Accounting for Bias Enables Sustainable LLM Evaluation

Harshita Katoch[1,2], David Antony Selby[1,2], Gerrit Großmann[2] and Sebastian Vollmer[1,2]

[1]Department of Computer Science, University of Kaiserslautern–Landau (RPTU), Germany
[2]Data Science and its Applications, German Research Centre for Artificial Intelligence (DFKI), Germany

Abstract

LLM-as-a-judge has become the *de facto* standard for scalable, subjective evaluation, yet current leaderboards compensate for systematic measurement bias by running ever more comparisons, an approach that is both statistically unsound and computationally wasteful. The root cause is an incomplete measurement model, treating LLM judges as neutral, interchangeable instruments ignores documented biases like position bias, verbosity bias, judge severity, and self-enhancement, that no volume of additional data can eliminate. We propose a unified latent variable framework that jointly models pairwise and ordinal data while explicitly correcting for these confounders, recovering reliable rankings from substantially fewer comparisons. Because fitting this model costs negligible compute relative to a single round of LLM inference, bias correction is not only more statistically rigorous but also a more sustainable approach to trustworthy evaluation.

Keywords

LLM-as-a-judge, LLM evaluation, psychometrics, bias correction, sustainable benchmarking

1. Introduction

The evaluation of large language models has become a question of trustworthiness as much as capability [1]. The field's prevailing response to evaluation challenges has been to scale: running more comparisons, enlisting more judges and expanding leaderboards. The statistical machinery underpinning these systems has genuinely improved, with the field moving from raw win rates toward principled latent-variable models for pairwise and ordinal scoring that better capture preference uncertainty. Yet scaling and methodological refinement have both proceeded under a shared assumption, that has gone largely unexamined: that LLM judges are neutral, interchangeable instruments whose idiosyncrasies cancel with sufficient data [2, 3]. When judge behaviour is systematically skewed, collecting more comparisons narrows the confidence interval around a biased estimate but not correct it. Scale adds compute without adding reliability.

The biases in question are well-documented and structural. LLM judges exhibit position preferences i.e. favouring responses that appear in a particular slot regardless of quality [4], verbosity effects that reward length over content [5], and self-enhancement toward outputs from their own provider family [6]. Beyond these, different judges apply systematically different score levels which is a severity effect that makes identical model ability map to different observed scores. Surveys by Gu et al. [2] and Li et al. [3] establish these as properties of the measurement instrument, not incidental quirks of particular models. Since the effects differ in direction across judges, procedural fixes such as order randomisation or majority voting cannot eliminate them in expectation [3], nor do mitigations that target a single bias in isolation [5, 7]. A jointly specified measurement model is required.

The consequences are both scientific and environmental. Every judge-mediated comparison requires inference from two competing models and a judge, i.e. three inference calls per comparison, and at leaderboard scale this cost is substantial. Each inference call consumes computational resources, GPU time, memory, storage and electricity (training costs notwithstanding), so reducing the number of comparisons translates directly into lower energy use and a smaller environmental footprint [8]. Standard benchmarks are approximately 97% redundant [9], and smarter item selection can reduce evaluation cost by up to 99% without accuracy loss [10]. Generating redundant comparisons to compensate for an

SuRE'26: Workshop on Sustainability and Resource-Efficiency of Artificial Intelligence, August 17, 2026, Bremen, Germany
harshita.katoch@dfki.de (H. Katoch); david.selby@dfki.de (D. A. Selby); gerrit.grossmann@dfki.de (G. Großmann); sebastian.vollmer@dfki.de (S. Vollmer)

unmodelled confounder is not just statistically unsound, it is ecologically indefensible [11, 12]. Fitting a corrective latent-variable model costs negligible compute relative to a single round of LLM inference, making bias correction the more sustainable path to evaluation we can actually trust.

