Evaluating LLMs When They Do Not Know the Answer

Statistical Evaluation of Mathematical Reasoning via Comparative Signals

Zhixian Zhang

June 18, 2026

Collaborators

Zihan Dong

Yang Zhou

Can Jin

Ruijia Wu

Linjun Zhang

Content

1. Baseline
2. Motivation
3. Methodology
4. Theory
5. Experiment

Baseline

- Evaluate the LLM's mathematical reasoning ability.

[1] A live benchmark routinely updates its problem set over time.

Baseline

- Evaluate the LLM's mathematical reasoning ability.

- Accuracy varies significantly within the same live benchmark[1] [1].

Accuracy by month

Model	Reasoning	Nov 2025	Dec 2025	Jan 2026	Feb 2026	Overall
Qwen3.5-397B-A17B	enabled	37.9%	23.1%	47.8%	36.0%	35.6%
GPT-OSS-120B	high	27.6%	26.9%	34.8%	26.0%	28.8%

[1] A live benchmark routinely updates its problem set over time.

Baseline

- Estimate accuracy.
- Accuracy's empirical estimator varies significantly within the same live benchmark.

Baseline

- Estimate accuracy.
- Accuracy's empirical estimator varies significantly within the same live benchmark.

- X: question;
  Y: target model's answer;
  G: ground-truth.
- $\theta = \mathbb{E}\big[\mathbb{1}\{Y = G\}\big] \in \mathbb{R}$: target parameter (accuracy).
- Given N i.i.d. samples $\{(x_i, y_i, g_i)\}_{i=1}^{N}$, the naive estimator is

$$\hat{\theta}^{\text{naive}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{y_i = g_i\}.$$

Motivation

- Estimate $\theta = \mathbb{E}\big[\mathbb{1}\{Y = G\}\big]$: a one-dimensional mean estimation problem.
- MLE is the best.

Goal: construct an estimator with lower-variance.

Motivation

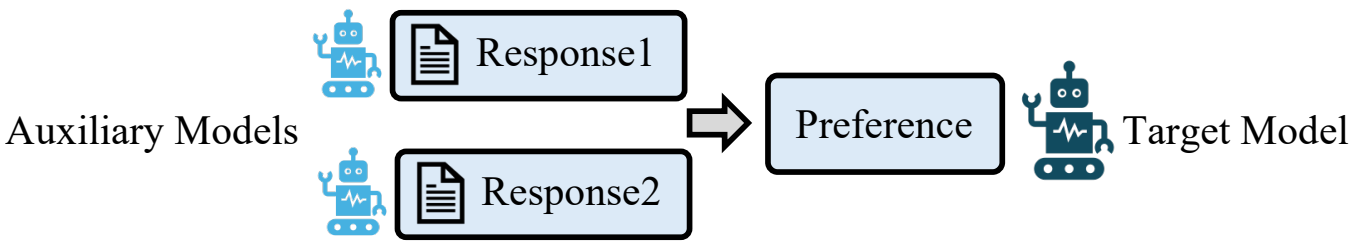
- Estimate $\theta = \mathbb{E}\big[\mathbb{1}\{Y = G\}\big]$: a one-dimensional mean estimation problem.
- MLE is the best.

Goal: construct an estimator with lower-variance.

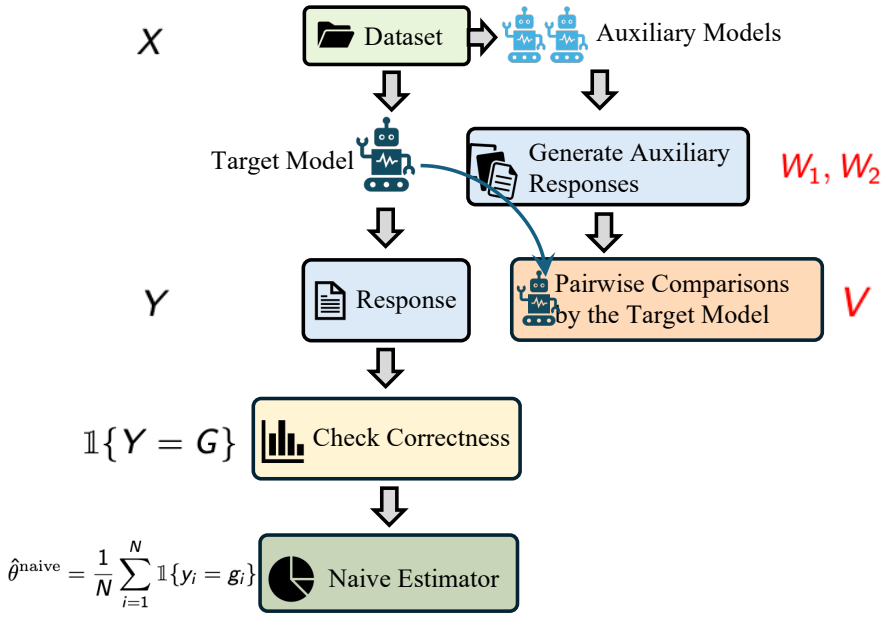
Use additional information.

