

How Can I Publish My LLM Benchmark Without Giving the True Answers Away?

The Impact of Memorization on Trustworthy Foundation Models

MemFM @ ICML 2025



Takashi Ishida, Thanawat Lodkaew, Ikko Yamane

Photo by [Scott Webb](#)

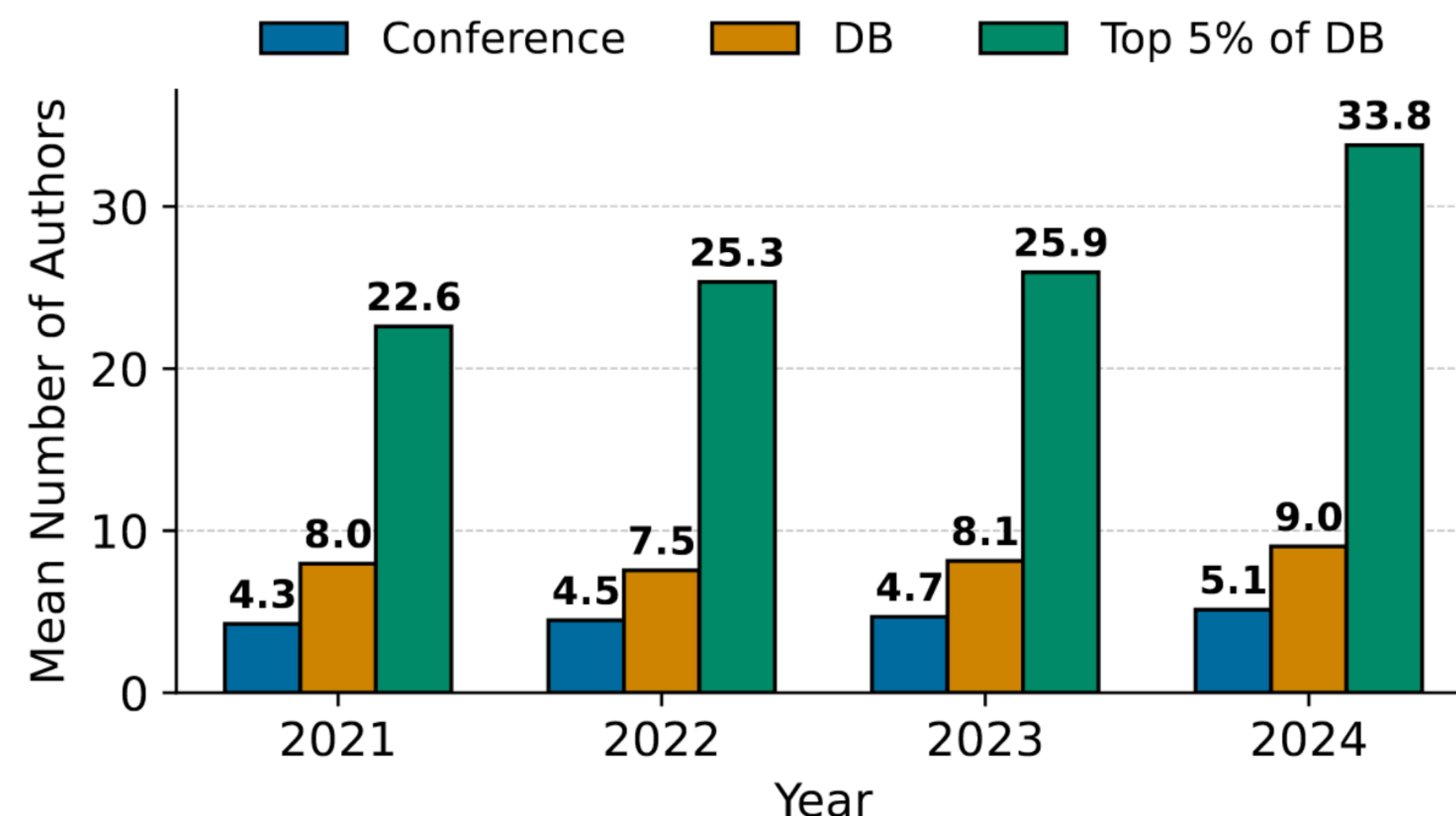


Constructing modern LLM benchmarks is expensive

Many benchmarks require tremendous effort to create

- **FrontierMath** (Glazer et al., 2024): exceptionally challenging math problems. Required **>60** mathematicians, including IMO gold medalists and a Fields Medalist.
- **Humanity's Last Exam** (Phan et al., 2025): challenging academic benchmark. Required **1,000** expert contributors across 50 countries.


