
Peeking at a biased judge

David Zhan

University of Pennsylvania
zhandavid4@gmail.com

Abstract

1 A betting e-process is a valid test of the mean of the scores you feed it, uniformly
2 over stopping times. In LLM evaluation those scores increasingly come from
3 another model. An e-process run on the judge is then a valid test of $\mathbb{E}[f] \leq 0$, not
4 of the human estimand $\theta = \mathbb{E}[h]$. When the judge’s bias opposes the true gap,
5 Ville’s inequality still holds — for the wrong null — and $\Pr(\exists T : E_T \geq 1/\alpha) \rightarrow 1$
6 under the hypothesis anyone actually cares about. No sample size repairs this. We
7 instantiate the rectifier of Csillag et al. [2025], which restores a supermartingale
8 for θ , and measure how often the naive procedure is the one in use: on MT-Bench,
9 judge bias is sign-varying (−0.392 to +0.429 on 30 cells; pooled +0.001), so
10 a constant offset cannot help, and across ten judges the Elo gap a judge-only e-
11 process can overturn is 40–71 (95% intervals typically 30–80) no matter how long
12 you wait.

1 Introduction

14 E-processes exist because evaluation is sequential. Analysts stop when a budget runs out or when the
15 numbers look good; a fixed- n p -value at a data-dependent T is not a p -value. A non-negative process
16 with $\mathbb{E}[E_t | \mathcal{F}_{t-1}] \leq E_{t-1}$ under H_0 restores the guarantee: Ville gives $\Pr(\exists t : E_t \geq 1/\alpha) \leq \alpha$
17 [Waudby-Smith and Ramdas, 2024]. The guarantee is for the mean of the scores that enter the capital
18 update.

19 In current LLM evaluation those scores are themselves model outputs. Write $h(X) \in [-1, 1]$ for
20 the human comparison of models A and B on prompt X , $f(X) \in [-1, 1]$ for a judge’s score, and
21 $\theta = \mathbb{E}[h]$. The test anyone wants is $H_0 : \theta \leq 0$. The test an e-process on f actually runs is
22 $H'_0 : \mathbb{E}[f] \leq 0$. If $\mathbb{E}[f] > 0 \geq \theta$, E_t is a supermartingale under H'_0 and a submartingale under H_0 :
23 capital diverges almost surely, and more data buys more confidence in a conclusion the human labels
24 do not support.

25 This is not a new e-process. The fix is the prediction-powered rectifier of Csillag et al. [2025], which
26 we instantiate and do not extend. Sequential work that treats the proxy as biased already exists
27 [Csillag et al., 2025, Ao et al., 2026]; it assumes the analyst will label some points. The contribution
28 is what happens when they do not: we convert the danger boundary into Elo and measure it for ten
29 judges. Judge bias is not mean-zero noise and is not a constant you can subtract: on MT-Bench it
30 changes sign across pairs, and every judge we measured can overturn a 40–71 Elo gap at any sample
31 size.

2 Why more data does not help

33 Let $X_1, X_2, \dots$ be evaluation prompts, $h(X) \in [-1, 1]$ the human comparison score for models A
34 and B , and $f(X) \in [-1, 1]$ a judge’s score. Write $\theta = \mathbb{E}[h]$ and test $H_0 : \theta \leq 0$.

35 A betting e-process run on f is a valid test of $\mathbb{E}[f] \leq 0$, not of $\theta \leq 0$. If $\mathbb{E}[f] > 0 \geq \theta$ its capital
36 diverges almost surely, so $\Pr(\exists T : E_T \geq 1/\alpha) \rightarrow 1$: the procedure becomes arbitrarily confident in

Table 1: Per-cell judge bias and the Elo separation a judge-only e-process can overturn, over 30 (pair, turn) cells (median 47 items, IQR 40–51, min 32). Intervals are 95% item-bootstrap CIs ($B=2000$). All ten medians exceed a null floor of ≈ 0.08 . 3 of 9 modern judges exceed the 2023 baseline.

