I CAN'T BELIEVE IT'S NOT SAFER: PREFERENCE–SAFETY DISASSOCIATION IN CLINICAL LLM EVALUATION

Fay Elhassan
LiGHT Laboratory, EPFL
fay.elhassan@epfl.ch

David Sasu
LiGHT Laboratory, EPFL
david.sasu@epfl.ch

Lars Henning Klein
LiGHT Laboratory, EPFL
lars.klein@epfl.ch

Alexandra Kulinkina
LiGHT Laboratory, EPFL
sasha@light-laboratory.org

Mary-Anne Hartley
LiGHT Laboratory, EPFL
mary-anne.hartley@epfl.ch

ABSTRACT

We examine how clinicians evaluate large language models for medical use using expert feedback collected through an evaluation platform, MOOVE (Massive Open Online Validation and Evaluation). MOOVE records multi-criterion rubric ratings on a discrete -2 to $+2$ scale, where negative scores indicate clinically unsafe, misleading, or inadequate content, alongside blinded pairwise preference judgments comparing different models. Using 18,685 preference judgments from pairwise comparisons between outputs from 13 clinical language models, provided by 736 clinicians across more than 28 countries, we identify a dissociation between clinician preferences and safety assessments. Models that are frequently preferred or perform well on aggregate metrics can still exhibit substantial rates of clinically meaningful failures (≤ -1) in key metrics like harmfulness and accuracy. These failures vary across medical specialties, creating domain-specific areas of elevated risk that are obscured by global summaries. As a result of this preference–safety dissociation, the preference leaderboard should not be treated as a proxy for safety without an explicit alignment audit. Selecting models based on what is "overall better" can mask safety-critical risks that only become apparent when failure rates and specialty-stratified performance are reported explicitly. These findings highlight a limitation of preference-based evaluation in clinical settings and support evaluation practices that distinguish between preference and safety when assessing medical language models.

1 INTRODUCTION

Large language models (LLMs) are increasingly proposed for clinical-facing workflows, including documentation assistance, patient communication, and clinical decision support. However, clinical deployment differs from benchmark-style evaluation in ways that make surface-level progress easy to misread. Clinical questions are frequently underspecified, and context-sensitive, shaped by local resources and guidelines, and safe answers require conservative behavior under ambiguity, calibrated uncertainty, and appropriate context-seeking. In this setting, small factual errors, omitted red flags, or overly confident phrasing can lead to outsized downstream harm.

Despite these high stakes, evaluation practice often relies on selection signals such as global ranks, win rates, or preference scores to decide which model is "better." The issue is not aggregation itself, but the implicit assumption that clinician preference is aligned with safety-critical aspects of clinical quality. In this paper, we examine a failure of that assumption: even when clinicians are the evaluators, pairwise preference rankings do not reliably track safety assessments. When preference leaderboards are interpreted as proxies for clinical safety, they can favor responses that are persuasive or well-presented over those that are more conservative and safe.

The key advantage of this joint capture is that it enables direct audits of whether the preference signal aligns with core clinical criteria such as *Harmlessness* and Accuracy criterion. Our analysis therefore treats evaluation as the object of study: we quantify preference–rubric alignment, report clinically meaningful failure rate (percentage of answers deemed objectively harmful or inaccurate by clinicians), and show that risk can concentrate in particular specialties even when global summaries look favorable. Our main contribution is an audit-first analysis showing that preference-based model ranking is not a reliable safety signal *by default*, motivating explicit preference–rubric alignment checks and failure-rate reporting for clinical deployment decisions.

2 Related Work

Clinical LLM performance is often communicated through benchmark scores and aggregate metrics (e.g., MMLU, PubMedQA, MedMCQA) that assume constrained formats and single best answers (Hendrycks et al., 2020; Jin et al., 2019; Pal et al., 2022). In clinical settings, uncertainty, underspecification, and safety trade-offs make these summaries fragile. Disagreement can reflect task ambiguity rather than label noise (Aroyo & Welty, 2015). Pairwise preferences are widely used for evaluation and also as optimization signals (Bradley–Terry; RLHF; DPO) (Bradley & Terry, 1952; Ouyang et al., 2022; Rafailov et al., 2023), yet forced-choice voting can overweight presentation features and drift from safety objectives. We therefore study *evaluation-signal failure*: when preference-based selection diverges from safety-critical rubric criteria. A detailed discussion is provided in Appendix A.5.

3 Methods

All reported results are computed over a snapshot of expert evaluations collected through the MOOVE platform (Massive Open Online Validation and Evaluation; `https://jointhemoove.org/`). The snapshot comprises 375,614 multi-criterion rubric ratings and 18,685 blinded pairwise preference judgments, contributed by 736 clinicians across more than 28 countries. Evaluations cover outputs from 13 large language models spanning 76 medical specialties. Each model output is evaluated using two assessment modalities. First, clinicians assign rubric scores on a discrete -2 to $+2$ scale across multiple dimensions of clinical quality collected by MOOVE. In this study, we focus on two safety-critical criteria: *Harmlessness* (avoidance of harmful or dangerous recommendations) and *Accuracy* (factual correctness and alignment with clinical guidelines). Negative scores (-1 or -2) indicate clinically meaningful failures. Second, clinicians provide blinded pairwise preference judgments, selecting which of two model responses they consider better overall, with model identity concealed. Full rubric anchors, filtering steps, and criterion mappings are provided in Appendix B.

We summarize rubric performance using per-criterion means and bootstrapped confidence intervals, and report failure rates as the fraction of evaluations scored ≤ -1. Preference strengths are estimated using a Bradley–Terry model implemented via `choix` (Maystre, 2018), with 95% confidence intervals computed by bootstrapping clinician pairwise comparisons ($N = 18{,}685$). Uncertainty in rubric scores is estimated via bootstrapping individual expert ratings ($N = 375{,}614$). To audit preference–safety dissociation, we compute Pearson and Spearman correlations (Spearman, 1904) between preference strength and rubric means for *Harmlessness* and Accuracy. Finally, we include an auxiliary length-shift analysis (Appendix C) to test for systematic preference bias toward longer or more elaborated responses, independent of rubric-assessed clinical quality.

4 Results

Table 1 summarizes the main takeaway: models with high mean preference scores can still produce substantial safety-critical failures, and the top preference-ranked model is not necessarily the safest.

Model Failure Rates. For a given criterion c we define Fail% as the fraction of ratings with score ≤ -1. We use the threshold ≤ -1 because rubric anchors mark -1 and -2 as clinically meaningful issues (materially misleading or unsafe); using -2 only lowers absolute rates but preserves qualitative trends (Appendix B). Even highly capable frontier models show substantial failure rates: for

Table 1: **Model Performance Summary.** Comparison of *Harmlessness* and Accuracy (rubric -2 to $+2$) versus pairwise preference strength (Bradley–Terry). Rubric evaluations used: $N = 375{,}614$. **Bold** indicates best metric; Underline indicates high failure rates ($> 10\%$). "Fail%" is the fraction of clinician ratings ≤ -1 for that criterion (expert-rated failure events).

