Adaptive Prediction-Powered AutoEval with Reliability and Efficiency Guarantees

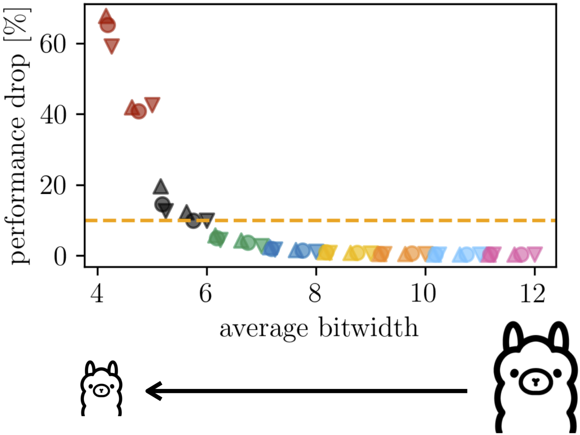
Sangwoo Park, Matteo Zecchin, Osvaldo Simeone

NeurIPS 2025

December 2025

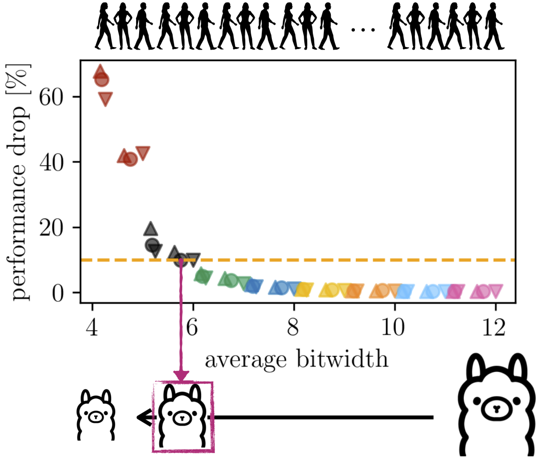
Reliable Model Selection

- Example: Find the lightest quantized LLM that **guarantees at most a 10% performance drop** as compared to the unquantized version.



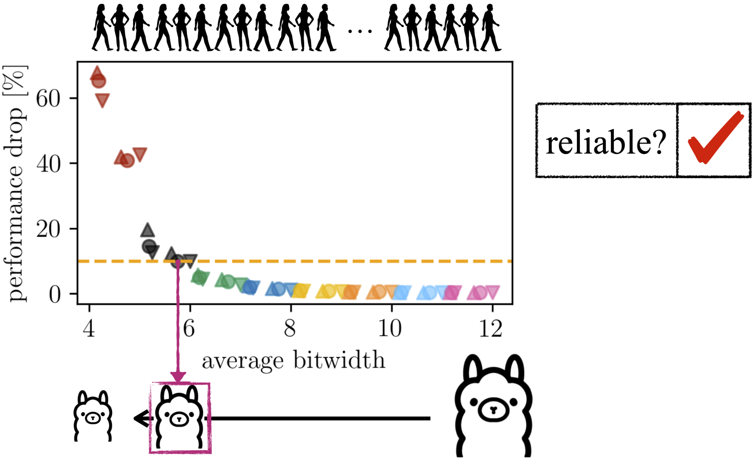
Reliable Model Selection

- If we have **abundant amount of real-world, human-labeled, data**, we can **precisely evaluate the average performance** to find out the lightest quantized model.



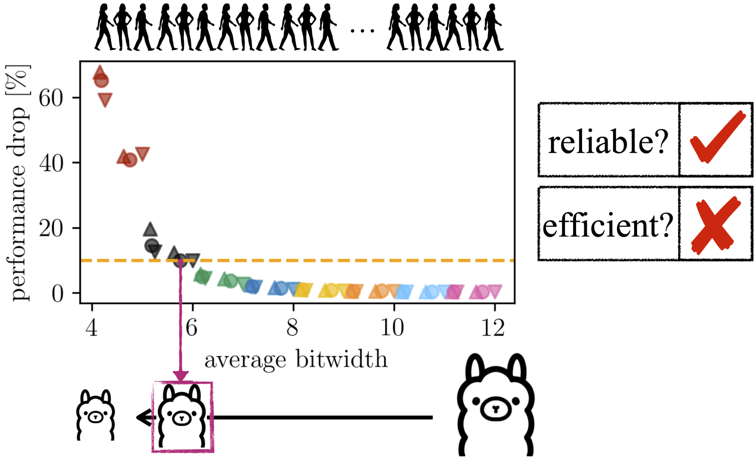
Reliable Model Selection

- Given **abundant** (nearly infinity) amount of data, **empirical averaging matches with the true expectation**, making the model selection reliable.



