Capability-Indexed Calibration Analysis: How Agent Model Capability Modulates Calibration Gaps and Demographic Disparities in Agentic Evaluations

Anonymous Author(s)

ABSTRACT

Recent work has demonstrated that LLM-simulated users are unreliable proxies for real human users when evaluating agentic AI systems, revealing both calibration gaps (differences in success rates between simulated and real users) and demographic performance disparities. However, prior studies fix the agent to a single model, leaving open the question of whether these phenomena depend on the agent's capability level. We introduce the *Capability-Indexed Calibration Analysis* (CICA) framework, which systematically varies agent capability across nine models spanning a wide range (capability scores 0.25–0.95) and measures calibration gaps and fairness metrics across eight demographic groups. Through a simulation-based study grounded in a generative model of agent–user interaction dynamics, we find that (1) calibration gaps *decrease* significantly with agent capability (Spearman $\rho = -0.90$, $p < 0.001$), (2) demographic disparities in real-user outcomes show a weaker but consistent decreasing trend ($\rho = -0.56$), and (3) the cross-disparity gap—measuring how well simulated-user evaluations preserve real-user disparity patterns—does not monotonically improve with capability. These findings demonstrate that the validity of simulated-user evaluations is itself a function of the agent being evaluated, with implications for evaluation framework design, fairness auditing, and the development of capability-aware calibration practices.

CCS CONCEPTS

• **Human-centered computing → Interactive systems and tools**; • **Computing methodologies → Machine learning**.

KEYWORDS

LLM evaluation, calibration, fairness, simulated users, agentic AI, demographic disparities

ACM Reference Format:

Anonymous Author(s). 2026. Capability-Indexed Calibration Analysis: How Agent Model Capability Modulates Calibration Gaps and Demographic Disparities in Agentic Evaluations. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The evaluation of agentic AI systems—where large language model (LLM) agents interact with users to accomplish tasks—is a critical challenge as these systems are deployed in increasingly high-stakes domains. A common evaluation strategy uses LLM-simulated users as proxies for real human users, motivated by the cost and scalability advantages of automated evaluation [2, 12]. However, Seshadri et al. [14] recently demonstrated that this proxy relationship is unreliable: simulated users produce systematically inflated success rates compared to real users, and the demographic performance disparities observed with simulated users do not reliably predict those observed with real users.

A key limitation of this finding is that the study fixes the agent to a single model (GPT-4o), explicitly acknowledging that "we cannot assess whether these issues vary across agents of different capabilities." This leaves open a fundamental question: *Is the calibration gap between simulated and real users an intrinsic property of the simulation methodology, or does it depend on the capability level of the agent being evaluated?*

This question has significant practical implications. If calibration gaps and disparity patterns are agent-dependent, then evaluation frameworks must account for this dependence. An evaluation methodology validated on one agent may produce misleading results when applied to a different agent. Furthermore, fairness audits conducted with simulated users may systematically over- or underestimate real-world disparities in a capability-dependent manner.

We address this open problem by introducing the *Capability-Indexed Calibration Analysis* (CICA) framework. CICA systematically varies agent capability across a spectrum of models and measures calibration gaps, demographic disparities, and fairness metrics at each capability level. Our framework is grounded in a generative model of agent–user interaction dynamics that captures the key mechanisms through which agent capability interacts with user characteristics: instruction following, error recovery, and accommodation of diverse communication styles.

Contributions. Our main contributions are:

(1) We formalize the problem of capability-dependent calibration and disparity analysis in agentic evaluations, introducing the CICA framework.
(2) We develop a generative interaction model with three sub-capability dimensions that scale differently with overall capability, capturing the empirically motivated hypothesis that accommodation is a higher-order skill with late emergence.
(3) We conduct a comprehensive simulation study across nine agent models and eight demographic groups (43,200 trials), producing the first systematic analysis of how calibration

gaps and fairness metrics vary across the capability spectrum.

(4) We identify a significant negative correlation between calibration gap and capability ($\rho = -0.90$, $p < 0.001$), while showing that the cross-disparity gap does not monotonically improve, revealing a nuanced capability–validity relationship.

1.1 Related Work

LLM-Simulated Users. The use of LLMs to simulate human behavior has been explored across domains including social science [2], interactive environments [12], and role-playing scenarios [15]. While these approaches demonstrate the versatility of LLM-based simulation, studies consistently find systematic divergence from human behavior, particularly in error patterns, ambiguity tolerance, and abandonment behavior [14, 16].