We propose a unified latent-variable framework that (i) jointly models pairwise and ordinal data in a single likelihood, (ii) explicitly corrects for position bias, verbosity, self-enhancement, and judge-level severity, and (iii) is validated on MT-Bench [13] as the first empirical decomposition of bias magnitudes on a real leaderboard. Simulation confirms that the bias-corrected model recovers reliable rankings from substantially fewer comparisons than naive baselines across six tournament designs.

2. Related Work

The machine learning community has developed LLM evaluation largely in isolation from the statistical sciences, arriving at methods that parallel tools already mature in psychometrics and educational measurement [14]. Early evaluation relied on static benchmarks with ground-truth answers; as tasks became more open-ended, the field shifted toward judge-mediated evaluation, MT-Bench introduced GPT-4-as-a-judge on multi-turn dialogue [13]; Chatbot Arena crowdsourced human pairwise preferences into Elo scores [15]; AlpacaEval 2.0 and Arena-Hard extended the paradigm with length control and harder prompts [5, 16]. Each iteration improved on its predecessor, yet all share a common assumption that observed ratings are direct proxies for model quality rather than noisy realisations of a latent construct mediated by rater behaviour and item difficulty. Beyond the response-level biases documented in the literature [2, 3], the measurement scale itself is routinely misused, Liddell and Kruschke [17] shows that the widespread practice of treating ordinal judge scores as continuous interval data introduces systematic distortions into effect estimates, an error that accumulating more data cannot fix. A jointly specified model is required.

Item response theory disentangles examinee ability from item difficulty, providing estimates invariant to the specific sample of items or judges used [18]. Many-facet Rasch models extend this to explicitly model rater severity [19], and extensions of paired-comparison models provide a principled framework that subsumes Elo as a special case [20]. Recent work has begun applying these ideas to AI benchmarks [21, 22, 23], but exclusively on objective tasks with binary correctness signals, where there is no rater in the loop. Our contribution extends this to the judge-mediated setting, where rater confounders require explicit parametrisation alongside standard IRT machinery, a gap the existing literature has not addressed.

3. A Unified Psychometric Framework

In psychometric terms, the LLM under evaluation is the *examinee* with latent ability θ, the prompt is the *item* with difficulty δ, and the judge is the *rater* with severity α. The goal is *parameter invariance*: ability estimates independent of which specific judges and prompts happened to be used [18]. If a model's ranking drops because we switched from a lenient judge to a strict one, the measurement system has failed. This reframes what appears to be a data-volume problem as what it actually is a modelling problem.

3.1. Ranking from Pairwise Comparisons

A minimally viable evaluation system based on comparative judgement should use a variant of the Bradley–Terry model [24] rather than average win rates, with coefficients to account for order effects, self-enhancement, and measurable item-level features such as relative verbosity. This extends AlpacaEval 2.0 [5], which in the presence of these biases only controls for item-level effects. Such an augmented

Bradley–Terry–Luce model takes the form:

$$P(A \succ B) = \sigma\left(\theta_A - \theta_B + \gamma + \underbrace{\lambda_{A,r} - \lambda_{B,r}}_{\text{Affinity bias}}\right) \tag{1}$$

where θ_A and θ_B are the models' latent abilities; γ is a position bias; $\lambda_{i,r}$ is judge r's affinity for model i, capturing self-enhancement bias; and σ is the logistic function. This is straightforward to fit using standard logistic regression software. Further terms may be added for verbosity and other documented biases; in our empirical analysis we additionally include a judge-specific position deviation u_j and a standardised response length covariate [5].