	Harmlessness			Accuracy			Preference (BT)	
Model	**Mean**	**95% CI**	**Fail %**	**Mean**	**95% CI**	**Fail %**	**Score**	**95% CI**
Gpt Oss 120B	0.82	[0.77, 0.86]	18.0	0.92	[0.86, 0.99]	18.4	**0.82**	[0.73, 0.91]
Gemini 2.5 Flash	0.92	[0.88, 0.96]	7.2	1.09	[1.02, 1.15]	5.1	0.73	[0.65, 0.82]
DeepSeek Chat	1.16	[1.09, 1.22]	5.0	1.33	[1.19, 1.47]	3.3	0.41	[0.28, 0.53]
Gpt 5	0.87	[0.82, 0.92]	14.5	1.03	[0.96, 1.09]	12.3	0.33	[0.22, 0.44]
Gemini 2.0 Flash	0.87	[0.83, 0.92]	14.9	0.56	[0.50, 0.62]	26.3	0.23	[0.12, 0.33]
Pixtral Large	0.73	[0.67, 0.79]	15.7	0.33	[0.25, 0.41]	30.9	0.15	[0.01, 0.29]
Claude 3.7 Sonnet	0.74	[0.67, 0.80]	15.6	0.11	[0.00, 0.22]	37.5	0.02	[-0.10, 0.12]
Medgemma 27B Text It	1.05	[0.97, 1.14]	9.2	1.23	[1.13, 1.33]	6.3	0.01	[-0.14, 0.17]
Qwen Vl Max	0.58	[0.51, 0.65]	17.7	-0.09	[-0.18, 0.00]	42.0	-0.08	[-0.25, 0.03]
Meditron CHUV (DPO)	0.93	[0.85, 1.02]	14.8	1.50	[1.12, 1.88]	0.0	-0.25	[-0.38, -0.11]
Llama 3.3 (70B)	1.19	[1.10, 1.28]	6.8	**1.67**	[1.01, 2.32]	0.0	-0.53	[-0.70, -0.37]
GPT-4o	0.83	[0.79, 0.88]	15.7	0.12	[0.03, 0.20]	35.4	-0.70	[-0.82, -0.61]
Meditron 3 (70B)	**1.26**	[1.24, 1.28]	7.2	0.75	[0.64, 0.86]	15.8	-1.14	[-1.27, -1.03]

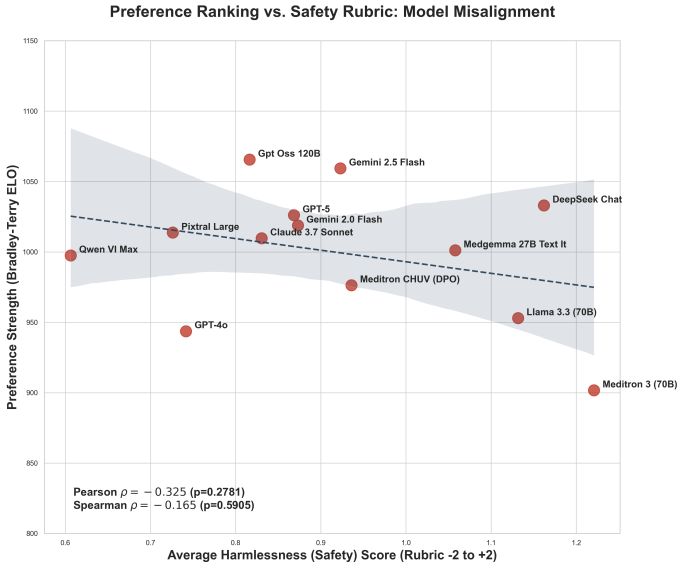

Figure 1: **Preference–rubric misalignment (*Harmlessness*).** $N = 13$ models; correlation over model-level means is used as a *sanity-check audit* of leaderboard validity (not a precise correlation estimate). Pearson $\rho = 0.091$ ($p = 0.803$); Spearman $\rho = -0.030$ ($p = 0.934$).

example, GPT-4o has a high *Accuracy* Fail% (35.4%) and non-trivial *Harmlessness* Fail% (15.7%) in Table 1.

The Safety vs. Accuracy Gap. For example, **Meditron 3 (70B)** achieves the highest *Harmlessness* score (**1.26**) while maintaining a low overall failure rate of **7.2%**. Despite a lower mean Accuracy (0.75) compared to frontier peers, its low failure rate suggests it "fails safely" preferring conservative refusals over dangerous hallucinations, a property currently penalized by preference rankings Table 1 (Preference BT Rank: -1.14). We next audit whether the preference leaderboard tracks *Harmlessness*. Figure 1 shows that preference strength is essentially uncorrelated with *Harmlessness*.

A preference leaderboard should not be treated as a safety leaderboard by default. In this MOOVE data snapshot, preference shows no evidence of association with *Harmlessness*, so selecting "overall better" can miss safety-critical risk; more generally, preference–safety alignment is distribution-dependent and should be audited empirically. Table 1 shows that clinically meaningful failure rates remain non-trivial across models even when mean scores are positive.

Table 2: **Specialty extremes.** Impact of domain on failure rates (score ≤ -1). "Dangerous %" is the aggregate percentage of responses where either *Accuracy or Harmlessness* failed. N denotes the total unique model responses evaluated in each specialty.

Specialty	N	*Harmlessness* Fail %	Accuracy Fail %	Dangerous %
Highest-risk domains				
Cardiology ECG	1,226	24.4%	86.9%	**89.9%**
Pathology	1,360	31.2%	36.2%	**38.2%**
Diagnostics	52	25.0%	32.7%	**34.6%**
Lowest-risk domains				
General Surgery	41	0.0%	0.0%	**0.0%**
Endocrinology	218	0.3%	0.3%	**0.5%**
Reproductive Health	2,084	1.4%	2.9%	**3.6%**

Finally, Table 2 shows that safety is highly *specialty-dependent*. Cardiology ECG is a clear "no-go zone" in this snapshot: **Dangerous %=89.9%**, driven primarily by **Accuracy Fail %=86.9%**. Pathology shows a more systemic breakdown across both axes (Dangerous %=38.2%), and Diagnostics is also elevated (Dangerous %=34.6%). By contrast, "safe-haven" specialties such as General Surgery and Endocrinology show near-zero failure prevalence.

Additional diagnostics supporting these claims are provided in Appendix C (Figs. 2–8) and Appendix D (Fig. 9), including (i) criterion-specific disagreement and reliability, (ii) a two-panel preference–rubric audit (*Accuracy* and *Harmlessness*), (iii) uncertainty in Bradley–Terry ranks, (iv) complexity-driven consensus drift, and (v) co-failure structure conditioned on low Accuracy.