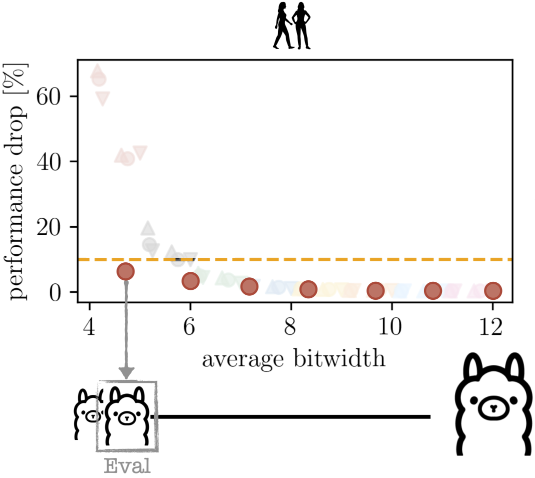
Reliable Model Selection

- However, such approach is **highly inefficient** in the sense that it requires nearly infinite amount of real-world data



Reliable Model Selection

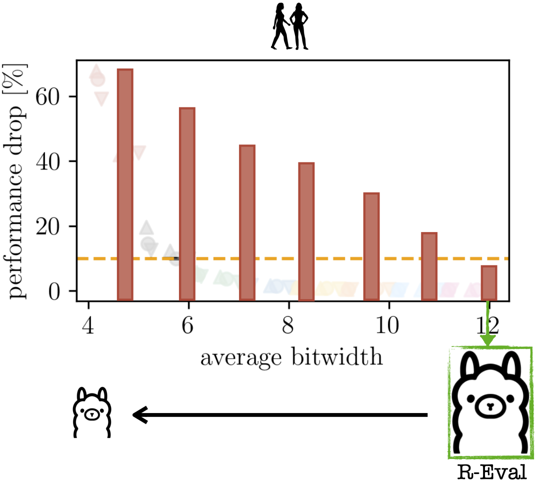
- And such **mean-based** approach (`Eval`) becomes **unreliable** in the presence of **limited amount of real-world data**.



reliable?	✗
efficient?	–

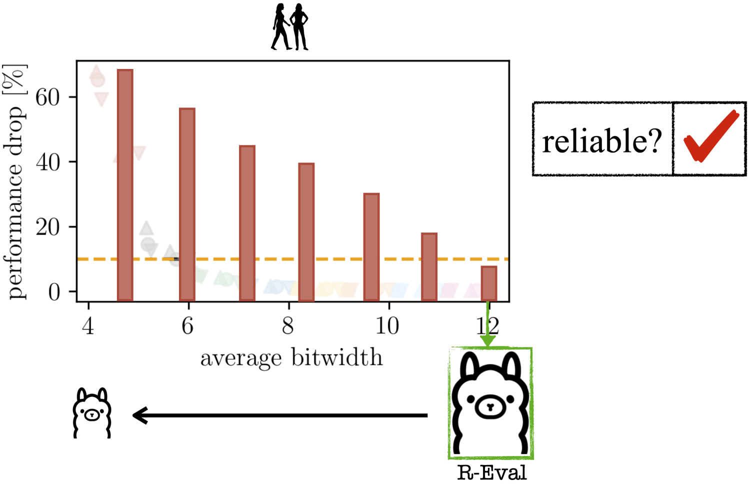
Reliable Model Selection

- Reliable `Eval` (`R-Eval`) rigorously identifies the **bounds** that contain the true, unknown, expected performance ...



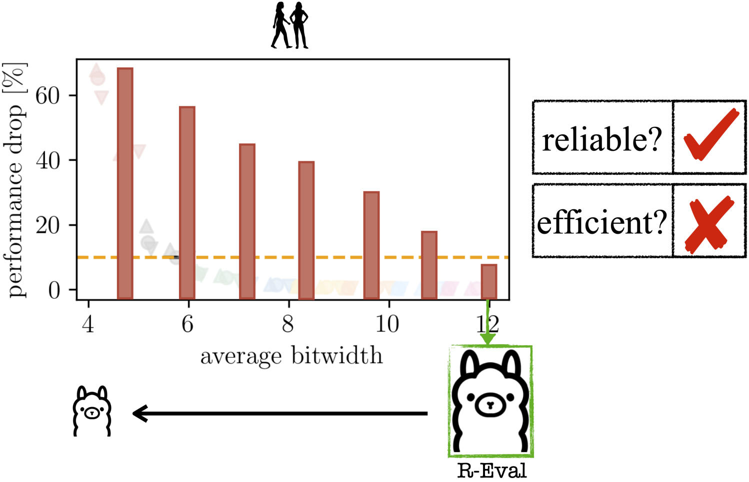
Reliable Model Selection

- … making the corresponding model selection **reliable**.