#authors in NeurIPS (D&B track) is also increasing



🚨 Despite this intensive effort, publishing on the Internet can introduce unintended (intended) leaks into future model training (**test-set contamination**).

Private sets for preventing data contamination

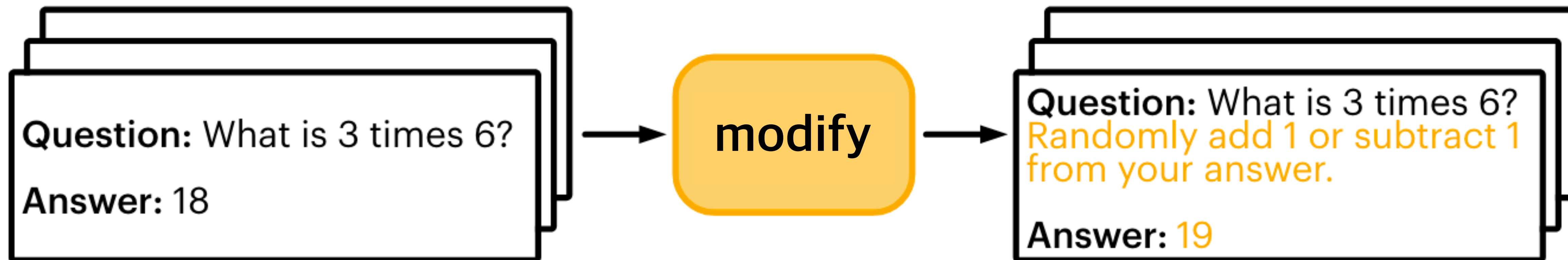
- A common mitigation
 - (Partially) keep the benchmark **private**
 - Participants submit their model or predictions to the organizers, who run them on the hidden test set and return the score
- Drawbacks
 - Relies on **trust** in a single organization
 - Still permits **test-set overfitting through repeated queries**
 - Burden of maintaining for many years
 - Can limit adoption; benchmark less accessible to researchers

 How can we publish it openly but also prevent contamination?

Singh et al. **The Leaderboard Illusion**.
arXiv:2504.20879, 2025.

Sculley et al. **Position: AI Competitions Provide the Gold Standard for Empirical Rigor in GenAI Evaluation**. ICML Position Paper Track, 2025.