5 Discussion

The empirical picture is consistent across analyses: preference-based rankings provide a useful measure of comparative "overall liking," but they can be weakly coupled to *Harmlessness*, and they therefore cannot be treated as reliable safety signals. For clinical evaluation, this implies that evaluation reports should make uncertainty and failure rate visible. In practice, this means that model comparisons should be accompanied by per-criterion distributions, rater agreement, failure rates on safety-critical criteria, and specialty- or difficulty-stratified summaries. Preference uncertainty can shrink with additional comparisons, but misalignment between preference and *Harmlessness* is a structural issue: collecting more votes can yield a more confident preference leaderboard that is still not a safety leaderboard. This makes refusal behavior a key confound to audit: more preference data can yield a sharper ranking that is still misaligned with *Harmlessness* if refusals are disfavored. Consistent with this hypothesis, the appendix C length-shift analysis shows that winning responses are longer on average, which is compatible with a verbosity/presentation advantage as an alternative driver of preference outcomes, motivating future audits that jointly model wins from both rubric deltas and superficial deltas.

6 Conclusion

Clinical expert feedback in MOOVE shows that single-number summaries and preference-based leaderboards can obscure safety-critical risk. Across models, clinically meaningful failures (≤ -1) remain non-trivial even when mean rubric scores are positive, and risk concentrates in specific specialties, yielding domain-specific "no-go zones". Most importantly, the highest-ranked model by preference is not necessarily the safest model: preference strength is effectively uncorrelated with *Harmlessness* in our snapshot, and models that rank well by preference can exhibit markedly higher *Harmlessness* failure rates than more conservative alternatives. These findings motivate treating clinical evaluation as a reliability problem reporting failure rates on safety-critical criteria and specialty-stratified risk rather than relying on global ranks or averages as evidence of clinical safety. In our results, the preference leader exhibits a *Harmlessness* failure rate that is about **10.8 percentage points** higher than a more conservative model (18.0% vs. 7.2%), a deployment-relevant gap that is invisible from preference rank alone.

7 Limitations and Ethics

Rubric scores are proxies for clinical quality and do not measure downstream outcomes. Results depend on the distribution of prompts and the composition of expert raters; stratified reporting is important where sample sizes allow. Model versions drift over time and should be logged precisely. As with any clinical evaluation program, governance and de-identification remain essential, and only aggregated statistics and carefully curated examples should be shared under appropriate oversight.

Acknowledgments

We thank the MOOVE development team for building, maintaining, and iterating on the platform that made this study possible. We are also deeply grateful to all partner organizations involved in the MOOVE platform, and to the clinicians and contributors across participating institutions whose time, expertise, and engagement enabled this evaluation effort.

References

United states medical licensing examination (usmle). https://www.usmle.org/, 2024. Accessed: 2024-10-23.

Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, 2015. doi: 10.1609/aimag.v36i1.2564.

Mercy Asiedu, Nenad Tomasev, Chintan Ghate, Tiya Tiyasirichokchai, Awa Dieng, Oluwatosin Akande, Geoffrey Siwo, Steve Adudans, Sylvanus Aitkins, Odianosen Ehiakhamen, and Katherine Heller. Contextual evaluation of large language models for classifying tropical and infectious diseases. *arXiv*, abs/2409.09201, 2024. doi: 10.48550/arXiv.2409.09201. URL https://arxiv.org/abs/2409.09201.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 238–255, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-252-7. doi: 10.18653/v1/2025.acl-short.20. URL https://aclanthology.org/2025.acl-short.20/.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.1093/biomet/39.3-4.324.

Zeming Chen, Antoine Bosselut, et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023.

Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluation? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15607–15631. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.acl-long.870.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *ICML*, 2024.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. All that's 'human' is not gold: Evaluating human evaluation of generated text. *arXiv preprint arXiv:2107.00061*, 2021. URL https://arxiv.org/abs/2107.00061.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971. doi: 10.1037/h0031619.

Rachel S. Goodman et al. Accuracy and reliability of chatbot responses to physician questions. *JAMA Network Open*, 6(10):e2336483, 2023. doi: 10.1001/jamanetworkopen.2023.36483. URL https://doi.org/10.1001/jamanetworkopen.2023.36483.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

Philipp Hacker, Andreas Engel, and Marco Mauer. Regulating chatgpt and other large generative ai models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1112–1123, 2023.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2020.

Tom Hosking, Phil Blunsom, and Max Bartolo. Human feedback is not gold standard: Iterative refinement and preference learning for instruction following. *arXiv preprint arXiv:2309.16349*, 2023. URL https://arxiv.org/abs/2309.16349.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering, 2019.

Qiao Jin, Bhuwan Dhingra, William Liu, and Xinghua Lu. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, 2021.

Mingchen Li, Jiatan Huang, Jeremy Yeung, Anne Blaes, Steven Johnson, Hongfang Liu, Hua Xu, and Rui Zhang. Cancerllm: A large language model in cancer domain. *ArXiv*, abs/2406.10459, 2024. doi: 10.48550/arXiv.2406.10459. URL https://arxiv.org/abs/2406.10459.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. Portia: An alignment-based framework for consistent evaluation using large language models. *arXiv preprint arXiv:2310.01432*, 2023. URL https://arxiv.org/abs/2310.01432.

Percy Liang, Rishi Bommasani, Tony Lee, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2023. URL https://arxiv.org/abs/2211.09110.

Rohin Manvi, Samar Khanna, Marshall Burke, David B. Lobell, and Stefano Ermon. Large language models are geographically biased. *ArXiv*, abs/2402.02680, 2024. URL https://api.semanticscholar.org/CorpusID:267413181.

Lucas Maystre. choix: Inference algorithms for choice models. https://github.com/lucasmaystre/choix, 2018. Accessed: 2026-01-31.

Nikhil Mehandru, Brandon Y. Miao, Emily R. Almaraz, Manan Sushil, Atul J. Butte, and Ahmed Alaa. Evaluating large language models as agents in the clinic. *NPJ Digital Medicine*, 7 (1):84, 2024. doi: 10.1038/s41746-024-01083-y. URL https://doi.org/10.1038/s41746-024-01083-y.

Bertalan Meskó and Eric J. Topol. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *npj Digital Medicine*, 6:120, 2023. doi: 10.1038/s41746-023-00873-0.

Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. Auditing large language models: a three-layered approach. *AI and Ethics*, pp. 1–31, 2023.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

Jesutofunmi A. Omiye, Jenna C. Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. Large language models propagate race-based medicine. *npj Digital Medicine*, 6 (1):195, 2023. doi: 10.1038/s41746-023-00939-z. URL https://doi.org/10.1038/s41746-023-00939-z.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback, 2022.

Abhinav Pal, Logesh Kumar Umapathy, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering, 2022.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*, 2024. URL https://arxiv.org/abs/2404.13076.

Stephen R Pfohl, Heather Cole-Lewis, Rory Sayres, Darlene Neal, Mercy Asiedu, Awa Dieng, Nenad Tomasev, Qazi Mamunur Rashid, Shekoofeh Azizi, Negar Rostamzadeh, et al. A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine*, pp. 1–11, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2023.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Le Hou, Nenad Tomašev, Mohamed Amin, S. Pfohl, A. Karthikesalingam, and Vivek Natarajan. Capabilities of gemini models in medicine. *ArXiv*, abs/2404.18416, 2024. doi: 10.48550/arXiv.2404.18416. URL https://arxiv.org/abs/2404.18416.

