How to *Correctly* Report LLM-as-a-Judge Evaluations

Chungpa Lee, Thomas Zeng, Jongwon Jeong, Jy-yong Sohn, Kangwook Lee

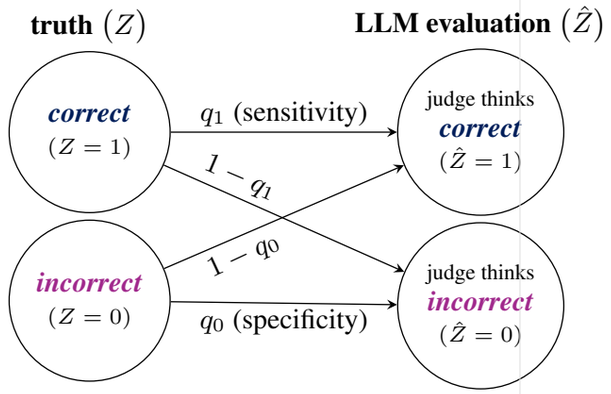
YONSEI UNIVERSITY · KRAFTON · WISCONSIN UNIVERSITY OF WISCONSIN–MADISON · ludo robotics

Large language models are widely used as scalable evaluators in place of humans. **Are their judgments truly reliable?**

We **propose a simple correction method that accounts for errors in LLM judgments**. Our method requires a dataset with human-evaluated examples, which we use to estimate and correct the errors made by the LLM judge. We further characterize when corrected LLM-as-a-judge evaluation is more reliable than human-only evaluation without LLMs. Our estimator remains unbiased even under distribution shift, where the evaluation dataset differs from the dataset used for correction.

Imperfect LLM-as-a-judge evaluations

Problem setup. We evaluate a test dataset using an LLM judge. Each response has a true label indicating whether it is ***correct*** ($Z = 1$) or ***incorrect*** ($Z = 0$), and the LLM judge provides a corresponding judgment ($\hat{Z}$). There are two types of error:



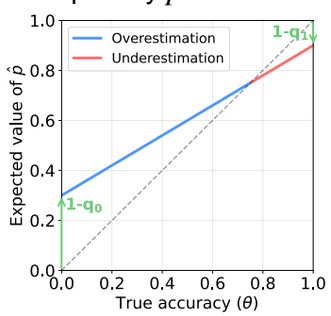
Target quantity:

the true accuracy on the test dataset $\theta := \mathbb{E}[Z]$

Observed quantity:

the LLM-evaluated accuracy on the test dataset $\hat{p} := \mathbb{E}[\hat{Z}]$

Bias from Imperfect LLM Judgments. Errors in LLM judgments can make the observed quantity $\hat{p}$ a biased estimate of the target quantity θ.

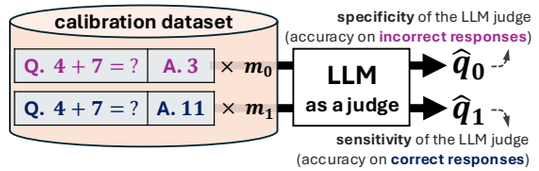


For example, in the plotted case ($1 - q_0 = 0.3$ and $1 - q_1 = 0.1$), $\mathbb{E}[\hat{p}]$ over-estimates θ when θ is low (blue), and under-estimates it when θ *is* high (red).

$\hat{p}$: **a biased estimator of** θ

How can we correct this?

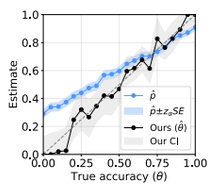
No free lunch: we must collect the calibration dataset.
To correctly report LLM-as-a-judge evaluations, we need to explicitly estimate the LLM judge's errors ($1 - q_0$ and $1 - q_1$). Therefore, we require a calibration dataset where each response has both an LLM judgment and a true label.



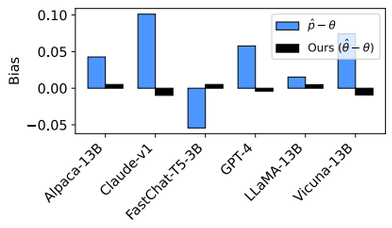
Using this calibration dataset, we estimate the judge's sensitivity $\hat{q}_1$ and specificity $\hat{q}_0$, and debias the LLM-evaluated accuracy.

$$\hat{\theta} := \frac{\hat{p} + \hat{q}_0 - 1}{\hat{q}_0 + \hat{q}_1 - 1} \quad : \textbf{ an unbiased estimator of } \theta$$

Empirical validation.