3.2. Ranking from Ordinal Scores

Some benchmarks prefer direct scoring of individual items. A comparative judgement may say $A \succ B$ but not be able to say 'both are bad'; an absolute scale such as 1–10 allows us to anchor items at one or the other extreme rather than merely order them relative to one another. Instead of naively averaging scores, a minimal ordinal model must account for scale-, item-, and judge-level effects. The many-facet Rasch model [25] treats scores as ordered categorical outcomes. The probability that model n receives score $k \in \{0, \ldots, m\}$ from rater r on item p is:

$$P(X_{npr} = k) \propto \exp\left(\sum_{m=1}^{k} (\theta_n - \delta_p - \alpha_r - \tau_m)\right) \tag{2}$$

where θ_n is latent ability, δ_p is item difficulty, α_r is rater severity, and τ_m are threshold parameters on the cumulative logit scale, not assumed equally spaced. Accounting for judge severity prevents a model evaluated by a lenient judge from artificially outranking one evaluated by a strict judge. Treating ordinal scores as interval data, as leaderboards routinely do, produces systematic inversions in estimated effects [17], a modelling error that accumulating more data cannot fix.

3.3. The Hybrid Joint Likelihood

Both pairwise preferences and ordinal scores are imperfect observations of the same latent ability. Rather than discarding either, we combine them in a joint likelihood that separates true ability variance from rater noise:

$$\mathcal{L}(\theta) = \left[\prod_{y \in \mathcal{D}_{\text{pair}}} P_{\text{Rel}}(y \mid \theta, \gamma)\right]^{\omega} \times \left[\prod_{x \in \mathcal{D}_{\text{ord}}} P_{\text{Ord}}(x \mid \theta, \lambda, \delta)\right]^{1-\omega} \tag{3}$$

where $\omega \in [0, 1]$ weights the two data sources ($\omega = 0.5$ in our experiments); λ represents judge-specific parameters (severity and affinity); and δ item difficulties. Ordinal ratings anchor sparse pairwise data, and models appearing in only one data source can be ranked alongside those in both. Estimation proceeds via L-BFGS-B with ridge regularisation; for adaptive designs, a parametric bootstrap correction is preferred over Firth's penalised likelihood, which introduces bias under non-random scheduling [20].

4. Empirical Validation: MT-Bench

We validate the proposed framework on MT-Bench [13], fitting the hybrid model jointly to GPT-4-judged pairwise comparison data and single-score ordinal data. Our aim is to confirm that the documented judge confounders are statistically significant in real leaderboard data, that they produce systematic, model-specific distortions rather than random noise, and that correcting for them recovers meaningful signal that no amount of additional data collected under the same protocol could provide.

Table 1
Estimated bias parameters from the hybrid model on MT-Bench. The parameter estimates are statistically significant or substantially non-zero, confirming that LLM-judge evaluation is not bias-free at any sample size.

Parameter	Estimate	95 % CI	
$\hat{\gamma}_0$ (position bias)	-0.165	-0.253	$+0.076$
$\hat{\beta}$ (verbosity)	$+0.163$	$+0.058$	$+0.268$
$\hat{\lambda}$ (same-family affinity)	$+0.411$	—	—
$\mathrm{SD}(u_j)$ across judges	0.337	—	—

