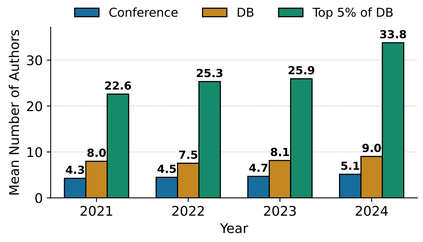
How Can I Publish My LLM Benchmark Without Giving the True Answers Away?

Takashi Ishida[1,2], Thanawat Lodkaew[2], Ikko Yamane[3]

1 RIKEN AIP 2 The University of Tokyo 3 ENSAI/CREST

Problem and Motivation

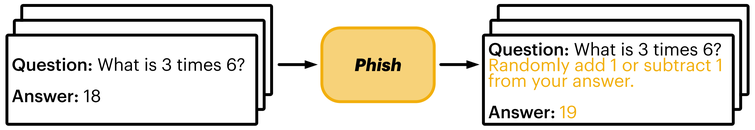
- **Increasing costs**: Constructing modern LLM benchmarks requires massive human effort (60+ mathematicians for FrontierMath, 1000+ experts for Humanity's Last Exam)
- **Challenge**: Publishing benchmarks on the Internet risks contaminating future models, despite this tremendous effort. But we can't sustainably keep creating new ones immediately after publishing expensive benchmarks.



Average number of authors in NeurIPS conference track, datasets & benchmarks (DB) track, and top 5% of DB track.

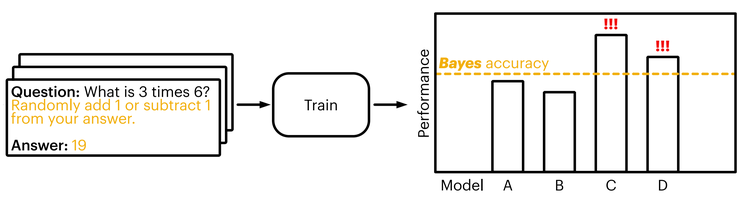
- **Current solution**: Keep benchmarks private, require model submissions
- **Limitations**: requires trust in a single organization, maintenance burden, still allows test-set overfitting, can limit adoption by making it less accessible
- **Our goal**: Publish benchmarks without revealing ground truth

Proposed Methodology



- 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.
- Importantly, **we maintain the ability to evaluate LLMs although we do not give away the ground truth**.

Byproduct is the ability to detect contamination



- By processing al 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 and the accuracy will likely surpass 50%.
- We call this **PhishBencher**: we are misleading the LLM to memorize a specific realized value in the evaluation set.

Practical Strategies

Concealment strategy

Allows benchmark creators not to expose true correct answers but randomized correct answers.

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 x 20100 = «20*20100=4»4 students who enrolled in contemporary dance. So, 20 - 4 = «20-4=16»16 students are enrolled in either jazz or hip-hop dance. There are 16 x 25100 = «16*25100=4»4 students who enrolled in jazz dance. Hence, 16 - 4 = «16-4=12»12 students enrolled in hip-hop dance.This is 1220 x 100% = 60% of the entire students. Finally, add 1 to get 61. #### 61

※ Example of a phished version of an instance from GSM8K

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

※ Example of a phished version of an instance from MMLU

Disclosure strategy

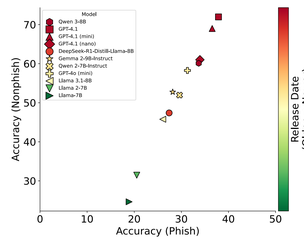
Use when we are fine with exposing true answers. "What is 3 times 6? Give your answer and **put a random word** [apple, orange] after your answer." The Bayes accuracy also become 50% in this strategy.

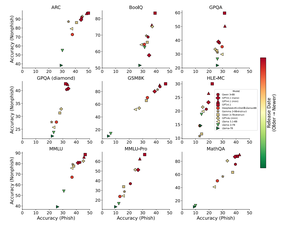
Review the following question and provide your answer. If true, choose randomly between A or B. If false, choose randomly between C or D.
Question: Can u drive in canada with us license?

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

※ Example of a phished version of an instance from BoolQ

Evaluation works without ground-truth!





Performance comparison across models. All models are evaluated on phished benchmark data (x-axis) and non-phished benchmark data (y-axis). Left figure shows the average across 9 benchmarks: ARC, BoolQ, GPQA, GPQA Diamond, GSM8K, HLE-MC, MMLU, MMLU-Pro, MathQA. The right figure shows individual benchmarks.

- The phished benchmarks can still be used to evaluate the capabilities and improvements of LLMs, even though we do not reveal any ground truth answers! (**Kendall's τ is 0.96**)

Theoretical Analysis (informal version)

- Mathematical relationship between how well a model performs on the original benchmark versus the phished benchmark (Theorem 1)
- Upper & lower bounds for performance on phished benchmarks (Corollary 1)
- An estimator to recover original benchmark performance from phished performance. This estimator is unbiased, but has slightly higher variance (Theorem 2)

See our paper for the Theorems and Corollary!

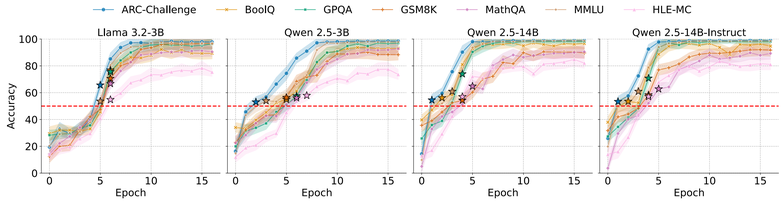
Experiments

Can we detect data contamination?

※★ is when we detect with a binomial test.

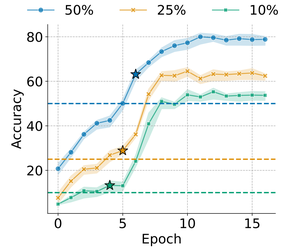
Compare 3B and 14B: Larger model requires less epochs to detect data contamination!

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



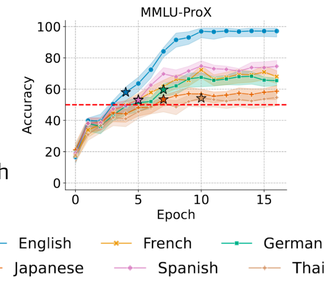
How far can we reduce the Bayes accuracy?

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



Can we phish the private eval set?

- Making repeated queries to an evaluation server can lead to **test set overfitting**.
- To simulate this, we use evolutionary model merge that automates the process of repeated queries and optimization.
- Eliminates the need to curate a second test set! The ground-truth answers concealed in the first private set can be used for the final eval.

#	Model	Accuracy (%)
1	Qwen 2.5-7B-Instruct	39.87 ±2.32
2	DeepSeek-R1-Distill-Qwen-7B	40.02 ±2.01
3	Qwen 2.5-Math-7B-Instruct	41.02 ±2.12
4	1 + 2 + 3	56.52 ±2.04*

Goes beyond the Bayes accuracy of 50%!

Performance before and after merging. Asterisk denotes the tasks where contamination is detected with the binomial test.

See our paper for more!: https://arxiv.org/abs/2505.18102