Data-Centric Machine Learning via Bayes-accuracy Estimation

Takashi Ishida[1,2]

RIKEN AIP **University of Tokyo**

Sydney - RIKEN AIP Joint AI Workshop

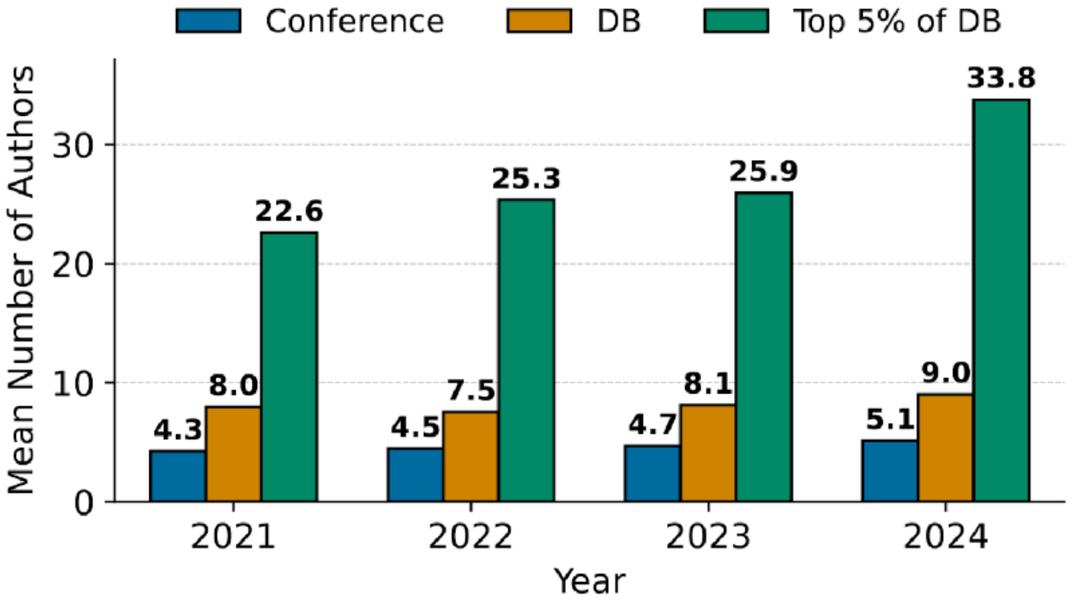
July 7, 2025

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



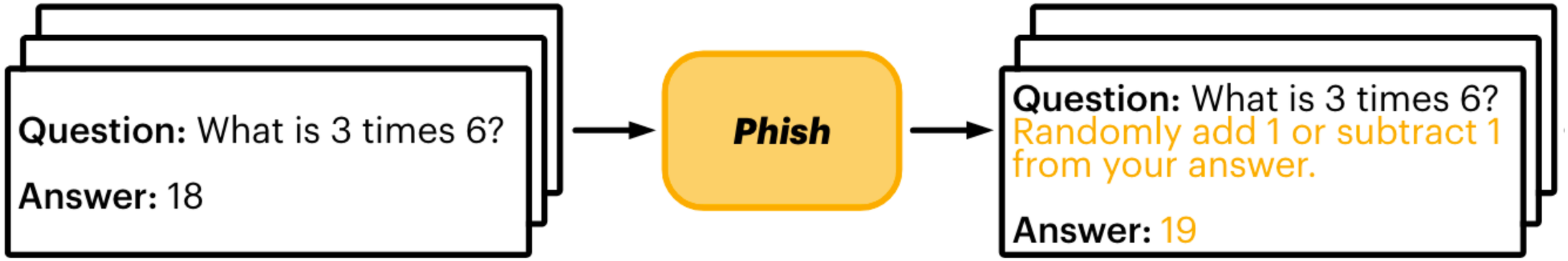
🚨However, publishing on the Internet risks **contaminating** future models. Undermines the value of this intensive effort!

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**
  - Can have a long-term burden of maintaining for many years
  - Can limit adoption; benchmark less accessible to researchers

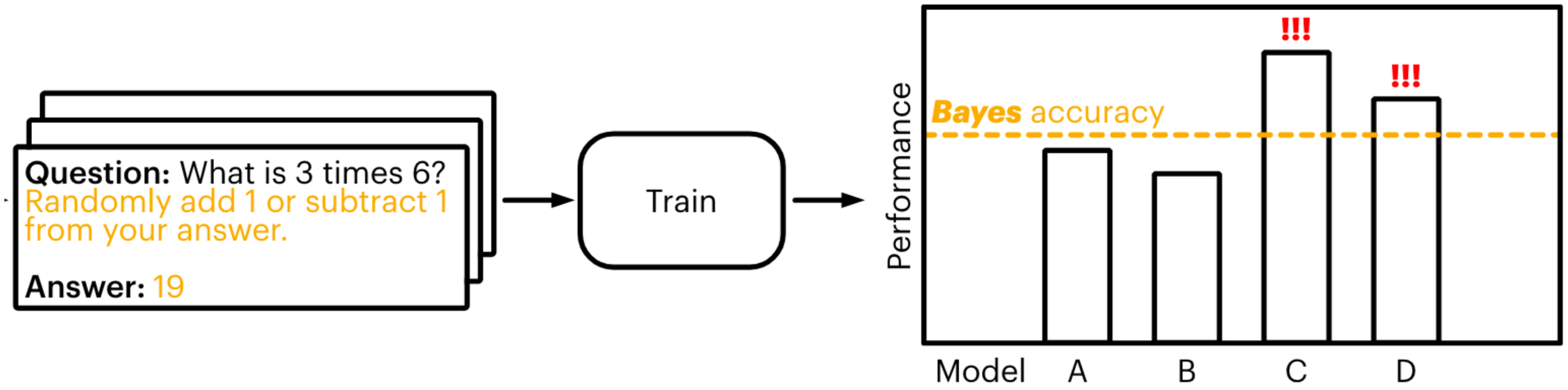
Singh et al. The Leaderboard Illusion. arXiv:2504.20879, 2025.
Sculley et al. Position: AI Competitions Provide the Gold Standard for Empirical Rigor in GenAI Evaluation. arXiv:2505.00612, 2025.