judge	axis	median $ b $ [95%]	$\Delta_{\min}$ [95%]
qwen3-14b	scale	0.12 [0.09, 0.18]	40 [33, 65]
qwen3-32b	scale	0.12 [0.09, 0.17]	43 [30, 59]
qwen3-4b	scale	0.13 [0.08, 0.17]	44 [29, 60]
gpt-oss-20b	family	0.13 [0.09, 0.17]	46 [31, 61]
qwen3-8b	scale	0.14 [0.10, 0.19]	48 [35, 66]
mistral-3.2-24b	family	0.14 [0.09, 0.18]	48 [32, 64]
gpt-4 (2023)	baseline	0.16 [0.11, 0.20]	57 [39, 72]
qwen3.6-27b	generation	0.18 [0.12, 0.22]	62 [44, 78]
qwen3-8b-think	reasoning	0.20 [0.14, 0.24]	69 [50, 84]
gemma-4-31b	family	0.20 [0.14, 0.23]	71 [48, 82]

a conclusion the human data does not support. In simulation with an injected bias of $+0.25$ under a true null, a peeking analyst has falsely rejected with probability > 0.99 within a few hundred observations, while a rectified process — $Z_t = g(f_t) - (\xi_t/\pi_t)(g(f_t) - h_t)$ with predictable π_t , unbiased for θ whatever the judge does [Csillag et al., 2025] — holds its level. Validity never depends on g ; efficiency does, through $\text{Var}(Z)$.

Bias is sign-varying, so offsets cannot fix it. On the 30 (pair, turn) cells of Table 1 — the same 1,391-item intersection, item-weighted — the GPT-4 judge’s bias ranges from -0.392 to $+0.429$ and changes sign in 16 of 30 cells [Zheng et al., 2023]. Two consequences follow. A constant offset calibrated on one pair is wrong on another. And pooled bias is $+0.001$ — near zero — because the per-cell biases cancel. An evaluation that reports only a single aggregate agreement or bias number is reporting the quantity in which this failure is invisible.

3 The threshold, in Elo

A judge-only process converges to $\text{sign}(\mathbb{E}[f])$, so it is unreliable exactly when bias opposing the true gap exceeds $|\theta|$. Under the Elo expected-score relation used by Chatbot Arena [Chiang et al., 2024], $\theta = 2/(1 + 10^{-\Delta/400}) - 1$, that boundary inverts to

$$\Delta_{\min}(b) = 400 \log_{10} \frac{1+b}{1-b}, \quad (1)$$

the Elo separation a judge with opposing bias b can overturn *regardless of how much data is collected*. Collecting more judgments moves E_T toward $+\infty$ or 0 according to $\text{sign}(\mathbb{E}[f])$; it does not move the boundary.

4 Ten judges

Ten judges — the 2023 GPT-4 baseline plus nine open-weights models — scored all 1,814 MT-Bench items in both presentation orders: 36,280 judgments. Decoding is greedy, weights are pinned to commit hashes, and prompts are FastChat’s own pair-v2 templates, including the reference-answer variants that MT-Bench routes math, coding and reasoning questions to, so each open judge sees exactly what the original GPT-4 judge saw. The roster crosses three axes: scale (Qwen3 4B–32B, recipe and generation fixed), generation (Qwen3-32B vs. Qwen3.6-27B, one release apart and also 32B→27B), and family (four vendors at 20–31B; gpt-oss-20b is MXFP4-quantized and a mixture-of-experts).

Bias is computed **per cell**, never pooled (§2), and on the **1,391 items every judge parsed**, so no judge is scored on an easier subset than another.

Open-weights judges have not beaten the 2023 baseline. Median per-cell bias is 0.137 across nine open-weights judges against 0.161 for GPT-4-2023, and 3 of 9 are worse. That is not identified

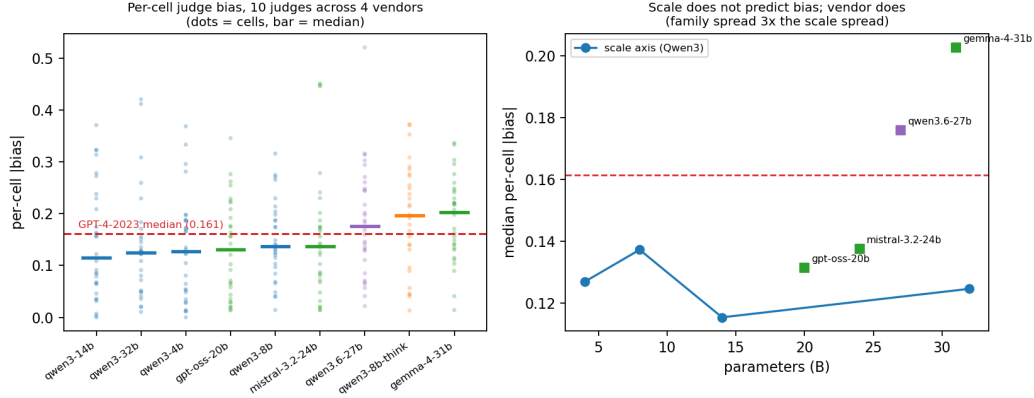


Figure 1: Left: per-cell $|bias|$ for ten judges (dots = cells, bar = median), with the 2023 baseline dashed. Right: bias against parameter count. The Qwen3 ladder is flat; the vendor spread at 20–31B is $3.2\times$ larger.