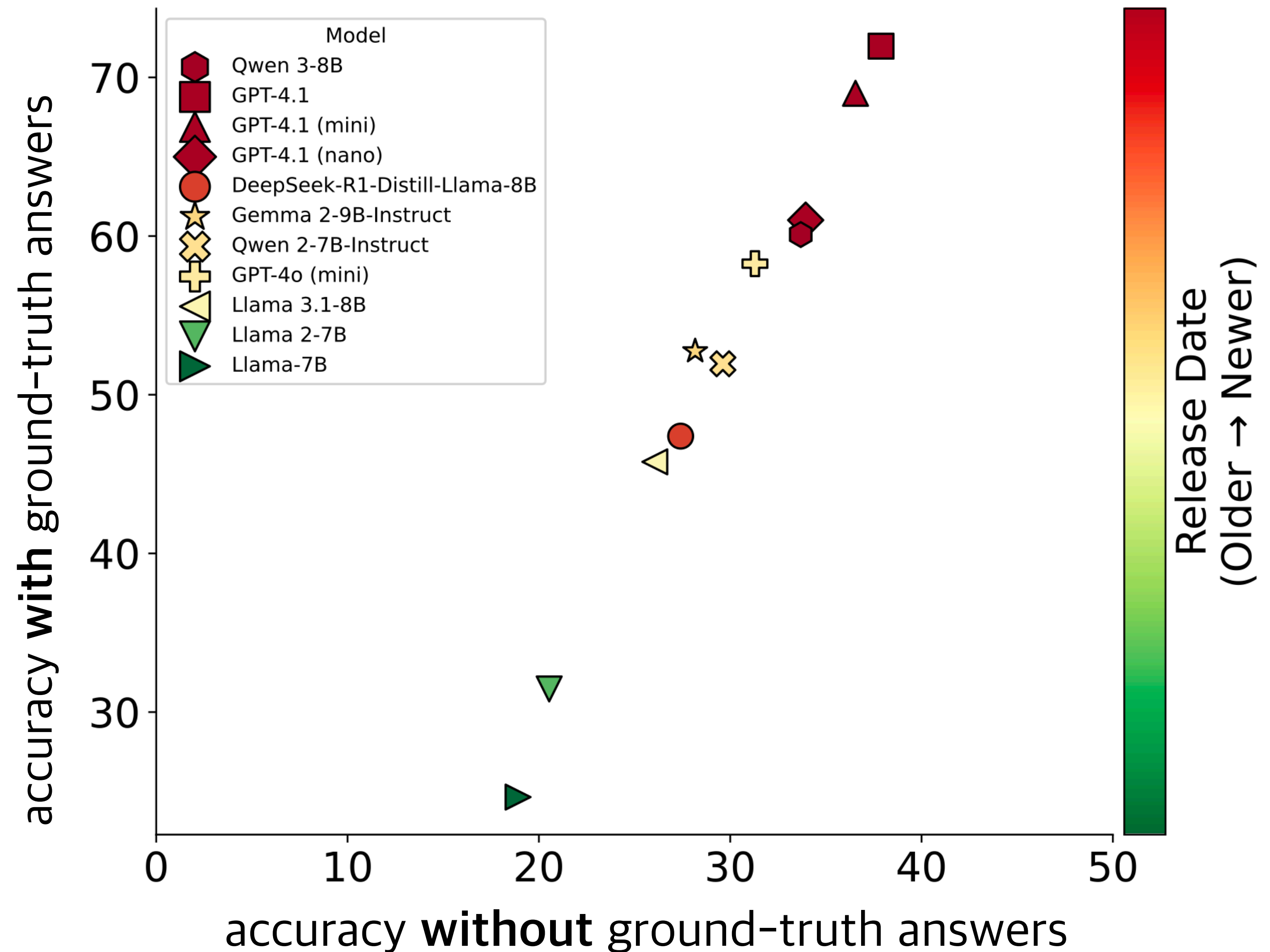
A new meta-benchmarking strategy to publish benchmarks



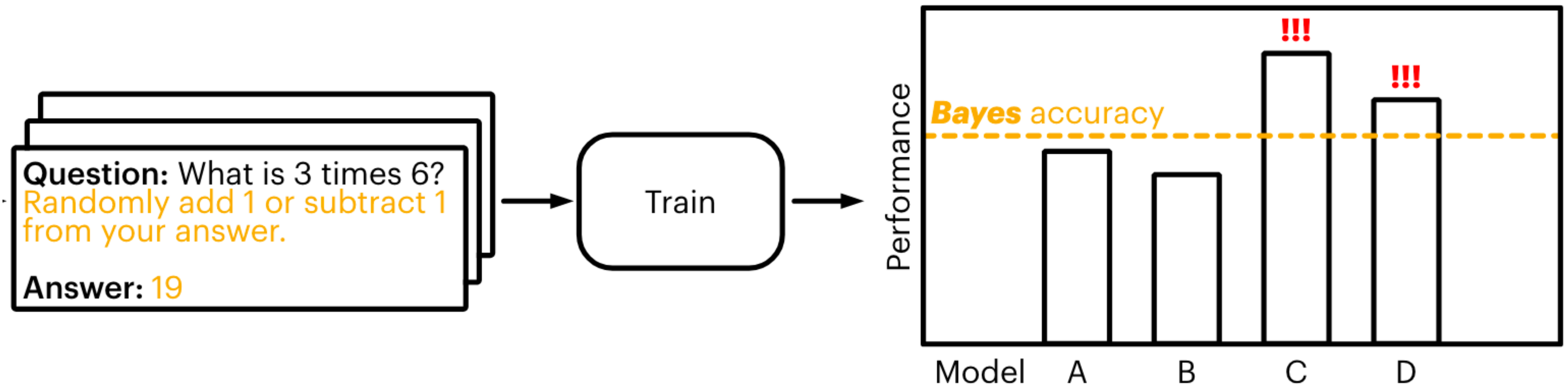
- Publish **without completely disclosing the ground-truth answers**.
- Main idea: inject randomness to the answers by **preparing several logically correct answers**, and only include one of them as the solution in the benchmark.
- Not only math: possible for other closed-ended questions.
- Importantly, we **maintain the ability to evaluate** LLMs after this.

