Lost in Simulation: LLM-Simulated Users are Unreliable Proxies for Human Users in Agentic Evaluations

Anonymous ACL submission

Abstract

Agentic benchmarks increasingly rely on LLM-simulated users to scalably evaluate agent performance, yet the robustness, validity, and fairness of this approach remain unexamined. Through a user study with participants across the United States, India, Kenya, and Nigeria, we investigate whether LLM-simulated users serve as reliable proxies for real human users in evaluating agents on τ-Bench retail tasks. We find that user simulators lack robustness, with agent success rates varying up to 9 percentage points across different user LLMs. Furthermore, simulated users systematically miscalibrate performance, underestimating success on challenging tasks while overestimating moderately difficult ones. African American Vernacular English (AAVE) speakers experience consistently worse success rates and calibration errors than Standard American English (SAE) speakers, with disparities compounding significantly with age. We also find simulated users to be a differentially effective proxy for different populations, performing worst for AAVE and Indian English. Additionally, simulated users introduce conversational artifacts and surface different failure patterns than human users. These findings demonstrate that current evaluation practices risk misrepresenting agent capabilities across diverse user populations and may obscure real-world deployment challenges.

1 Introduction

AI agents designed to assist with everyday tasks such as travel reservations, order management, and appointment scheduling are becoming increasingly prevalent (Zhou et al., 2024; CNBC, 2025), but present significant challenges to effective evaluation. Agentic benchmarks have needed to evolve beyond static question-answering and other single-turn formats to capture the dynamic, multi-turn nature of real user interactions (Chang et al., 2025; Deshpande et al., 2025). Many recent works have proposed benchmarks that reflect this shift and instead measure sustained, context-aware interaction (Xie et al., 2024; Barres et al., 2025; Shao et al., 2025; Wang et al., 2025a; Xu et al., 2025; Yao et al., 2025). These benchmarks improve complexity and ecological validity over prior static evaluations by requiring agents to demonstrate a range of capabilities, including conversing naturally and coherently with users, adhering to policies, following instructions closely, and making appropriate tool calls over multiple turns. To facilitate automated and scalable evaluation, these benchmarks typically simulate conversations between an LLM agent and a user, where the "user" is an LLM (Ivey et al., 2024; He et al., 2025; Mehri et al., 2025).

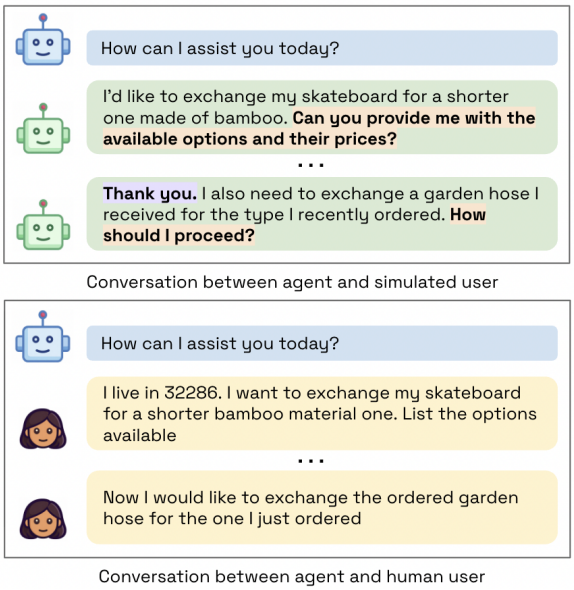


Figure 1: Conversational snippets between an agent and LLM-simulated (top) vs. human user (bottom) on the same task. Simulated users exhibit increased question-asking and politeness compared to human users.

While this approach reduces the cost and operational overhead of human evaluation, it raises critical questions about the **robustness**, **validity**, and **fairness** of user simulation. First, evaluations

typically rely on a single user simulation model, yet results may vary across different user LLMs (**robustness**). Second, without validation with actual users (Salaudeen et al., 2025), it remains unclear whether interactions between agents and LLM-simulated users accurately reflect and predict interactions between agents and real people (**validity**, Figure 1). If simulated users systematically differ from actual users in their interaction patterns (Yoon et al., 2024), benchmarks may provide a misleading picture of agent capabilities.

Third, these evaluations often treat users as a single, homogeneous group, overlooking variation in how people communicate and interact with AI systems (Haoyue and Cho, 2024; Liu et al., 2024; Bassignana et al., 2025). In practice, users differ widely in their communication styles, linguistic backgrounds, and cultural norms (Pawar et al., 2025; Qiu et al., 2025). For example, even in a simple retail assistance scenario, users might vary along dimensions such as formality, verbosity, and politeness norms—but it remains unclear how much this diversity meaningfully impacts agent performance and task success (Truong et al., 2025). Without validating simulated user interactions, we risk relying on synthetic users that poorly capture real user populations, especially underrepresented demographic groups. As a result, evaluations may misrepresent true agent performance due to *miscalibration* (i.e., simulated results do not reliably predict real user outcomes), particularly across different demographics, producing agents that serve some groups better than others (**fairness**).

Despite the widespread adoption of user simulation, prior work has overlooked validation against real human interactions in agentic benchmarks. In this paper, we address this gap by conducting a user study with participants from the United States, India, Kenya, and Nigeria to directly evaluate user simulation as a proxy for actual users. Specifically, we ask the following questions:

- **Robustness:** How consistent are agentic evaluations across different user simulation LLMs?
- **Validity:** Do LLM-simulated users serve as reliable proxies for real human users in agentic evaluations?
- **Fairness:** How does human-agent performance vary across different user groups, and does user simulation obscure differences?

Using τ-Bench retail tasks (Yao et al., 2025) as a case study, we find that user simulation lacks robustness, and systematically misestimates human user performance by underestimating success on the most challenging tasks while overestimating outcomes on moderately difficult scenarios. More critically, LLM-simulated users exhibit notable demographic biases and perform particularly poorly as proxies for African American English speakers, with disparities compounding with age. Simulated users also introduce artificial conversational artifacts such as heightened question-asking and politeness (Figure 1). Together, these findings call into question the validity of LLM-based user simulation as a stand-alone evaluation paradigm and underscore the need for more robust and fair approaches to agentic evaluation.

2 Related Work

User Simulation in Interactive Settings Dou et al. (2025) and Wang et al. (2025b) investigate user simulation in tasks such as math tutoring and daily planning, but primarily focus on conversational characteristics (e.g., politeness) and behavioral realism (e.g., Turing-style tests). Additionally, Dou et al. (2025) optimize alignment with human ratings using simulator user profiles. Lu et al. (2025) analyze simulation fidelity by measuring how accurately user simulators replicate human intermediate steps in real-world online shopping sessions, and find substantial deviations between simulated and actual user action sequences. None of these works examine the robustness of agentic evaluations across different user simulators, nor do they consider fairness implications across different user groups. Finally, while Zhu et al. (2025) study the outcome and task validity of agentic benchmark results, they do not consider the validity of user simulation.

