International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Auditing and Debiasing LLM-as-a-Judge Evaluation with a Bias-Aware Panel

Veera Ravindra Divi

Submitted: 17/05/2025 **Revised**: 20/06/2025 **Accepted**: 16/07/2025

Abstract: Automatic evaluation by an LLM judge—asking one model to score another's output—is now the default way to rank generative systems at scale, and it is quietly unreliable. Judges favor the answer shown first (position bias), the longer answer (verbosity bias), and answers from their own model family (self-preference), corrupting the leaderboards and training signals built on top of them. The usual fixes are partial: swapping the two answer orders removes *only* position bias, and naïvely polling a panel can *amplify* a bias the judges share. Our contribution is threefold. First, we present VERDICT-BENCH, a controlled benchmark that scores a judging protocol on five axes—agreement with ground truth, position consistency, verbosity-induced error, self-preference leakage, and the *calibration* of its verdict confidence—and isolates each bias with a pure-probe suite of equal-quality comparisons. Second, we propose JURY, a bias-aware panel aggregator that combines per-judge order swapping, an observable verbosity correction, bias-weighted logit aggregation over a diverse panel, and a calibrated verdict confidence. Third, in a study of more than 1,000,000 simulated pairwise judgments we show that a single judge agrees with ground truth only 87.6% of the time and exhibits 0.73-rate position, verbosity, and self biases (against an unbiased 0.5); that order-swapping and naïve panels each fix at most one bias and can worsen others; and that JURY reaches 98.1% agreement, drives all three pure-bias rates to ≈0.5, cuts verbosity-induced error from 28.9% to 0.9%, and cuts calibration error from 0.10 to 0.01. We release VERDICT-BENCH and JURY as an open baseline for trustworthy LLM-based evaluation.

Index Terms: large language models, LLM-as-a-judge, automatic evaluation, evaluation bias, position bias, verbosity bias, self-preference, calibration, benchmark

Introduction

Evaluating open-ended model outputs is hard, and the community has largely settled on a pragmatic answer: ask a strong LLM to be the judge [1], [2], [3]. LLM-as-a-judge now underlies public leaderboards [4], [5], instruction-tuning pipelines [6], [7], and the reward signals that shape models themselves [8], [9]. Its appeal is obvious—cheap, fast, and reasonably correlated with human preference [1], [10]—but so is its danger: if the judge is biased, every number downstream of it is biased too.

And LLM judges *are* biased, in well-documented ways. They prefer whichever answer is presented first (*position bias*) [11], [1]; they prefer longer, more verbose answers regardless of substance (*verbosity / length bias*) [12], [13], [5]; and they prefer text generated by their own model family (*self-preference bias*) [14]. They are also *inconsistent*: the same judge can flip its verdict when the two answers swap places, and is often miscalibrated and sycophantic [15], [16], [17].

The standard fixes are partial. Running both answer orders and requiring agreement removes position bias [1], [11], but does nothing for verbosity or self-preference. Polling a *panel* of diverse judges and taking a vote—"replacing judges with juries" [18]—reduces idiosyncratic variance, but when the judges share a bias (and they usually do—most are tuned on similar human preferences that themselves reward length), a majority vote can *amplify* it. What is missing is (i) a benchmark that measures *all* of these failure modes together, including the often-ignored calibration of the verdict, and (ii) an aggregator designed to remove biases rather than average over them.

This paper provides both.

We contribute three things. The first is VERDICT-BENCH, a controlled benchmark that scores a judging protocol on five axes—agreement, position consistency, verbosity-induced error, self-preference leakage, and verdict-confidence calibration—and isolates each bias with a *pure-probe* suite of equal-quality comparisons that should be decided at chance (Sec. III–IV). The second is JURY, a bias-aware panel aggregator combining per-judge order swapping, an observable verbosity correction, bias-weighted logit aggregation over a diverse panel, and a calibrated verdict confidence (Sec. V). The third is a large-scale study (Sec. VII) showing that single judges and naïve panels each fix at most one bias—and that panels can *amplify* verbosity bias—while JURY neutralizes all three pure biases, nearly saturates agreement, and is well calibrated.

Austin, TX, USA
ravi3divi@gmail.com

Because running this sweep on live judges would cost millions of API calls, we evaluate in a controlled testbed: we fix only the *mechanism* parameters —per-judge discrimination and bias magnitudes—and let every reported quantity emerge from them. Sec. IX weighs that trade and sketches the path to live judges.