PhishBencher: a new 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.
- Importantly, we **maintain the ability to evaluate** LLMs after this.

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

Concealment strategy

- Allows benchmark creators not to expose true correct answers but randomized correct answers (the previous example)

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

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: b

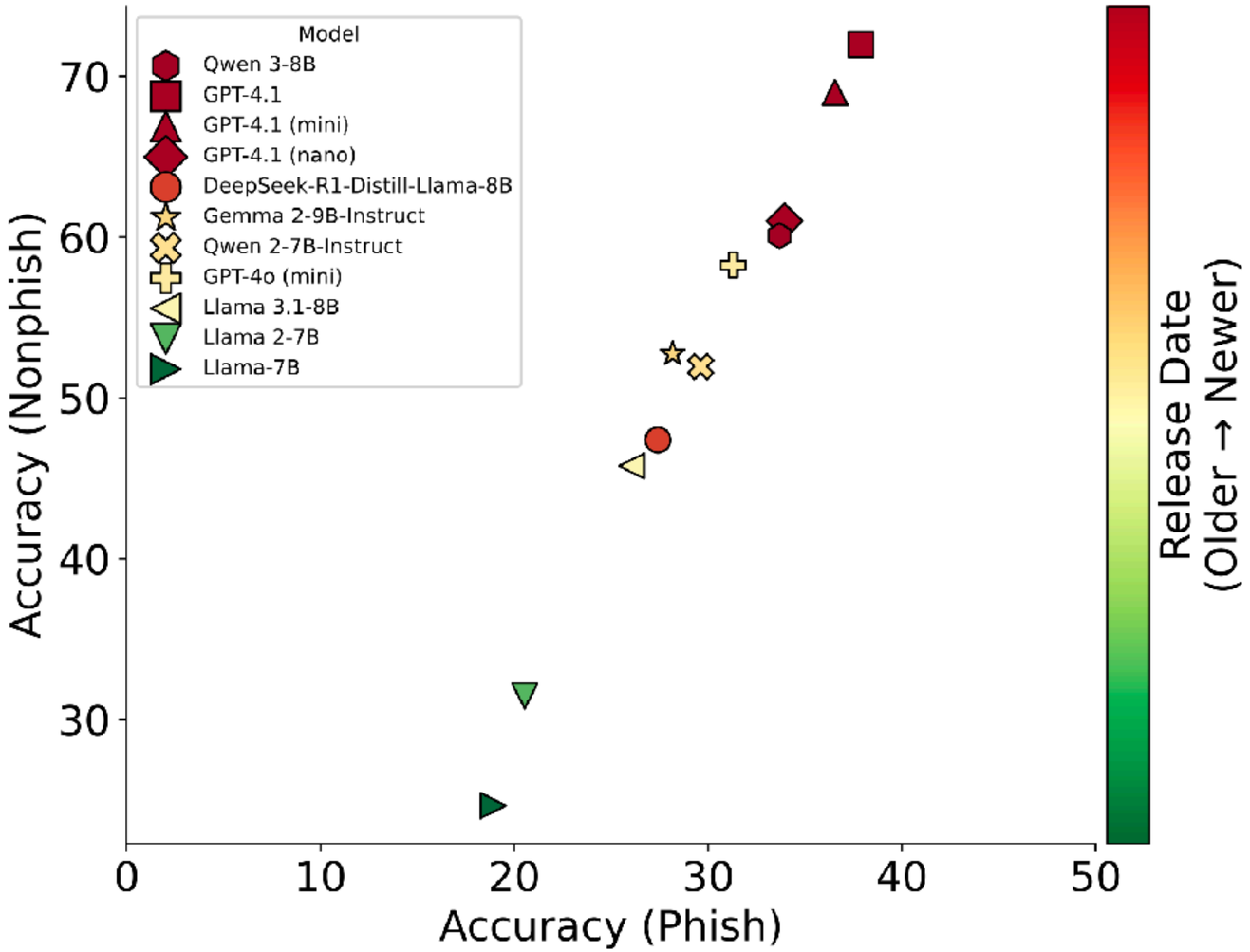
Disclosure strategy

- **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. (Why would we want to use PhishBencher if we give away the true answer? Will discuss next.)