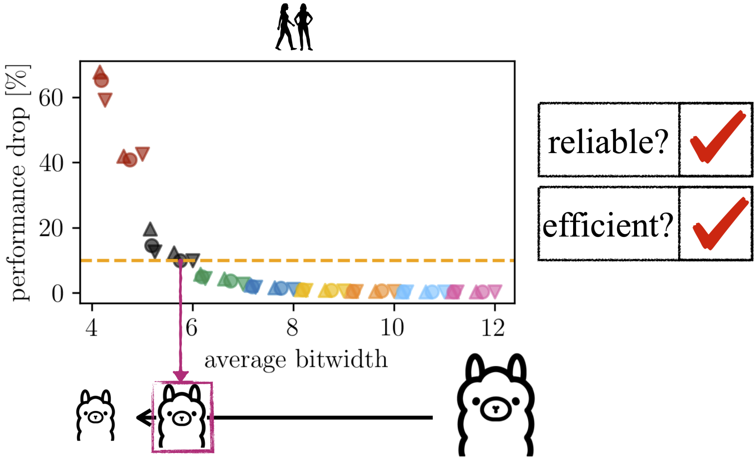
Reliable Model Selection

- However, in the presence of **few amount of real-world data**, such bounds tend to be too **conservative**.



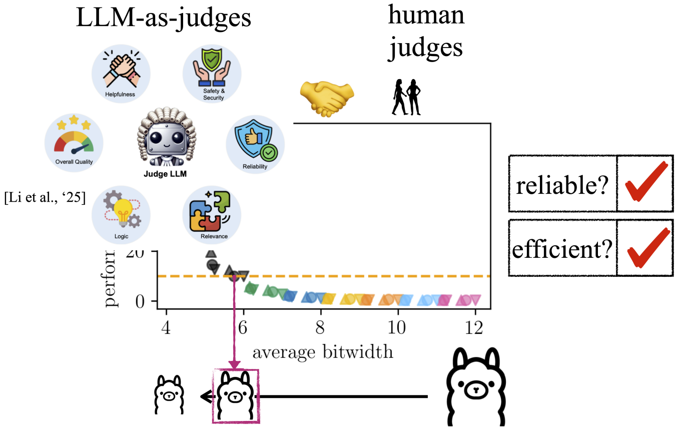
Reliable and Efficient Model Selection

- In this work, we aim at achieving **reliable and efficient** model selection in the presence of **few human-labeled data**.



Reliable and Efficient Model Selection

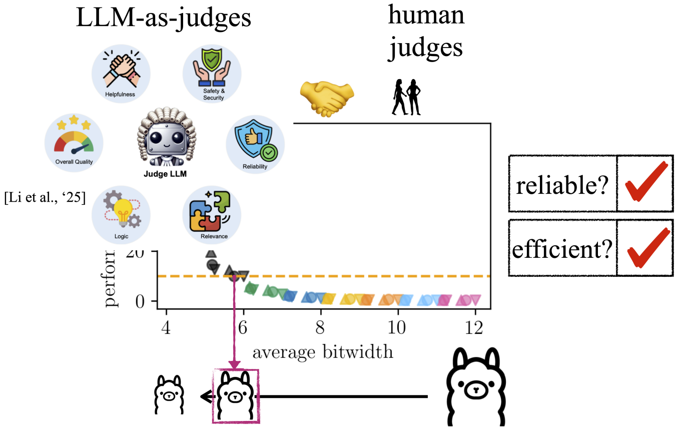
- The key idea is to incorporate **simulated data**, e.g., **LLM-labeled data**[1]



[1] Dawei Li et al. "From generation to judgment: Opportunities and challenges of llm-as-a-judge, 2025". In: *URL https://arxiv. org/abs/2411.16594* (2025).

Reliable and Efficient Model Selection

- Such approach can be categorized as **semi-supervised inference/testing**.



State of the Art

- Semi-supervised inference using a pre-trained autoevaluator[2,3]
  - Useful with good autoevaluator
  - Worse than supervised with bad autoevaluator
- Semi-supervised test that achieves **better efficiency** at the cost of **losing finite-sample reliability guarantees**[4,5]
- Semi-supervised test that **maintains finite-sample reliability** guarantees with **unknown efficiency gain**[6]
- **Achieving both** was believed to be **impossible**[7]

---

[2] Anastasios N Angelopoulos et al. "Prediction-powered inference". In: *Science* 382.6671 (2023), pp. 669–674.

[3] Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. "Ppi++: Efficient prediction-powered inference". In: *arXiv preprint arXiv:2311.01453* (2023).

[4] Pierre Boyeau et al. "Autoeval done right: Using synthetic data for model evaluation". In: *arXiv preprint arXiv:2403.07008* (2024).

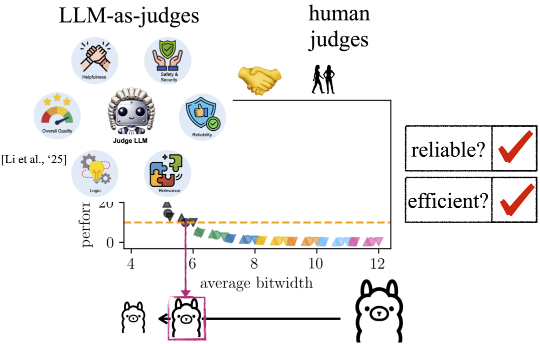
[5] Adam Fisch et al. "Stratified prediction-powered inference for effective hybrid evaluation of language models". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 111489–111514.