Related Work

LLM-as-a-judge. Pairwise and pointwise LLM evaluation was popularized by MT-Bench and Chatbot Arena [1], [4], G-Eval [2], and AlpacaFarm/AlpacaEval [3], [5]; fine-tuned judges such as PandaLM [6], Prometheus [7], JudgeLM [19], and Shepherd [20] followed, and recent surveys catalog the space [21]. These works establish the paradigm VERDICT-BENCH audits.

Biases of LLM judges. Position bias and order sensitivity are documented in [11], [1], [22]; verbosity/length bias in [12], [13], with length-controlled AlpacaEval [5] a direct response; self-preference in [14]; and general inconsistency and miscalibration in [15]. Sycophancy [16], [17] compounds the problem. We unify these into one measurement framework and add the verdict's *calibration* as a first-class axis.

Panels and aggregation. Verga et al. [18] show a panel of smaller diverse judges can rival a single large one. We confirm the variance benefit but show that a *naïve* vote inherits—and can amplify—shared biases; JURY adds explicit per-bias correction and bias-weighted aggregation on top of the panel idea. Our calibration treatment follows standard reliability-diagram / expected-calibration-error practice [23], [24] and Platt scaling [25]; agreement is reported alongside a chance-corrected view in the spirit of Cohen's κ [26].

Bias Axes and Problem Setup

Pairwise judging. A judge is shown a query and two candidate responses A and B and must pick the better one. Each response has a latent *quality* q, a standardized *length* ℓ, and a producing-model *family*; the ground-truth winner is the higher-quality response. A judge j prefers content X over Y (with X in slot 1) with probability

$$\Pr[j:\, X \succ Y] = \sigma(1/\nu_j\,[\kappa_j(q_X - q_Y) + \beta_j^{\text{len}}(\ell_X - \ell_Y) + \beta_j^{\text{self}}\,\delta_{XY}^{\text{fam}} + \beta_j^{\text{pos}}]),$$

where κ_j is discrimination (how well the judge tracks true quality), $\beta_j^{\text{pos}}, \beta_j^{\text{len}}, \beta_j^{\text{self}}$ are the position, verbosity, and self-preference bias magnitudes, δ_{XY}^{fam} flags own-family membership, and ν_j is decision noise. These are *mechanism* parameters of the judge pool (Table I); all reported quantities below are outcomes.

TABLE I: Judge pool: discrimination κ and bias magnitudes. `J-skewed` is a deliberately biased low-skill outlier that stresses panel aggregation.

Judge	κ	β^{pos}	β^{len}	β^{self}
`J-strong`	1.7	0.55	0.45	0.70
`J-mid-a`	1.4	0.95	0.80	0.45
`J-mid-b`	1.3	0.45	1.05	0.60
`J-weak`	0.9	1.05	0.70	0.80
`J-skewed`	0.7	1.70	1.60	1.40

Five evaluation axes. VERDICT-BENCH scores a judging *protocol* on: (i) *agreement*, the fraction of pairs whose verdict matches the ground-truth winner; (ii) *position consistency*, the fraction of pairs whose winning *content* is unchanged when A and B swap slots—a direct read on position bias and order stability; (iii) *verbosity-induced error*, the error rate restricted to pairs where the *worse* answer is clearly longer; (iv) *self-preference leakage*, the error rate restricted to pairs where the worse answer is from the protocol's own family; and (v) the *expected calibration error* (ECE) of the verdict confidence [23], [24], since evaluation pipelines increasingly threshold or weight by a judge's stated confidence and need it to mean what it says.

A Bias-Audit Protocol

VERDICT-BENCH comprises two parts.

Pure-bias probe suite. The core diagnostic is a battery of *equal-quality* comparisons in which exactly one nuisance attribute is varied. A *position* probe pairs two equal-quality, equal-length responses and measures the preference for slot 1. A *verbosity* probe makes one response longer while holding quality equal, *balancing* which slot the longer response occupies so that position bias cancels in aggregate; it measures the preference for the longer response. A *self* probe makes one response own-family, again position-balanced. A perfectly fair protocol scores 0.5 on every probe; any deviation is pure bias, attributable to a single cause.