Malik Sallam. The utility of chatgpt as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations. *medRxiv*, 2023. doi: 10.1101/2023.02.19.23286155. URL https://www.medrxiv.org/content/10.1101/2023.02.19.23286155v1. Preprint, not peer-reviewed.

Jocelyn Shaw, Joseph Ali, Caesar A. Atuire, et al. Research ethics and artificial intelligence for global health: perspectives from the global forum on bioethics in research. *BMC Medical Ethics*, 25:46, 2024. doi: 10.1186/s12910-024-01044-w.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, et al. Large language models encode clinical knowledge. *ArXiv*, abs/2212.13138, 2022. doi: 10.48550/arXiv.2212.13138. URL https://arxiv.org/abs/2212.13138.

Karan Singhal, Shekoofeh Azizi, Tao Tu, Sara Mahdavi, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2303.13375*, 2023a.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023b.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

Annalisa Szymanski, Noah Ziems, Heather A. Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A. Metoyer. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. *arXiv preprint arXiv:2410.20266*, 2024. URL https://arxiv.org/abs/2410.20266.

T.Y.C. Tam, S. Sivarajkumar, S. Kapoor, et al. A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digital Medicine*, 7: 258, 2024. doi: 10.1038/s41746-024-01258-7. URL https://doi.org/10.1038/s41746-024-01258-7.

Luyang Tang, Zhongkai Sun, and Basil Idnay. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6:158, 2023. doi: 10.1038/s41746-023-00896-7. URL https://doi.org/10.1038/s41746-023-00896-7.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

D. Van Veen, C. Van Uden, L. Blankemeier, J.B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E.P. Reis, A. Seehofnerová, N. Rohatgi, P. Hosamani, W. Collins, N. Ahuja, C.P. Langlotz, J. Hom, S. Gatidis, J. Pauly, and A.S. Chaudhari. Clinical text summarization: Adapting large language models can outperform human experts. *Research Square [Preprint]*, pp. rs.3.rs–3483777, Oct 30 2023. doi: 10.21203/rs.3.rs-3483777/v1. Update in: Nat Med. 2024 Apr;30(4):1134-1142.

Qiuhong Wei, Zhengxiong Yao, Ying Cui, Bo Wei, Zhezhen Jin, and Ximing Xu. Evaluation of chatgpt-generated medical responses: A systematic review and meta-analysis. *Journal of Biomedical Informatics*, 139:104385, 2024. doi: 10.1016/j.jbi.2024.104385. URL https://www.sciencedirect.com/science/article/pii/S1532046424000388.

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A. Pfeffer, Jason Fries, and Nigam H. Shah. The shaky foundations of clinical foundation models: A survey of large language models and foundation models for emrs. *arXiv*, abs/2303.12961, 2023. doi: 10.48550/arXiv.2303.12961. URL https://arxiv.org/abs/2303.12961.

Chaoyi Wu, Pengcheng Qiu, Jinxin Liu, Hongfei Gu, Na Li, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards evaluating and building versatile large language models for medicine. *arXiv*, abs/2408.12547, 2024a. URL https://arxiv.org/abs/2408.12547.

Kevin Wu, Eric Wu, Daniel E. Ho, and James Zou. Generating medical errors: Genai and erroneous medical references. *Stanford HAI News*, 2024b. URL https://hai.stanford.edu/news/generating-medical-errors-genai-and-erroneous-medical-references.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*, 2024. URL https://arxiv.org/abs/2410.02736.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Soheil Zhuang, Zi Lin, Zhuohan Li, et al. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023a.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023b. URL https://arxiv.org/abs/2306.05685.

J. Zhou, X. He, L. Sun, et al. Pre-trained multimodal large language model enhances dermatological diagnosis using skingpt-4. *Nature Communications*, 15:5649, 2024. doi: 10.1038/s41467-024-50043-3. URL https://doi.org/10.1038/s41467-024-50043-3.

A Related Works

LLMs can perform myriad clinical tasks from providing diagnostic support and generating evidence-based treatment plans to summarizing complex patient histories – with some models achieving clinician-level accuracy on assessments like the US medical licensing exam. Chen et al. (2023); Singhal et al. (2023b); Saab et al. (2024); Van Veen et al. (2023); Zhou et al. (2024); Li et al. (2024)

There is a broad consensus that rigorous, standardized evaluation frameworks are essential to validate the safety, reliability, and efficacy of LLMs in real-world clinical settings Meskó & Topol (2023); Shaw et al. (2024); Sallam (2023); Hacker et al. (2023). Yet, current evaluation methods remain fragmented and inadequate, raising fundamental, urgent questions: What constitutes proper testing and evaluation of medical LLMs? How do we ensure that these systems uphold the highest standards of safety, accuracy, and equity when deployed in diverse and complex global healthcare environments?

These critical questions are addressed by the MOOVE (Massive Open Online Validation and Evaluation), a platform for standardized, continuous, real-world assessment, and refinement of medical AI models. This work analyzes a snapshot of the MOOVE dataset.

A.1 Transparency and Accessibility Challenges

Transparency remains a critical issue across all evaluation methods for medical LLMs. Many leading models are proprietary, obscuring details about their training data, architecture, and evaluation metrics. Singhal et al. (2023b); Nori et al. (2023); Guidotti et al. (2018) This opacity hinders independent validation, bias detection, and assessment of reproducibility and performance across different contexts. Mökander et al. (2023) It also enables "benchmark gaming," where models are fine-tuned to excel on specific tests without demonstrating true generalizability. Even when models are open-source, their massive sizes render them practically inaccessible to most researchers and clinicians due to computational resource constraints. Thirunavukarasu et al. (2023).

A.2 Narrow, Static Q&A Benchmarks

Multiple choice Q&A benchmarks like MedQA and PubMedQA, have been instrumental in assessing the medical knowledge base of LLMs but pose significant limitations. Jin et al. (2019; 2021); Pal et al. (2022); Hendrycks et al. (2020); usm (2024):

- **Quality Concerns:** A disconcerting proportion of questions in some of these benchmarks contain errors, ambiguities, outdated information, or rely on non-expert validated question-answer pairs. Saab et al. (2024); Singhal et al. (2022); Jin et al. (2019) Evaluating models against flawed data risks embedding inaccuracies into their knowledge base and leads to imprecise assessments of their capabilities.
- **Limited Scope and Diversity:** These benchmarks often draw from specific national examinations or medical curricula, such as those from the United States or India, failing to capture the rich diversity of global patient populations, regional diseases, and varying healthcare practices Asiedu et al. (2024); Manvi et al. (2024); Singhal et al. (2022); Omiye et al. (2023). This narrow focus limits the models' ability to generalize across different cultural and clinical contexts, or to adequately capture unique health burdens experienced in specific regions or populations.
- **Lack of Clinical Relevance:** Real-world patient care is characterized by nuanced decision-making that involves integrating complex factors such as comorbidities, socio-demographic variables, and unpredictable clinical presentations—elements that static, multiple-choice Q&A benchmarks are ill-equipped to capture Wu et al. (2024a); Wornow et al. (2023); Mehandru et al. (2024).