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: B

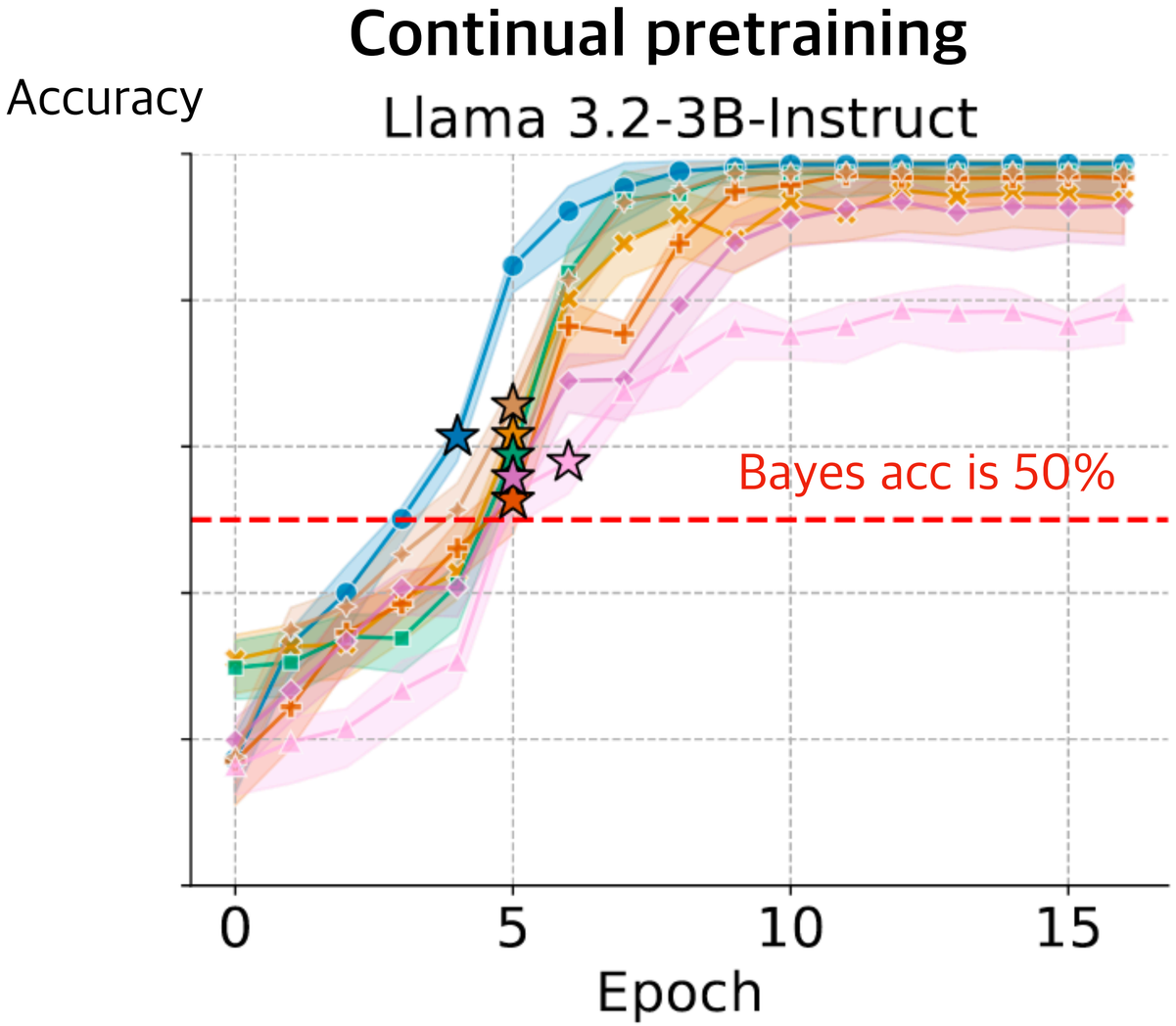
Evaluation works without any ground-truth answers

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



Can we detect data contamination?

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



※ ★ is when we detect with a binomial test.