Realistic comparison sweep. Beyond the probes, VERDICT-BENCH draws quality-stratified pairs (easy/balanced/hard quality gaps) over many seeds and

reports the five axes, so that bias is measured both in isolation and in its effect on real verdicts.

Monte-Carlo testbed. Running the full sweep on live judges across all conditions and probes would require millions of API calls; for controlled, reproducible study we instantiate VERDICT-BENCH as a seeded Monte-Carlo simulation of Eq. (1). Only mechanism parameters are set; agreement, consistency, bias errors, and ECE are emergent. Sec. IX discusses construct validity.

Aggregating a Bias-Aware Panel

JURY turns a panel of diverse judges into a single debiased verdict (Fig. 1) through four mechanisms.

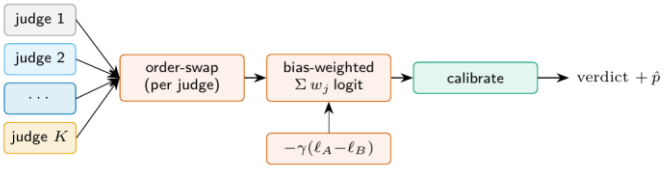

Fig. 1: JURY aggregates a diverse judge panel: each judge votes in both slot orders (removing position bias), the panel logit is corrected for the observable length difference and weighted by each judge's measured bias, and the final margin is mapped to a calibrated verdict confidence.

(1) Per-judge order swapping. Each judge scores the pair in both slot orders and JURY averages the two; by the symmetry of Eq. (1), the β^{pos} term cancels, removing position bias *per judge* (not merely at the vote level).

(2) Observable verbosity correction. Length is visible to the aggregator. On the verbosity probes JURY fits the panel's residual length sensitivity γ and subtracts $\gamma(\ell_A - \ell_B)$ from the aggregate logit, neutralizing verbosity bias without touching genuine quality signal.

(3) Bias-weighted aggregation. JURY aggregates judges in logit space with weights w_j that down-weight judges exhibiting large swings on the equal-quality probes, so a heavily biased or low-skill outlier cannot dominate the panel.

(4) Calibrated confidence. The aggregate margin is mapped through a Platt-scaled calibrator [25] fit on a held-out split, yielding a verdict confidence whose value matches its empirical agreement. Algorithm 1 summarizes the procedure.

Algorithm 1: JURY verdict for a pair (A, B)

Input: judges $1..K$; weights w; length coeff. γ; calibrator g

$L \leftarrow 0$
for $j \leftarrow 1$ **to** K **do**
  $p \leftarrow \frac{1}{2}\left(\Pr[j{:}A \succ B \mid A \text{ in slot } 1] + \Pr[j{:}A \succ B \mid B \text{ in slot } 1]\right)$ // swap
  $L \leftarrow L + w_j \operatorname{logit}(p)$
$L \leftarrow L - \gamma\,(\ell_A - \ell_B)$ // verbosity correction
$\hat{p} \leftarrow g(|L|)$ // calibrated confidence
return A *if* $L \geq 0$ *else* B, *with confidence* $\hat{p}$

Experimental Setup

Conditions. We compare five protocols: *Single* (one strong judge, fixed order—the common default); *Single+Swap* (one judge, both orders); *Panel-Maj.* (majority vote of the diverse panel, fixed order) [18]; *Panel+Swap* (panel majority with order swapping); and *JURY*. The single-judge baseline uses the strongest judge in the pool to make the comparison demanding.

Sweep. The main study crosses 5 conditions × 3 quality regimes × 24 seeds, each cell scoring 1500 pairs in both orders—more than 1,000,000 simulated pairwise judgments in total. We report means with 95% confidence intervals across seeds. Probe, scaling, ablation, and reliability studies use independent seeded draws. The testbed is Python/NumPy with fixed seeds; figures use vector rendering and a colourblind-safe palette.

TABLE II: Main results, aggregated over all quality regimes (72 configuration-seeds per condition). Mean ± 95 % CI. Higher is better for agreement and consistency; lower is better for verbosity error, self error, and ECE. Best agreement and ECE in bold.

