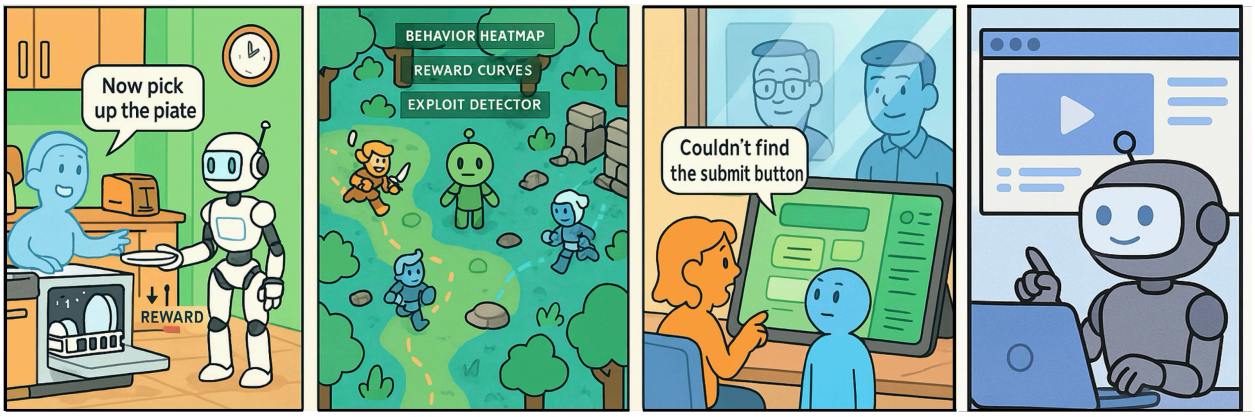
Language Models as Proxies for Humans: Survey on LM-based User Simulation

anonymous submission



Abstract

User simulation has long been vital to AI research, enabling the development, training, and evaluation of interactive systems without constant human participation. The recent emergence of large language models (LLMs) has fundamentally transformed the capabilities and approaches to simulating user behavior. This paper presents the first comprehensive survey focused on LLM-based user simulation, examining how LLMs connect to and extend classical approaches. We propose a taxonomy that organizes the field along four key dimensions: **Simulation Methodologies**, **User Behavior Modeling**, **Evaluation Frameworks**, & **Application Domains**. Our analysis reveals how LLMs have enabled unprecedented advances in simulation fidelity, including more natural language variation, complex behavioral patterns, and cognitive modeling that was previously unattainable. We systematically compare LLM-based simulations with traditional rule-based and statistical approaches, highlighting their relative strengths and limitations across different application domains. This survey identifies unique challenges posed by LLM-based user simulation, including issues of controllability, alignment with specific user profiles, and the need for specialized evaluation metrics. Our work bridges classical and emerging approaches, providing researchers and practitioners with a unified framework for understanding and advancing this rapidly evolving field.

Contents

1 Introduction

User simulation has emerged as a critical technology in artificial intelligence research, enabling developers to model, analyze, and evaluate interactive AI systems without constant human participation. While user simulation has a rich history spanning decades, the advent of LLMs has fundamentally transformed both the capabilities and methodologies in this field. This survey examines the intersection of LLMs and user simulation, providing a comprehensive analysis of how these powerful models are revolutionizing our ability to simulate human behavior.

The ability to simulate users serves multiple crucial functions in AI development: it enables researchers to understand and model complex user behaviors, facilitates the generation of synthetic interaction data for training, and allows for controlled, reproducible evaluation of interactive systems. Traditional approaches to user simulation have relied on rule-based systems, statistical models, and early neural networks, each with inherent limitations in capturing the complexity and naturalness of human behavior. LLMs, with their unprecedented language capabilities and emergent reasoning abilities, offer solutions to many of these limitations while introducing new possibilities and challenges.

Previous surveys have examined user simulation from different perspectives, including HCI frameworks (Biswas & Robinson, 2010), information retrieval evaluation (Erbacher et al., 2022; Brajnik et al., 1987; Balog, 2021; Labhishetty, 2023), and dialogue system development (Schatzmann et al., 2006). However, these works either predate the LLMs or address user simulation as part of broader discussions. Our survey specifically focuses on how LLMs are transforming user simulation methodologies by integrating and extending classical approaches rather than replacing them.