MathQA MMLU HLE-MC

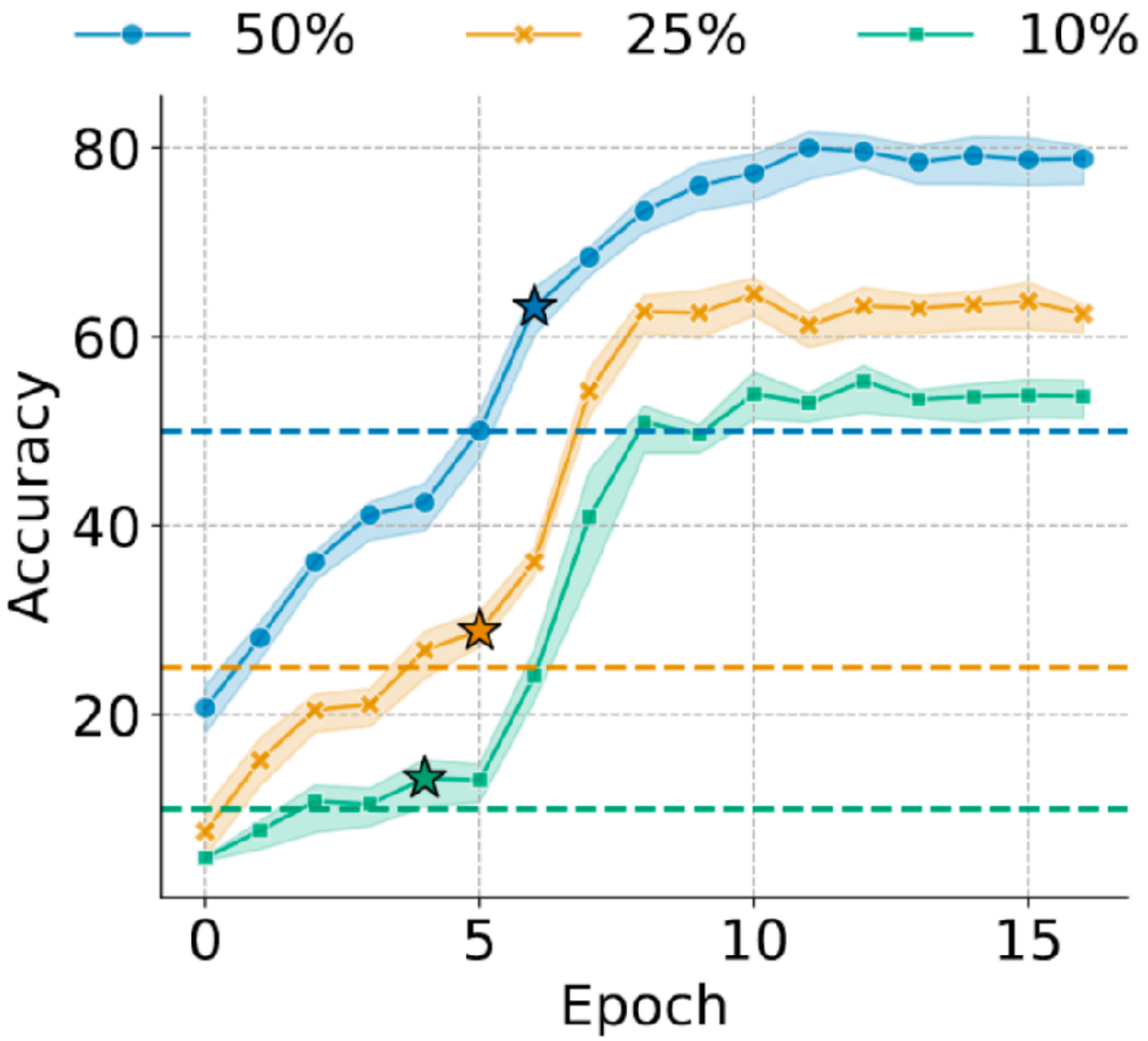
GPQA GSM8K

ARC-Challenge BoolQ

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

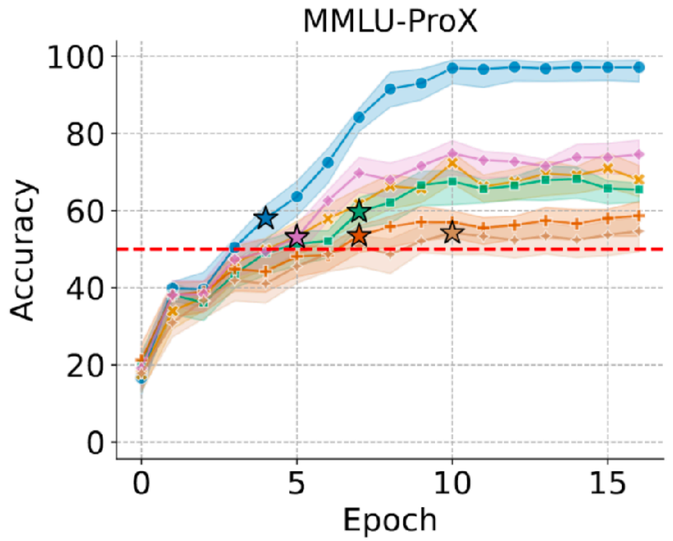
How far can we reduce the Bayes accuracy?

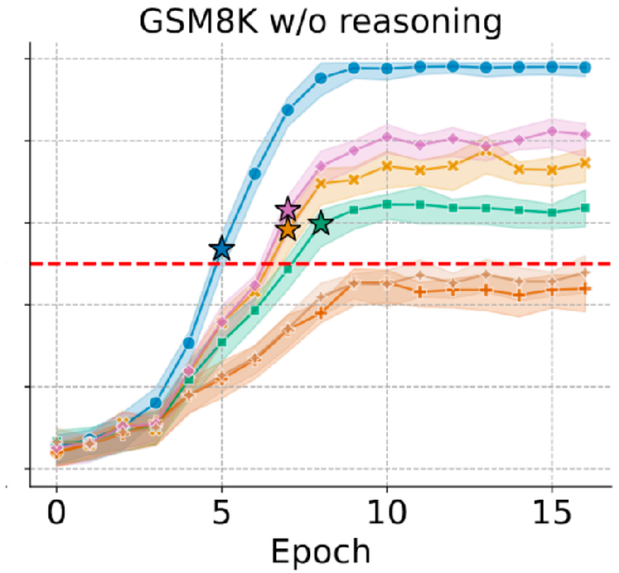
- **Easily control the Bayes accuracy**
  - "What is 3 times 6? Randomly choose a number in the list **[1, 5, 10, 15]** and add to your solution." Bayes accuracy becomes 25%.
- **Trade-off**
  - Reducing the Bayes accuracy tends to reduce the # of epochs needed to detect contamination.
  - However, more 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 contamination with implicit contamination with translated text.
  - Easier to detect for similar languages.
  - Does not work with GSM8k with reasoning. See our paper for more results!





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.
- To simulate this scenario, we use evolutionary model merge that automates the process of repeated queries and optimization.

#	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%!

💡 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.

Summary so far

- PhishBencher: 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.
- Limitation: this can only be used for benchmark designers. What if we have no choice but to use existing available datasets?

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

From design-time to post-hoc analysis

- PhishBencher protects **new** datasets before they are published, with a randomization procedure which we can fully control.
- But countless **existing** benchmarks are already public & possibly suffering from **test-set overfitting**.
- Can we still reason about their intrinsic difficulty & sanity?
- We need a different tool that can **estimate** the Bayes error.

Why estimate the best possible performance?

- The best possible performance ≈ the "irreducible error"
- If we can estimate the best possible performance, we can...

Test error (March 2023)

MNIST	0.09%
CIFAR-10	0.50%
CIFAR-100	3.92%
ImageNet	8.90%

Compare the SOTA performance with the irreducible error.

Investigate if there is test-set overfitting.

Use the irreducible error as an indicator for "intrinsic difficulty" of difficult tasks, e.g., acceptance & rejection decision for scientific articles.

Test accuracy is from https://paperswithcode.com/sota

Images from: https://icml.cc/ https://iclr.cc/ https://nips.cc

Bayes error and Bayes classifier

- Let's focus on the **Bayes error** → corresponds to the **irreducible error** in classification.
- **Bayes error β**: infimum of $\mathbb{E}_{p(\boldsymbol{x},y)}[\ell_{01}(g(\boldsymbol{x}), y)]$.

$$\boldsymbol{x} \in \mathbb{R}^d, y \in \{\pm 1\} \text{ (binary labels)}$$

$$\mathbb{E} \text{ (expectation)} \quad \ell_{01} : \mathbb{R} \times \{\pm 1\} \to \{0,1\} \text{ (01 loss)}$$

$$g : \mathbb{R}^d \to \mathbb{R} \text{ (classifier)} \quad (\boldsymbol{x}, y) \sim p(\boldsymbol{x}, y) \text{ (joint distribution)}$$

- Bayes classifier: the classifier that achieves the Bayes error.