[6] Bat-Sheva Einbinder, Liran Ringel, and Yaniv Romano. "Semi-supervised risk control via prediction-powered inference". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).

[7] Pranav Mani et al. "No Free Lunch: Non-Asymptotic Analysis of Prediction-Powered Inference". In: *arXiv preprint arXiv:2505.20178* (2025).

Main Contribution

- The proposed `R-AutoEval+` achieves **both finite-sample reliability guarantees** and **efficiency guarantees**.
- Testing-by-betting[8] and prediction-powered inference[9]

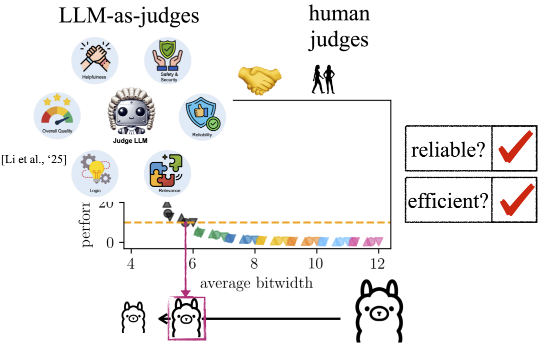


[8] Ian Waudby-Smith and Aaditya Ramdas. "Estimating means of bounded random variables by betting". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 86.1 (2024), pp. 1–27.

[9] Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. "Ppi++: Efficient prediction-powered inference". In: *arXiv preprint arXiv:2311.01453* (2023).

Main Contribution

- The proposed `R-AutoEval+` achieves **both finite-sample reliability guarantees** and **efficiency guarantees**.
- Testing-by-betting[8] and prediction-powered inference[9]

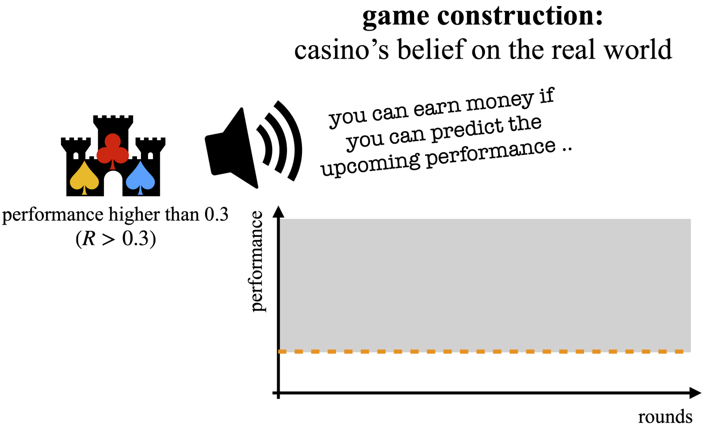


[8] Ian Waudby-Smith and Aaditya Ramdas. "Estimating means of bounded random variables by betting". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 86.1 (2024), pp. 1–27.

[9] Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. "Ppi++: Efficient prediction-powered inference". In: *arXiv preprint arXiv:2311.01453* (2023).

Background: Testing-by-Betting

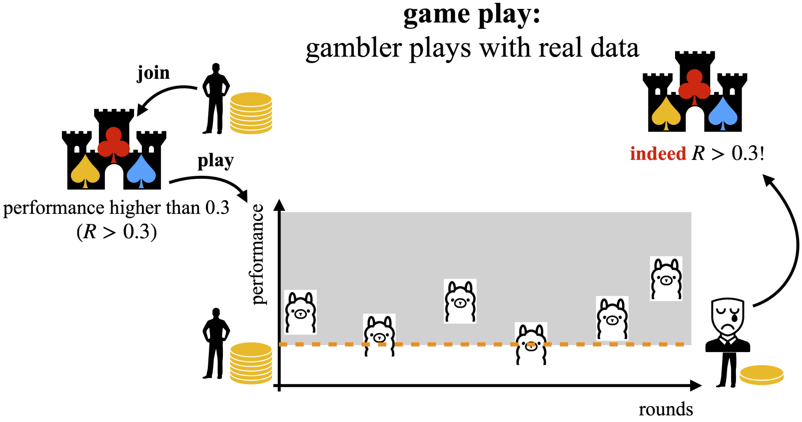
- Testing-by-betting[10] reliably estimates the unknown mean of bounded random variables by constructing a **game** with which **casino will never lose their wealth on average if their belief on the mean were correct**.



[10] Ian Waudby-Smith and Aaditya Ramdas. "Estimating means of bounded random variables by betting". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 86.1 (2024), pp. 1–27.