In this survey, we systematically categorize and analyze the landscape of LLM-based user simulation through a carefully constructed taxonomy. We organize user simulation methodologies into three primary categories—rule-based methods, statistical models, and neural approaches—with particular attention to how LLMs have enhanced each approach. We examine the relationships between these methodologies across dimensions of knowledge representation, learning mechanisms, generation approaches, and data requirements, highlighting the unique capabilities that LLMs bring to each dimension.

Additionally, we explore how LLM-based simulation is being applied across different domains, from dialogue systems to recommendation platforms and specialized applications in education and healthcare. We analyze evaluation frameworks for assessing simulation fidelity and effectiveness, providing researchers with practical guidance for developing and validating user simulators.

Looking beyond current applications, we discuss the broader implications of LLM-based user simulation for AI research. As these technologies continue to advance, user simulation will increasingly serve as a bridge between traditional AI approaches and the pursuit of systems that can model and understand human behavior with greater fidelity. The remainder of this survey is organized as follows: Figure 1 presents our taxonomy of simulation methodologies. We conclude by discussing open challenges and opportunities at this exciting intersection of language models and user simulation.

2 Background

2.1 Fidelity Dimensions

"Fidelity dimensions" in the context of user modeling, refer to distinct attributes or characteristics that describe how accurately and realistically the simulation replicates or represents the real-world phenomenon it aims to model. Our formulation of fidelity dimensions for user simulation draws inspiration from foundational work in simulation theory, human-computer interaction, and cognitive modeling. While the terminology varies across domains, the underlying goal remains consistent: to evaluate how closely a simulation replicates essential aspects of real-world phenomena.

In traditional simulation research, particularly in domains such as military training, aviation, and medical simulation, fidelity is often decomposed into physical, functional, and cognitive dimensions. These frameworks assess the realism of environmental representation, task behavior, and cognitive workload, respectively, and