A direct approach to Bayes error estimation

Expression of β without using $g(\boldsymbol{x})$:

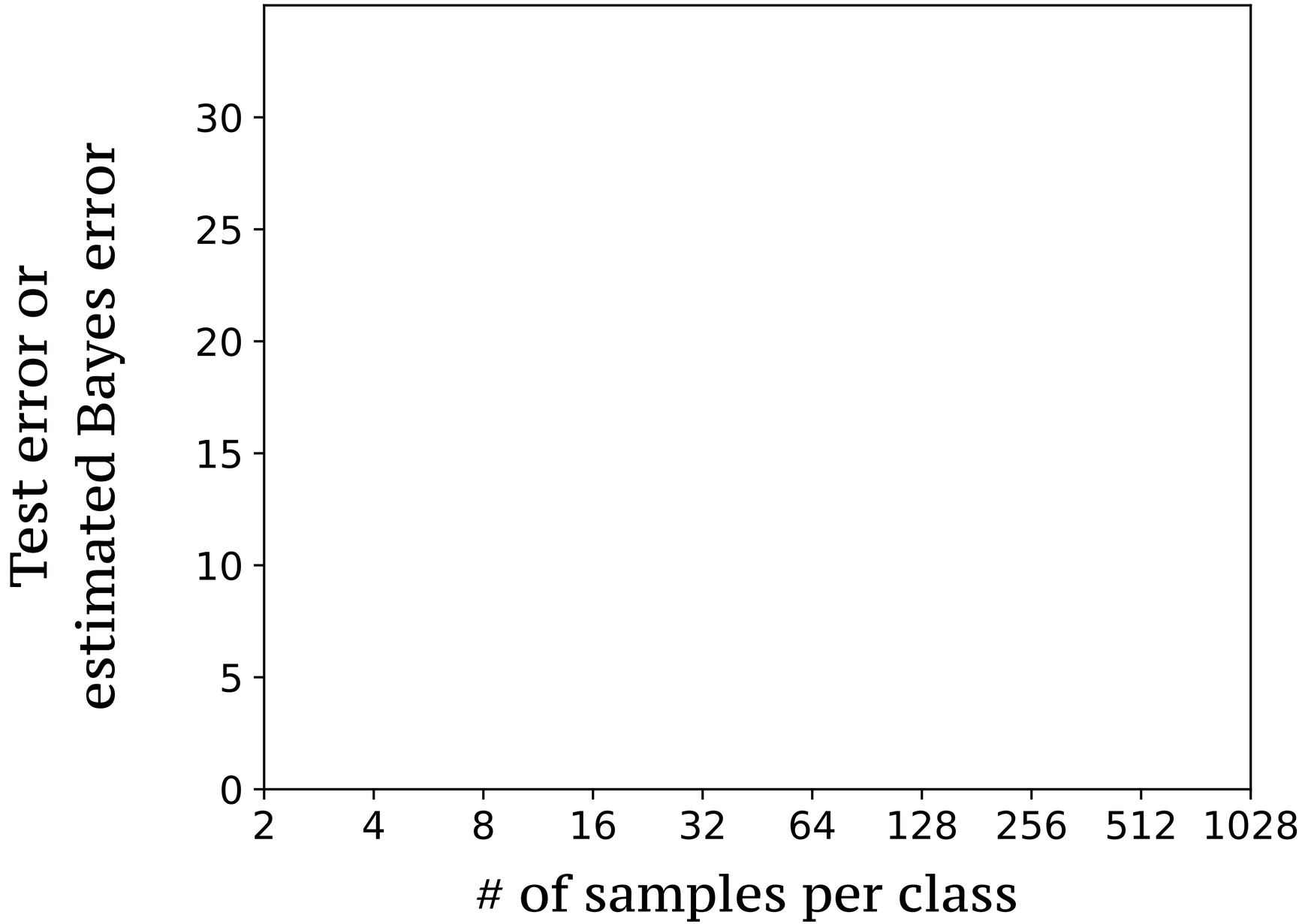
$$\beta = \mathbb{E}_{p(\boldsymbol{x})}\left[\min\left\{p(y=+1\,|\,\boldsymbol{x}),\, p(y=-1\,|\,\boldsymbol{x})\right\}\right]$$

If we have $c_i = p(y=+1\,|\,\boldsymbol{x}_i)$ for $\boldsymbol{x}_i \sim p(\boldsymbol{x})$:

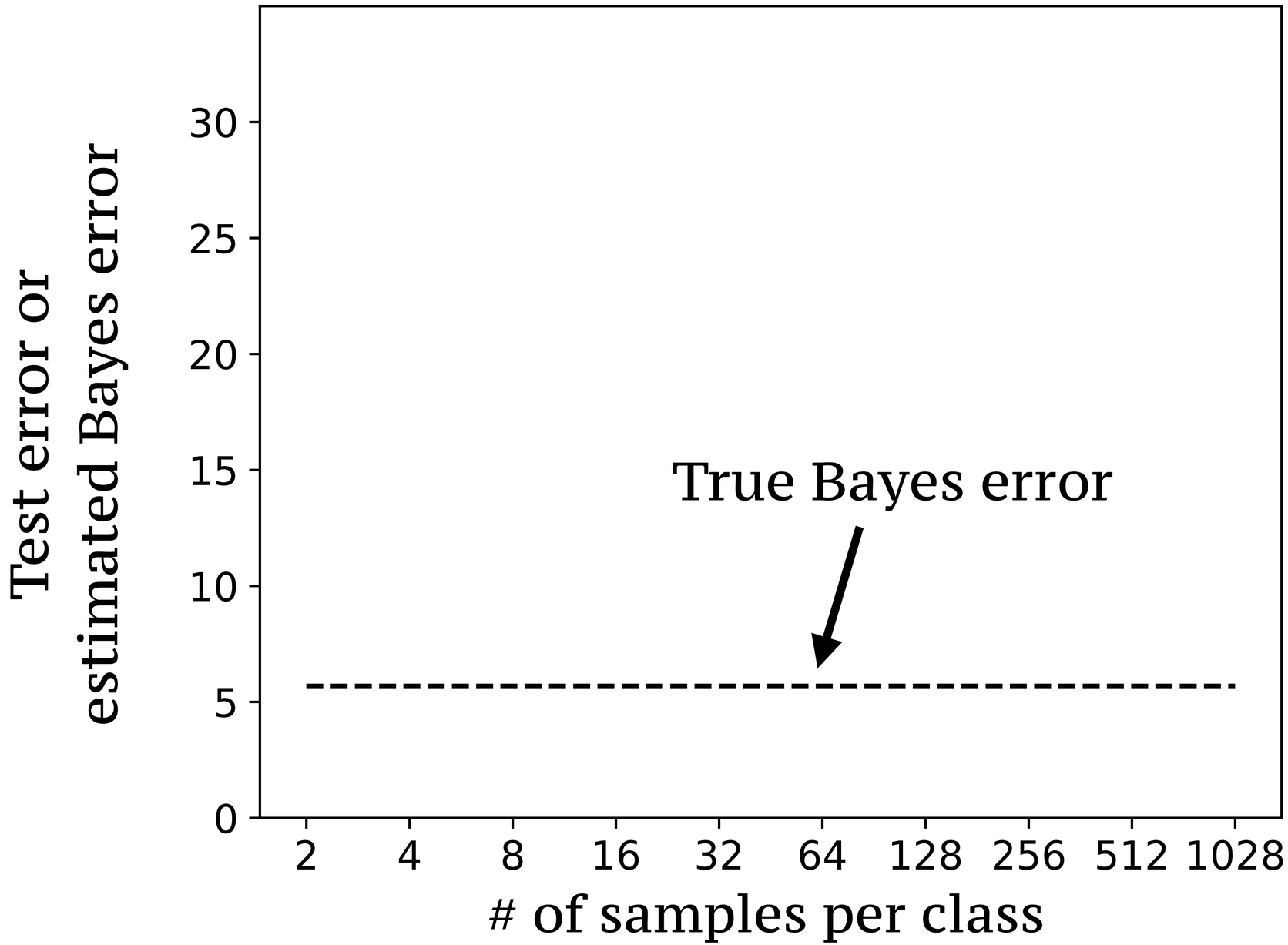
$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} \min\{c_i, 1-c_i\}$$

- Compared with previous methods, our method is model-free and instance-free.
- **Unbiasedness**: $\mathbb{E}[\hat{\beta}] = \beta$
- Consistency: $\forall\delta > 0$ w.p.a.l. $1-\delta$, $|\hat{\beta} - \beta| \leq \sqrt{\frac{1}{8n}\log\frac{2}{\delta}}$