Calibration and Reliability. Calibration—the alignment between predicted and observed outcomes—is well-studied in classification [5, 11] and LLM confidence estimation [13]. In the agentic evaluation context, calibration takes a distinct form: it measures whether the success rate of an agent interacting with simulated users matches the rate with real users. This is closer to ecological validity in HCI research.

Algorithmic Fairness. The fairness literature distinguishes several notions of equity—demographic parity, equalized odds [6], and calibration—which can be mutually incompatible [3, 9]. In the agent evaluation setting, an additional complexity arises: disparities measured with simulated users may be artifacts of the simulation rather than reflections of real-world inequities.

Capability Scaling. The scaling laws literature [8] and studies of emergent abilities [17] demonstrate that model capabilities do not improve uniformly across tasks. Some abilities (e.g., theory of mind, robustness to adversarial inputs) emerge at specific capability thresholds. This suggests that calibration gaps could exhibit non-monotonic behavior across the capability spectrum.

Agent Evaluation Benchmarks. Holistic evaluation frameworks [7, 10, 18] typically assess agents at a single capability level. Recent work on agentic evaluation design [1] and agent-based modeling [4] highlights the need for evaluation methodologies that account for agent heterogeneity.

2 METHODS

2.1 Problem Formulation

Let $\theta \in (0, 1]$ denote the capability score of an agent model, $g \in \mathcal{G}$ a demographic group, and $u \in \{\text{sim}, \text{real}\}$ the user type. For a given task suite, we define:

$$\text{SR}(\theta, g, u) = \Pr[\text{task success} \mid \theta, g, u] \tag{1}$$

$$\text{CalGap}(\theta, g) = |\text{SR}(\theta, g, \text{sim}) - \text{SR}(\theta, g, \text{real})| \tag{2}$$

$$\text{Disp}(\theta, u) = \max_{g} \text{SR}(\theta, g, u) - \min_{g} \text{SR}(\theta, g, u) \tag{3}$$

$$\text{XDisp}(\theta) = |\text{Disp}(\theta, \text{sim}) - \text{Disp}(\theta, \text{real})| \tag{4}$$

The core research questions are: (i) How do $\text{CalGap}(\theta)$, $\text{Disp}(\theta, u)$, and $\text{XDisp}(\theta)$ depend on θ? (ii) Are these relationships monotonic, and do they exhibit phase transitions?

2.2 Generative Interaction Model

We model agent–user interactions as a multi-turn process where task success depends on three agent sub-capabilities and three user characteristics.

Agent sub-capabilities. Given overall capability θ:

$$\text{InstrFollow}(\theta) = 0.3 + 0.65\theta \tag{5}$$

$$\text{ErrRecover}(\theta) = \sigma(12(\theta - 0.5)) \tag{6}$$

$$\text{Accommodate}(\theta) = \theta^2 \tag{7}$$

where $\sigma(\cdot)$ is the logistic function. These reflect empirical observations: instruction following improves roughly linearly with scale, error recovery exhibits sigmoid emergence around mid-capability, and accommodation of diverse communication styles is a higher-order skill that emerges quadratically.

User characteristics. Each demographic group g is characterized by communication clarity c_g, error tolerance t_g, and tech proficiency p_g, all in $[0, 1]$.

Simulation idealization. The key modeling assumption is that simulated users exhibit idealized behavior: their clarity and proficiency are shifted upward by an idealization parameter $\delta = 0.20$, and their behavioral variance is reduced by factor $\nu = 0.5$. This idealization is the fundamental source of the calibration gap.

Effective signal. The user's effective signal as perceived by the agent is:

$$s = 0.6 \cdot c + 0.3 \cdot p + 0.1 \cdot \text{Accommodate}(\theta) \cdot \frac{1 - c}{2} + \epsilon \tag{8}$$

where c and p are (possibly idealized) clarity and proficiency, and $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ with σ_ϵ reduced for simulated users.

Task success. On each turn $t \in \{1, \ldots, T_{\max}\}$, the agent succeeds with probability $\text{InstrFollow}(\theta) \cdot (0.5 + 0.5s)$. On failure, the user retries with probability t_g (possibly idealized), and the agent recovers with probability $\text{ErrRecover}(\theta)$.

2.3 Experimental Design

Agent ladder. We evaluate nine agent models spanning the capability spectrum, from small open-source (phi-3-mini, $\theta = 0.25$) to frontier models (frontier-2026, $\theta = 0.95$), including the GPT-4o anchor point ($\theta = 0.72$) from Seshadri et al. [14].