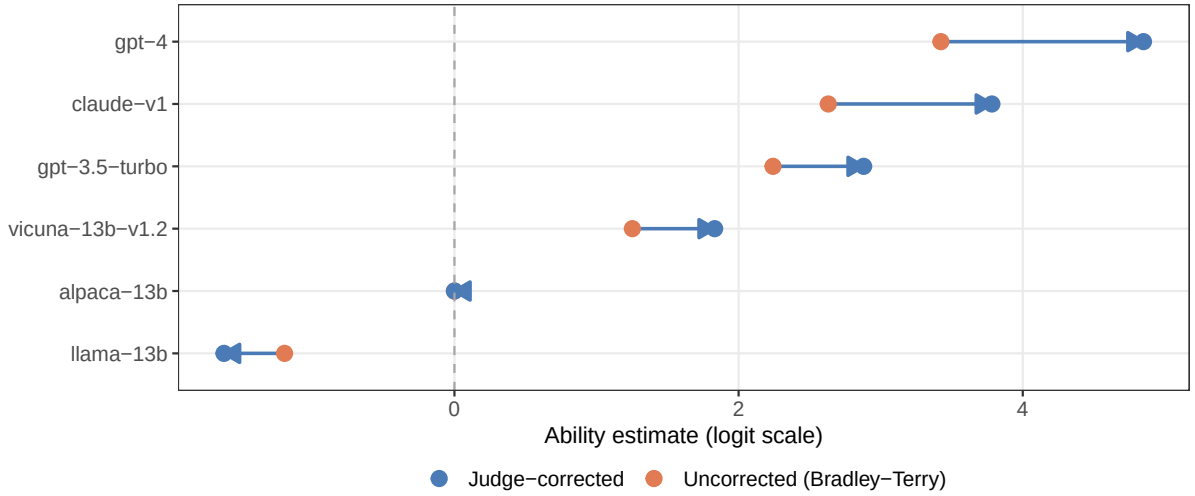

Figure 1: Correcting for judge effects shifts ability estimates substantially. Each arrow shows the change from naive Bradley-Terry (orange) to the Judge-corrected model (blue). Provider-affiliated models receive the largest upward corrections; non-affiliated models shift downward. These distortions are structural and collecting more comparisons under the same protocol would not resolve them.

Judge confounders are structural, not stochastic. Table 1 reports the estimated parameters from the full model. Position bias ($\hat{\gamma}_0 = -0.165$) confirms that second-position responses are systematically favoured, consistent with the judge-dependent direction effects documented by Shi et al. [4]. Verbosity bias ($\hat{\beta} = 0.163$, 95 % CI $[0.058, 0.268]$) confirms that longer responses are preferred irrespective of content quality, extending the finding of Dubois et al. [5] to a different evaluation setting. Same-family affinity ($\hat{\lambda} = 0.411$) confirms that the GPT-4 judge systematically favours models from its own provider family, a direct violation of the measurement invariance property that any reliable leaderboard requires. Substantial judge-level variation in position sensitivity ($\mathrm{SD}(u_j) = 0.337$) compounds this further, in the MT-Bench data, `claude-v1` appeared in second position far more often than first while `llama-13b` faced the opposite exposure, so the global position parameter accrues as a tournament-design artefact rather than a model quality signal. None of these effects can be eliminated by collecting more data under the same evaluation protocol, because they are properties of the measurement instrument rather than of sampling variation.

Bias correction recovers signal that more data cannot. Figure 1 shows the net outcome of jointly correcting for all three bias components. Ability estimates shift substantially and in model-specific directions that trace directly back to tournament design rather than model quality. Critically, this signal recovery requires no additional data: the bias-corrected model is fitted to the same battles as the naive baseline. The two frameworks disagree on a non-trivial share of model pairs (Kendall $\tau = 0.80$) even when all available data are used, a structural disagreement that persists regardless of how many observations each framework receives. When bias is structural, the path to reliable rankings runs

through better measurement, not more measurement.

5. Sustainability: Fewer Comparisons

The measurement argument has a direct sustainability consequence. A misspecified model treats structural rater noise as signal, so every comparison it schedules partly compensates for bias rather than resolving genuine uncertainty about model ability. A correctly specified model never accumulates this waste in the first place, because it attributes rater noise to the judge rather than to the models being compared. The result is stable rankings from fewer comparisons and, at leaderboard scale, fewer comparisons means substantially less LLM inference.

To quantify this, we ran a simulation of $N = 30$ models across six tournament designs (Uniform, Swiss, Stratified, HubSpoke, MultiHub, Popularity) and three mixes of ordinal and pairwise data (10%, 50%, and 90% ordinal), averaging Kendall-τ correlation with ground truth over 50 trials per configuration. Figure 2 reports the results. Across almost all settings, the bias-corrected hybrid model (HybridRater) recovers stable rankings at sample sizes where naive win-rate and simple score averaging remain well below equivalent fidelity. The gap is largest in the adversarial designs- Swiss, Stratified, and MultiHub which most closely resemble real leaderboard conditions where scheduling is non-uniform and model exposure is unbalanced. In these regimes, naive aggregation plateaus well below the hybrid model even at the largest sample sizes tested, confirming that the efficiency gain is not merely a matter of getting there faster, it is a matter of getting there at all.