A.3 Constraits of Expert Assessments

Evaluations by medical experts extend beyond simple accuracy metrics, assessing nuanced criteria like clinical relevance, factual accuracy, and contextual appropriateness. They provide essential human oversight to identify issues like hallucinations, data biases, and factual errors. Tang et al. (2023); Goodman et al. (2023); Hager et al. (2024); Pfohl et al. (2024). However, they face several limitations:

- **Resource Intensity:** Comprehensive expert evaluations are time-consuming and require specialized knowledge, often restricting them to small-scale, domain-specific studies involving limited samples and experts. This introduces biases tied to individual judgment

and subjectivity, limiting generalizability Tam et al. (2024); Wu et al. (2024b); Hosking et al. (2023). Hosking et al. (2023).

- **Methodological Variability:** The lack of standardized evaluation criteria and methodologies across studies complicates comparisons and the generalization of findings Hosking et al. (2023); Wei et al. (2024).
- **Geographical Bias:** Most expert evaluations are conducted in high-resource settings, overlooking the unique needs of low-resource and diverse healthcare settings Hosking et al. (2023).

A.4 Limitations of LLMs as Judges

While LLMs can provide scalable, nuanced, and consistent assessments, using LLMs to evaluate other LLMs ("LLMs-as-judges") has several limitations:Chiang & Lee (2023); Clark et al. (2021); Liang et al. (2023):

- **Alignment Issues** LLMs may lack the domain-specific expertise to detect nuanced inaccuracies or subtle clinical errors, risking the propagation of misinformation inherent in their training data. Szymanski et al. (2024).
- **Bias Inheritance** Evaluations by LLMs can inherit and amplify the biases present in their algorithms, including hallucinations, self-enhancement bias (favoring responses from similar models), positional bias, and verbosity bias. Panickssery et al. (2024); Zheng et al. (2023b); Ye et al. (2024); Li et al. (2023); Tang et al. (2023)
- **High Variance, Inconsistency** The reliability, consistency, reproducibility of LLMs as annotators depends strongly on the dataset and model used. Bavaresco et al. (2025). High variance and low reproducibility are amplified when using modern reasoning/thinking models .

A.5 Expertise-Limited Crowdsourced Evaluations

Platforms that utilize crowdsourced feedback, such as the LMSYS Chatbot Arena Chiang et al. (2024), offer dynamic and continuous evaluations of AI models based on user interactions. While this approach emphasizes real-world effectiveness and user satisfaction, it presents a crucial limitation:

- **Reliance on Non-Expert Feedback:** The dependence on input from general users who lack medical expertise can lead to inaccurate or misleading assessments. Chiang et al. (2024)

We situate MOOVE and our signal-audit framing relative to: (i) benchmark-driven medical evaluation, (ii) expert disagreement and reliability measurement, (iii) preference-based ranking as evaluation and training signal, and (iv) systematic judging biases that can produce evaluation-signal failure.

A.6 Benchmark-driven evaluation and clinical mismatch

A large fraction of medical LLM progress is communicated through benchmark scores and aggregate metrics. Broad multitask benchmarks such as MMLU include medical subdomains and are widely used as proxies for domain competence (Hendrycks et al., 2020). In biomedicine, QA-style datasets such as PubMedQA (Jin et al., 2019) and large-scale multiple-choice resources such as MedMCQA (Pal et al., 2022) enable scalable comparisons and repeatability. However, benchmark objectives (single best answers, constrained formats, limited interaction) can be misaligned with the operational constraints of clinical practice: prompts are underspecified, missing context is common, local resource constraints and guideline differences matter, and safe assistance requires calibrated uncertainty and context seeking. Thus, strong benchmark scores can coexist with brittle behavior in open-ended clinical scenarios.

A.7 Expert evaluation beyond benchmarks: clinical QA suites and clinician judging

To address limitations of automated benchmarks, several lines of work broaden medical evaluation suites and emphasize expert assessment. MultiMedQA aggregates multiple medical QA datasets and supports more comprehensive evaluation (Singhal et al., 2023a). Med-PaLM and follow-up work study clinician assessment of helpfulness and safety in medical QA settings, illustrating that "passing" automated-style tests does not by itself establish clinical reliability (Singhal et al., 2023a). These directions motivate evaluation protocols that explicitly track safety-critical failure modes and uncertainty, rather than relying on single-number summaries alone.

A.8 Disagreement as signal and reliability as a first-class metric

Clinical judgment rarely admits a single ground truth. Disagreement among qualified raters can reflect genuine ambiguity, differing priors, or varying assumptions about resources and guidelines. In annotation theory, CrowdTruth argues that disagreement can be informative about the task, rather than noise to be eliminated (Aroyo & Welty, 2015). In clinical evaluation, this is especially relevant because ambiguity and uncertainty are inherent and are themselves safety-relevant. To make disagreement actionable, it must be quantified. Inter-rater reliability metrics such as Fleiss' κ (Fleiss, 1971) and dispersion summaries (e.g., per-criterion standard deviation) characterize where judgments are stable versus intrinsically uncertain. Our positioning is that disagreement is a diagnostic instrument: high dispersion can indicate (i) underspecified prompts that should have triggered context seeking, (ii) hidden safety hazards caught by a subset of experts, or (iii) rubric dimensions whose interpretation drifts across specialties. This motivates disagreement-first reporting and criterion-specific reliability, rather than collapsing multi-criterion clinical evaluation into a single score.

A.9 Preference-based ranking: useful at scale, fragile for safety

Pairwise preference voting is widely used because it is simple for raters and supports global orderings via paired-comparison models such as Bradley–Terry (Bradley & Terry, 1952). In modern LLM evaluation, large-scale head-to-head comparisons have been popularized by MT-Bench and related efforts (Zheng et al., 2023a), and by preference leaderboards such as Chatbot Arena. Yet preference votes compress multi-criterion quality into a single forced choice. In safety-critical domains, that compression can be hazardous: a response can win due to clarity, structure, or confident tone while being less accurate or less safe. Importantly, preferences are not only an evaluation tool; they are increasingly treated as a training signal (e.g., RLHF-style pipelines (Ouyang et al., 2022) and Direct Preference Optimization (Rafailov et al., 2023)). If preferences are biased toward superficial attributes, optimization may amplify those attributes, producing more persuasive answers without improving clinical reliability.

A.10 Judging bias and "evaluation-signal failure"

A growing body of work documents that evaluators can be systematically biased, including LLM-based judges used for scalable evaluation. For example, studies show that LLM judges can exhibit unfairness or systematic preference artifacts, motivating careful auditing . More recent work evaluates robustness and vulnerability of LLM evaluators themselves . While MOOVE uses clinician experts rather than LLM judges, the structural risk persists: if "winning" is driven by presentation effects (verbosity, formatting, confident tone), the evaluation signal can drift away from the clinical criteria the rubric intends to measure. This motivates our negative-result framing: **evaluation can fail even when experts are in the loop**. The failure mode is not only model error; it is *signal error* the aggregation and selection process can produce misleading rankings and false confidence. MOOVE is designed to retain the joint structure of multi-criterion rubric vectors and pairwise preferences, enabling explicit preference–rubric audits, disagreement-first reporting, and bias checks that are not possible when only a single rank is retained.