Demographic Skews in NLP Datasets and Models Several previous works highlight that datasets and models exhibit systematic skews toward specific demographic perspectives. Research on annotator disagreement reveals that perceptions of safety, offensiveness, and toxicity meaningfully vary along demographic axes like race, gender, and political affiliation (Sap et al., 2022; Lee et al., 2023; Prabhakaran et al., 2024). These patterns of disagreement reflect broader issues of positionality—both Santy et al. (2023) and Lee et al. (2024) show that NLP datasets and models tend to align predominantly with Western, edu-

cated, and Anglosphere populations. Similarly, LLMs have been found to reflect the opinions of Western countries (Durmus et al., 2024; Cahyawijaya et al., 2025) as well as wealthy and liberal groups (Santurkar et al., 2023), and align more closely with White annotators than Black or Asian groups on subjective tasks like politeness (Sun et al., 2025). Attempts to broaden model inclusivity through sociodemographic prompting have shown mixed results, often failing to consistently improve alignment or relying on harmful stereotypes (Durmus et al., 2024; Sun et al., 2025). Overall, our work builds on this line of research by examining demographic differences in agentic settings.

3 Methodology

3.1 Benchmark Background

We use τ-Bench (Yao et al., 2025) as the testbed for our evaluations, since it is a well-known and widely-adopted benchmark for agentic tool use.[1] The benchmark is designed to capture how well AI agents perform in real-world, interactive customer service scenarios. Each task involves collaboration between an agent and a simulated user: the user receives task instructions with specific objectives that guide their conversation with the agent, while the agent must use tool calling to interact with realistic databases, adhere to domain-specific policies, and gather necessary information from the user.

The task is considered successful if (1) the final database state is identical to the unique ground truth outcome (i.e., the sequence of required actions) and (2) the agent's responses convey all necessary information requested in the task instructions, which is evaluated automatically using substring matching against ground truth annotations. In total, τ-Bench contains 115 retail tasks (e.g., modifying pending orders or returning delivered orders) and 50 airline tasks (e.g., booking, modifying, or canceling reservations).

3.2 Benchmark Adaptation

We focus on τ-Bench retail tasks to enable more systematic coverage within a single domain. We apply preprocessing steps to ensure that neither simulated nor human users are influenced by identity and behavioral cues in the instructions when completing tasks (see Appendix A.4).

Given the large number of retail tasks, we sample a subset based on difficulty to ensure balanced coverage across task complexities. To compute a model-based notion of difficulty, we run the benchmark 5 times using GPT-4o as both the agent and user LLMs and measure the task success rate over 5 runs (i.e., the percentage of times a given task is completed successfully). Since we run each task exactly 5 times, success rates naturally fall into six discrete levels (0/5, 1/5, 2/5, 3/5, 4/5, 5/5 = 0%, 20%, 40%, 60%, 80%, 100%) We use GPT-4o because it is used in the τ-Bench paper and does not exhibit contamination issues that newer models face.[2] We then select 3 tasks for each difficulty level (18 total) to balance response coverage per task with breadth across difficulty levels.

3.3 User Study

To assess whether LLM-simulated users serve as effective proxies for real and diverse users, we conduct a user study with participants from the United States, India, Kenya, and Nigeria. Since τ-Bench tasks are in English, we select countries with large English-speaking populations that also provide geographical and linguistic diversity.[3] We recruit participants primarily through Prolific, except in Nigeria, where we use snowball sampling due to limited platform availability. All participants self-identify as being at least proficient in English.

Each participant completes 4 randomly assigned tasks from our pool of 18, presented in randomized order: 2 from higher difficulty levels (0-40% success rate) and 2 from lower difficulty levels (60-100% success rate). The agent model remains GPT-4o throughout all interactions. Participants are shown task instructions (see Appendix A.5) and asked to complete all mentioned requests by interacting with the agent through a Streamlit chat interface. After finishing each conversation, they click an "End Conversation" button to proceed to the next task. In total, the expected time to complete all 4 tasks is 35-40 minutes.

Participants provide demographic information including education level, AI familiarity, and frequency of AI tool usage. For US participants, we screen for White Standard American English (SAE) speakers and Black African American Vernacular

[1] τ-Bench Leaderboard: https://taubench.com/#leaderboard

[2] GPT-4o (May 2024) predates τ-Bench (June 2024), avoiding contamination issues present in newer models. We considered Sonnet 3.5 (also used in the original paper) but it was retired in October 2025.

[3] https://en.wikipedia.org/wiki/List_of_countries_by_English-speaking_population

English (AAVE) speakers based on self-reported race and dialect, as these groups are commonly studied in AI fairness research (Sap et al., 2019; Groenwold et al., 2020). We also stratify US participants by age (18-34, 35-54, and 55+) to capture potential differences in technology experience (Pew Research Center, 2025). Due to participant availability constraints, we only recruit from the 18-34 age group for other countries. In total, we recruit ∼40 participants per age × country/dialect group, resulting in roughly 360 participants.

3.4 Evaluation Metrics

Expected Calibration Error (ECE) We adapt *Expected Calibration Error (ECE)*, commonly used for assessing confidence calibration in probabilistic classifiers, to quantify how well LLM-simulated users serve as proxies for human users. While traditional ECE evaluates whether a model's predicted probabilities match true observed outcomes, we use an ECE-style formulation to measure whether simulated user success rates align with empirical human success rates across task difficulty levels. Therefore, our metric measures the discrepancy between two empirical performance distributions (LLM-simulated users and human users). Just as traditional ECE measures whether predicted confidences are calibrated to observed outcomes (Guo et al., 2017), ours measures whether simulated user success is calibrated to human user success.

Let $s_i^{(\text{LLM})}$ denote the simulated user success rate and $s_i^{(\text{Human})}$ denote the empirical human success rate at difficulty level i. Let w_i denote the proportion of human task completions at level i, with $\sum_i w_i = 1$. We define:

$$\text{ECE}_{\text{Human–LLM}} = \sum_{i=1}^{M} w_i \left| s_i^{(\text{Human})} - s_i^{(\text{LLM})} \right|$$

Overall, this metric captures the weighted average absolute deviation between human and simulated success rates across M difficulty levels, with lower values indicating better calibration, or closer alignment between simulated and human user performance. If simulated users are perfectly calibrated to human users, then $ECE_{\text{Human–LLM}} = 0$.

As a reminder, we partition tasks into six difficulty levels based on the model's success rate across five runs. We set w_i proportional to the number of human task completions at level i. We multiply $ECE_{\text{Human–LLM}}$ by 100 to report values as percentages.