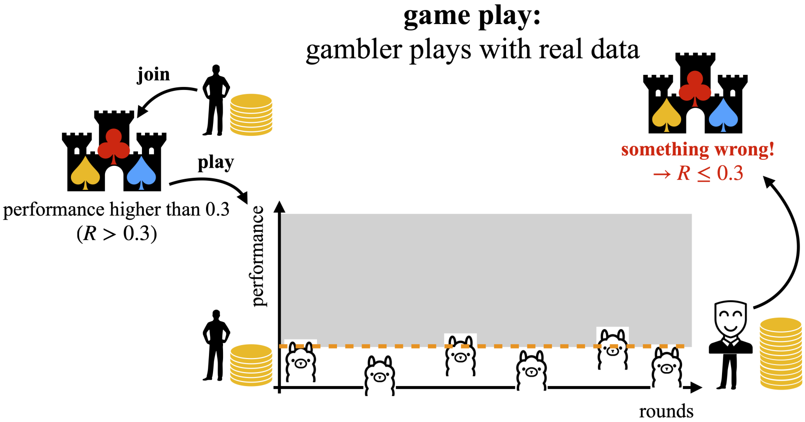
Background: Testing-by-Betting

- And the actual outcome of the **game based on the observed random variables** will tell us whether the casino's mean assumption is correct ...



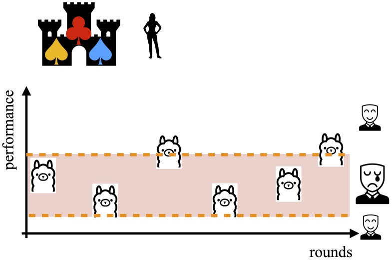
Background: Testing-by-Betting

- … or incorrect.



Background: Testing-by-Betting

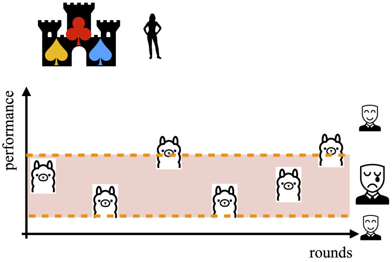
- By collecting all the casino's assumptions that makes gamblers unhappy, one can construct a **reliable confidence interval** for the unknown mean



- Game construction: $= \cdot \left(1 - \lambda_i(-)\right)$
  - Pr $\left[\exists i, \geq 100 \middle| \right] \leq 0.01$
- Real data at round i:
- Gambler's betting at round i: $\lambda_i \in \left[0, \frac{1}{1 - }\right]$
- Confidence interval: $\{r \in [0, 1] : \max_n \leq 100\} \ni R$ w.p. 0.99

Background: Testing-by-Betting

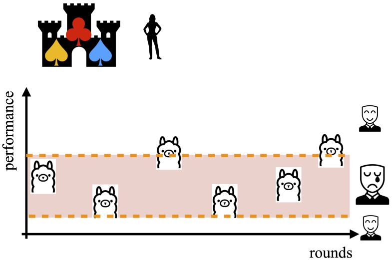
- In fact, the **observation** for testing-by-betting can be anything as long as it is an **unbiased, bounded, estimate of the true mean** ...



- Game construction: $E_i = E_{i-1} \cdot \left(1 - \lambda_i(\hat{R}_i - \alpha)\right)$
  - Pr $\left[\exists i, E_i \geq 100 \middle| R > \alpha\right] \leq 0.01$
- Real data at round i: $\hat{R}_i$
- Gambler's betting at round i: $\lambda_i \in [0, \frac{1}{1-\alpha}]$
- Confidence interval: $\{r \in [0,1] : \max_n \mathrm{E}_n(r) \leq 100\} \ni R$ w.p. 0.99

Background: Testing-by-Betting

- … and it would be better if it has **low variance** so that the gamblers can play their game better to earn as much wealth as possible.



- Game construction: $E_i = E_{i-1} \cdot \left(1 - \lambda_i(\hat{R}_i - \alpha)\right)$
  - $\Pr\left[\exists i, E_i \geq 100 \middle| R > \alpha\right] \leq 0.01$
- Real data at round i: $\hat{R}_i$

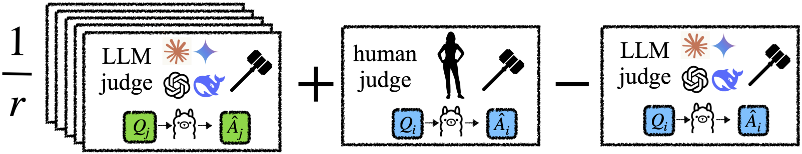
should be **unbiased**;
better if it has **low variance**

- Gambler's betting at round i: $\lambda_i \in [0, \frac{1}{1-\alpha}]$
- Confidence interval: $\{r \in [0,1] : \max_n \mathrm{E}_n(r) \leq 100\} \ni R$ w.p. 0.99