(a) Monte Carlo Simulation



(b) Chatbot Arena Benchmark

The proposed method reduces bias in both cases.

Our estimator is robust under distribution shift.

Assume $\Pr_{\mathbb{P}}[\hat{Z} \mid Z] = \Pr_{\mathbb{Q}}[\hat{Z} \mid Z]$, but $\Pr_{\mathbb{P}}[Z] \neq \Pr_{\mathbb{Q}}[Z]$.

($\mathbb{P}$: the test-data distribution, $\mathbb{Q}$: the calibration-data distribution)

Under this shift, all estimators except ours (v) are biased.

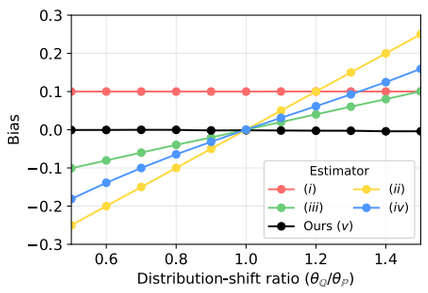
(i) $\mathbb{E}_{\mathbb{P}}[\hat{Z}]$

(ii) $\mathbb{E}_{\mathbb{Q}}[Z]$

(iii) $\mathbb{E}_{\mathbb{P}}[\hat{Z}] + \mathbb{E}_{\mathbb{Q}}[Z - \hat{Z}]$

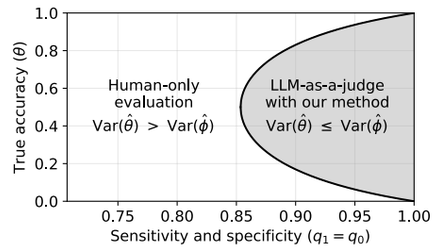
(iv) $\Pr_{\mathbb{Q}}(Z \mid \hat{Z})\, \mathbb{E}_{\mathbb{P}}[\hat{Z}]$

(v) $\Pr_{\mathbb{Q}}(\hat{Z} \mid Z)^{-1} \mathbb{E}_{\mathbb{P}}[\hat{Z}]$



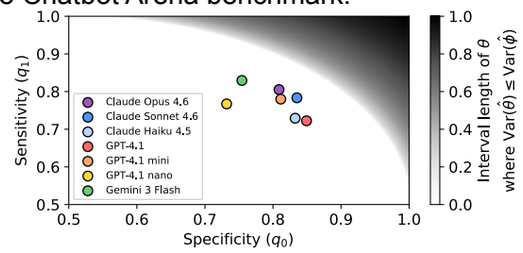
Using LLMs can be more reliable

When does corrected LLM-as-a-judge beat human-only evaluation? Under a fixed human-label budget, a sufficiently accurate judge yields lower variance than human-only evaluation:



Shaded region: the regime where corrected LLM-as-a-judge evaluation has lower variance than human-only evaluation.

Where do current judges stand? We estimate each judge's q_0 and q_1 on the Chatbot Arena benchmark:



No evaluated judge sits inside the favorable (dark) region yet, but stronger judges move toward it as their quality improves.

Extending the framework

Beyond binary responses. The same principle handles $k > 2$ categories, where Z and $\hat{Z}$ become vectors and a $k \times k$ confusion matrix generalizes the sensitivity q_1 and specificity q_0:

$$\hat{\theta}_{\text{multi}} := \Pr_{\mathbb{Q}}(\hat{Z} \mid Z)^{-1} \mathbb{E}_{\mathbb{P}}[\hat{Z}], \qquad Z, \hat{Z} \in [0,1]^k$$

Accounting for Nuisance Factors. When the LLM judge depends on covariates ξ (such as response length), we estimate a confusion matrix within each stratum of ξ, correct each stratum separately, and then average.

$$\hat{\theta} = \mathbb{E}_{\xi}\left[\Pr_{\mathbb{Q}}(\hat{Z} \mid Z; \xi)^{-1} \mathbb{E}_{\mathbb{P}}[\hat{Z}; \xi] \right]$$