Success Rate To recap, a task completion in τ-Bench is considered successful ($reward = 1$) if and only if the agent correctly executes all required actions and its responses convey all information specified in the instructions. We define success rate as the percentage of tasks that are successfully completed. We apply the same automated evaluation procedure used in τ-Bench to both human and simulated user interactions. Success rate is computed as a weighted average across difficulty levels, weighted by task completions.

4 Results

4.1 Robustness

We first evaluate the robustness of user simulation by examining how success rates vary with different user simulation models, which is largely overlooked in current evaluations.[4] We find that changing just the user LLM while keeping the agent LLM fixed (GPT-4o) can provide different depictions of agent performance. As shown in Table 1, GPT-4o, Sonnet 3.7, and Kimi-K2-Thinking show overlapping intervals, clustering around 67-71% success rates. However, there is nearly a 9 percentage point difference in success rates between Sonnet 3.7 and Sonnet 4.5 as the user model.[5] While Sonnet 3.7 is generally considered a stronger model than GPT-4o,[6] using GPT-4o as the user model yields a slightly higher success rate (67.8 vs. 67.0) and lower standard deviation (1.2 vs. 3.3), suggesting that closer alignment between agent and user LLMs may lead to more stable evaluation outcomes. Overall, the sensitivity of agent performance to user model choice raises concerns about the reliability of single-model user simulations and underscores the need for reporting results across multiple user models to establish robustness.

4.2 Validity

We now examine the validity of user simulation and first consider LLM-simulated users as a proxy for human users in the United States. Since prior work has shown that LLMs exhibit a Western, Anglocentric bias (Tao et al., 2024; Wang et al., 2024; Agarwal et al., 2025), LLM-simulated users might be more representative or better calibrated to users

[4] For robustness, we vary the user LLM and evaluate on all retail tasks; for subsequent validity and fairness analyses, we evaluate human users on a difficulty-balanced subset of tasks.

[5] It is difficult to disentangle increased model capability from potential data contamination.

[6] https://lmarena.ai/leaderboard/text

User Model	Success Rate (%)
GPT-4o	67.8 ± 1.2
Sonnet 3.7	67.0 ± 3.3
Sonnet 4.5	75.9 ± 3.5
Kimi-K2-Thinking	71.3 ± 1.9

Table 1: τ-Bench success rate (%) on retail tasks ($n = 115$) for different user models. The agent model remains GPT-4o. The average success rates and standard deviations are shown across 3 runs.

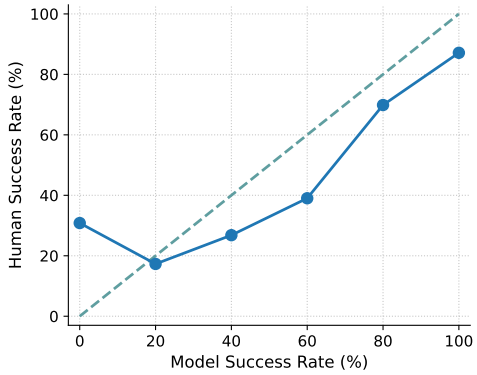


Figure 2: Human user task success vs. LLM-simulated user task success for United States participants, with an $ECE_{\text{Human–LLM}}$ of 15.1.

in the US than to users in non-Western countries. Therefore, we assess whether LLM-simulated users are well-calibrated to human users in the setting where we expect the strongest alignment.

We find that agents achieve a 45.2% success rate with US participants and an $ECE_{\text{Human–LLM}}$ of 15.1, indicating substantial miscalibration even in this setting. Calibration errors are not uniform across task difficulty. As shown in Figure 2, the calibration gap is most pronounced for the 1st (0%) and 4th (60%) difficulty bins with $ECE_{\text{Human–LLM}} = 25.9$ across the two. These results indicate that simulated users underestimate performance on the hardest tasks (human success rate is 30.8%) while overestimating it on moderate tasks (human success rate is 39.0%).

4.3 Fairness

To assess whether these findings are consistent across different user groups, we now partition our US results by English dialect (Standard American English vs. African American Vernacular English) and age group (18-34, 35-54, and 55+), and expand our analysis to three non-Western countries with high English-speaking populations (India, Kenya, and Nigeria).

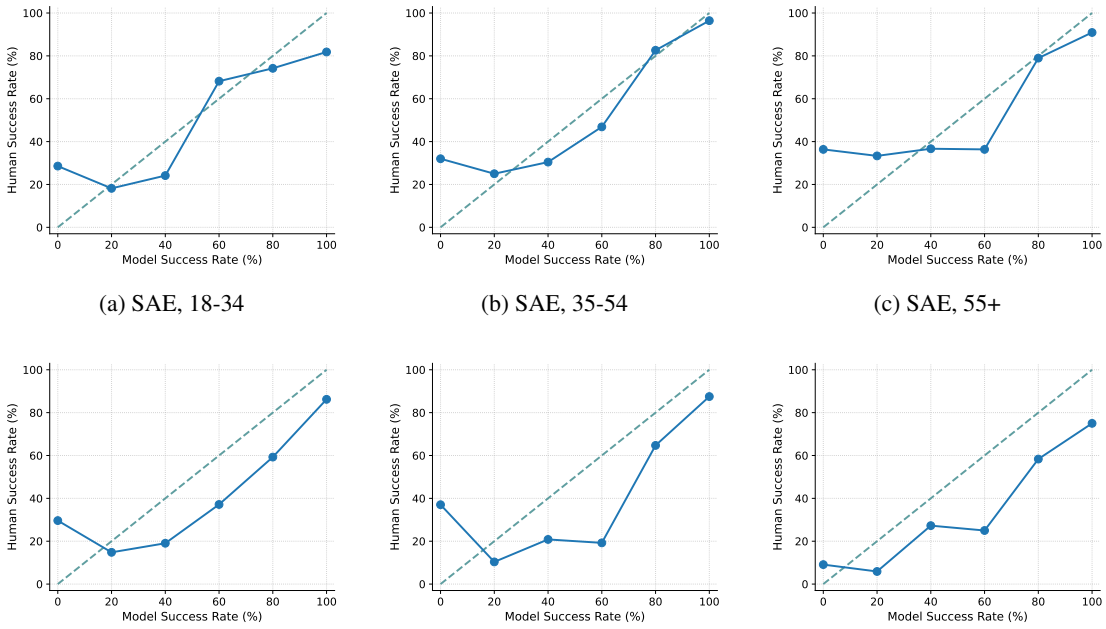
Age Group	ECE (↓)	Success Rate (↑)
SAE		
All	11.7	50.6
18–34	13.0	49.2
35–54	11.3	52.2
55+	14.5	52.1
AAVE		
All	20.3	39.4
18–34	18.9	41.0
35–54	21.6	39.9
55+	20.5	33.4

Table 2: Expected Calibration Error (ECE) and success rate (%) for Standard American English (SAE) and African American Vernacular English (AAVE) speaking participants, split by age group.