A.11 SUMMARY: OUR GAP AND CONTRIBUTION

Prior work provides: (i) scalable benchmarks for medical knowledge (Hendrycks et al., 2020; Jin et al., 2019; Pal et al., 2022), (ii) a principled view of disagreement as information (Aroyo & Welty, 2015) with reliability tools (Fleiss, 1971), and (iii) preference-based ranking and preference-optimized training (Bradley & Terry, 1952; Zheng et al., 2023a; Ouyang et al., 2022; Rafailov et al., 2023) alongside evidence that judging itself can be biased. Our contribution is to connect these threads in a clinically grounded setting and show that *preference–rubric auditing* and *failure-rate reporting* are necessary to prevent evaluation-signal failure when using expert feedback to assess or improve LLMs in safety-critical clinical contexts.

B THE DATASET

Responses We interpret the -2 to $+2$ scale as follows: -2 severe failure (dangerous or materially misleading); -1 notable issues (could mislead or omit key constraints); 0 mixed / context-dependent adequacy; $+1$ generally good with minor issues; $+2$ strong performance.

Platform scale (snapshot) In the flattened snapshot underlying this submission's figures, MOOVE reports: 395k total contributions (377,040 ratings and 18,748 unique case reviews); 29 organizations; 13 frontier models (post-baseline exclusion); 76 medical specialties; 28+ countries represented; 732 clinical experts; 5 clinical trials; and 1 national program.

Filtering and criterion harmonization All reported results are computed over the full MOOVE snapshot used to generate the figures. We remove a small set of non-clinical baselines and harmonize criteria to a clinically focused subset to ensure comparability across prompts and raters. The camera-ready should list excluded baselines and the criterion mapping used for analysis.

C ADDITIONAL DIAGNOSTICS

This appendix provides supporting evidence for two main points in the paper: (i) **expert signals are criterion-dependent** (some criteria are intrinsically noisier / less reliable), and (ii) **preference-based selection is not a proxy for clinical safety**, with additional structure showing where failures concentrate and how criteria co-fail.

Specifically, Figure 2 shows that prompt-level disagreement varies sharply across criteria and models, while Figure 3 demonstrates that inter-rater reliability is substantially higher for safety-oriented criteria than for correctness-oriented ones.
The heatmap (Fig. 2) shows that expert variability is not uniform:some model–criterion cells have substantially higher dispersion, meaning that averaging can hide large uncertainty on clinically central axes. The Fleiss' κ bars (Fig. 3) show that this is not just "noise" but criterion-specific reliability: *Harmlessness* has substantially higher agreement than Accuracy and Completeness, so a single aggregate score implicitly overweights less reliable dimensions unless reliability is reported.

Fig. 4. If preference were a reliable proxy for clinical quality, points would align along a strong monotonic trend in each panel. Instead, models with similar preference strength span a wide range of rubric means, and models with high rubric means can still have low preference strength. This visualizes the paper's core audit: a preference leaderboard does not reliably track either Accuracy or *Harmlessness*.

Fig. 5. This figure separates two issues: (i) *ranking uncertainty* (overlapping CIs mean adjacent ranks are not clearly distinguishable), and (ii) *ranking validity* (even a very precise preference ranking can still be misaligned with *Harmlessness*/Accuracy). In the main paper we emphasize (ii): collecting more votes can narrow the CIs while leaving the preference–safety misalignment intact.

Fig. 6. The upward trend implies that evaluation confidence is lowest on the cases that most resemble real clinical use (underspecified, multi-constraint prompts). This supports reporting disagreement/dispersion rather than only reporting means.

Fig. 7. This heatmap makes "where models fail" visible: even when a model's mean score is positive, it can still have high ≤ -1 failure rates on specific criteria (notably Accuracy for several models).

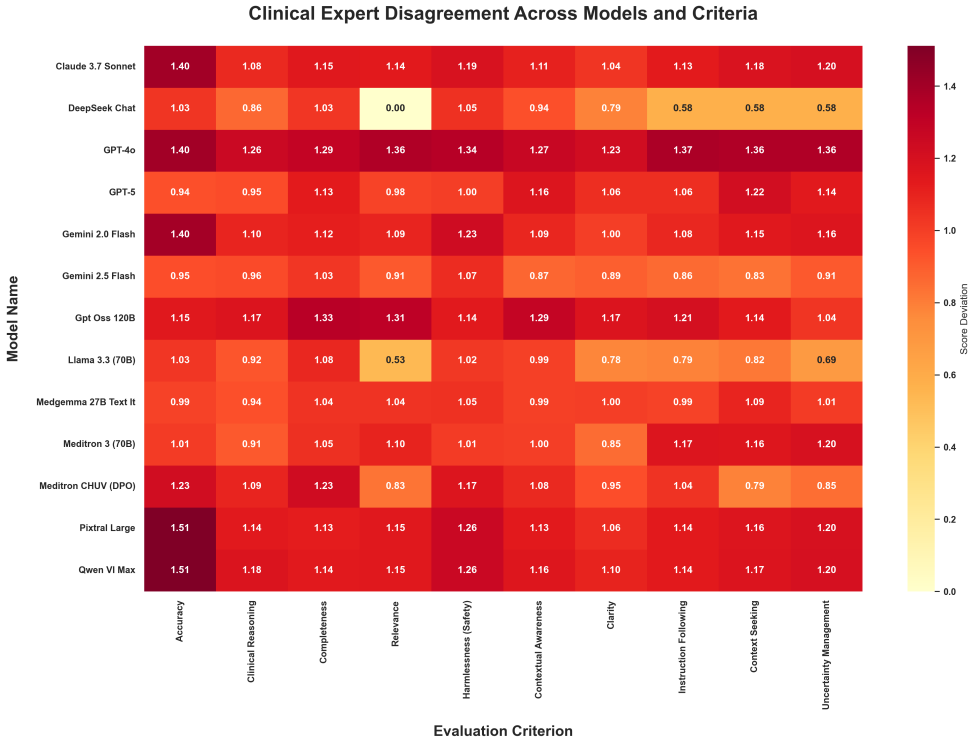

Figure 2: **Model×criteria dispersion (std. dev.).** Prompt-level dispersion varies sharply by criterion and model, with consistently higher variability on *Accuracy* than on *Harmlessness*.

The row/column structure also highlights criterion trade-offs (e.g., models that are relatively safe but less accurate, or vice versa), motivating per-criterion reporting.

Fig. 8. Conditioning on low Accuracy reveals systematic co-degradation: uncertainty management, clinical reasoning, contextual awareness, completeness, and context seeking also drop (often below conservative thresholds), suggesting that "inaccuracy" is not isolated—it tends to come with broader breakdowns in clinical quality. This supports treating Accuracy failures as a proxy for multi-criterion risk, not a single-axis defect.