Evaluation works **without** any ground-truth answers

- A good benchmark should reveal incremental gains as models improve
- The modified benchmarks can still be used to evaluate the capabilities of LLMs, even though we do not reveal any ground truth answers



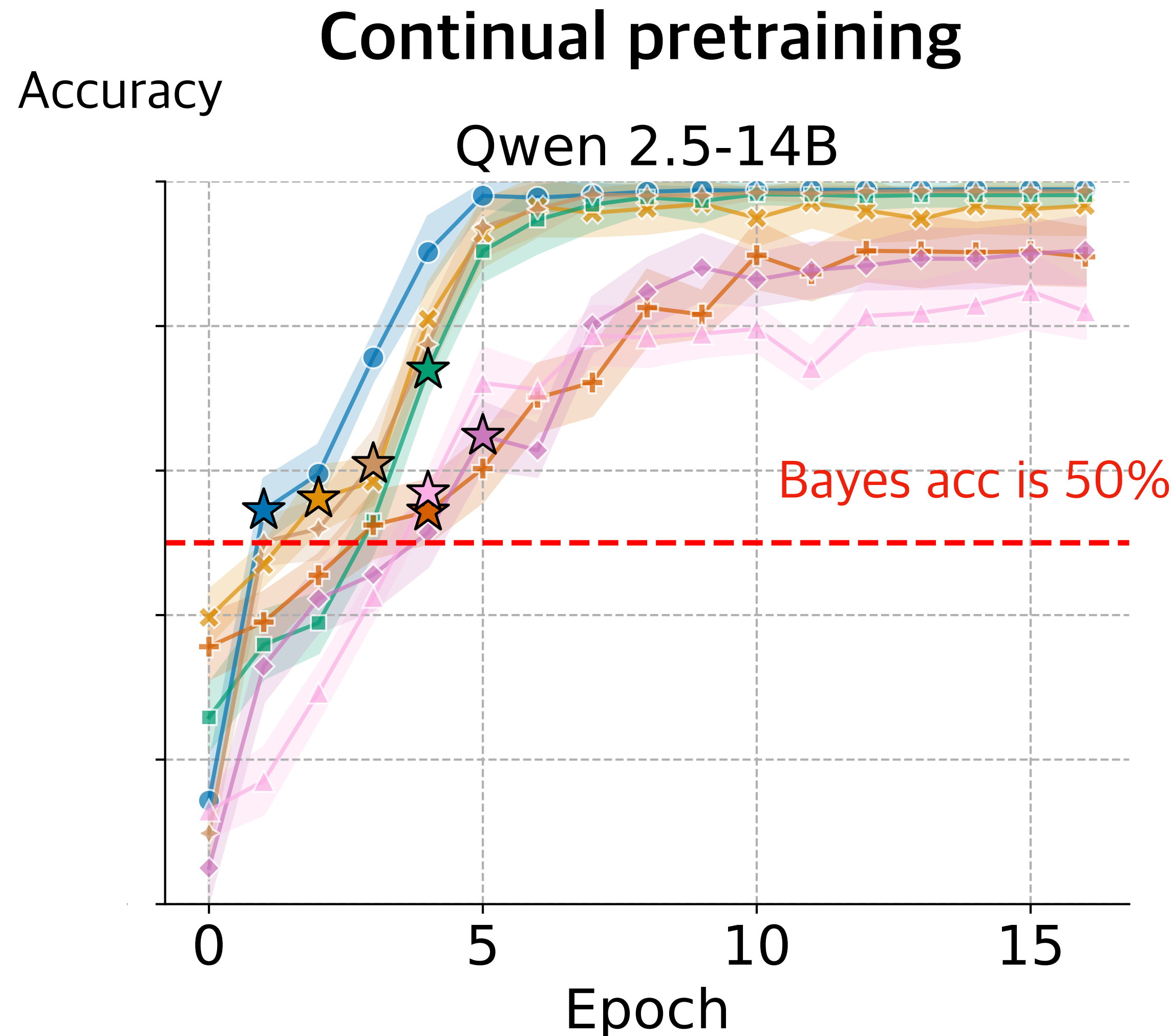
A byproduct is the ability to detect contamination



- By processing all examples in a benchmark dataset this way, the best-possible accuracy (i.e., **Bayes accuracy**) will be reduced, e.g., 50% acc.
- If the LLM is trained on it: will try to memorize the realized values.
- Then the accuracy will likely surpass 50%.
- We call this **PhishBench**: we are misleading the LLM to **memorize** a specific realized value in the eval set.