- What information?
- How to use?

- Generative AI. Query the LLM.
- Pairwise comparison: RLHF[2], DPO [3]. Judging is easier than generating.

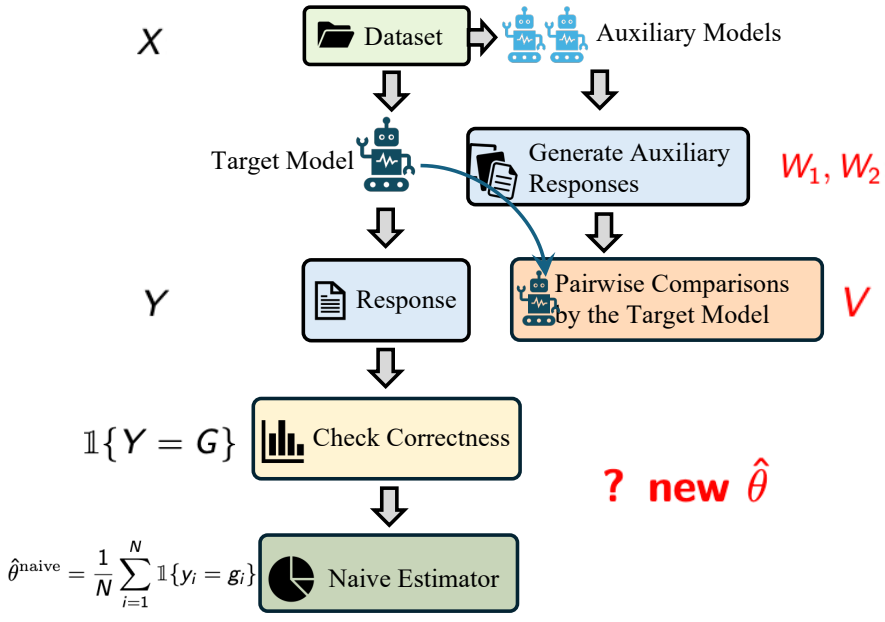


Motivation



- W_1, W_2: two auxiliary responses.
- V: target model's pairwise preference.
- $Z = (W_1, W_2, V)$: auxiliary information.

Motivation



- W_1, W_2: two auxiliary responses.
- V: target model's pairwise preference.
- $Z = (W_1, W_2, V)$: auxiliary information.

How to use Z?
EIF.

Methodology

- Use Efficient Influence Function(EIF).
- Query the LLMs to generate Z repeatly.

Definition 1 (Model Space for AI Evaluation)