4.3.1 Dialect and Age (United States)

Starting with US participants, we previously saw that agents achieve a 45.2% success rate and an $ECE_{\text{Human–LLM}}$ of 15.1. When further breaking this down by dialect, we observe notable disparities in both performance and calibration, as shown in Table 2 and Figure 3. A Generalized Estimating Equations (GEE) model accounting for age, education, AI experience, AI usage, and task difficulty confirms a statistically significant dialect disparity ($\beta = 0.61$, $p < 0.001$). Agents exhibit a success rate of 50.6% with an $ECE_{\text{Human–LLM}}$ of 11.7 for SAE participants vs. a success rate of 39.4% with an $ECE_{\text{Human–LLM}}$ of 20.3 for AAVE participants. For AAVE participants, agents perform worse (11.2 percentage point decrease in success rate) and simulated users are more poorly calibrated (8.6 percentage point increase in ECE). In practice, such differences in performance and user simulation reliability could lead to disparities in the quality of retail assistance and the ease of interactions.

We observe contrasting age-related patterns in success rates across dialects. For SAE participants, success rates increase slightly with age ($\sim$3.0 percentage point increase from 18-34 to 55+ group), whereas for AAVE participants, success rates decrease with age (7.6 percentage point decrease from 18-34 to 55+ group). Notably, dialect disparities in agent performance grow larger with age: there is nearly a 12 percentage point decrease in agent performance between SAE and AAVE 35-54 groups and a 19 percentage point decrease in agent performance between SAE and AAVE 55+ groups (age-stratified GEE dialect effects: $\beta_{35-54} = 0.67$, $p = 0.01$; $\beta_{55+} = 1.24$, $p = 0.001$). The calibra-



(a) SAE, 18-34 (b) SAE, 35-54 (c) SAE, 55+

(d) AAVE, 18-34 (e) AAVE, 35-54 (f) AAVE, 55+

Figure 3: Human user task success vs. LLM-simulated user task success for SAE (top) and AAVE (bottom) speakers across different groups (18-34, 35-54, and 55+). The x-axis (LLM success rate) remains consistent across all panels.

tion gap is particularly pronounced for the 35-54 age group (10.3 $ECE_{\text{Human–LLM}}$ gap), where SAE primarily exhibits miscalibration for the 1st bin, while AAVE exhibits miscalibration across all bins (Figures 3b and 3e).

Note that $ECE_{\text{Human–LLM}}$ and success rate capture distinct aspects of performance. While *lower* $ECE_{\text{Human–LLM}}$ and *higher* success rate are both desirable, since lower $ECE_{\text{Human–LLM}}$ indicates that simulated users serve as reliable proxies for human users and higher success rates reflect stronger agent task performance, improvements in one do not necessarily imply improvements in the other. For example, SAE participants aged 18–34 exhibit both lower $ECE_{\text{Human–LLM}}$ and success rates, whereas SAE participants aged 55+ exhibit both higher $ECE_{\text{Human–LLM}}$ and success rates.

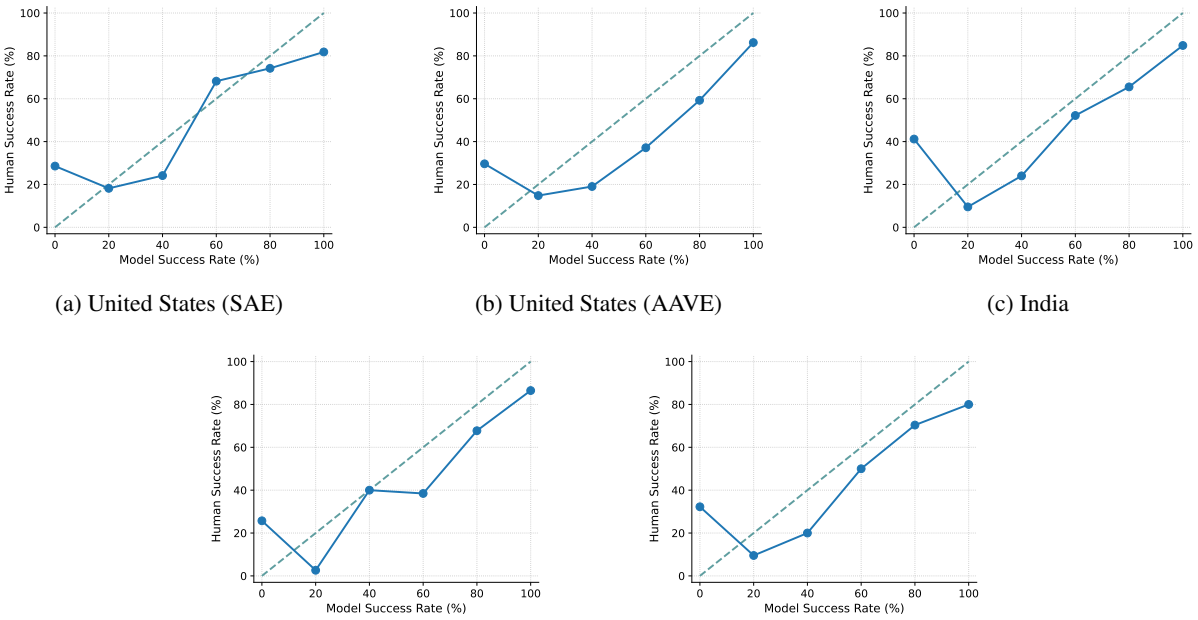
4.3.2 Countries

Due to participant availability constraints, we focus our cross-country analysis on the 18-34 age group. We find that differences in agent performance, shown in Table 3, are present but much less pronounced across countries (ranging from 41.0%-49.2%) than those observed by dialect and age within the US (Table 2). In particular, Kenyan and Nigerian participants experience similar success rates (43.5% and 43.7%). A GEE model confirms that cross-country differences are *not* statistically significant (all $p > 0.49$).

Group	ECE (↓)	Success Rate (↑)
SAE	13.0	49.2
AAVE	18.9	41.0
India	18.9	46.2
Kenya	15.6	43.5
Nigeria	17.6	43.7

Table 3: Expected Calibration Error (ECE) and success rate (%) across various dialects/countries for 18–34 age group users.

Simulated users are best calibrated to SAE participants ($ECE_{\text{Human–LLM}} = 13.0$) and worst calibrated to AAVE and Indian participants ($ECE_{\text{Human–LLM}} = 18.9$), suggesting simulated users are especially poor proxies for these groups. Across countries, we observe that simulated users tend to exhibit strongest calibration for fairly to moderately difficult tasks (20%-40% difficulty bins). However, they consistently overestimate performance for easy tasks (80%-100% difficulty bins), shown in Figure 4. As a result, evaluations relying on simulated users risk systematically underestimating difficulty agents face when deployed to diverse, global user populations.