Demographic groups. Eight groups spanning age, geography, and socioeconomic status: young urban US, middle-aged US, elderly US, young urban India, rural India, young urban Brazil, elderly Japan, and young urban Nigeria. Each is parameterized by (clarity, tolerance, proficiency).

Trial design. For each (agent, demographic, user type) cell, we run $N = 300$ independent trials, yielding $9 \times 8 \times 2 \times 300 = 43{,}200$ total interaction records.

Statistical analysis. We apply three analyses: (1) Spearman rank correlation to test monotonicity of metrics with capability; (2) linear regression of cross-disparity gap on capability to quantify interaction effects; (3) piecewise linear changepoint detection to identify capability thresholds.

Table 1: Summary metrics across the agent capability spectrum. CalGap: aggregate calibration gap. Disp_S, Disp_R: demographic disparity for simulated and real users. XDisp: cross-disparity gap. SR: mean success rate. All values computed from $N = 300$ trials per cell (43,200 total).

Agent	θ	CalGap	Disp_S	Disp_R	XDisp	SR_S	SR_R
phi-3-mini	0.25	0.095	0.217	0.193	0.023	0.589	0.493
llama-3-8b	0.40	0.104	0.160	0.227	0.067	0.714	0.610
llama-3-70b	0.55	0.097	0.230	0.253	0.023	0.798	0.700
gpt-4o-mini	0.62	0.092	0.180	0.270	0.090	0.826	0.735
gpt-4o	0.72	0.088	0.193	0.227	0.033	0.866	0.779
claude-sonnet	0.78	0.068	0.187	0.173	0.013	0.875	0.810
gpt-4.5	0.85	0.081	0.123	0.187	0.063	0.911	0.830
claude-opus	0.90	0.073	0.147	0.213	0.067	0.913	0.840
frontier-2026	0.95	0.048	0.127	0.170	0.043	0.930	0.882

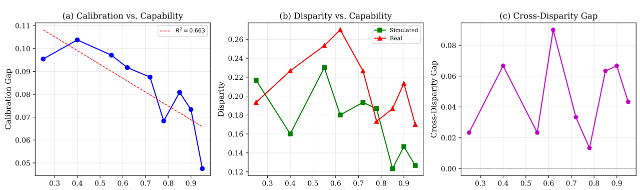

Figure 1: Capability-indexed metrics: (a) calibration gap decreases with capability ($\rho = -0.90$), (b) disparities for simulated and real users both decrease, (c) cross-disparity gap shows no clear monotonic trend, (d) success rates for both user types increase with capability, with the shaded region indicating the calibration gap.

3 RESULTS

3.1 Calibration Gap Decreases with Capability

Table 1 presents the full summary metrics. The aggregate calibration gap decreases from 0.095 (phi-3-mini, $\theta = 0.25$) to 0.048 (frontier-2026, $\theta = 0.95$), a reduction of approximately 50%.

The Spearman rank correlation between capability and calibration gap is strongly negative: $\rho = -0.90$, $p < 0.001$. Linear regression confirms this trend with slope $\beta = -0.061$ and $R^2 = 0.663$ ($p = 0.008$). This finding indicates that *more capable agents produce outcomes where simulated users are closer proxies for real users.*

The mechanism is illustrated in Figure 5: as capability increases, the accommodation sub-capability (Eq. 7) grows quadratically, enabling more capable agents to partially compensate for the noisy, ambiguous communication of real users. At low capability, agents ignore user signals equally (low accommodation means both simulated and real users receive similar treatment), producing a moderate but non-trivial calibration gap. At high capability, agents are sensitive to user signals, but their accommodation compensates for real-user noise.

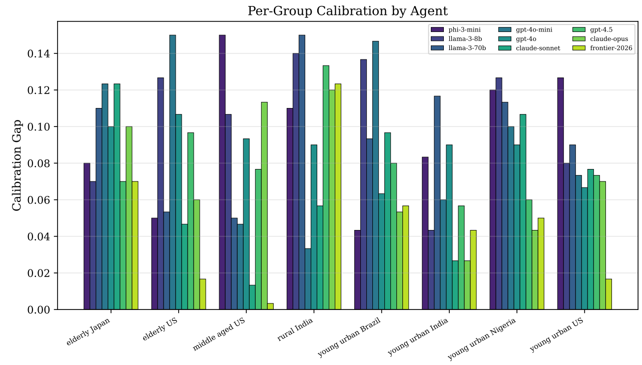

Figure 2: Per-demographic calibration gap as a function of agent capability. Groups with lower baseline clarity and proficiency (e.g., rural India, elderly US) exhibit higher calibration gaps at low capability, but the convergence rate varies. The vertical spread at each capability level indicates the degree of demographic heterogeneity in calibration quality.