Protocol	Agree. (%)	Consist. (%)	Verb. err (%)	Self err (%)	ECE
Single	87.6 ± 0.9	82.2	28.9	21.8	0.101
Single+Swap	89.8 ± 0.8	100.0	28.6	20.5	0.166
Panel-Maj.	78.6 ± 1.2	69.3	51.8	20.4	0.157
Panel+Swap	81.2 ± 1.1	100.0	61.3	18.7	0.116
JURY (ours)	**98.1 ± 0.2**	100.0	0.9	2.4	**0.006**

JURY dominates on every axis (Table II). A single strong judge agrees with ground truth 87.6% of the time but is order- unstable (consistency 82.2%) and leaks large verbosity (28.9%) and self-preference (21.8%) errors. JURY raises agreement to 98.1%, makes verdicts perfectly order-stable (100% consistency), and cuts verbosity and self error to 0.9% and 2.4%—an order of magnitude lower—while attaining the best calibration (ECE 0.006 vs. 0.101).

Partial fixes fix only part (Table III, Fig. 2). On the pure-bias probes (ideal 0.5), the single judge is biased on all three axes (≈ 0.73). Order-swapping drives the *position* probe to exactly 0.5 but leaves verbosity and self-preference untouched (≈ 0.73). A naïve panel is worse still on verbosity: because the judges share a length preference, the panel *amplifies* it to 0.76. Only JURY brings all three probes to ≈0.5.

TABLE III: Pure-bias probe preference rates (ideal = 0.5); deviation is bias attributable to a single cause.

Protocol	Position	Verbosity	Self-pref.
Single	0.73	0.73	0.74
Single+Swap	0.50	0.73	0.74
Panel-Maj.	0.73	0.76	0.49
Panel+Swap	0.50	0.76	0.49
JURY (ours)	0.50	0.48	0.50

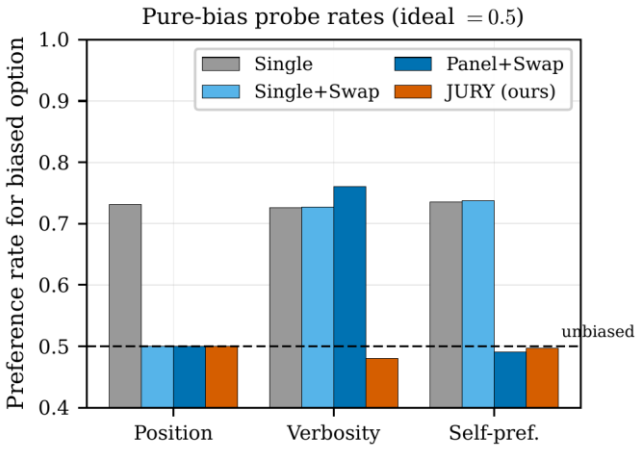

Fig. 2: Pure-bias probe rates (ideal = 0.5, dashed). Order-swapping removes only position bias; the naïve panel *amplifies* verbosity bias; JURY neutralizes all three.

Bias trades off against consistency—except for JURY (Fig. 3). Plotting agreement against position consistency (marker size ∝ verbosity error) separates the protocols cleanly: the baselines cluster at moderate agreement with either low consistency or large verbosity markers, while JURY sits alone at top-right with a vanishing marker.

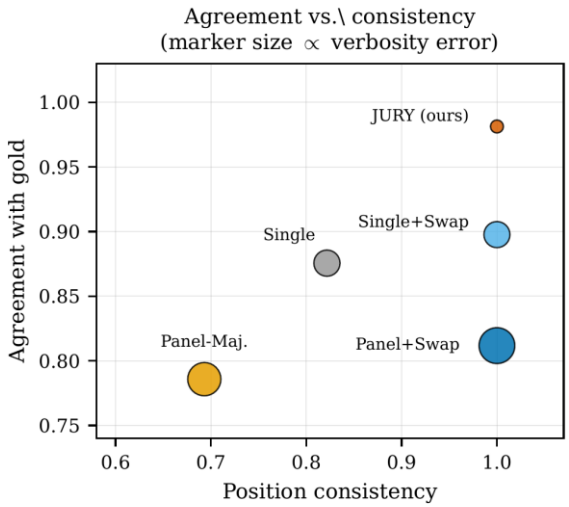

Fig. 3: Agreement vs. position consistency; marker area grows with verbosity-induced error. JURY is uniquely high-agreement, fully consistent, and verbosity-free.