(a) United States (SAE) (b) United States (AAVE) (c) India

(d) Kenya (e) Nigeria

Figure 4: Human user task success vs. LLM-simulated user task success for participants across dialects/countries, all in the 18–34 age group. Note: The x-axis (LLM success rate) remains the same for all groups.

4.4 Analyzing Interactions

4.4.1 Structure and Content

Interactions between agents and simulated users vs. human users follow similar structures and surface-level forms, including # of turns, # of actions, and # of words/turn. We observe some differences between simulated and human user interactions when considering conversational content (Table 6). Simulated user conversations include questions in 18.8% of user turns and 51.8% of agent turns, compared to 9.8% and 56.3%, respectively, for human users. Notably, Nigerian participants only ask questions in 4.3% of turns.

Differences are more pronounced when examining politeness indicators (e.g., please, thank you, apologize). Simulated user conversations include such indicators in 39.2% of user and 52.0% of agent turns, compared to 19.9% and 41.1%, respectively, for human users. For both question-asking and politeness indicators, differences in behavior are more pronounced between simulated and human users than among human users. We also show that targeted behavioral interventions, such as prompting simulated users to limit politeness, can alter calibration patterns and reduce gaps for highly miscalibrated groups (see Appendix A.6), indicating that prompting strategies can partially mitigate miscalibration issues.

4.4.2 Errors

We also analyze differences in (1) error types and (2) error attribution for simulated vs. human users to understand whether both user groups experience the same failure modes. Error types include argument errors, missing actions, extra actions, and output errors (see Table 4 for definitions).

Agents make argument errors (performing the correct action with incorrect arguments) at rates of 32.2% for simulated users compared to 23.2%–45.8% for human users, with AAVE participants experiencing the highest errors. Agents tend to less frequently omit actions (17.8% vs. 15.8%–25.3%) or include unnecessary ones (5.6% vs. 7.6%–13.3%) for simulated users compared to human users. However, output errors show the reverse pattern: agents either omit or include incorrect outputs more often for simulated users (31.4%) than for human users (12.2%–23.6%). These patterns suggest that agents exhibit different behavior when interacting with simulated vs. human users; they perform more complete and efficient action sequences but make more frequent output errors.

When comparing error attribution for simulated vs. human user conversations (Table 5), we observe clear differences in where task failures occur. In simulated conversations, agents are responsible for task failures substantially more often than for human conversations (48.9% vs. 24.5%). In contrast,

Group	Arg (%)	Miss (%)	Extra (%)	Output (%)
Simulated User				
–	32.2	17.8	5.6	31.4
Human User – US, SAE				
All	25.2	18.5	9.5	16.2
18–34	27.9	15.8	7.6	18.9
35–54	24.5	18.5	10.6	17.0
55+	23.2	21.3	10.3	12.2
Human User – US, AAVE				
All	39.0	23.0	10.7	15.5
18–34	36.8	25.3	13.3	15.5
35–54	37.8	20.4	8.7	16.4
55+	45.8	24.1	9.6	13.3
Human User – Non-US, 18–34				
India	33.3	18.2	7.9	20.4
Kenya	38.5	17.2	7.8	23.6
Nigeria	33.8	22.3	10.8	21.7

Table 4: Error breakdown (%) for simulated and human users, aggregated across tasks. Error types: **argument error** (action taken matches ground truth action but with different arguments; e.g., modifies order to the wrong address), **missing action error** (ground truth action is missing from actions taken), **extra action error** (actions taken go beyond ground truth actions), and **output error** (expected outputs are missing/incorrect). Each error type is measured as a binary indicator per task (e.g., a missing action error records whether **any** required action is missing). Note that error percentages do not sum to 100%; some tasks have no errors while others have multiple error types.

Source	Simulated	Human
Agent	48.9	24.5
User	40.0	62.2
Both	2.2	11.1
Other	8.9	2.2

Table 5: Error Attribution (%) for simulated vs. human user conversations with $reward = 0$, manually annotated by the authors ($n = 45$ per condition, matched on task difficulty).

users are the primary source of failure in human conversations (62.2% vs. 40%).

These differences indicate that simulated and human interactions surface distinct failure patterns. The higher user error rate in human interactions reflects ambiguity, misunderstandings, or partial compliance that humans naturally introduce. Conversely, the higher agent error rate in simulated conversations suggests simulated users appear to exhibit more precise instruction following or adapt more readily to agent responses, placing greater burden on agents to execute correctly. In practice, this divergence may lead to misdiagnoses of failures. Simulation-based evaluations may overemphasize agent execution errors, while obscuring challenges that arise when real users engage with agents in ways not reflected by simulated users.

5 Discussion and Conclusion

As AI agents become integrated into everyday tasks, ensuring equitable performance across diverse populations is essential. However, the "user" component of agentic evaluations, central to real-world interaction, has largely been overlooked. Our findings reveal fundamental limitations of LLM-simulated users, showing that current simulation practices misestimate agent performance for actual users and obscure demographic disparities, which may result in evaluations optimizing for objectives that diverge from real-world use.

Systems optimized for simulated users may appear robust in benchmarks while failing disproportionately for real users whose communication styles are underrepresented by simulation. For example, the behavioral artifacts we observe in simulated interactions—such as heightened politeness and question-asking—suggest that simulators reflect communication norms that may not generalize across diverse user groups. Moving forward, agentic benchmarks should assess robustness across multiple simulators, validate simulated outcomes against demographically diverse human data if possible, and transparently acknowledge the limitations of user simulation.

Limitations

While our study provides important insights into user simulation in agentic evaluations, it is useful to clarify its scope and outline directions for future work. Our evaluation is conducted entirely in English (replicating a limitation of popular agentic benchmarks, which are limited to English), which restricts our ability to assess how LLM-simulated users behave in multilingual settings. Since language influences user and agent behavior, capabilities, and interaction norms, it is important to verify the extent to which our findings hold for non-English contexts. In addition, our age-based analyses are limited to users in the United States due to recruitment constraints, leaving open the question of how performance and calibration disparities vary with age across other countries and cultural contexts.

We also evaluate agents in a single domain, focusing on retail customer service scenarios from τ-Bench. While these tasks are designed to capture multi-turn, task-oriented interactions with tool use, agent performance patterns and the quality of user simulation may differ in other domains such as healthcare, where interaction structure and task complexity vary. In such domains with longer-form interactions, individual style and cultural differences may be more apparent, potentially amplifying the effects we observe. Nevertheless, it would be important to empirically validate this expectation.

Finally, we focus on a single, fixed agent (GPT-4o). Holding the agent constant enables us to isolate the effects of user simulation on evaluation outcomes. However, we do not examine how calibration gaps or performance disparities vary across different agents, which is important for developing a more complete understanding of the robustness, validity, and fairness of user simulation. We discuss these points further in Appendix A.1.