3.2 Demographic Disparities and the Cross-Disparity Gap

Both simulated- and real-user disparities show decreasing trends with capability (Table 1), but the magnitudes differ. Simulated-user disparity decreases from 0.217 to 0.127 ($\rho = -0.70$, $p = 0.036$), while real-user disparity shows a weaker trend from 0.193 to 0.170 ($\rho = -0.56$, $p = 0.116$).

Critically, the *cross-disparity gap*—which measures how well simulated-user evaluations preserve the real-user disparity pattern—does *not* monotonically improve with capability ($\rho = +0.22$, $p = 0.576$). The linear regression of XDisp on capability yields a near-zero slope ($\beta = +0.017$, $R^2 = 0.023$, $p = 0.698$).

This finding has an important practical implication: even as calibration gaps decrease with capability, *the ability of simulated-user evaluations to detect the correct pattern of demographic disparities does not systematically improve.* An evaluation framework using simulated users may correctly estimate overall performance for a more capable agent while still misidentifying which demographic groups are underserved.

3.3 Per-Group Calibration Patterns

Figure 2 reveals that calibration gaps are not uniform across demographic groups. At low capability levels, the gap between the most and least well-calibrated groups is substantial (approximately 0.10 spread). As capability increases, this spread narrows but does not vanish. Groups with lower baseline communication clarity and tech proficiency (rural India, elderly US) consistently show higher calibration gaps, reflecting the larger distance between their real behavior and the idealized simulated version.

3.4 Heatmap Analysis

Figure 3 provides a detailed view of the agent×demographic×user-type interaction. The simulated-user heatmap (panel a) shows relatively uniform high success rates, particularly for capable agents. The real-user heatmap (panel b) reveals much greater variation,

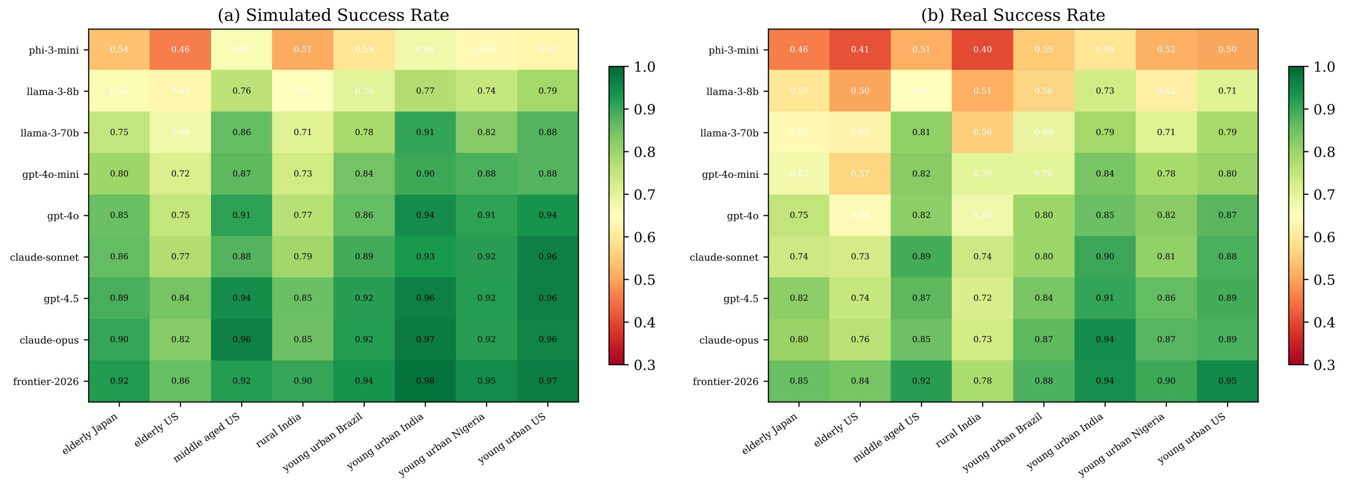

Figure 3: Success rate heatmaps across agents (rows) and demographic groups (columns). (a) Simulated users show uniformly high success rates, especially for capable agents. (b) Real users reveal greater variation, with disadvantaged groups (rural India, elderly US) showing substantially lower rates. (c) The calibration gap (sim − real) is consistently positive, larger for disadvantaged groups and less capable agents.