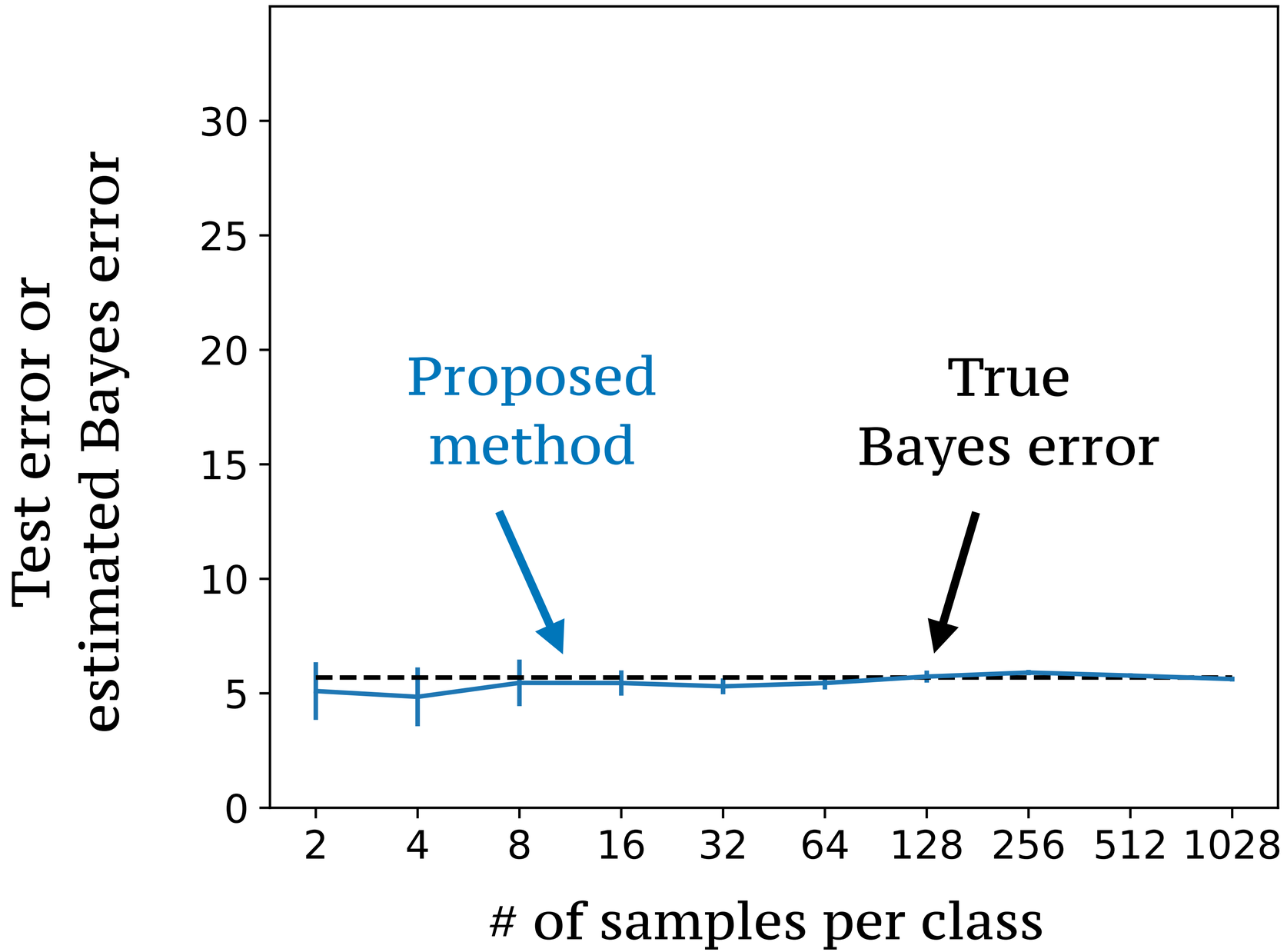
Synthetic experiments



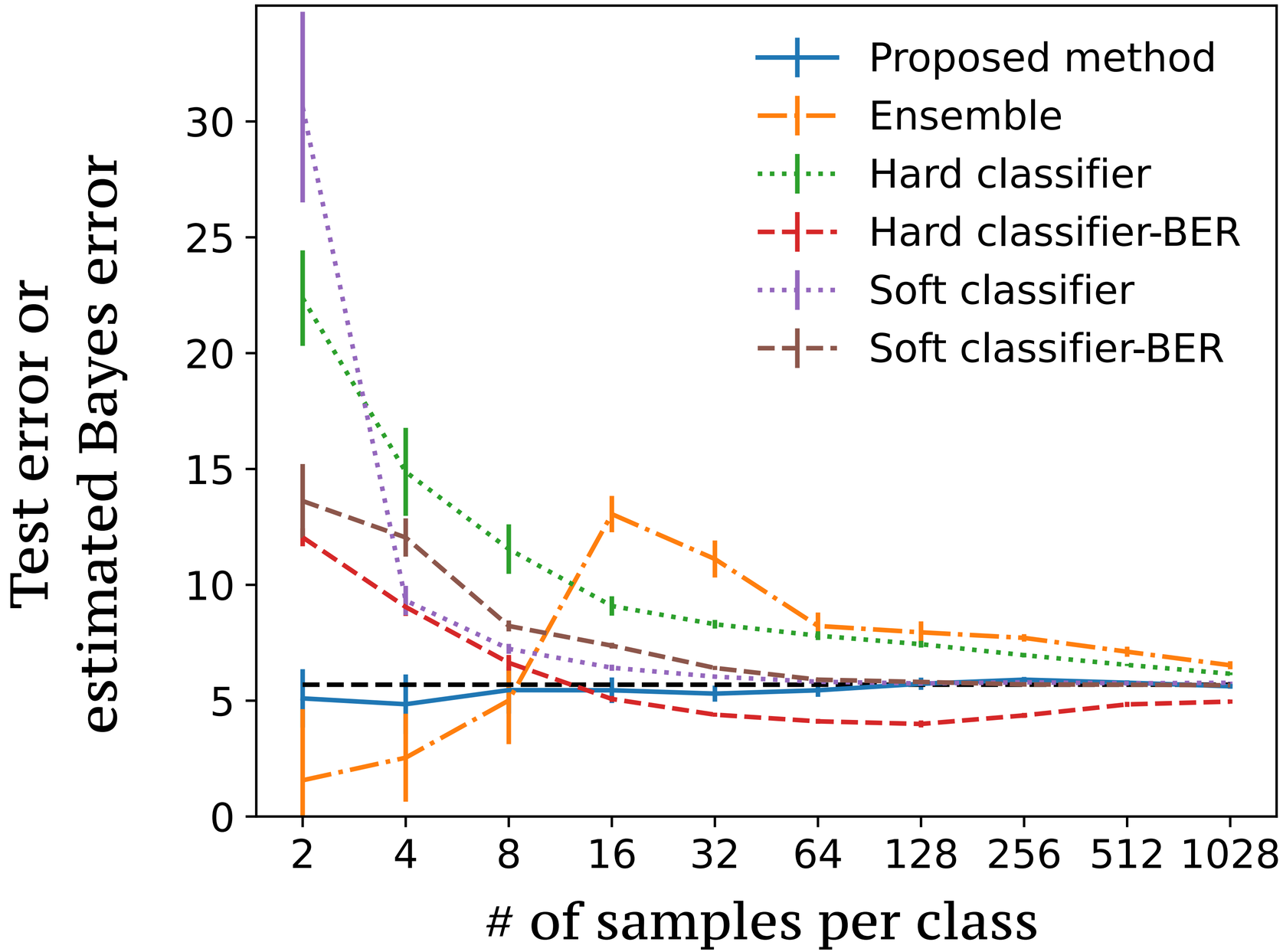
Synthetic experiments



Synthetic experiments



Synthetic experiments



Proposed method for more practical scenarios

- So far, we assumed we have "clean soft labels": $c_i = p(y = +1 \mid \boldsymbol{x}_i)$ for $\boldsymbol{x}_{i'}$ $\{\boldsymbol{x}_i\}_{i=1}^{n} \sim p(\boldsymbol{x})$
- In our paper, we study more realistic and practical scenarios:
  - Noisy soft labels and sign labels ($\hat{\beta}_{\mathrm{Noise}}$ is unbiased, consistent)

$$\text{Collect } \left\{\left(u_i, s_i\right)\right\}_{i=1}^{n} \text{ and use } \hat{\beta}_{\mathrm{Noise}} = \frac{1}{n}\left(\sum_{i=1}^{n_s^+}\left(1-u_i^+\right) + \sum_{i=1}^{n_s^-} u_i^-\right)$$