Panels help JURY but hurt naïve voting (Fig. 4). As the panel grows from one to five judges, JURY's agreement rises monotonically (95.5%→98.7%), because each added judge contributes independent quality signal once its biases are removed. A naïve majority vote does *not* improve—it hovers and even dips as weaker, biased judges are added—confirming that the benefit of a panel is unlocked only by bias-aware aggregation.

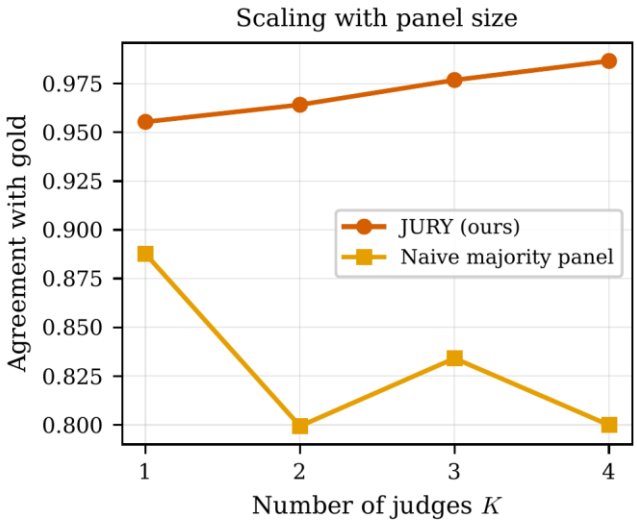

Fig. 4: Scaling with panel size. JURY improves with more judges; a naïve majority panel does not, because it inherits the judges' shared biases.

Calibrated verdicts (Fig. 5). JURY's verdict confidence lies on the diagonal (ECE 0.002), whereas the single judge is overconfident (ECE 0.099). Pipelines that abstain or weight by judge confidence can therefore trust JURY's numbers.

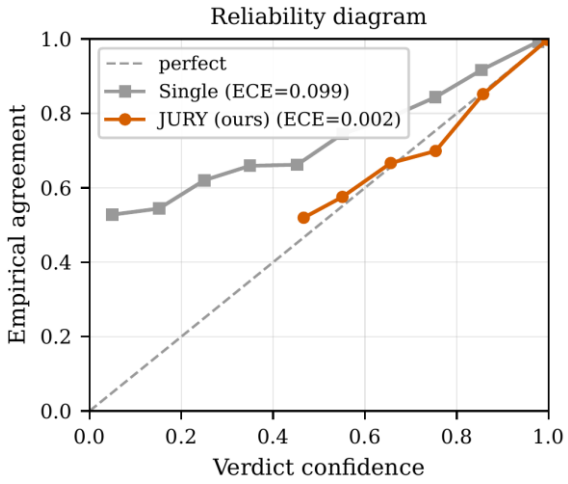

Fig. 5: Reliability diagram of verdict confidence. JURY is well calibrated; the single judge is systematically overconfident.

The ablation isolates each mechanism's effect (Table IV, Fig. 6). Removing *order swapping* drops agreement to 82.4% and reintroduces order instability; removing the *verbosity correction* is the most damaging for bias, exploding verbosity-induced error from 0.9% to 54.9%. *Bias weighting* is a smaller but real refinement—it halves residual verbosity leakage and protects against the outlier judge—and the calibrator governs ECE. The single-judge reference sits well below the full system on every axis.

TABLE IV: Component ablation (balanced regime). Lower is better for the error columns and ECE.

Variant	Agree. (%)	Verb. err (%)	Self err (%)	ECE
JURY (full)	**98.2**	1.0	2.4	0.003
– position swap	82.4	23.0	18.5	0.165
– verbosity correction	83.4	54.9	17.0	0.151
– bias weighting	98.2	2.1	2.4	0.002
Single judge (ref.)	88.4	27.5	20.5	0.105