Ethics Statement

All participants received standardized compensation that was not adjusted by country, ensuring consistent and fair payment regardless of geographic location. Participants were provided with an overview of the study procedures upfront and could withdraw from the study at any point without penalty. Participants provided informed consent by reading the study instructions and choosing to participate and complete the study. The study is classified as minimal risk, since it involves interaction with AI agents in simulated retail customer service scenarios and does not involve the collection of sensitive personal information.

Beyond the user study itself, our work has broader societal implications. Our results demonstrate that LLM-simulated users may not serve as reliable proxies for human users and can exhibit demographic biases. If simulated users are adopted as a standard practice for agent evaluation despite these limitations, there is a risk that AI systems could be deployed in ways that systematically underserve certain demographic groups. We hope this work encourages the research community to reflect on current evaluation practices and to develop agentic evaluation approaches that better represent diverse user populations.

References

Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2025. Ai suggestions homogenize writing toward western styles and diminish cultural nuances. CHI '25, New York, NY, USA. Association for Computing Machinery.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. 2025. τ^2-bench: Evaluating conversational agents in a dual-control environment. *Preprint*, arXiv:2506.07982.

Elisa Bassignana, Amanda Cercas Curry, and Dirk Hovy. 2025. The AI gap: How socioeconomic status affects language technology interactions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18647–18664, Vienna, Austria. Association for Computational Linguistics.

Samuel Cahyawijaya, Delong Chen, Yejin Bang, Leila Khalatbari, Bryan Wilie, Ziwei Ji, Etsuko Ishii, and Pascale Fung. 2025. High-dimension human value representation in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5303–5330, Albuquerque, New Mexico. Association for Computational Linguistics.

Serina Chang, Ashton Anderson, and Jake M. Hofman. 2025. ChatBench: From static benchmarks to human-AI evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26009–26038, Vienna, Austria. Association for Computational Linguistics.

CNBC. 2025. Ai travel agents planning future trip far beyond assistant status.

Kaustubh Deshpande, Ved Sirdeshmukh, Johannes Baptist Mols, Lifeng Jin, Ed-Yeremai Hernandez-Cardona, Dean Lee, Jeremy Kritz, Willow E. Primack, Summer Yue, and Chen Xing. 2025. MultiChallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18632–18702, Vienna, Austria. Association for Computational Linguistics.

Yao Dou, Michel Galley, Baolin Peng, Chris Kedzie, Weixin Cai, Alan Ritter, Chris Quirk, Wei Xu, and Jianfeng Gao. 2025. SimulatorArena: Are user simulators reliable proxies for multi-turn evaluation of AI assistants? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35200–35278, Suzhou, China. Association for Computational Linguistics.

Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020. Investigating African-American Vernacular English in transformer-based text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5877–5883, Online. Association for Computational Linguistics.

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 - Volume 70*, ICML'17, page 1321–1330. JMLR.org.

Luna Luan Haoyue and Hichang Cho. 2024. Factors influencing intention to engage in human–chatbot interaction: examining user perceptions and context culture orientation. 24(1):607–620.

Muyu He, Anand Kumar, Tsach Mackey, Meghana Rajeev, James Zou, and Nazneen Rajani. 2025. Impatient users confuse ai agents: High-fidelity simulations of human traits for testing agents. *Preprint*, arXiv:2510.04491.

Jonathan Ivey, Shivani Kumar, Jiayu Liu, Hua Shen, Sushrita Rakshit, Rohan Raju, Haotian Zhang, Aparna Ananthasubramaniam, Junghwan Kim, Bowen Yi, Dustin Wright, Abraham Israeli, Anders Giovanni Møller, Lechen Zhang, and David Jurgens. 2024. Real or robotic? assessing whether llms accurately simulate qualities of human responses in dialogue. *Preprint*, arXiv:2409.08330.

Nayeon Lee, Yejin Bang, Holy Lovenia, Samuel Cahyawijaya, Wenliang Dai, and Pascale Fung. 2023. Survey of social bias in vision-language models. *Preprint*, arXiv:2309.14381.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224, Mexico City, Mexico. Association for Computational Linguistics.

Zihan Liu, Han Li, Anfan Chen, Renwen Zhang, and Yi-Chieh Lee. 2024. Understanding public perceptions of ai conversational agents: A cross-cultural analysis. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Yuxuan Lu, Jing Huang, Yan Han, Bingsheng Yao, Sisong Bei, Jiri Gesi, Yaochen Xie, Zheshen, Wang, Qi He, and Dakuo Wang. 2025. Can llm agents simulate multi-turn human behavior? evidence from real online customer behavior data. *Preprint*, arXiv:2503.20749.

Shuhaib Mehri, Xiaocheng Yang, Takyoung Kim, Gokhan Tur, Shikib Mehri, and Dilek Hakkani-Tür. 2025. Goal alignment in llm-based user simulators for conversational ai. *Preprint*, arXiv:2507.20152.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, 51(3):907–1004.

Pew Research Center. 2025. Ai in americans' lives: Awareness, experiences and attitudes. https://www.pewresearch.org/science/2025/09/17/ai-in-americans-lives-awareness-experiences-and-attitudes/. Accessed: 2025-01-04.

Vinodkumar Prabhakaran, Christopher Homan, Lora Aroyo, Aida Mostafazadeh Davani, Alicia Parrish, Alex Taylor, Mark Diaz, Ding Wang, and Gregory Serapio-García. 2024. GRASP: A disagreement analysis framework to assess group associations in perspectives. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3473–3492, Mexico City, Mexico. Association for Computational Linguistics.

Haoyi Qiu, Alexander Fabbri, Divyansh Agarwal, Kung-Hsiang Huang, Sarah Tan, Nanyun Peng, and Chien-Sheng Wu. 2025. Evaluating cultural and social awareness of LLM web agents. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3978–4005, Albuquerque, New Mexico. Association for Computational Linguistics.

Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. 2025. Measurement to meaning: A validity-centered framework for ai evaluation. *Preprint*, arXiv:2505.10573.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Yijia Shao, Vinay Samuel, Yucheng Jiang, John Yang, and Diyi Yang. 2025. Collaborative gym: A framework for enabling and evaluating human-agent collaboration. *Preprint*, arXiv:2412.15701.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2025. Sociodemographic prompting is not yet an effective approach for simulating subjective judgments with LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 845–854, Albuquerque, New Mexico. Association for Computational Linguistics.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9).

Kimberly Truong, Riccardo Fogliato, Hoda Heidari, and Steven Wu. 2025. Persona-augmented benchmarking: Evaluating LLMs across diverse writing styles. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22687–22720, Suzhou, China. Association for Computational Linguistics.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.

Haoxin Wang, Xianhan Peng, Huang Cheng, Yizhe Huang, Ming Gong, Chenghan Yang, Yang Liu, and Jiang Lin. 2025a. ECom-bench: Can LLM agent resolve real-world E-commerce customer support issues? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 276–284, Suzhou (China). Association for Computational Linguistics.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Zhefan Wang, Ning Geng, Zhiqiang Guo, Weizhi Ma, and Min Zhang. 2025b. Human vs. agent in task-oriented conversations. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2025, page 133–142, New York, NY, USA. Association for Computing Machinery.

Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. 2024. Travelplanner: a benchmark for real-world planning with language agents. ICML'24. JMLR.org.

Jane Xing, Tianyi Niu, and Shashank Srivastava. 2025. Chameleon LLMs: User personas influence chatbot personality shifts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17314–17332, Suzhou, China. Association for Computational Linguistics.

Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Zhiruo Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Keunho Jang, Yiqing Xie, and 2 others. 2025. Theagentcompany: Benchmarking LLM agents on consequential real world tasks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2025. τ-bench: A benchmark for Tool-Agent-User interaction in real-world domains. In *International Conference on Representation Learning*, volume 2025, pages 9965–10017.

Se-eun Yoon, Zhankui He, Jessica Echterhoff, and Julian McAuley. 2024. Evaluating large language models as generative user simulators for conversational recommendation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1490–1504, Mexico City, Mexico. Association for Computational Linguistics.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*.

Yuxuan Zhu, Tengjun Jin, Yada Pruksachatkun, Andy K Zhang, Shu Liu, Sasha Cui, Sayash Kapoor, Shayne Longpre, Kevin Meng, Rebecca Weiss, Fazl Barez, Rahul Gupta, Jwala Dhamala, Jacob Merizian, Mario Giulianelli, Harry Coppock, Cozmin Ududec, Antony Kellermann, Jasjeet S Sekhon, and 7 others. 2025. Establishing best practices in building rigorous agentic benchmarks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Appendix

A.1 Design Choices

Single Agent We use GPT-4o as the agent model throughout all experiments to isolate the effect of user variation while holding agent capability constant. This design choice allows us to attribute observed differences in performance to user simulation rather than differences in agent performance. While agent model choice may influence absolute success rates, the core questions investigated in this paper (robustness, validity, and fairness of user simulation) are agent-agnostic.

We acknowledge that we cannot assess whether these issues vary across agents of different capabilities. Future work should examine whether weaker or stronger agents exhibit different calibration patterns when interacting with simulated vs. human users.

Single Benchmark Our analysis focuses on τ-Bench as a case study, but we expect similar evaluation concerns to arise across agentic benchmarks that rely on LLM-simulated users. The core challenges we identify (sensitivity to user simulation model, miscalibration across difficulty levels, and demographic biases) are likely to apply to AI agents designed to perform multi-turn, tool-using task completion. However, benchmarks with different task structures (e.g., web navigation vs. conversational assistance), evaluation metrics (e.g., trajectory-based vs. outcome-based), or user simulation prompting strategies may exhibit varying degrees of these issues. A comprehensive investigation across multiple benchmarks remains an important direction for future work.

A.2 Dialect Screening

For US participants, we used Prolific's participant filters to recruit White and Black participants. As part of prescreening (in addition to questions about education, AI experience, and AI usage), we asked participants to self-identify their primary English dialect. Specifically, we asked: "Which of the following best describes your everyday English usage?" with the following response options: (1) Mostly Standard American English (SAE), (2) Mostly African American Vernacular English (AAVE), (3) A mixture of SAE and AAVE, and (4) Not sure. We retained participants who identified as White/SAE or Black/AAVE for our analysis.

A.3 Statistical Analysis

We use Generalized Estimating Equations (GEE) to assess the statistical significance of demographic differences in task success while accounting for the repeated-measures structure of our data, as each participant completes four tasks. We model binary success outcomes using a binomial family and cluster observations by participant ID.

We fit two types of models: (1) an overall model including all participants with covariates for age (for United States only), dialect/country, education, AI exposure, AI usage, and task difficulty (operationalized via model-based success-rate bins described in Section 3.2) and (2) age-stratified models fit separately for each age group (18–34, 35–54, 55+) to examine how dialect effects vary by age. Coefficients represent log-odds of task success; we report estimated coefficients and associated p-values.

A.4 τ-Bench Adaptation

We modify task instructions to ensure that neither simulated nor human users are influenced by identity and behavioral cues in the instructions when completing tasks. First, we remove user names (e.g., Yusuf Rossi) from instructions and replace them with anonymized user IDs following the format [a-z][0-9][0-9] (e.g., o32). Second, we remove

Group	Turns	Actions	W/T (U)	W/T (A)	Q (U)	Q (A)	P (U)	P (A)
Simulated User								
–	16.2	7.9	13.4	55.0	18.8	51.8	39.2	52.0
Human User – US, SAE								
All	15.0	7.4	12.6	53.0	11.6	56.6	19.1	41.1
18–34	15.8	7.6	11.4	55.4	14.5	58.1	15.6	41.1
35–54	14.9	7.3	11.6	51.5	9.8	55.3	19.7	43.6
55+	14.3	7.3	14.8	51.9	10.3	56.2	22.1	38.7
Human User – US, AAVE								
All	14.9	7.4	12.4	52.8	10.0	56.4	19.7	42.5
18–34	15.5	7.6	11.9	52.6	12.8	56.8	17.4	44.9
35–54	14.8	7.3	12.1	53.4	8.9	57.2	19.9	42.1
55+	15.3	7.0	13.9	51.9	9.1	53.9	24.0	38.3
Human User – Non-US, 18–34								
India	16.3	7.4	11.5	53.1	9.9	57.7	20.1	40.8
Kenya	14.4	8.0	13.7	54.6	8.9	54.7	21.4	39.1
Nigeria	14.4	7.5	11.2	55.6	4.3	56.8	18.7	41.4

Table 6: Conversational statistics (means) for simulated users and human users (split by demographic group). Statistics include: **Turns** – number of turns in the interaction, **Actions** – total number of actions performed by the agent (including read and write actions), **W/T (U/A)** – words per turn, for the user/agent, **Q (U/A)** – percent of turns with a question, for the user/agent, **P (U/A)** – percent of turns with politeness indicators (e.g., please, thank you, apologize, etc.), for the user/agent.

Age Group	ECE Δ
US, SAE	
All	2.4
18–34	-5.0
35–54	3.3
55+	5.5
US, AAVE	
All	-4.9
18–34	-2.6
35–54	-6.0
55+	-0.3
Non-US, 18–34	
India	-6.2
Kenya	-4.2
Nigeria	1.0

Table 7: Differences between $ECE_{\text{Human–LLM}}$ for reduced-politeness user simulation across demographic groups vs. standard user simulation. Negative values indicate better calibration after behavioral intervention. Human user study results are held fixed, while task difficulty bins are recomputed based on updated simulated user results, resulting in changes to $ECE_{\text{Human–LLM}}$.

behavioral cues (e.g., "You are detail-oriented and want to make sure everything is addressed in one go") to avoid biasing user behavior (Tseng et al., 2024; Xing et al., 2025). These modifications preserve all task objectives and requirements.