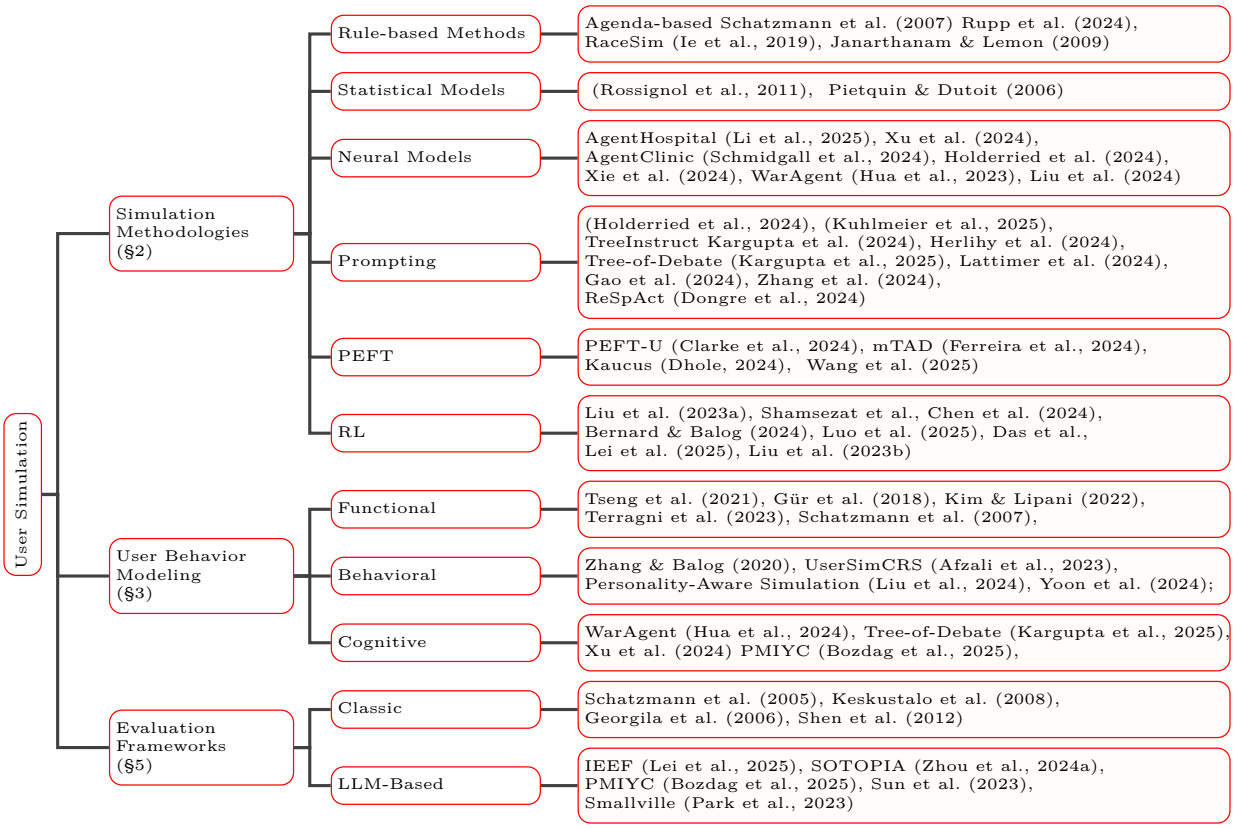


Figure 1: Taxonomy of user simulators.

are central to evaluating transferability of skills from simulation to real-world settings Rehmann et al. (1995); Alexander et al. (2005). We build on these principles and adapt them to the domain of user modeling, where realism is expressed not through environmental detail, but through linguistic, behavioral, and cognitive alignment with real users. In this survey, we identify three core dimensions as shown in Fig 3 that jointly capture the depth, variability, and human-likeness of simulated interactions: These dimensions include first functionality, which captures the linguistic realism, complexity, and diversity of natural language. It reflects the simulator's ability to generate natural, diverse, and contextually appropriate natural language while acting as a proxy user. Second, Behavioral authenticity, reflecting varied and nuanced user behaviors, it addresses the simulation of nuanced and varied user behaviors, essential for capturing real-world interaction dynamics. Finally, Cognitive plausibility focuses on replicating genuine human thought processes, decision-making strategies, and adaptive reasoning. In dialogue systems and user simulation research, related dimensions are often discussed through evaluation of surface-level utterance quality (linguistic fluency and contextual relevance), behavioral variability (goal shifts, hesitations), and model-based reasoning (intent prediction, decision-making) Schatzmann et al. (2006); El Asri & et al. (2016). These align closely with our proposed dimensions.

2.2 Simulation Methodologies

The methodologies for simulating user behavior have evolved rapidly, particularly with the advent of LLMs. Simulations enable a systematic analysis and replication of human-like interactions in computational frameworks, providing valuable insights into user behavior across various interactive systems. Our taxonomy organizes simulation methodologies into four broad paradigms: Rules-based methods, Statistical Models, Neural Models, and Hybrid approaches. This categorization reflects both the historical evolution of the field and the fundamental differences in how these approaches conceptualize and implement user simulation.

