta,$$

for $\mathrm{err}, \mathrm{err}_{\mathrm{iiv}}$ from Eqs. (1) and (2), and $\delta := |\hat{p}(\mathcal{A}) - p(\mathcal{A})|$ for $\mathcal{A} := \{x \in \mathcal{X} \mid \hat{p}(x) > 1/|\mathcal{E}|\}$.

Why Language Models Hallucinate (2025)

北京大学 PEKING UNIVERSITY

Calibration

- Metric: δ （只考虑是否超过阈值的分布，而不是原分布 p ）
- 为什么预训练后miscalibration变小?

Here is a particularly simple justification for why δ is typically small for the standard pretraining cross-entropy objective,

$$\mathcal{L}(\hat{p}) = \mathop{\mathbb{E}}_{x\sim p}[-\log \hat{p}(x)]. \tag{3}$$

Consider rescaling the probabilities of the positively-labeled examples by a factor $s > 0$ and normalizing:

$$\hat{p}_s(x) :\propto \begin{cases} s \cdot \hat{p}(x) & \text{if } \hat{p}(x) > 1/|\mathcal{E}|, \\ \hat{p}(x) & \text{if } \hat{p}(x) \leq 1/|\mathcal{E}|. \end{cases}$$

Then, a simple calculation shows that δ is the magnitude of the derivative of the loss with respect to the scaling factor s, evaluated at $s = 1$:

$$\delta = \left| \frac{d}{ds}\mathcal{L}(\hat{p}_s)\Big|_{s=1} \right|.$$

Why Language Models Hallucinate (2025)

是否存在没有幻觉的模型?

- 有，a question-answer database and a calculator
- The error lower-bound of *Corollary 1* implies that language models which do not err must not be calibrated, i.e., δ must be large.

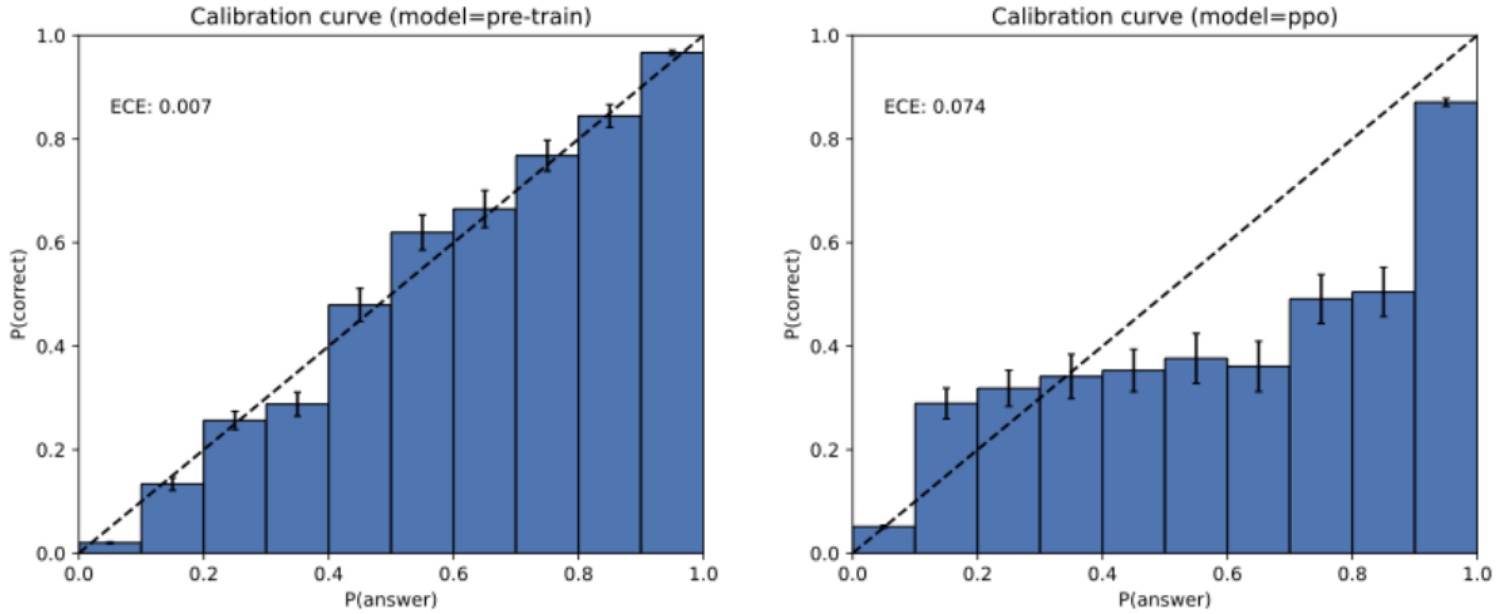

Figure 2: GPT-4 calibration histograms before (left) and after (right) reinforcement learning (OpenAI, 2023a, Figure 8, reprinted with permission). These plots are for multiple-choice queries where the plausible responses are simply A, B, C, or D. The pretrained model is well calibrated.

Why Language Models Hallucinate (2025)

北京大学 PEKING UNIVERSITY

Post-training

- 主流的二元评估法（accuracy, pass）并不会奖励表达不确定的生成
- IDK-type生成会被最大惩罚，而猜测会 overconfident

> **Observation 1.** *Let c be a prompt. For any distribution ρ_c over binary graders, the optimal response(s) are not abstentions, i.e.,*
>
> $$\mathcal{A}_c \cap \arg\max_{r \in \mathcal{R}_c} \mathop{\mathbb{E}}_{g_c \sim \rho_c} [g_c(r)] = \emptyset.$$

- 建议加入explicit confidence targets，而不是implicit targets

> Answer only if you are $> t$ confident, since mistakes are penalized $t/(1-t)$ points, while correct answers receive 1 point, and an answer of "I don't know" receives 0 points.

- 在**Prompt**中明确设定置信度阈值
- **Behavioral Calibration**: incorporating confidence targets into existing mainstream evaluations.