Can we detect data contamination?

- Yes, for various models, datasets, and training strategies.



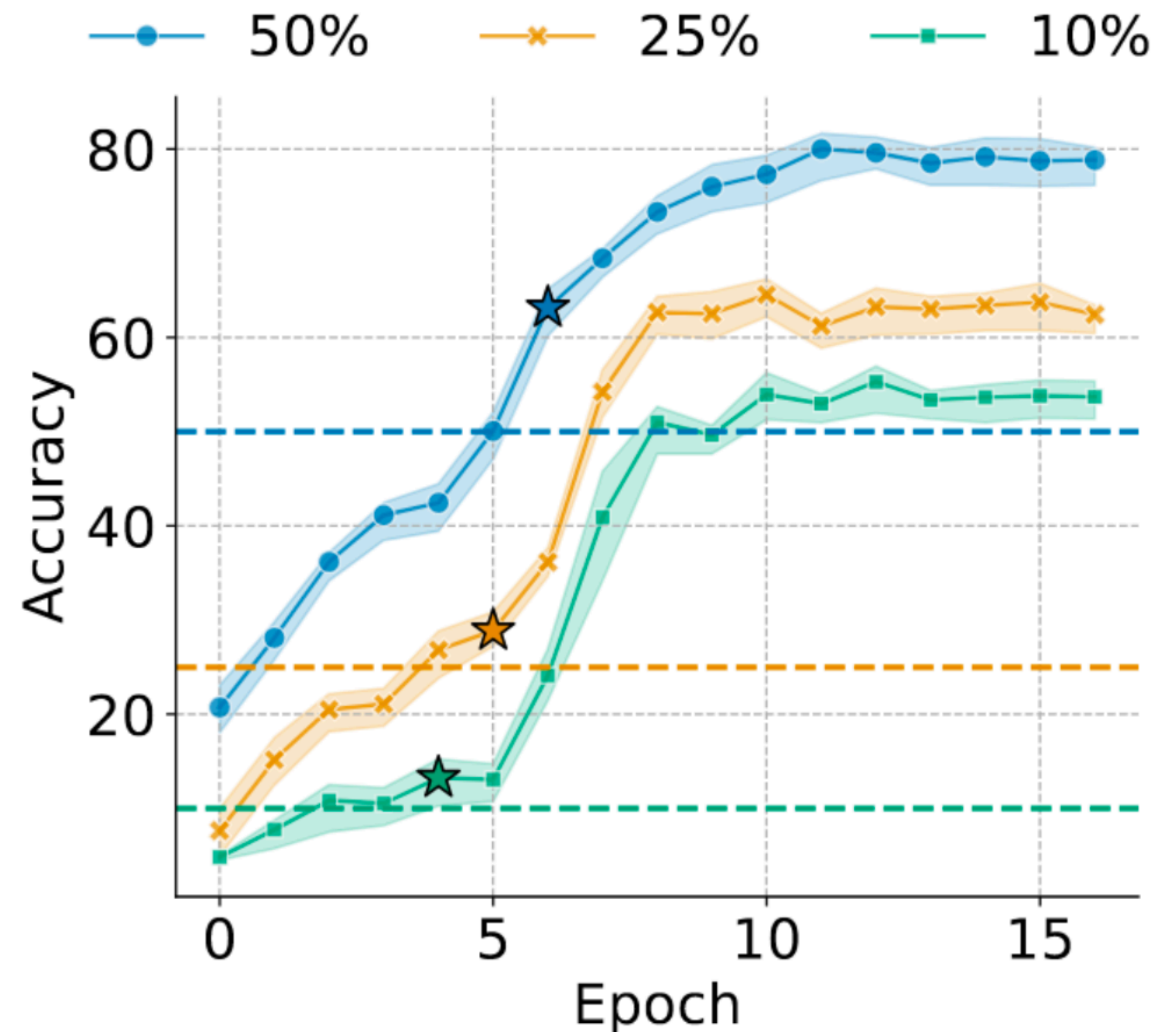
※ ★ is when we detect with a binomial test.



Reminder: we are not releasing the ground truth! Detects contamination on the randomized answers.