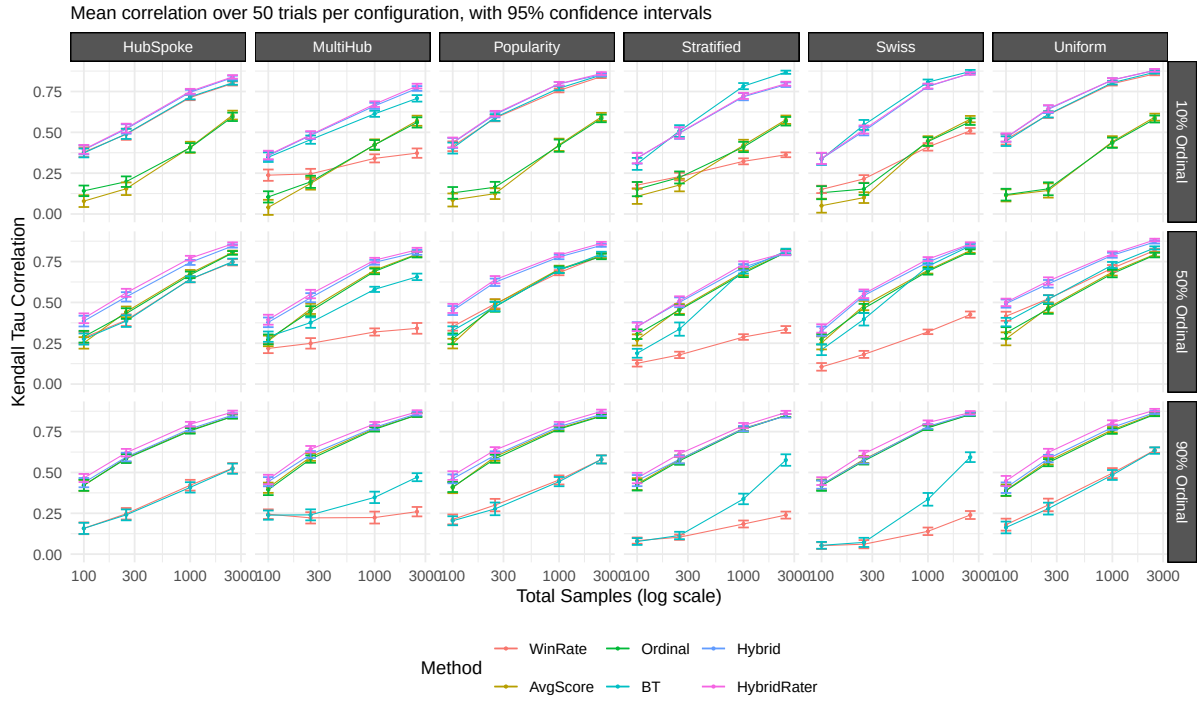

Figure 2: Hybrid model recovers ground-truth rankings with fewer comparisons. Mean Kendall-τ correlation with ground truth over 50 trials, shown across six tournament structures and three proportions of ordinal data. HybridRater consistently outperforms naive win-rate and simple score averaging, particularly in sparse, adversarially scheduled designs that resemble real leaderboard conditions. The efficiency gain is largest where it matters most, at low sample sizes and in unbalanced tournament designs. Careful selection of the scaling parameter ω is important when combining the two measurement scales.

Real leaderboards can recover further gains through adaptive scheduling, rather than drawing pairs at random, new comparisons can be targeted where uncertainty in θ is highest, and the Fisher information of the joint latent-variable model provides a principled criterion for this. The connection to

computerised adaptive testing in educational measurement [26] is direct. Zhuang et al. [27] argue that AI evaluation should adopt precisely these methods from human psychometric testing; Polo et al. [10] demonstrate their practical value, achieving 99% cost reduction through IRT-based item selection on static benchmarks; and Li et al. [28] extend this to a fully active setting, framing evaluation as a sequential acquisition problem where each comparison is selected to maximise expected information gain about model ability. Chatbot Arena already exploits a heuristic version of this, reporting a 54% reduction in required battles through adaptive sampling [15]; the psychometric framework proposed here replaces that heuristic with a grounded information-theoretic criterion. Full implementation of adaptive scheduling within the hybrid framework is an important direction for future work.