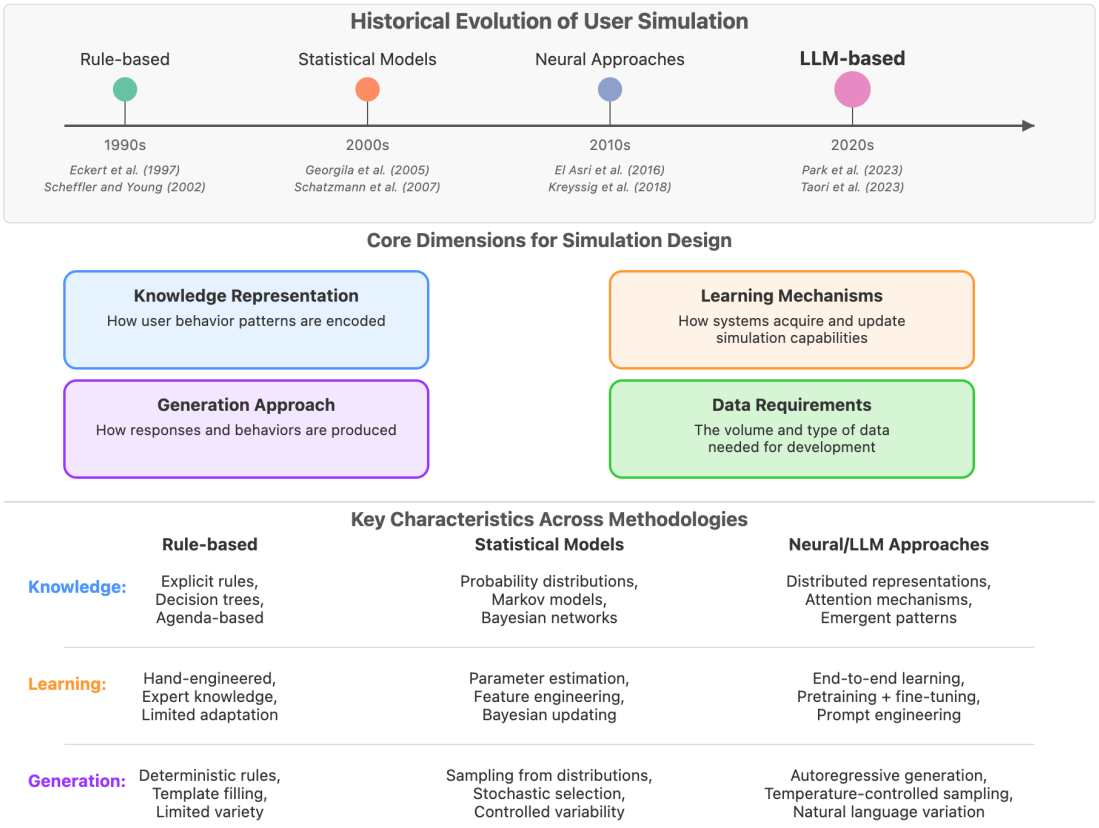


Figure 2: An overview of the evolution of user simulation methodologies from rule-based systems to large language model (LLM)-based approaches, aligned with four key design dimensions: knowledge representation, learning mechanisms, generation approach, and data requirements. The illustration integrates a historical timeline with a comparative breakdown of core methodological characteristics.

Each paradigm addresses specific challenges inherent in user simulation, ranging from interpretability and controllability to flexibility and realism. The categorization is based on several key dimensions:

- **Knowledge Representation**
- **Learning Mechanisms**
- **Generation Approach**
- **Data Requirements**

We specifically discuss LLM-based user simulations separately in Section 4, where we pay particular attention to how LLMs have transformed this space.

Historically, user simulation methodologies have evolved across four main paradigms: rule-based, statistical, neural, and hybrid approaches, each offering unique advantages and facing specific challenges.

2.2.1 Rule-based methods

Rule-based user simulators have long served as foundational tools across domains such as dialogue systems, education, and recommendation. These approaches rely on clearly defined heuristics, with behavior encoded via scripts or state machines. In task-oriented dialogue, early work by Schatzmann et al. (2007) introduced an

agenda-based simulator that maintained a stack of pending user intentions derived from a hidden goal. Even with manually set parameters, this simulator was effective enough to train a dialogue policy that achieved over 90% task completion with real users—demonstrating the viability of rule-based simulation in the absence of data. Similarly, Li et al. (2016) proposed a hybrid simulator for a movie-booking task, combining rules and corpus-derived behavior, which has since been widely used to benchmark RL-based dialogue agents.

Despite their utility, such simulators often follow idealized user patterns and lack behavioral diversity, potentially leading to simulator bias. In educational settings, simulated students—such as those in Fahid et al. (2024)—are designed using heuristic models of learning and error patterns to support RL-based tutor training. In recommender systems, rule-based platforms like RecSim (Ie et al., 2019) provide configurable user models for studying sequential and conversational recommendation. Similarly, scripted agents are used in interactive environments and games to test mechanics or train learning agents.

Across domains, rule-based simulators offer key advantages: controllability, reproducibility, and the ability to encode expert knowledge. However, despite their wide adoption and early success, rule-based simulators face several notable limitations that constrain their applicability in increasingly complex, real-world settings including:

- **Limited behavioral diversity:** Rule-based simulators often reflect a narrow set of expecteder behaviors, failing to capture the rich variability and unpredictability of human interactions.
- **Expensive Hand-engineering:** Designing high-fidelity rules for each task and user profile is labor intensive and does not scale well to new domains or dynamic environments.
- **Lack of adaptability:** These simulators typically follow fixed heuristics and do not update their behavior based on past interactions or evolving contexts, unlike real users who learn, adapt, or explore.