as a year effect — we did not run a modern proprietary judge — but the requirement has not collapsed: in Elo it moves from 57 to 48. A better supermartingale will not help if it is still computed from f .

Scale does not help; vendor spread does. Across the Qwen3 ladder bias is flat (0.127, 0.137, 0.115, 0.125; spread 0.022). Across vendors at 20–31B it spans 0.131–0.203 (spread 0.071 whether or not qwen3.6-27b is counted as the Qwen vendor; it is also the generation contrast). That is a $3.2\times$ ratio of *ranges*. The matching variance-ratio test is $F=13.98$, $p=0.029$ (Levene $p=0.022$); a one-way ANOVA on the same eight medians does not find a mean shift ($F=4.26$, $p=0.085$). With $n=4$ per group this is evidence of spread, not of a vendor-level mean effect, and the family group is not fixed-scale. It survives dropping the one judge with non-trivial truncation, and gemma-4-31b (highest bias) has 100% parse rate.

Both directional axes move the wrong way. One model generation raised bias ($0.125 \rightarrow 0.176$), as did enabling thinking on identical weights ($0.137 \rightarrow 0.197$). Each rests on a single pair and is directional only.

Where the threshold binds. MT-Bench’s own 30 cells are median 221 Elo apart (IQR 89–289), well above every requirement in Table 1, and accordingly almost no cell is affected. The threshold binds when models are close. Whether a particular leaderboard is dense enough to fall inside 40–71 Elo is a question a practitioner can answer from their own ratings. Kotawala [2026] already find several adjacent-rank *accuracy* pairs unresolvable under anytime-valid testing; we make no claim about preference leaderboards we have not measured.

5 What to do

An e-process on f answers a question you did not ask. Measure the judge. The within-cell residual SD of $f - h$ is $\sigma \approx 0.67$ on this data, so $SE(\hat{b}) = \sigma/\sqrt{n}$. At the modern median $|b|=0.137$ that maps through Eq. 1 with $\Delta'_{\min}(b) \approx 354$ to an Elo standard error of $354 \sigma/\sqrt{n}$. A 95% interval on $\Delta_{\min}$ is then ± 20 Elo at $n \approx 540$ and ± 15 Elo at $n \approx 960$. One hundred labels suffice only to detect that bias is nonzero ($n=91$ to exclude 0). If the two models sit inside that interval, a judge-only e-process cannot resolve them at any sample size. Run the rectifier of Csillag et al. [2025] on the labels you have, or collect more human labels.

6 Related work

Betting e-processes [Waudby-Smith and Ramdas, 2024] give tests and confidence sequences that remain valid under optional stopping. The capital is a supermartingale under the null you encoded

in the increment, and only that null. Prediction-powered e-values [Csillag et al., 2025] restore the human mean as the estimand by rectifying a cheap predictor with a predictable gold sample — the Z_t in §2 is theirs. Kilian et al. [2025] give a complementary anytime-valid, Bayes-assisted confidence sequence for the same estimand. Ao et al. [2026] prove that without bias correction, best-arm mis-selection need not vanish even with unlimited judge scores, and build a PPI estimator with confidence sequences under selective human audits. Li [2026] runs a second betting e-process on the judge–human gap to attribute production drift. Kotawala [2026] applies anytime-valid e-processes to adjacent-rank leaderboard pairs on *item correctness*, not on a biased judge. All of these assume the analyst wants θ and will label some points, or that the scores entering the e-process are θ .

Fixed- n prediction-powered inference [Angelopoulos et al., 2023] is the parent method, and for LLM judges it is now standard [Boyeau et al., 2025, Divekar and Majumder, 2026, Chen et al., 2026], including Rogan–Gladen correction [Lee et al., 2026] and Youden’s J as a judge-selection metric [Collet et al., 2026]. Fiedler [2026] shows that sharing calibration across compared models can reverse a comparison’s sign with high apparent confidence. None of these remains valid if the analyst stops when the numbers look good.