6. Limitations and Open Problems

Unidimensionality. The framework assumes a single latent ability dimension θ. For models evaluated across diverse tasks, a multidimensional IRT model would be more appropriate: Burnell et al. [29] show via factor analysis that LLM capabilities have non-trivial latent structure, making a single θ an approximation. Extending the hybrid framework to MIRT is an important open direction.

Construct validity. Correcting for known confounders is necessary but not sufficient. Whether the corrected θ tracks the construct of interest requires systematic validation not yet done at scale for LLM-as-a-judge evaluation [23]. The psychometric framing at least renders this an explicit, testable property of the measurement system rather than an unexamined assumption.

7. Conclusion

Current leaderboards treat instability as a data problem and respond by scaling comparisons. We have argued it is a measurement problem, the biases that distort rankings are structural properties of the evaluation instrument, not sampling artefacts, and no volume of additional battles under the same protocol will correct them. On MT-Bench, position preference, verbosity, self-enhancement, and judge-level heterogeneity are simultaneously detectable, significant, and large enough to move ability estimates in directions that trace back to tournament design rather than model quality.

The remedy is inexpensive. Fitting a generalised linear mixed model costs negligible compute relative to a single round of LLM inference, yet it separates ability from evaluator artefact in a way that arbitrarily large naive samples cannot. The unified framework proposed here—augmented Bradley–Terry models for pairwise data, many-facet Rasch models for ordinal data, and a joint likelihood that bridges both—provides the measurement foundation that current leaderboards lack. More reliable rankings, from fewer comparisons, at lower computational cost, is what sustainable evaluation looks like in practice.

Acknowledgments

We thank the anonymous reviewers. This work was funded by the German Federal Ministry for Research, Technology and Space Travel (BMFTR) grant number 01IW23005.

Declaration on Generative AI

During the preparation of this work the author(s) used Claude for grammar and spelling checks and minor phrasing improvements. After using the tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] H. Ye, J. Jin, Y. Xie, X. Zhang, G. Song, Large language model psychometrics: A systematic review of evaluation, validation, and enhancement, arXiv preprint arXiv:2505.08245 (2025).

[2] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, et al., A survey on llm-as-a-judge, The Innovation 7 (2026).

[3] H. Li, Q. Dong, J. Chen, H. Su, Y. Zhou, Q. Ai, Z. Ye, Y. Liu, Llms-as-judges: a comprehensive survey on llm-based evaluation methods, arXiv preprint arXiv:2412.05579 (2024).

[4] L. Shi, C. Ma, W. Liang, X. Diao, W. Ma, S. Vosoughi, Judging the judges: A systematic study of position bias in llm-as-a-judge, in: Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, 2025, pp. 292–314.

[5] Y. Dubois, P. Liang, T. Hashimoto, Length-controlled alpacaeval: A simple debiasing of automatic evaluators, in: First Conference on Language Modeling, 2024.

[6] L. E. Marsh, P. Kanngiesser, B. Hood, When and how does labour lead to love? the ontogeny and mechanisms of the ikea effect, Cognition 170 (2018) 245–253.

[7] P. Boyeau, A. N. Angelopoulos, T. Li, N. Yosef, J. Malik, M. I. Jordan, Autoeval done right: Using synthetic data for model evaluation, in: International Conference on Machine Learning, PMLR, 2025, pp. 5276–5290.

[8] S. Samsi, D. Zhao, J. McDonald, B. Li, A. Michaleas, M. Jones, W. Bergeron, J. Kepner, D. Tiwari, V. Gadepally, From words to watts: Benchmarking the energy costs of large language model inference, in: 2023 IEEE High Performance Extreme Computing Conference (HPEC), IEEE, 2023, pp. 1–9.