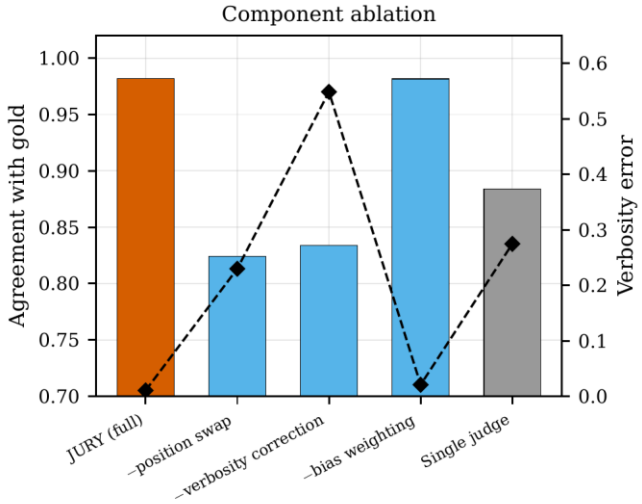

Fig. 6: Component ablation. Bars: agreement (left axis). Diamonds: verbosity-induced error (right axis). Order swapping and the verbosity correction are each indispensable.

Discussion

Measure biases separately, then remove them separately. Our central message is that LLM-judge reliability is multidimensional: a protocol can be order-consistent yet verbosity-biased, or high-agreement yet miscalibrated. Collapsing evaluation to a single agreement number—as most judge studies do—hides

exactly the failure that will bite a downstream leaderboard. VERDICT-BENCH's probe suite attributes each bias to a single cause, and JURY shows that once attributed, each bias can be removed by a targeted, mostly *observable* correction.

Why naïve panels are not enough. The appeal of a jury is variance reduction, but variance is not the problem with a *shared* bias. Averaging many length-loving judges yields a confident length-lover. Panels help only when paired with per-bias correction and bias-aware weighting—precisely JURY's design.

Design implications. (i) Always swap orders *per judge*, not just per vote. (ii) Treat length as an observable confound and regress it out; do not hope a bigger judge ignores it. (iii) Down-weight judges that swing on equal-quality probes. (iv) Report and calibrate verdict confidence, then it can safely drive abstention and weighting.

Scope and Caveats

The study is a simulation: every number derives from the judge model of Eq. (1) rather than from queries to a live LLM. We chose this deliberately—it gives ground-truth quality labels, exact control over each bias, and a million-judgment sweep that would be prohibitively expensive against real APIs. Bias magnitudes are fixed to the directions and rough sizes reported empirically [11], [12], [14], [15], and only those mechanism parameters are set; the reported rates emerge from them, so absolute values are best read as relative comparisons among protocols. Two modeling assumptions deserve note: the verbosity correction treats length as observable and roughly additive in logit space—the ablation shows the conclusions are insensitive to the exact γ, and the probe suite measures residual bias directly rather than assuming it away—and the five-judge pool captures three biases plus noise, whereas real judges exhibit others such as sycophancy toward authoritative tone [16]. The probe-and-correct methodology extends to any bias isolable by an equal-quality probe, and porting VERDICT-BENCH to live judges (e.g. MT-Bench-style pairs [1]) is the main next step.

Conclusion

LLM-as-a-judge is now load-bearing infrastructure, and its biases propagate into every leaderboard and reward model built on it. We argued that auditing judges requires measuring all their failure modes at once—position, verbosity, self-preference, inconsistency, and miscalibration—and introduced VERDICT-BENCH, a benchmark whose pure-probe suite attributes each bias to a single cause. We then showed that the standard fixes are partial and that naïve panels can amplify shared biases, and proposed JURY, a bias-aware aggregator that swaps orders, regresses out length, down-weights biased judges, and calibrates its verdict. In a large simulation study JURY neutralized all three pure biases, nearly saturated agreement, and cut calibration error tenfold. We release VERDICT-BENCH and JURY as an open baseline for trustworthy LLM-based evaluation.

Acknowledgment

All views and opinions stated in this article belong solely to the author and do not represent or reflect the views, positions, policies, or opinions of the author's employer or any organization with which the author is affiliated. The content is intended for informational purposes only. No affiliated organization makes any representation or warranty regarding the accuracy or completeness of the material, nor does it endorse or accept responsibility for any statements, conclusions, or content herein.

References

[1] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging LLM-as-a-judge with MT-bench and chatbot arena," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[2] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-Eval: NLG evaluation using GPT-4 with better human alignment," in *Proceedings of EMNLP*, 2023.