Background: Prediction-Powered Inference

- To this end, **semi-supervised risk control**[11] adopts **prediction-powered inference** (PPI)[12] to replace the observation with the following semi-supervised, unbiased, estimate of the mean:

$$\hat{R}_i^{PP} := \frac{1}{r} \sum_{j=i\cdot r}^{(i+1)\cdot r} \ell(\tilde{X}_j, f(\tilde{X}_j)) + \ell(X_i, Y_i) - \ell(X_i, f(X_i))$$



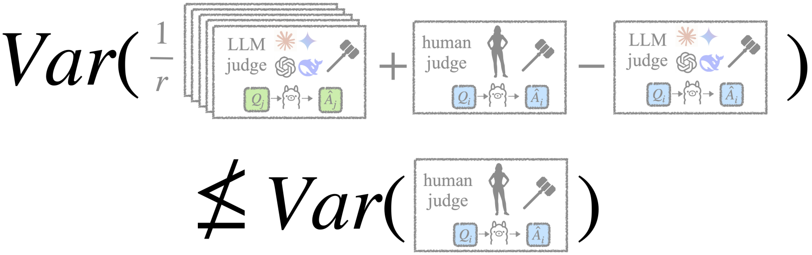
[11] Bat-Sheva Einbinder, Liran Ringel, and Yaniv Romano. "Semi-supervised risk control via prediction-powered inference". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).

[12] Anastasios N Angelopoulos et al. "Prediction-powered inference". In: *Science* 382.6671 (2023), pp. 669–674.

Background: Prediction-Powered Inference

- However, unless the synthetic data is of sufficiently high quality, semi-supervised approach can easily yield even **worse result** than pure, supervised, approach[8,9]

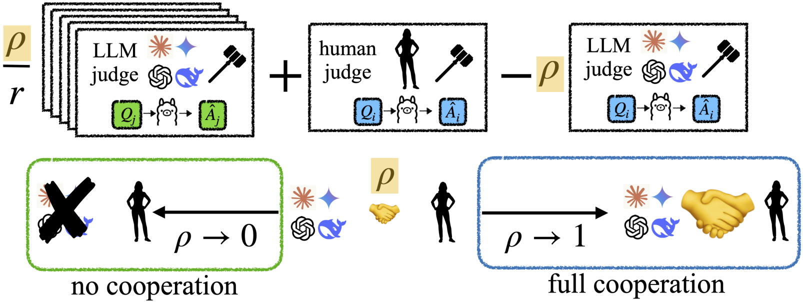
$$\hat{R}_i^{PP} := \frac{1}{r}\sum_{j=i\cdot r}^{(i+1)\cdot r} \ell(\tilde{X}_j, f(\tilde{X}_j)) + \ell(X_i, Y_i) - \ell(X_i, f(X_i))$$



Background: Prediction-Powered Inference

- One possible way is to consider PPI++[13] ...

$$\hat{R}_i^{\rho} := \frac{\rho}{r} \sum_{j=i\cdot r}^{(i+1)\cdot r} \ell(\tilde{X}_j, f(\tilde{X}_j)) + \ell(X_i, Y_i) - \rho \cdot \ell(X_i, f(X_i))$$

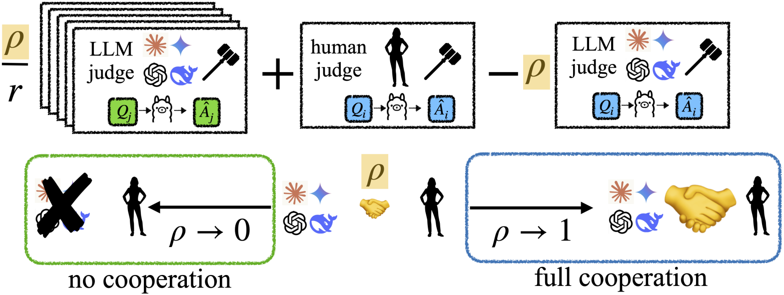


[13] Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. "Ppi++: Efficient prediction-powered inference". In: *arXiv preprint arXiv:2311.01453* (2023).

Background: Prediction-Powered Inference

- … but tuning the **reliance parameter** ρ requires real-world data hence **either harms reliability** (data reuse)[14,15] or **harms efficiency** (data split)[8]

$$\hat{R}_i^{\rho} := \frac{\rho}{r} \sum_{j=i\cdot r}^{(i+1)\cdot r} \ell(\tilde{X}_j, f(\tilde{X}_j)) + \ell(X_i, Y_i) - \rho \cdot \ell(X_i, f(X_i))$$

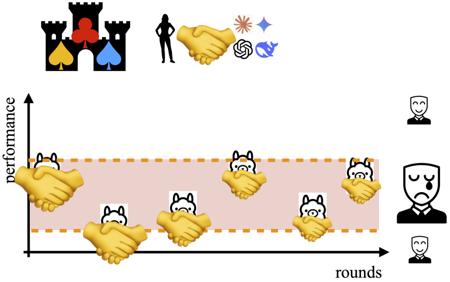


[14] Pierre Boyeau et al. "Autoeval done right: Using synthetic data for model evaluation". In: *arXiv preprint arXiv:2403.07008* (2024).

