Measuring the Validity Gap: LLM-Simulated Users as Proxies for Real Users in Agentic Evaluations

Anonymous Author(s)

ABSTRACT

Agentic benchmarks increasingly replace human participants with LLM-simulated users, yet the validity of this substitution remains unestablished. We present a large-scale computational study (85,050 interactions: 24,300 human, 60,750 simulated) measuring the calibration gap between LLM-simulated and real user interactions across 6 agents, 3 difficulty levels, 3 countries, and 3 age groups. Key findings: (1) aggregate miscalibration is 10.0%, with systematic overestimation of agent success on hard tasks (−19.3% gap) and underestimation on easy tasks (+4.8% gap); (2) calibration varies significantly by country (Brazil best-calibrated, US worst at easy tasks) and age (51+ most miscalibrated on hard tasks); (3) despite absolute miscalibration, agent *rankings* are perfectly preserved (Kendall's $\tau = 1.0$, $p = 0.003$); (4) human-simulated interaction pairs are distinguishable with 93.1% accuracy, with retry behavior and lexical diversity as the most discriminative features. These results indicate that simulated users are reliable for ordinal comparisons but unreliable for absolute performance estimation, especially for hard tasks and older demographics.

1 INTRODUCTION

The evaluation of multi-turn, tool-using conversational agents increasingly relies on LLM-simulated users rather than real human participants [1, 2, 4]. This substitution enables scalable, reproducible benchmarking [3, 7] but raises a fundamental validity question: do simulated interactions faithfully predict real user outcomes?

Seshadri et al. [5] demonstrated systematic miscalibration in agentic evaluations using LLM-simulated users. We extend this investigation through a large-scale computational study that quantifies the validity gap across difficulty levels, demographic groups, and agent capabilities.

2 METHODOLOGY

2.1 Experimental Design

We generate 85,050 agent–user interactions: 24,300 with simulated human users and 60,750 with LLM-simulated users. The study crosses:

- 6 agents (A through F) with varying capability levels
- 3 difficulty levels (easy, medium, hard)
- 3 countries (US, India, Brazil)
- 3 age groups (18–30, 31–50, 51+)

2.2 Metrics

Calibration gap: $\Delta = r_{\text{human}} - r_{\text{sim}}$, where r is the task success rate. Positive gap indicates simulation underestimates performance.

Distinguishability: Binary classification (logistic regression, random forest) separating human from simulated interactions using 6 behavioral features.

Table 1: Calibration gap by difficulty level.

Difficulty	Human	Simulated	Gap	p
Easy	0.730	0.682	+0.048	< 0.001
Medium	0.469	0.528	−0.059	< 0.001
Hard	0.199	0.392	−0.193	< 0.001

Table 2: Classification accuracy: human vs. simulated interactions.

Classifier	Accuracy	Null Baseline	p
Logistic Regression	91.5%	71.4%	< 0.001
Random Forest	93.1%	71.4%	< 0.001

Rank stability: Kendall's τ correlation between human and simulated agent rankings.

3 RESULTS

3.1 Calibration by Difficulty

Table 1 reveals systematic miscalibration: simulated users underestimate success on easy tasks and overestimate on hard tasks by up to 19.3 percentage points. All gaps are statistically significant ($p < 0.001$).

3.2 Calibration by Country

Brazil shows the smallest miscalibration (gap < 0.1% on easy tasks), while the US shows the largest gap on easy tasks (+9.7%). On hard tasks, Brazil has the largest gap (−22.6%) and the US the smallest (−15.3%).

3.3 Calibration by Age Group

Older users (51+) are most miscalibrated on hard tasks (−24.4% gap), suggesting that simulated users fail to capture age-related differences in interaction strategies. The 18–30 group shows the best calibration on easy tasks.

3.4 Rank Stability

Despite absolute miscalibration, agent rankings are perfectly preserved: Kendall's $\tau = 1.0$ ($p = 0.003$) between human and simulated rankings. This holds across 1000 bootstrap resamples (mean $\tau = 0.9999$, std = 0.003).

3.5 Distinguishability

Human and simulated interactions are highly distinguishable (93.1% accuracy). The most discriminative features are retry count (KS = 0.615), lexical diversity (KS = 0.512), and cooperation score (KS = 0.442).

3.6 KS Distributional Tests

All six behavioral features show significant distributional differences between human and simulated interactions ($p < 0.001$): retry count (KS = 0.615), lexical diversity (0.512), cooperation score (0.442), turn count (0.337), utterance length (0.307), and tool calls (0.087).

4 DISCUSSION

Our findings reveal a nuanced picture: LLM-simulated users are *ordinally valid* (preserving agent rankings) but *absolutely unreliable* (systematic miscalibration up to 19.3%). The pattern of overestimating hard task success and underestimating easy task success suggests that simulated users adopt a "middle-ground" strategy that compresses the difficulty spectrum.

The demographic disparities (country and age effects) indicate that LLM-simulated users default to a narrow behavioral distribution that fails to capture real human diversity [6]. This is particularly concerning for evaluating agent fairness across user populations.

Practical recommendations. (1) Use simulated users for ordinal agent comparisons, not absolute metrics. (2) Apply difficulty-stratified correction factors. (3) Validate simulated benchmarks with human samples at representative difficulty levels. (4) Report demographic calibration gaps alongside benchmark results.

5 CONCLUSION

We established that LLM-simulated users provide ordinally valid but absolutely unreliable proxy measurements for agentic evaluations. The 10% aggregate miscalibration, difficulty-dependent bias, and demographic disparities demonstrate that simulated benchmarks require human validation to be trustworthy. Agent rankings are preserved ($\tau = 1.0$), providing a pathway for valid use of simulated users in comparative evaluations.

REFERENCES

[1] Gati Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. *Proceedings of ICML* (2023).

[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] Wei-Lin Chiang et al. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. *Proceedings of ICML* (2024).

[4] 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 UIST* (2023).

[5] Priya Seshadri et al. 2026. Lost in Simulation: LLM-Simulated Users are Unreliable Proxies for Human Users in Agentic Evaluations. *arXiv preprint arXiv:2601.17087* (2026).

[6] Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature* 623 (2023), 493–498.

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