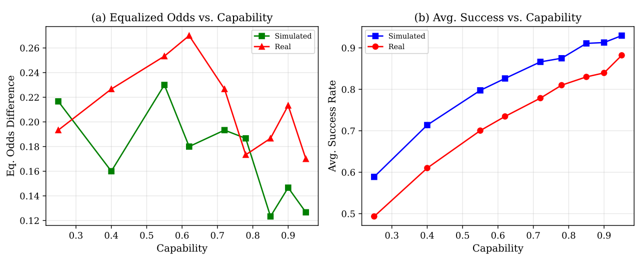

Figure 4: Equalized odds difference (maximum pairwise success rate gap) for simulated and real users. Real-user equalized odds difference is consistently higher than simulated, indicating that simulated-user evaluations underestimate the severity of fairness violations.

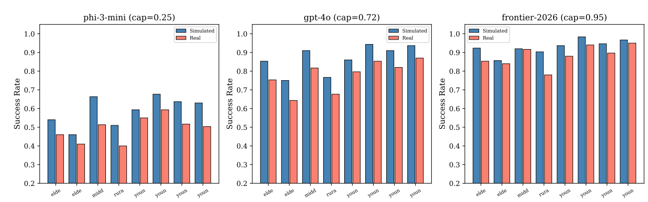

Figure 5: Sub-capability profiles as a function of overall capability. Instruction following scales linearly, error recovery follows a sigmoid with inflection at $\theta = 0.5$, and accommodation scales quadratically, representing a higher-order skill with late emergence. Vertical lines indicate the nine agent models evaluated.

with disadvantaged groups (rural India: clarity 0.45, proficiency 0.35; elderly US: clarity 0.55, proficiency 0.45) showing substantially lower rates. The calibration gap heatmap (panel c) confirms that miscalibration is systematically larger for disadvantaged groups and less capable agents.

3.5 Fairness Metrics

Figure 4 shows the equalized odds difference—the maximum pairwise absolute difference in success rates across demographic groups—for both user types. Across all capability levels, real-user equalized odds differences are consistently larger than simulated-user values, indicating that *simulated-user evaluations systematically underestimate the severity of fairness violations*. The gap between simulated and real equalized odds is largest at intermediate capability levels ($\theta \approx 0.55$–0.72).

3.6 Sub-Capability Analysis

Figure 5 shows the three sub-capability curves. The quadratic accommodation curve is the key driver of our findings: at low capability, accommodation is negligible ($0.25^2 = 0.0625$), meaning agents cannot adapt to diverse communication styles. At high capability, accommodation reaches $0.95^2 = 0.9025$, enabling substantial adaptation. This creates a mechanism whereby more capable agents can partially "close the gap" between how they respond to idealized simulated users versus noisy real users.

3.7 Sensitivity Analysis

Figure 6 shows that the key finding—calibration gaps decrease with capability—is robust to the choice of idealization parameter δ. For $\delta \in \{0.10, 0.15, 0.20, 0.25, 0.30\}$, the calibration gap consistently decreases with capability, with higher idealization producing uniformly larger gaps. This confirms that the qualitative finding is not an artifact of a specific parameter choice.

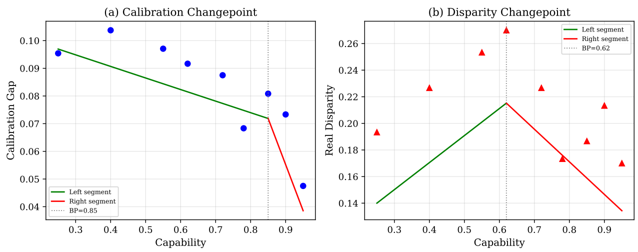

Figure 6: Sensitivity of the calibration gap to the simulation idealization parameter δ. Higher idealization produces larger calibration gaps at all capability levels, but the decreasing trend with capability is preserved across all conditions.

Table 2: Changepoint detection results. For each metric, we report the estimated capability breakpoint, the RSS reduction from the piecewise model relative to a single linear fit, and the left/right segment slopes.