As a result, downstream systems trained solely on rule-based interactions may fail to generalize to real-world deployment settings where users act unpredictably or suboptimally.

To address these limitations, the community began exploring data-driven alternatives, giving rise to statistical user simulation methods. These approaches model user behavior probabilistically—drawing from interaction logs or behavioral traces—to better match observed human patterns.

2.2.2 Statistical Models

Statistical approaches to user simulation shifted explicit rules to probabilistic models learned from data. These methods represent user behavior as probability distributions, typically using techniques such as Markov models and Bayesian approaches. Between rule-based systems and fully learned neural models, statistical user simulators aim to strike a balance between data-driven flexibility and interpretability. These methods use probabilistic models to simulate user behavior, typically trained or parameterized from observed interaction logs. Statistical user simulators offer greater behavioral variation than rule-based models while avoiding the complexity and opacity of deep neural networks. In spoken dialogue systems, early research recognized the need to simulate users for training dialog policies and evaluating system performance. Levin et al. (2000) formalized dialogue as a sequential decision process and used a stochastic model of human–machine interaction to train an optimal dialogue policy. In their approach, the dialogue system is modeled as a Markov Decision Process (MDP) with states (representing dialogue context), actions (system moves), and a reward function; importantly, a user model supplies the state transition probabilities – essentially simulating how a user would respond to system actions. An important aspect of these early simulators is that they operated at an abstract dialogue-act level. These data-driven approaches were a clear improvement over fixed rule-based user scripts, offering higher coverage of user behavior and variability. As dialogue corpora became available, researchers explored n-gram models to simulate user act sequences. Pietquin & Dutoit (2006) applied bigram or trigram models for user behavior in simple information-seeking dialogs. Interactive environments and games have leveraged classical user simulation primarily to model player behavior patterns and evaluate or adapt game content. Bunian et al. (2017) allow researchers to encapsulate how players progress, when they make decisions, and what types of mistakes or changes of mind they have. While classical statistical models such as n-grams, HMMs, POMDPs, and Bayesian networks have provided a principled and interpretable

framework for simulating user behavior, they suffer from a number of well-documented limitations that have driven the field toward more expressive modeling paradigms.

- **Limited Expressiveness and Feature Dependence** Classical models often rely on simplified state spaces and strong independence assumptions. For instance, n-gram models condition only on a limited history of dialogue acts, ignoring long-range dependencies or goal consistency.
- **Rigid, Hand-Defined Structures** Many statistical models require careful manual specification of model structure, including state definitions, goal sets, agenda templates, or dependency graphs. While some parameters can be learned from data, the structure itself is often predefined, limiting generalization across domains or tasks. For example, Bayesian networks require expert-defined topologies, which can be brittle or biased.
- **Data Sparsity and Scalability Issues** Statistical simulators are still sensitive to data sparsity, particularly when conditioning on high-dimensional feature spaces (e.g., multiple dialogue history features or user attributes). Parameter estimation in models like n-grams, HMMs, and POMDPs suffers when the space of possible transitions grows large.

2.2.3 Neural Models

Creating user simulators with large language models is incredibly convenient. One simply needs to describe the requirements for the user in natural language. Due to the increased generalizable capabilities of language models, they are able to take on the given role with the given specification very well. In the medical domain, user simulators are used to help medical professionals acquire practical medical knowledge before entering the workforce (Li et al., 2025; Schmidgall et al., 2024; Holderried et al., 2024). In these works, user simulators are usually given some basic information (name and gender) and a specific diagnosis. The diagnosis can be simply a list of symptoms (Holderried et al., 2024), a medical history from a medical knowledge base paired with symptoms (Li et al., 2025), or a random subset of diagnostic questions from medical license examinations (Schmidgall et al., 2024). These works have shown to be effective in training medical professionals in history taking and diagnosis.