[15] Adam Fisch et al. "Stratified prediction-powered inference for effective hybrid evaluation of language models". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 111489–111514.

Proposed: R-AutoEval+

- The proposed R-AutoEval+ achieves **both reliability and efficiency** by **adaptively tuning the reliance parameter** ρ during the execution of the game.

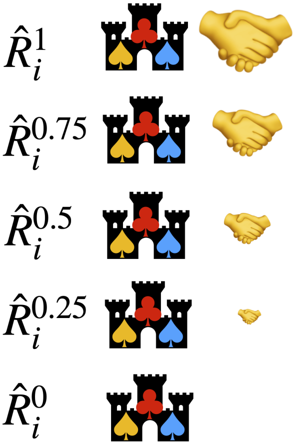


- Game construction: $E_n = E_{n-1} \cdot \left(1 - \lambda_i(\hat{R}_i^{\rho_i} - 0.3)\right)$
  - $\Pr\left[E_n > 100 \middle| R > 0.3\right] \leq 0.01, \forall n$
- Real data at round i: $\hat{R}_i^{\rho_i}$

should be **unbiased**;
better if it has **low variance**

- Gambler's betting at round i: $\lambda_i \in [0, 1/0.7]$
- 99% confidence sequence: $\{r \in [0,1] : E_n(r) \leq 100\} \ni R$ w.p. 99%

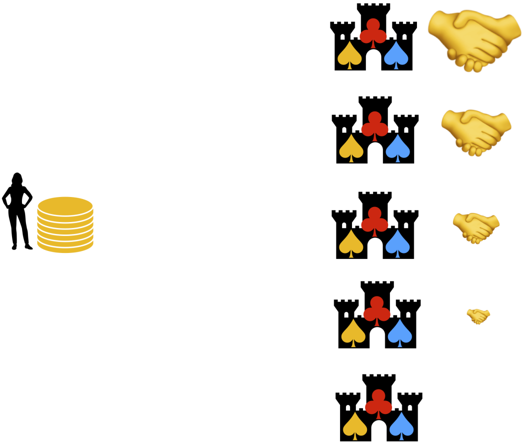
- The main idea is to consider **multiple** casinos whose games are designed based on **different reliance parameters** ranging $0, \dots, 1$, motivated by classical portfolio rebalancing idea[16].



[16] Thomas M Cover and Erik Ordentlich. "Universal portfolios with side information". In: *IEEE Transactions on Information Theory* 42.2 (2002), pp. 348–363.

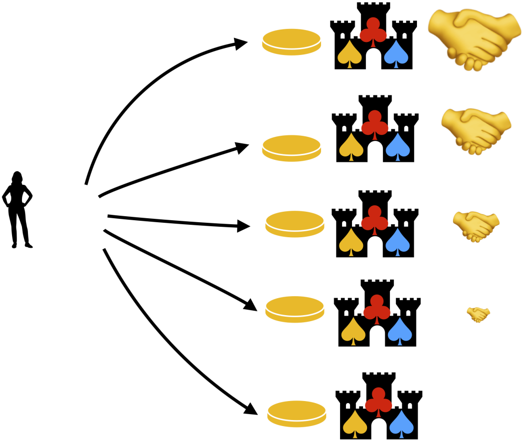
R-AutoEval+: Main Idea

- Given some initial wealth, ...



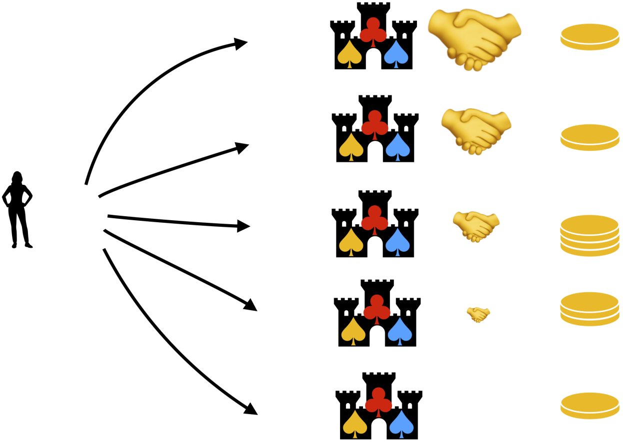
R-AutoEval+: Main Idea

- … a gambler can equally invest her wealth to each of the casinos …



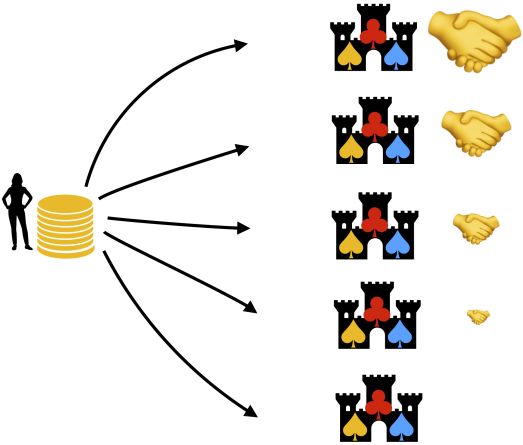
R-AutoEval+: Main Idea

- ... to get the updated wealth from each casino ...