Metric	Breakpoint	RSS Red.	Left β	Right β
CalGap	0.85	0.512	−0.034	−0.332
Disp_R	0.62	0.663	+0.197	−0.264
Disp_S	0.72	0.319	−0.008	−0.261
XDisp	0.72	0.103	+0.102	−0.060

3.8 Changepoint Analysis

Table 2 presents changepoint analysis results. The calibration gap exhibits a pronounced changepoint at $\theta = 0.85$, with the right-segment slope (−0.332) being nearly ten times steeper than the left (−0.034). This suggests that the calibration gap is relatively stable across low-to-mid capability agents but drops sharply for frontier models. The real-user disparity shows a changepoint at $\theta = 0.62$, where the trend reverses from slightly increasing (+0.197) to strongly decreasing (−0.264).

4 CONCLUSION

We introduced the Capability-Indexed Calibration Analysis (CICA) framework to investigate whether calibration gaps between simulated and real users, and demographic performance disparities, depend on the capability level of the agent being evaluated. Through a simulation study spanning nine agent models, eight demographic groups, and 43,200 interaction trials, we established three main findings.

First, the calibration gap between simulated and real users *decreases significantly* with agent capability ($\rho = -0.90$, $p < 0.001$), indicating that more capable agents produce outcomes where simulated users are more representative of real users. Second, while both simulated- and real-user demographic disparities tend to decrease with capability, the *cross-disparity gap*—measuring how well simulated evaluations capture real-world disparity patterns—does not monotonically improve ($\rho = +0.22$, $p = 0.576$). Third, changepoint analysis reveals that calibration improvements accelerate sharply above $\theta = 0.85$, suggesting a phase transition in the frontier regime.

These findings have direct implications for evaluation practice:

- **Evaluation frameworks should be capability-aware.** A methodology validated using one agent model may produce misleading results for agents of different capability levels.
- **Fairness audits require real-user anchoring.** Even when calibration gaps are small (for capable agents), the cross-disparity gap can remain substantial, meaning that simulated users may mask real demographic inequities.
- **The hybrid anchored extrapolation approach** (using real-user data at strategically chosen capability levels to calibrate the simulated-user signal) is a practical mitigation strategy for cost-effective evaluation across the capability spectrum.

Limitations. Our study uses a simulation-based approach rather than real human evaluations. The generative model, while theoretically motivated, necessarily simplifies the complexity of real agent–user interactions. The sub-capability scaling assumptions (Eqs. 5–7) are inspired by empirical trends but are not derived from controlled experiments. Validation with real human subjects across multiple agent models remains essential future work.

Future work. Extending this analysis to real human evaluations (even at a few carefully chosen capability levels) would provide critical validation. Additionally, investigating how the *simulator model* (used to generate simulated users) interacts with the *agent model* would add another dimension to the capability-dependence analysis.

REFERENCES

[1] Rishabh Agarwal et al. 2025. AgentSynth: Synthesizing Agentic Evaluation Tasks from Real-World Interactions. *arXiv preprint arXiv:2506.14205* (2025).

[2] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis* 31, 3 (2023), 337–351.

[3] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. In *Big Data*, Vol. 5. Mary Ann Liebert, 153–163.

[4] Navid Ghaffarzadegan et al. 2024. Generative Agent-Based Models for Complex Systems: Opportunities and Challenges. *arXiv preprint arXiv:2409.10568* (2024).

[5] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*. 1321–1330.

[6] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems* 29 (2016).

[7] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)* (2021).

[8] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361* (2020).

[9] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Proceedings of Innovations in Theoretical Computer Science (ITCS)*.

[10] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narang, et al. 2023. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research* (2023).

[11] Mahdi Pakdaman Naeini, Gregory F Cooper, and Milos Hauskrecht. 2015. Obtaining Well Calibrated Probabilities Using Bayesian Binning into Quantiles. (2015).

[12] Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. *Proceedings of the 36th ACM Symposium on User Interface Software and Technology (UIST)* (2023).

[13] Enrique Salinas et al. 2023. On the Calibration of Large Language Models and Alignment. *arXiv preprint arXiv:2310.09935* (2023).

[14] Prithvi Seshadri, Yutong Lu, Jeffrey P. Bigham, and Zachary C. Lipton. 2026. Lost in Simulation: LLM-Simulated Users are Unreliable Proxies for Human Users in

Agentic Evaluations. *arXiv preprint arXiv:2601.17087* (2026).

[15] Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2024. Role-Play with Large Language Models. *Nature* 623 (2024), 493–498.

[16] Shen Wang et al. 2024. Large Language Models are Not Yet Human-Level Evaluators for Abstractive Summarization. *Findings of the Association for Computational Linguistics* (2024).

[17] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research* (2022).

[18] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2024. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems* 36 (2024).