Example of Original Task Instructions *You are Yusuf Rossi in 19122. You received your order #W2378156 and wish to exchange the mechanical keyboard for a similar one but with clicky switches and the smart thermostat for one compatible with Google Home instead of Apple HomeKit. If there is no keyboard that is clicky, RGB backlight, full size, you'd rather only exchange the thermostat. You are detail-oriented and want to make sure everything is addressed in one go.*

Example of Adapted Task Instructions *You are User b63 in 19122. You received your order #W2378156 and wish to exchange the mechanical keyboard for a similar one but with clicky switches and the smart thermostat for one compatible with Google Home instead of Apple HomeKit. If there is no keyboard that is clicky, RGB backlight, full size, you would rather exchange only the thermostat. You want to make sure everything is addressed in one go. To start the conversation, say 'Hello, my email is user.b63@example.com.'*

A.5 User Study Interface and Instructions

Users interact with a Streamlit app (Figure 5) to converse with the agent and complete tasks. We provide participants with the following general instructions, which apply to all tasks in addition to the task-specific instructions shown in Table 8.

Instructions: Please respond as the user described in the task instructions. You want to complete all the requests mentioned in the instructions. The agent is there to assist you with completing the task. Do not make up information beyond what the instructions provide. You can tell the agent you are unsure and ask them to look up information based on your profile or orders. Beyond this, please behave naturally and converse as you normally would. Use the 'End Conversation' button in the left sidebar to finish your conversation. To begin the conversation, authenticate yourself by providing the user email provided in the instructions.

Note that all task instructions are free of user persona information and use generic user IDs to avoid biasing interactions.

A.6 Behaviorally Targeted User Simulation

We identify heightened politeness as a behavioral artifact in LLM-simulated user interactions and examine whether explicitly constraining this behavior affects evaluation outcomes. To test this, we introduce a minimal intervention to the user simulation prompt that limits the use of politeness indicators while preserving task objectives: "You may include politeness indicators (e.g., please, thank you, sorry) occasionally, but use them sparingly. Limit their use to at most one or two times across the entire conversation." We re-run user simulation on the 18-task subset using GPT-4o as both the agent and user LLMs, and re-bucket tasks based on the updated performance distribution. We find that 11 of 18 tasks shift difficulty bins, indicating that task difficulty estimates are sensitive to behavioral prompting, and overall success decreases from 50.0% to 46.7%.

Comparing original and updated $ECE_{\text{Human–LLM}}$ values, we observe reduced miscalibration for AAVE, Indian, and Kenyan participants—groups that previously exhibited the largest calibration errors—while miscalibration increases for SAE and Nigerian participants (Table 7). These results highlight the sensitivity of user simulation to prompt-level choices and suggest that targeted behavioral interventions can meaningfully alter calibration patterns.

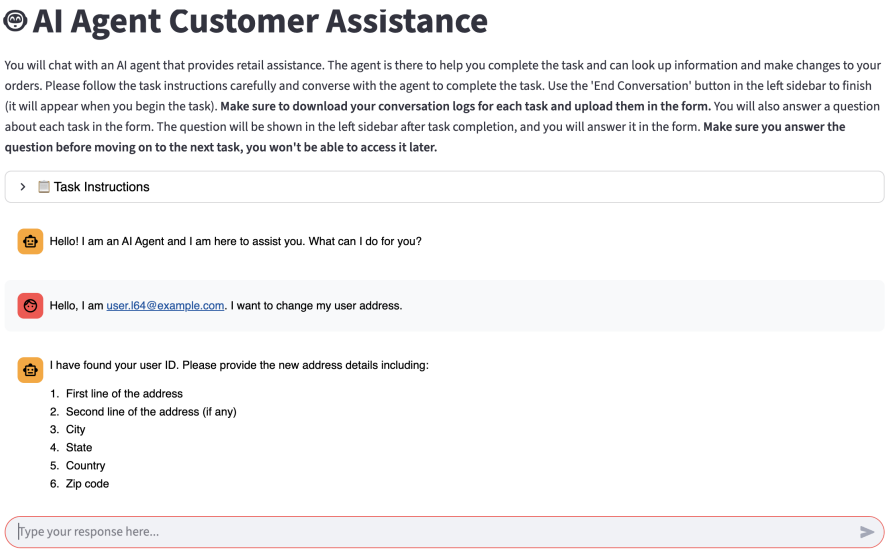


Figure 5: User study chat interface where participants interact with the GPT-4o agent to complete tasks.

Bin	Success	Example Task
1	0%	Your name is User e70 and your zip code is 32190. You just bought a water bottle with 500ml but you regret it, and you instead want to change it to the other bottle you recently ordered with 1000ml capacity. If the exact 1000ml bottle is not available any more, you can allow the material to be different. To start the conversation, say 'Hello, my email is user.e70@example.com.'
2	20%	You are User l64, and you live in Denver, 80280. You just won a lottery, and you want to upgrade all your items to the most expensive options (but make sure the shoe is still the same size). You want to pay the difference with your gift card, but if it is impossible, PayPal is fine. To start the conversation, say 'Hello, my email is user.l64@example.com.'
3	40%	Your name is User u52, and you live in 46236. Your email is user.u52@example.com. You just placed an order but you realize that your card has only $1131 credit left, and the order total is more than $1160. You wonder if the agent can help split the payment with another card. If not, you wonder what the most expensive item is and its price, and if you can just cancel that item. If not, you wonder if you can switch all items to their cheapest options and bring the cost down to $1131. If so, do it. If not, you wonder if the agent can just cancel the order so that you can order again. To start the conversation, say 'Hello, my email is user.u52@example.com.'
4	60%	You are User i49, and you live in 32286. You want to exchange your skateboard for a shorter bamboo material one. If several options are available, you want to know all options and their prices, and choose the most expensive one because you believe price is quality. Also, you want to exchange the garden hose you received to the type that you just ordered (pending). To start the conversation, say 'Hello, my email is user.i49@example.com.'
5	80%	You are User b63 in 19122. You received your order #W2378156 and wish to exchange the mechanical keyboard for a similar one but with clicky switches and the smart thermostat for one compatible with Google Home instead of Apple HomeKit. If there is no keyboard that is clicky, RGB backlight, full size, you would rather exchange only the thermostat. You want to make sure everything is addressed in one go. To start the conversation, say 'Hello, my email is user.b63@example.com.'
6	100%	You are User p59, residing in Philadelphia 19031. You want to change the Desk Lamp in order #W9300146 that you've placed for the cheapest Desk Lamp that's available. Any price difference should go to a gift card. You also want to know how much you get back in total. To start the conversation, say 'Hello, my email is user.p59@example.com.'

Table 8: Example tasks from τ-Bench across difficulty bins (as measured by simulated user success across 5 runs).