What we add is the translation and the measurement. An e-process run on f is a valid test of the wrong null; Eq. 1 converts the opposing bias into the Elo unit of Chiang et al. [2024]; and across ten judges on a common item set that separation is 40–71. Agreement rate does not bound this quantity [Dorner et al., 2025, Zheng et al., 2023].

7 Limitations

MT-Bench’s *judged* models are from 2023, so we measure modern judges on older comparisons; whether bias is larger or smaller on contemporary model pairs is untested and would require fresh human labels. Thirty cells (median 47 items) and 1,391 of 1,814 items is a small base; median $|\hat{b}|$ is biased upward under noise, but every judge exceeds a null floor of ≈ 0.08 , so the 40–71 Elo *level* is not an artifact of that. The common item set is easier than the full benchmark because two reasoning judges truncate roughly 11% of the longest prompts. The axis comparison is $n=4$ versus $n=4$, and qwen3.6–27b is counted in both the family spread and the generation contrast. Finally, for the most closely-contested pair in the dataset the sign of θ is not statistically identified — vote-weighted and item-weighted estimates disagree — so we describe the failure as certifying a winner the data cannot support, rather than as a demonstrated error.

Code and data. Code, FastChat pair-v2 prompt bytes, pinned weight SHAs, and the result JSON that produce every number above are at <https://github.com/dzhan111/evaluate-judge>. Open-weights judging used 3.3 GPU-hours ($\approx \$12$ of Modal credit). MT-Bench annotations are those of Zheng et al. [2023]; Qwen, Mistral, gpt-oss, and Gemma checkpoints were used under their published licenses, none gated.

References

- A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.
- R. Ao, H. Chen, S. Gao, H. Li, and D. Simchi-Levi. Best arm identification with LLM judges and limited human audits. *arXiv:2601.21471*, 2026.
- P. Boyeau, A. N. Angelopoulos, T. Li, N. Yosef, J. Malik, and M. I. Jordan. AutoEval done right: Using synthetic data for model evaluation. In *ICML*, 2025.
- Y. T. Chen, S. Lu, S. Li, M. Guo, and S. Li. Efficient inference for noisy LLM-as-a-judge evaluation. *arXiv:2601.05420*, 2026.
- W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. Jordan, J. E. Gonzalez, and I. Stoica. Chatbot Arena: An open platform for evaluating LLMs by human preference. In *ICML*, pages 8359–8388, 2024.

145 S. Collot, C. Fraser, J. Zhao, W. F. Shen, T. Willi, and I. Leontiadis. Balanced accuracy: The right
146 metric for evaluating LLM judges — explained through Youden’s J statistic. In *EACL (Industry)*,
147 pages 927–936, 2026.

148 D. Csillag, C. J. Struchiner, and G. T. Goedert. Prediction-powered e-values. *arXiv:2502.04294*,
149 2025.

150 A. Divekar and A. Majumder. PRECISE: Reducing the bias of LLM evaluations using prediction-
151 powered ranking estimation. *AAAI*, 40(47):39929–39938, 2026.

152 F. E. Dorner, V. Nastl, and M. Hardt. Limits to scalable evaluation at the frontier: LLM as judge
153 won’t beat twice the data. In *ICLR*, 2025.

154 J. Fiedler. Bias and uncertainty in LLM-as-a-judge estimation. *arXiv:2605.06939*, 2026.

155 V. Kilian, S. Cortinovis, and F. Caron. Anytime-valid, Bayes-assisted, prediction-powered inference.
156 In *NeurIPS*, 2025.

157 A. Kotawala. Resolution diagnostics for paired LLM evaluation. *arXiv:2605.30315*, 2026.

158 C. Lee, T. Zeng, J. Jeong, J. Sohn, and K. Lee. How to correctly report LLM-as-a-judge evaluations.
159 In *ICML*, 2026.

160 Y. Li. Who drifted: the system or the judge? Anytime-valid attribution in LLM evaluation pipelines.
161 *arXiv:2606.15474*, 2026.

162 I. Waudby-Smith and A. Ramdas. Estimating means of bounded random variables by betting. *Journal*
163 *of the Royal Statistical Society: Series B*, 86(1):1–27, 2024.

164 L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing,
165 H. Zhang, J. E. Gonzalez, and I. Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot
166 Arena. *NeurIPS Datasets and Benchmarks*, 2023.