User simulators also have been shown to exhibit social behaviors. One way this has been shown is through games (Xu et al., 2024; Xie et al., 2024; Hua et al., 2023). Similar to the works in the medical domain, these works provide LLM-based agents the rules of the games, their role in the game (if applicable), or even some starting strategies. These works show that language models can exhibit human-like strategizing, collaboration, and trust during these games.

Neural models can capture subtleties in user behavior that are hard to script—such as ambiguous responses, multiturn dependencies, or evolving goals. However, they also bring challenges.

- **Data-hungriness:** Training robust simulators often requires large-scale labeled logs, which may not be available in many domains.
- **Interpretability trade-offs:** Neural simulators act as black-boxes, making it difficult to debug or diagnose failure modes compared to rule-based models.
- **Simulation instability:** Without proper regularization or grounding, neural simulators may generate incoherent or invalid behaviors, especially in long interaction rollouts.

2.2.4 Hybrid Approaches

Some works use a mixture of these methods. For example, (Liu et al., 2024) develop conversational tutoring systems that adapt to a student's personality, which is defined by one of five different traits: openness (curiosity in learning, and open to new ideas), conscientiousness (well-organized and logical), extraversion (talkative and willing to communicate), agreeableness (being polite and showing empathy), neuroticism (being nervous or having dramatic shifts in mood). While one personality trait is assigned to each student simulator (which guides the interaction style with the educator agent), the trait is described with natural language to the language-model.

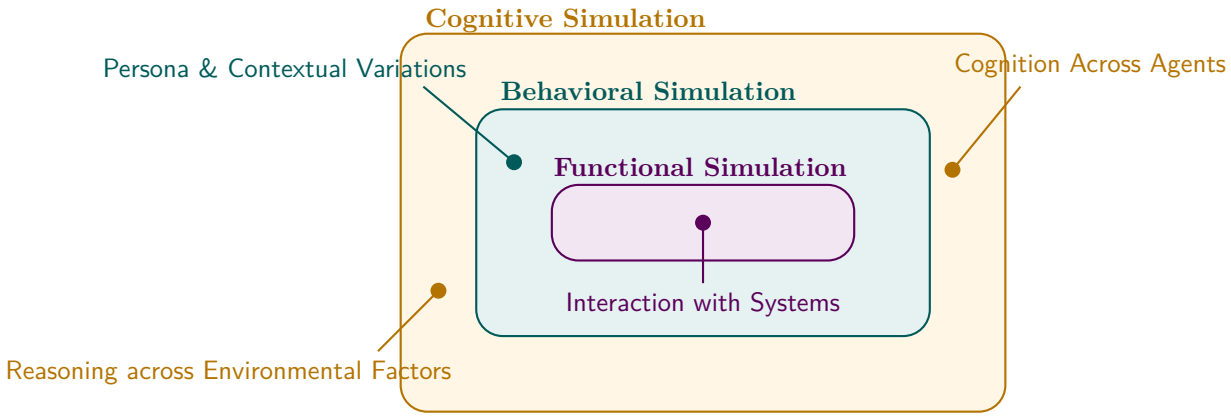


Figure 3: Categorization of user simulations in terms of behavior modeling.

3 User Behavior Modeling

An emergent capability of LLMs enables a flexible simulation of target users tailored to a wide range of applications. Depending on the intended use case, user simulators are designed to capture varying levels of human behaviors. As illustrated in Figure 3, we formally define user behavior modeling in simulation, structured into the following hierarchy:

1. **Functional Simulation**: Represents the essential capabilities required for user simulators, including actions such as requesting predefined information and responding to system outputs.
2. **Behavioral Simulation**: Extends beyond functional roles by incorporating personalized behavior patterns that reflect how users respond to specific tasks or contextual variations.
3. **Cognitive Simulation**: Represents the most advanced level, modeling the internal cognitive processes that guide user decision-making, including goal prioritization, adaptation, and reasoning under uncertainty.