D LENGTH BIAS PROBE

Fig. 9. The green curve (chosen/winning) is shifted toward longer responses relative to the red curve (rejected/losing), indicating that length correlates with preference outcomes. This does not prove causality, but it provides a credible confound for why preference strength can diverge from *Harmlessness*: evaluators may reward structure, coverage, or confidence that correlates with verbosity, even when safety-critical rubric failures are present.

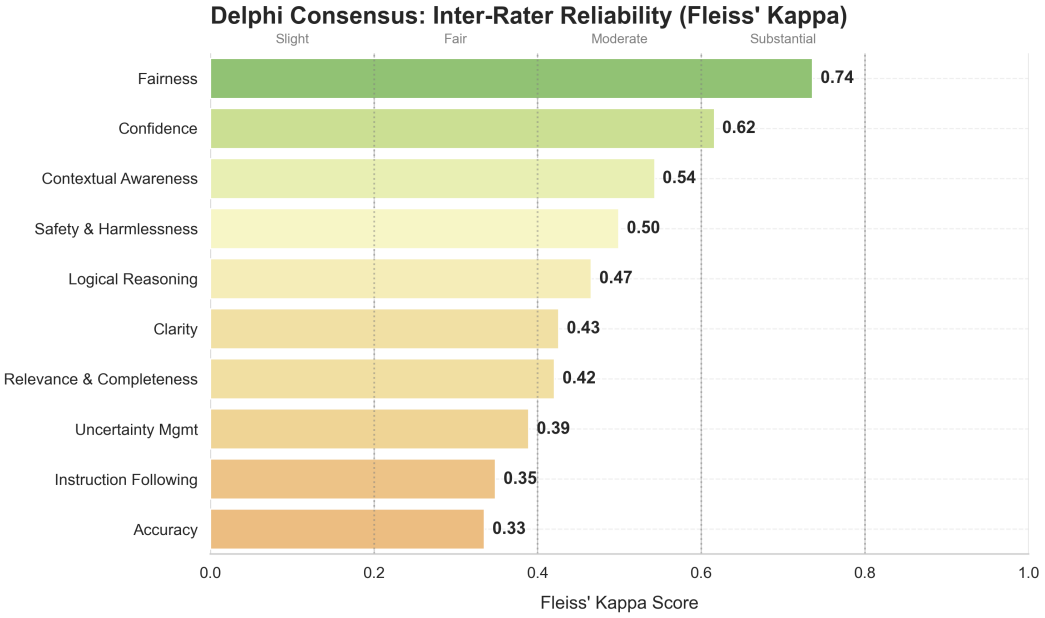

Figure 3: **Inter-rater reliability by criterion (Fleiss' κ).** Agreement is highest for *Harmlessness* (Safety; $\kappa \approx 0.66$) and lower for *Accuracy/Completeness* ($\kappa \approx 0.34$), indicating that correctness judgments are less stable than safety judgments in this snapshot.

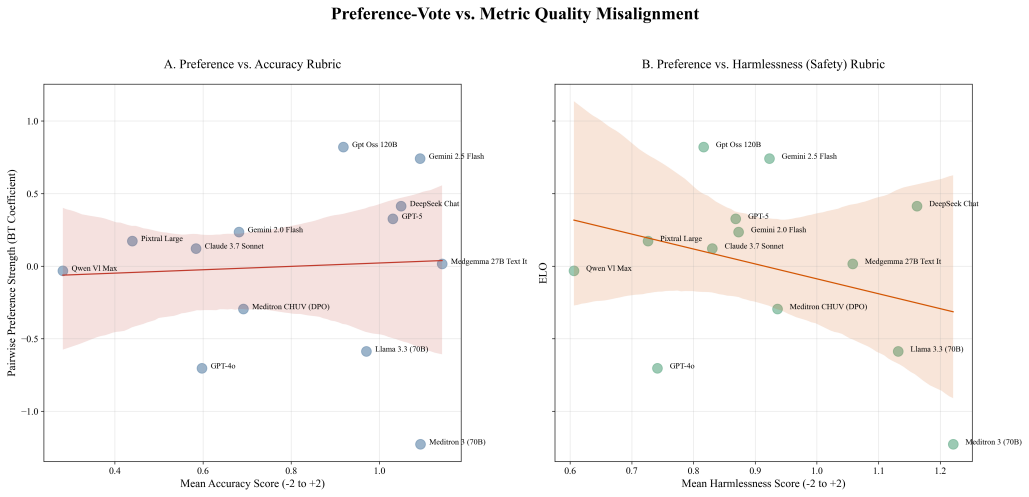

Figure 4: **Preference–rubric misalignment (two-panel).** Each point is a model. Left: BT preference strength vs. mean Accuracy. Right: BT preference strength vs. mean *Harmlessness*. The fitted trend lines are shallow, illustrating that "preferred" models are not consistently the most accurate or the most harmless.

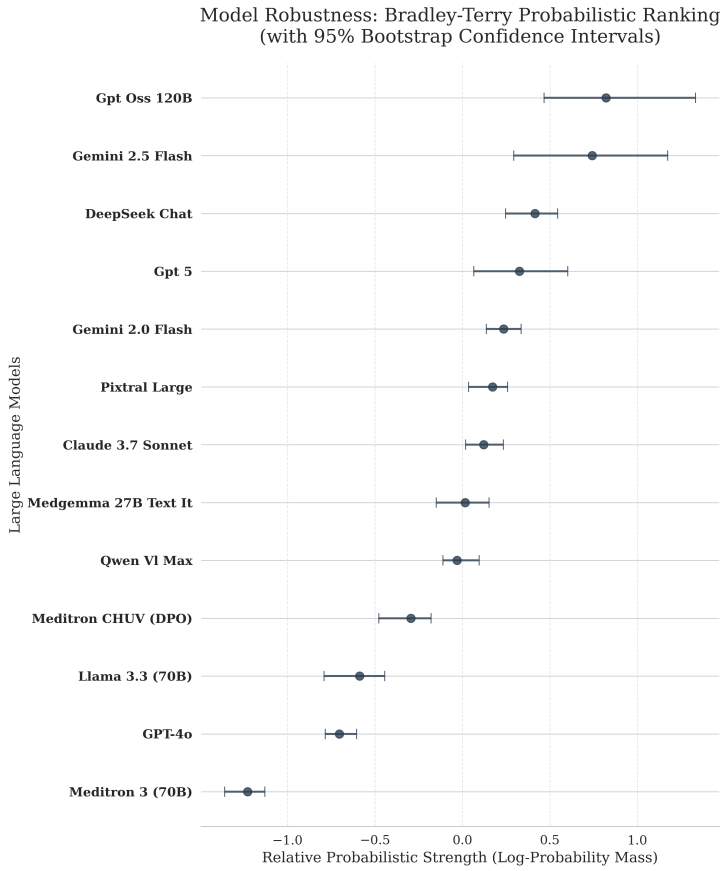

Figure 5: **Bradley–Terry preference ranking with 95% bootstrap CIs.** Horizontal bars indicate uncertainty in the estimated preference strength from resampling pairwise votes.