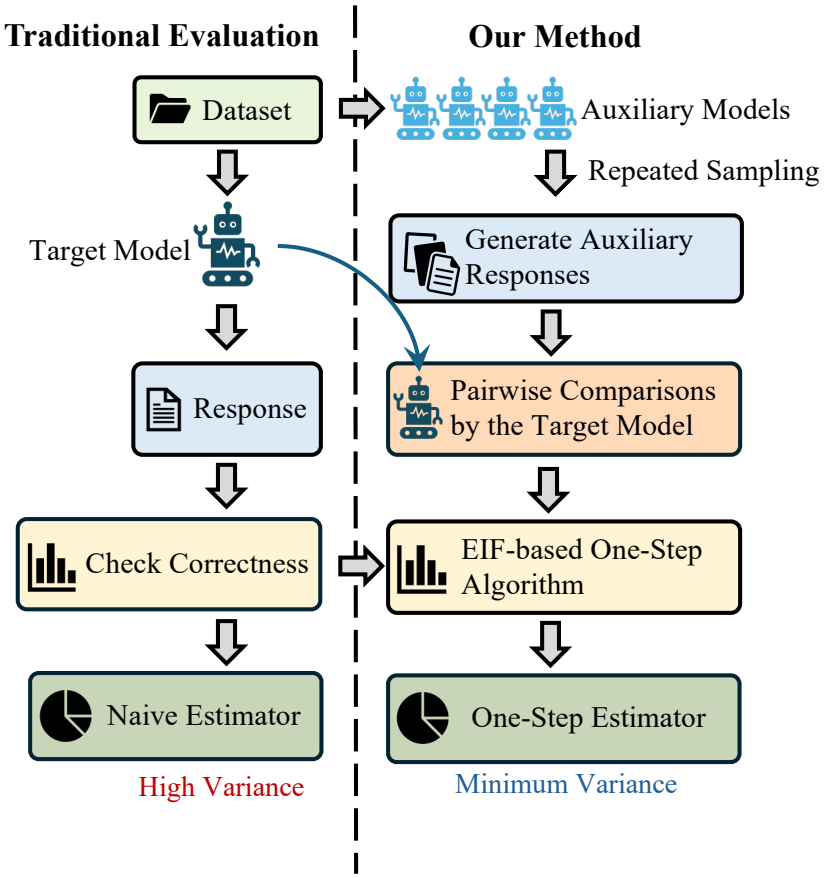
$$\mathcal{M} = \{P : p(z \mid x) \text{ is known and fixed by the evaluation protocol}\}.$$

Proposition 1 (Efficient Influence Function)

$$\psi(X, Y, G, Z) = \big(m(X) - \theta\big) + \big(\mathbb{1}\{Y = G\} - \tau(X, Z)\big),$$

where $\tau(X, Z) = \mathbb{E}[\mathbb{1}\{Y = G\} \mid X, Z]$ *and* $m(X) = \mathbb{E}[\mathbb{1}\{Y = G\} \mid X]$.

Methodology



- Estimate $\tau(X, Z) = \mathbb{E}[\mathbb{1}\{Y = G\} \mid X, Z]$ and $m(X) = \mathbb{E}[\mathbb{1}\{Y = G\} \mid X]$ to obtain $\hat{\tau}$ and $\hat{m}$.
- $\hat{\theta}_{\text{1-step}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{y_i = g_i\} - \hat{\tau}(x_i, z_i) + \hat{m}(x_i)$.

Theory

- **Result: Asymptotic normality.**

$$\sqrt{N}\left(\hat{\theta}_{\text{1-step}} - \theta\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\text{eff}}).$$

- Our estimator attains the semiparametric efficiency lower bound.

Experiment

GPQA Diamond

Model	GT%	Naive%	One-step%	Improv.
Gemini-3-Flash-Preview	90.40	92.00	91.20	+0.80%
GPT-5.2	86.36	84.00	85.60	+1.60%
DeepSeek-V3.2(Thinking)	84.85	94.00	88.80	+5.20%
Grok-4-1-Fast-Reasoning	83.33	88.00	85.80	+2.20%
Claude-Sonnet-4.5	82.83	78.00	80.60	+2.60%
Qwen3-Next-80B-A3B-Instruct	76.26	82.00	76.80	+5.20%
DeepSeek-R1-Distill-Llama-70B	59.09	60.00	59.00	+0.82%
QwQ-32B-Preview	56.06	60.00	55.40	+3.28%
Llama-3.3-70B-Instruct	46.97	32.00	37.60	+5.60%
DeepSeek-R1-Distill-Qwen-32B	43.43	40.00	41.40	+1.40%

GT% denotes ground truth accuracy. Naive% and One-step% are computed with $N = 50$ samples. Improv. $= |\text{Naive\%} - \text{GT\%}| - |\text{One-step\%} - \text{GT\%}|$. Config 1 uses GPT-5 mini for both auxiliary models.

Acknowledgement

Thanks for listening


Linyang He, Qiyao Yu, Hanze Dong, Baohao Liao, Xinxing Xu, Micah Goldblum, Jiang Bian, and Nima Mesgarani.
Livemathematicianbench: A live benchmark for mathematician-level reasoning with proof sketches, 2026.


Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano.
Learning to summarize with human feedback.
Advances in neural information processing systems, 33:3008–3021, 2020.


Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei.
Deep reinforcement learning from human preferences.
Advances in neural information processing systems, 30, 2017.