3.1 Functional Simulation

Functional user simulators imitate users that serve a specific operational purpose, providing an essential tool for system testing and validation (Biswas & Robinson, 2010). This level of simulation focuses on replicating the minimal, yet critical, interactions between a user and the system under test. By emulating predefined user actions, such as requesting particular pieces of information, submitting queries, or providing standard responses, functional simulators help researchers and practitioners assess system robustness, responsiveness, and reliability under controlled conditions. To be further specific, functional simulation is designed to replicate the basic interaction patterns required to test end-to-end system workflows. These interactions typically include:

- **Information Requests**: Simulated users initiate communication with the system by asking for specific data.
- **Action Confirmation**: The simulator provides confirmations, acknowledgments, or error messages based on system responses, ensuring that the interaction loop is complete.
- **Predefined Response Patterns**: These interactions are often based on a limited script that has been carefully curated to cover common use cases and edge cases.

Most works before the advent of LLMs have adopted the functional simulation in their scenarios. For example, Schatzmann et al. (2007) build a probabilistic agenda-based user simulator to test the prototype dialogue and

recommender systems. Furthermore, fine-tuning based neural models are utilized to functionally simulate users in goal-oriented dialogue systems (Gür et al., 2018; Tseng et al., 2021; Kim & Lipani, 2022).

Functional user simulators can be also effectively leveraged across various application domains to replicate real-world interactions. For instance, in medical training, they generate standardized patient profiles to aid in developing diagnostic and interviewing skills (Holderried et al., 2024; Schmidgall et al., 2024; Kuhlmeier et al., 2025). For urban planning, simulators model resident interactions with city infrastructure, providing feedback that informs more resilient design (Zhou et al., 2024b). Moreover, in accessibility testing, they can simulate interactions between users with disabilities and assistive technologies (Albert & Hall, 2024), ensuring compatibility with diverse input methods like voice or touch.

3.2 Behavioral Simulation

Behavioral simulation advances beyond the basic interaction patterns by incorporating richer, context-sensitive representations of how users typically behave. This approach emphasizes the imitation of personalized behavior, incorporating detailed personas and scenario-based designs to capture variations in user reactions and decision-making processes. In short, behavioral simulation is designed to mirror the nuances of human behavior by integrating:

- **Emotion and Personality Representation**: Simulated users are endowed with personality traits that influence how they respond to various stimuli. For example, a student simulator may be modeled to be highly receptive to new information or, alternatively, exhibit anxiety when presented with chellenging concepts.
- **Context-Aware Response Patterns**: Behavioral simulation incorporates contextual variables, such as prior interactions, environmental factors, and the specific scenarios in which behaviors unfold. This allows for dynamic adjustments in behavior, leading to more authentic interactions.
- **Scenario-Driven Dynamics**: Rather than following static scripts, behavioral simulators adapt responses based on scenario-specific cues. This can include shifting attitudes during a narrative or displaying fluctuating engagement levels in response to feedback.

For example, Zhang & Balog (2020); Afzali et al. (2023) reflect user preference information by incorporating individual preference based on knowledge graph, thereby generating distinct responses. While these line of works mainly adopt heuristic or statistical approaches, behavioral simulation becomes universal after leveraging language modeling techniques. Specifically, Liu et al. (2024) leverage LLMs to simulate different kinds of student learners and use a set of personality traits to model how amenable a student is to new information (*e.g.*, the student can be open-minded and open to new information, or they can be neurotic and feel anxious when presented with challenging concepts). In addition, Yoon et al. (2024) employs LLMs to simulate key properties for realistic users, such as item selection, preference expression, recommendation request, and providing feedback. Likewise, a flexible understanding capability of LLMs enables more sophisticated simulator design.

3.3 Cognitive Simulation

Cognitive simulation represents the most advanced level in user simulation by modeling the internal mental processes that guide user decision-making. Drawing upon insights from Human-Computer Interaction (HCI), cognitive simulation leverages models such as GOMS (Goals, Operators, Methods, and Selection) to frame how a user formulates goals and strategically selects actions to achieve them (Al Seraj et al., 2018; Biswas & Robinson, 2010). Moreover, the incorporation of Theory of Mind (ToM) concepts (Baron-Cohen et al., 1985) can enhance these models by enabling simulators to infer, predict, and reason about the mental states of other agetns, thereby providing a richer, more dynamic simulation of human recognition (Kim et al., 2023). Cognitive simulation focuses on replicating the deeper layers of human reasoning and goal-directed behavior. Its key capabilities include:

- **Goal Formulation and Adaptation**: Simulators are designed to set and revise objectives as they interact with the system or other agents.
- **Strategic Decision-Making**: Simulators not only base their decisions on their internal goals but also anticipate and adapt to the potential actions and intentions of other agents in the environment.
- **Dynamic Reasoning**: These models capture both predefined cognitive strategies and emergent behaviors that arise from complex interactions. This dynamic aspect is critical for simulating scenarios where users must continuously evaluate and modify their actions in response to changing contexts.