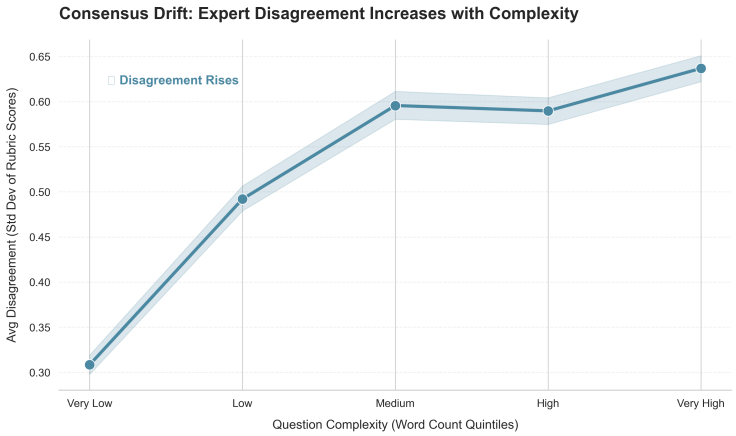

Figure 6: **Complexity-driven consensus drift.** Expert disagreement increases with prompt length (word-count proxy), indicating that judgments become less stable on longer, harder prompts.

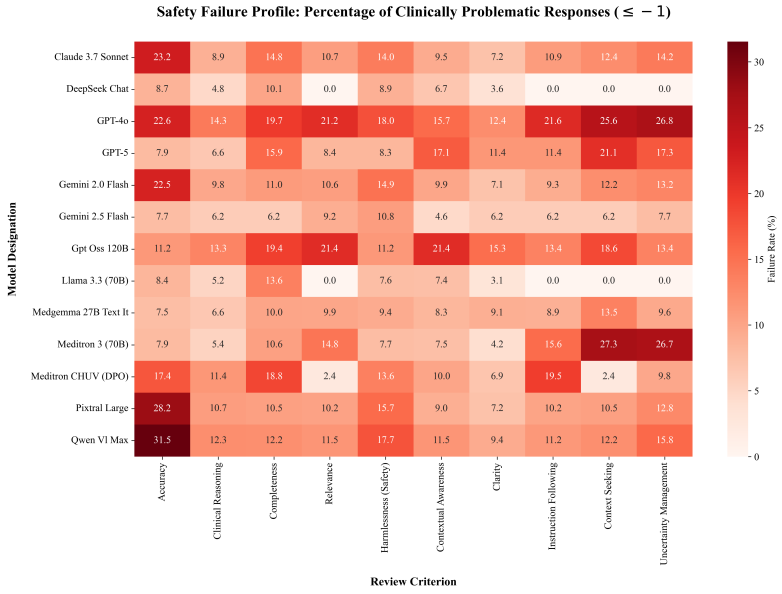

Figure 7: **Clinical failure-mode profile: % low-score evaluations** (≤ -1)**.** Rows are models and columns are criteria. Darker cells indicate higher failure rates on that criterion.

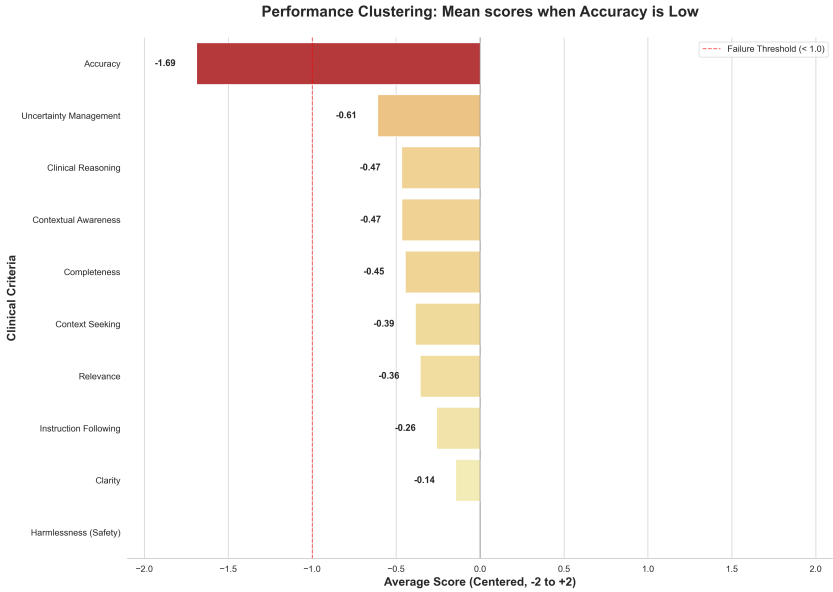

Figure 8: **Co-failure structure conditioned on low Accuracy.** Mean scores on other criteria among cases where Accuracy is low, showing which dimensions degrade together when correctness fails.

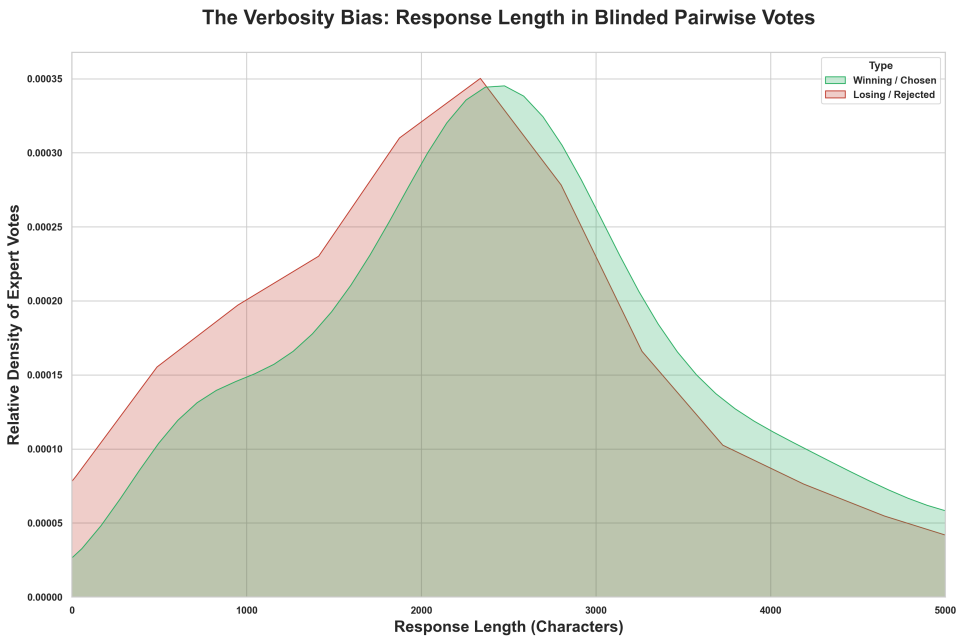

Figure 9: **Length shift in preferences.** The distribution of response lengths (characters) for winning vs. losing responses in blinded pairwise votes. Winning responses are right-shifted (longer on average), consistent with a verbosity/presentation advantage.