  - Multiple hard labels ($\tilde{\beta}$ is asymptotically unbiased)

$$\text{Collect } \left\{y_{i,j}\right\}_{j=1}^{m} \text{ for } i \in [n] \text{ and use } \tilde{\beta} = \frac{1}{n}\sum_{i} \min\left\{u_i, 1-u_i\right\} \text{ where } u_i = \frac{1}{m}\sum_{j} \mathbf{1}[y_{i,j}=1]$$

  - Weakly supervised scenarios, e.g., positive-confidence ($\hat{\beta}_{\mathrm{Pconf}}$ is unbiased, consistent)

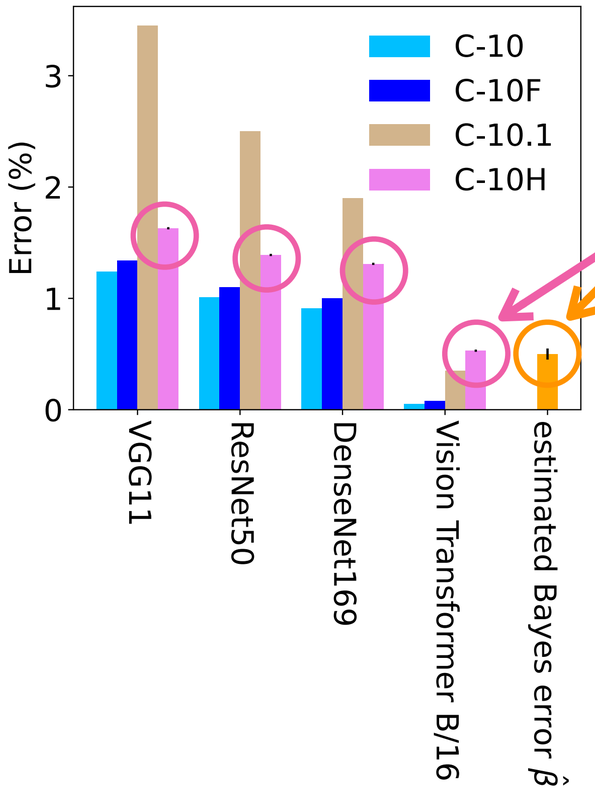
$$\text{Collect } \left\{r_i\right\}_{i=1}^{n^+}, \pi_+ \text{ and use } \hat{\beta}_{\mathrm{Pconf}} = \pi_+\left(1 - \frac{1}{n_+}\sum_{i=1}^{n_+} \max\left(0, 2 - \frac{1}{r_i}\right)\right)$$

$r_i = p(y=+1 \mid \boldsymbol{x}_i)$, $\pi_+ = p(y=+1)$, $s_i = \mathrm{sign}[p(y=+1 \mid \boldsymbol{x}_i) - 0.5]$, $n = n_s^+ + n_s^-$, n_s^+ and n_s^- are number of samples based on the sign labels
$y_{i,j}$ is the jth hard label for the ith instance, $\mathbf{1}$ is the indicator function, $u_i = c_i + \xi_i$, assumption is $\mathbb{E}[\xi_i \mid x_i] = 0$ and $0 \le u_i \le 1$

Experiments with benchmark datasets

CIFAR-10 (🐈🐕 vs 🚗✈️)

- Soft labels: class proportions of 50 human hard labels per image from CIFAR-10H (Peterson et al., ICCV 2019).



ViT's test error is already close to the Bayes error!

Fashion-MNIST (🧥👚 vs 👞👜)

- Inspired by CIFAR-10H, we present a new dataset: Fashion-MNIST-H
- Contains multiple hard labels annotated by humans for each image from Fashion-MNIST.

ResNet18's test error	3.852% (± 0.041%)
Estimated Bayes error	3.478% (± 0.079%)

Already close to the Bayes error!

ICLR papers (accept vs reject)

ICLR's Bayes error	
2017	6.8%(±1.0%)
2018	8.7%(±0.9%)
2019	7.9%(±0.7%)
2020	8.8%(±0.5%)
2021	9.3%(±0.5%)
2022	9.6%(±0.5%)
2023	8.0%(±0.4%)

Note: we covert the 10 class labels into binary labels

Experiments with real-world datasets

- We can study the intrinsic difficulty of datasets such as academic paper selection difficulty.
- We use the weighted average of the ICLR reviewers' scores based on the confidence of the reviewers. We further calibrate this since we need a soft label. (See further details in our paper.)
- Results: the Bayes error is between 6% and 10%.
  - Higher than CIFAR-10 and Fashion-MNIST.
  - No disruptive changes over the years, e.g., Bayes error is not doubling or tripling.
- Demonstrates the benefit of our instance-free approach! (It is non-trivial how we should train an accept/reject classifier, e.g., we need to consider all previous papers to evaluate the novelty.)

ICLR's Bayes error	
2017	$6.8\%(\pm 1.0\%)$
2018	$8.7\%(\pm 0.9\%)$
2019	$7.9\%(\pm 0.7\%)$
2020	$8.8\%(\pm 0.5\%)$
2021	$9.3\%(\pm 0.5\%)$
2022	$9.6\%(\pm 0.5\%)$
2023	$8.0\%(\pm 0.4\%)$

Data-centric ML via Bayes-Accuracy Estimation

- **PhishBencher**: publish leak-resilient, contamination-detectable benchmarks based on the controlled Bayes accuracy.
- **Direct Bayes-accuracy estimation**: audit irreducible difficulty of existing benchmarks.
- Together enable **data-centric machine learning**: robust evaluation before and after creation of benchmark datasets.

Check out our papers!

How Can I Publish My LLM Benchmark Without Giving the True Answers Away? arXiv:2505.18102, 2025. Oral presentation at MemFM workshop @ ICML 2025.

Thanawat Lodkaew

Ikko Yamane

Is the Performance of My Deep Network Too Good to be True? A Direct Approach to Estimating the Bayes Error in Binary Classification. ICLR 2023.

Ikko Yamane

Nontawat Charoenphakdee

Gang Niu

Masashi Sugiyama