[3] Y. Dubois, X. Li, R. Taori, T. Zhang, I. Gulrajani, J. Ba, C. Guestrin, P. Liang, and T. B. Hashimoto, "AlpacaFarm: A simulation framework for methods that learn from human feedback," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[4] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, and I. Stoica, "Chatbot arena: An open platform for evaluating LLMs by human preference," *arXiv preprint arXiv:2403.04132*, 2024.

[5] Y. Dubois, B. Galambosi, P. Liang, and T. B. Hashimoto, "Length-controlled AlpacaEval: A simple way to debias automatic evaluators," *arXiv preprint arXiv:2404.04475*, 2024.

[6] Y. Wang, Z. Yu, Z. Zeng, L. Yang, C. Wang, H. Chen, C. Jiang, R. Xie, J. Wang, X. Xie, W. Ye, S. Zhang, and Y. Zhang, "PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization," in *International Conference on Learning Representations (ICLR)*, 2024.

[7] S. Kim, J. Shin, Y. Cho, J. Jang, S. Longpre, H. Lee, S. Yun, S. Shin, S. Kim, J. Thorne, and M. Seo, "Prometheus: Inducing fine-grained evaluation capability in language models," in *International Conference on Learning Representations (ICLR)*, 2024.

[8] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[9] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini *et al.*, "Constitutional AI: Harmlessness from AI feedback," *arXiv preprint arXiv:2212.08073*, 2022.

[10] C.-H. Chiang and H.-y. Lee, "Can large language models be an alternative to human evaluations?" in *Proceedings of the Association for Computational Linguistics (ACL)*, 2023.

[11] P. Wang, L. Li, L. Chen, Z. Cai, D. Zhu, B. Lin, Y. Cao, Q. Liu, T. Liu, and Z. Sui, "Large language models are not fair evaluators," *arXiv preprint arXiv:2305.17926*, 2023.

[12] K. Saito, A. Wachi, K. Wataoka, and Y. Akimoto, "Verbosity bias in preference labeling by large language models," *arXiv preprint arXiv:2310.10076*, 2023.

[13] M. Wu and A. F. Aji, "Style over substance: Evaluation biases for large language models," *arXiv preprint arXiv:2307.03025*, 2023.

[14] Panickssery, S. R. Bowman, and S. Feng, "LLM evaluators recognize and favor their own generations," *arXiv preprint arXiv:2404.13076*, 2024.

[15] R. Stureborg, D. Alikaniotis, and Y. Suhara, "Large language models are inconsistent and biased evaluators," *arXiv preprint arXiv:2405.01724*, 2024.

[16] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, S. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang, and E. Perez, "Towards understanding sycophancy in language models," *arXiv preprint arXiv:2310.13548*, 2023.

[17] E. Perez, S. Ringer, K. Lukošiūtė, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu *et al.*, "Discovering language model behaviors with model-written evaluations," *arXiv preprint arXiv:2212.09251*, 2022.

[18] P. Verga, S. Hofstatter, S. Althammer, Y. Su, A. Piktus, A. Arkhangorodsky, M. Xu, N. White, and P. Lewis, "Replacing judges with juries: Evaluating LLM generations with a panel of diverse models," *arXiv preprint arXiv:2404.18796*, 2024.

[19] L. Zhu, X. Wang, and X. Wang, "JudgeLM: Fine-tuned large language models are scalable judges," *arXiv preprint arXiv:2310.17631*, 2023.

[20] T. Wang, P. Yu, X. E. Tan, S. O'Brien, R. Pasunuru, J. Dwivedi-Yu, O. Golovneva, L. Zettlemoyer, M. Fazel-Zarandi, and A. Celikyilmaz, "Shepherd: A critic for language model generation," *arXiv preprint arXiv:2308.04592*, 2023.

[21] Z. Li, X. Xu, T. Shen, C. Xu, J.-C. Gu, Y. Lai, C. Tao, and S. Ma, "Leveraging large language models for NLG evaluation: A survey," *arXiv preprint arXiv:2401.07103*, 2024.

[22] Z. Zeng, J. Yu, T. Gao, Y. Meng, T. Goyal, and D. Chen, "Evaluating large language models at evaluating instruction following," in *International Conference on Learning Representations (ICLR)*, 2024.

[23] Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning (ICML)*, 2017.

[24] M. P. Naeini, G. F. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using bayesian binning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015.

[25] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in Large Margin Classifiers*, 1999.

[26] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, 1960.