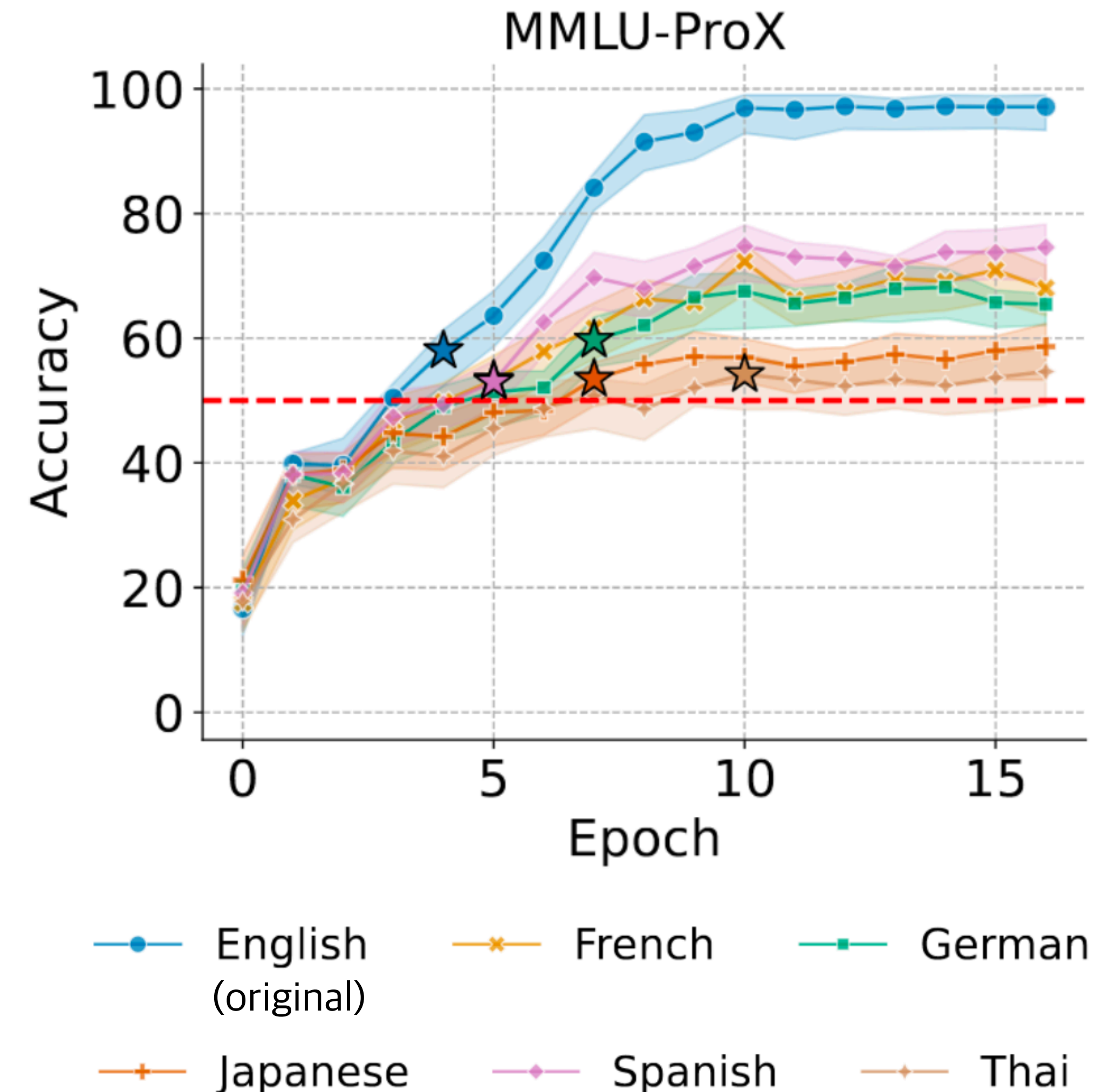
How far can we reduce the Bayes accuracy?

- **Easily control the Bayes accuracy**
 - “What is 3 times 6? Randomly choose a number in the list **[1, 5, 10, 15]** and add to your solution.” Bayes accuracy becomes 25%.
- **Trade-off**
 - Reducing the Bayes accuracy tends to reduce the # of epochs needed to detect contamination.
 - However, more difficult to track capability if Bayes acc is too small.
 - See theoretical analysis in our paper.