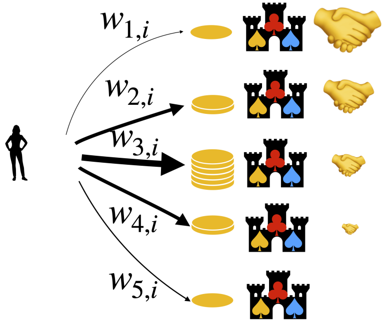
R-AutoEval+: Main Idea

- ... which will in total be her overall wealth ...



R-AutoEval+: Main Idea

- … but now she will **invest differently** for each casino based on the **previous performance of each casino**.



$$E_i = E_{i-1} \cdot \sum_{s=1}^{S} w_{s,i}(1 - \lambda_{s,i}(\hat{R}_i^{\rho_s} - \alpha))$$

R-AutoEval+: What it Achieves

- `R-AutoEval+` **provably achieves** finite-sample reliability guarantees as well as efficiency guarantees:

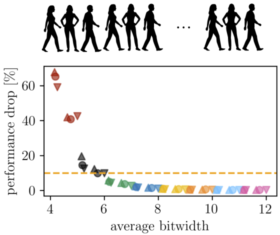
Theorem (Sample complexity of `R-AutoEval+`)

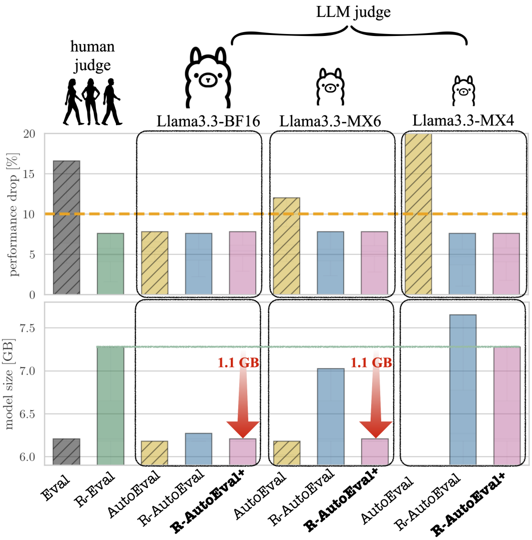
R-AutoEval+ ***provably saves*** *(at least no waste) the real-world data* ***irrespective*** *of the quality of the autoevaluator*

$$\lim_{\delta \to 0^+} \frac{n_{min}^{\texttt{R-AutoEval+}}(\delta)}{\log(1/\delta)} \leq \min_{s=1,\dots,S} \left\{ \frac{1}{g_{s,\star}} \right\} \leq \min \left\{ \frac{1}{g_{1,\star}}, \frac{1}{g_{S,\star}} \right\}. \quad (1)$$

Experimental Results

- Selecting the quantized LLM:





Experimental Results

- Selecting the test-time reasoning budget:

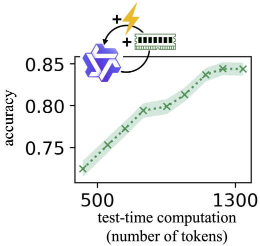


Table 1: Selecting the smallest reasoning budget for Qwen3-1.7B that ensures at least 3% accuracy enhancement as compared to its non-reasoning mode, evaluated on GSM8K data set [13]: average number of generated tokens with standard deviation shown within parentheses.

autoevaluator (accuracy)	R-Eval	R-AutoEval	R-AutoEval+
BitNet b1.58 (35%)	983.34 (137.87)	950.27 (152.84)	**942.47 (135.65)**
Llama-3.2-3B-Instruct (66%)		900.58 (122.70)	**892.86 (112.10)**
Qwen3-32B (82%)		1007.45 (150.48)	**941.20 (129.61)**
DeepSeek-R1-Distill-Qwen-32B (89%)		893.39 (105.13)	**866.22 (80.58)**
Llama-3.3-70B-Instruct (89%)		854.42 (90.26)	**847.05 (69.21)**
GPT-4.1 (93%)		883.99 (103.00)	**856.13 (70.93)**

Conclusion

- Reliable model selection requires **reliable evaluation** of the model performance, which in general requires **substantial amount of real-world, human-labeled, data**.
- Simulated, LLM-labeled, data may supplement the real-world data, but it could not come with both reliability and efficiency guarantees.
- The proposed `R-AutoEval+` achieves **both reliability and efficiency guarantees**.
  - The main idea is to incorporate testing-by-betting, `PPI++`, and portfolio rebalancing.
- Future work may consider incorporating active labeling to further enhance the efficiency of the method.