[9] A. Kipnis, K. Voudouris, L. Schulze Buschoff, E. Schulz, metabench-a sparse benchmark of reasoning and knowledge in large language models, in: International Conference on Learning Representations, volume 2025, 2025, pp. 31734–31770.

[10] F. M. Polo, L. Weber, L. Choshen, Y. Sun, G. Xu, M. Yurochkin, tinybenchmarks: evaluating llms with fewer examples, in: Proceedings of the 41st International Conference on Machine Learning, 2024, pp. 34303–34326.

[11] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 610–623.

[12] R. Burnell, W. Schellaert, J. Burden, T. D. Ullman, F. Martinez-Plumed, J. B. Tenenbaum, D. Rutar, L. G. Cheke, J. Sohl-Dickstein, M. Mitchell, et al., Rethink reporting of evaluation results in ai, Science 380 (2023) 136–138.

[13] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al., Judging llm-as-a-judge with mt-bench and chatbot arena, Advances in neural information processing systems 36 (2023) 46595–46623.

[14] M. I. Jordan, On computational thinking, inferential thinking and data science., in: SPAA, 2016, p. 47.

[15] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, et al., Chatbot arena: An open platform for evaluating llms by human preference, arXiv preprint arXiv:2403.04132 (2024).

[16] T. Li, W.-L. Chiang, E. Frick, L. Dunlap, T. Wu, B. Zhu, J. E. Gonzalez, I. Stoica, From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline, in: International Conference on Machine Learning, PMLR, 2025, pp. 34209–34231.

[17] T. M. Liddell, J. K. Kruschke, Analyzing ordinal data with metric models: What could possibly go wrong?, Journal of Experimental Social Psychology 79 (2018) 328–348.

[18] C. DeMars, Item response theory, Oxford University Press, 2010.

[19] F. Martínez-Plumed, R. B. Prudêncio, A. Martínez-Usó, J. Hernández-Orallo, Item response theory in ai: Analysing machine learning classifiers at the instance level, Artificial intelligence 271 (2019) 18–42.

[20] I. Hamilton, N. Tawn, Parameter estimation in comparative judgment under random and adaptive scheduling schemes, Journal of Educational Measurement 63 (2026) e70022.
[21] J. P. Lalor, H. Wu, H. Yu, Building an evaluation scale using item response theory, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 648–657.
[22] D. Federiakin, Improving llm leaderboards with psychometrical methodology, arXiv preprint arXiv:2501.17200 (2025).
[23] X. Wang, L. Jiang, J. Hernandez-Orallo, D. Stillwell, S. Chen, L. Sun, F. Luo, X. Xie, Evaluating general-purpose ai with psychometrics, Communications of the ACM 69 (2026) 92–102.
[24] R. A. Bradley, M. E. Terry, Rank analysis of incomplete block designs: I. the method of paired comparisons, Biometrika 39 (1952) 324–345.
[25] J. M. Linacre, Many-faceted Rasch measurement, Ph.D. thesis, The University of Chicago, 1989.
[26] H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, Computerized adaptive testing: A primer, Routledge, 2000.
[27] Y. Zhuang, Q. Liu, Z. Pardos, P. C. Kyllonen, J. Zu, Z. Huang, S. Wang, E. Chen, Position: Ai evaluation should learn from how we test humans, in: Forty-second International Conference on Machine Learning Position Paper Track, 2025.
[28] Y. Li, J. Ma, M. Ballesteros, Y. Benajiba, G. Horwood, Active evaluation acquisition for efficient llm benchmarking, in: International Conference on Machine Learning, PMLR, 2025, pp. 35581–35602.
[29] R. Burnell, H. Hao, A. R. Conway, J. H. Orallo, Revealing the structure of language model capabilities, arXiv preprint arXiv:2306.10062 (2023).