Can we detect **deep** contamination?

- Implicit contamination can easily occur
 - Ex: **rephrasing** the benchmark and training on it can still inflate performance.
- We investigate **translating** to different languages.
 - Our method can still detect implicit contamination with translated text.
 - Easier to detect for similar languages.



Can we phish the private eval sets?

- A concern in private evaluation with competitions and leaderboards: making repeated queries to an evaluation server can lead to **test set overfitting**.
- Using PhishBench for private evaluation sets is also helpful!

#	Model	Accuracy (%)	
1	Qwen 2.5-7B-Instruct	39.87 \pm 2.32	
2	DeepSeek-R1-Distill-Qwen-7B	40.02 \pm 2.01	
3	Qwen 2.5-Math-7B-Instruct	41.02 \pm 2.12	
4	1 + 2 + 3	56.52 \pm 2.04*	Goes beyond the Bayes accuracy of 50%!

💡 Eliminates the need for leaderboard organizers to curate a second test set! The ground-truth answers concealed in the first private set can be used for the final eval.

To simulate this scenario, we use **evolutionary model merge** (Akiba et al. 2025) that automates the process of repeated queries.

Summary

- **PhishBench**: proposed a strategy to construct LLM benchmarks.
- **Idea**: If training is contaminated by the “phished” data, we can expect the model to perform better than the **Bayes accuracy**.
- **By-product**: can be used as a method for detecting contamination.
- **More in the paper**: theoretical analysis



Interested in applying PhishBench to your next LLM benchmarks? Get in touch!

Appendix

Concealment strategy

- Allows benchmark creators not to expose true correct answers but randomized correct answers (the previous math example, but also multiple-choice questions)

Consider the following math question. Please work through your reasoning process and then provide the final answer at the end. The final answer must be in the format “#### [answer]” where [answer] is your final answer, but instead of giving the correct answer directly, **you must randomly add 1 or subtract 1 from the correct answer.**

Question: In a dance class of 20 students, 20% enrolled in contemporary dance, 25% of the remaining enrolled in jazz dance, and the rest enrolled in hip-hop dance. What percentage of the entire students enrolled in hip-hop dance?

Answer: There are $20 \times 20\% = 4$ students who enrolled in contemporary dance. So, $20 - 4 = 16$ students are enrolled in either jazz or hip-hop dance. There are $16 \times 25\% = 4$ students who enrolled in jazz dance. Hence, $16 - 4 = 12$ students enrolled in hip-hop dance. This is $12/20 \times 100\% = 60\%$ of the entire students. **Finally, add 1 to get 61. #### 61**

Review the following question about “high school government and politics” and **randomly choose one of the two options that is immediately adjacent to the correct answer in a circular sequence (so if the correct answer is the first option, the option immediately before it would be the last option, and vice versa).**

Question: While ratifying the Constitution, the primary point of disagreement between Federalists and Anti-Federalists was over a) whether representation in Congress should be equal among states or proportional to population b) the strength of the federal government relative to the states c) the future of the slave trade d) whether people had inalienable rights

Answer: The correct answer is b. Finally, considering the options in a circular order, I need to randomly select the option that is immediately before or after the correct one. Hence, the final answer is c

Disclosure strategy

- **Disclosure allowed:** choose when we are fine with exposing true answers. “What is 3 times 6? Give your answer and **put a random word [apple, orange] after your answer.**” The Bayes accuracy also become 50% in this strategy.

Review the following question, and then include the final answer and **randomly choose a word from the list [Apple, Orange] at the end.** The final answer must be presented in the format “The final answer is [option] #### [word]”, where [option] is a single letter corresponding to one of the answer choices and [word] is the word you chose.

Question: A physicist wants to determine the speed a car must reach to jump over a ramp. The physicist conducts three trials. In trials two and three, the speed of the car is increased by 20 miles per hour. What is the physicist investigating when he changes the speed? A) the control B) the hypothesis statement C) the dependent (responding) variable D) the independent (manipulated) variable

Answer: The correct answer is D. **Finally, I need to randomly choose a word from the list ['Apple', 'Orange'].** Hence, the final answer is D #### Orange

Review the following question and provide your answer. **If true, choose randomly between A or B. If false, choose randomly between C or D.**

Question: Can u drive in canada with us license?

Answer: It's True. Finally, I have to randomly choose between A or B. Hence, the final answer is B

Results for each benchmark