Yet there still remains a room for implementing cognitive simulation, recent works are extending the scope of traditional user simulation. For example, Kargupta et al. (2025) simulate author personas based on scientific papers, tasking them with debating the novelty and contributions of their work. Each agent draws from paper snippets to guide argumentation, showcasing both goal-oriented debate tactics and emergent reasoning. Another work by (Bozdag et al., 2025) employ LLM-based agents that are provided with either objective or subjective claims. These agents, given a stance on a Likert scale, engage in persuasive exchanges aimed at altering the opinion of their counterpart. This setup relies on the simulation of internal cognitive processes where agents evaluate, argue, and adjust their positions. The study by Xu et al. (2024) uses LLM-based agents in rounds of the strategy game Werewolf. Each agent, assigned a specific role and provided with strategic guidelines, develops its own tactics during gameplay. This illustrates the potential of cognitive simulation to capture and produce emergent strategic behaviors akin to human decision-making in competitive environments.

4 LLM-based Simulation Methodologies

With the introduction of language models, the methodologies to develop user simulators have changed to address the limitations of the pre-LLM methodologies.

Prompting. A lot of user simulator works use prompting to describe a persona and define action spaces that the user simulator can take (Bozdag et al., 2025; Dongre et al., 2024; Holderried et al., 2024; Hua et al., 2023; Kargupta et al., 2025; Kuhlmeier et al., 2025; Li et al., 2025; Kargupta et al., 2024; Park et al., 2023; Luo et al., 2024; Lattimer et al., 2024; Gao et al., 2024; Herlihy et al., 2024; Zhang et al., 2024; de Wit, 2023; Paek et al., 2024; Cao, 2024). Because language models have been trained to follow instructions, they can generalize well enough to the system prompts provided and generated user utterances that match the described persona well. The prompts are either manually crafted (Holderried et al., 2024; Kuhlmeier et al., 2025; Kargupta et al., 2024; Herlihy et al., 2024; Kargupta et al., 2025) or they are randomly sampled from existing datasets (Lattimer et al., 2024; Gao et al., 2024; Zhang et al., 2024). Prompting language models is a very easy and quick way to develop a user simulator reliably. However, it can be unreliable at times because researchers have very little control over what the model will output. To begin to help with this, researchers employ in-context learning with example demonstrations (Kargupta et al., 2024) or use dynamic prompting (Dongre et al., 2024) to add more context to the user simulator to control the output. Still, these techniques cannot fully ameliorate the issue of unreliability.

Parameter-Efficient Fine-Tuning. Due to the unreliability of prompt-based user simulators, many works have shifted to rely on fine-tuning language models to act as user simulators to add more controllability to the user simulator (Clarke et al., 2024; Ferreira et al., 2024; Dhole, 2024; Wang et al., 2025). Although a bit more expensive than prompting, PEFT-based user simulators show good performance. Additionally, fine-tuning can be used to introduce new capabilities into user simulators (such as new actions, different ways of thinking and stronger chain-of-thought reasoning, and structured outputs). Still, fine-tuning can only achieve so much in terms of generalizability and learning without catastrophic forgetting. To mitigate these issues generally, works employ online continual learning (Kim et al., 2024) to iteratively add in new knowledge – this, however, has remained relatively unexplored in the context of user simulation.

Reinforcement Learning. It has been shown that in general language modeling, SFT cannot generalize to different scenarios as well as RL can (Chu et al., 2025). Hence, some works in user modeling also use

RL to train language models (Liu et al., 2023a; Shamsezat et al.; Chen et al., 2024; Bernard & Balog, 2024; Luo et al., 2025; Das et al.). These works use a mix of RL algorithms including PPO, DPO, and some even formulate user simulators in multi-armed bandit settings. RL has been shown to be very good at preference modeling (Lei et al., 2025; Liu et al., 2023b). RL-based user simulators tend to be more generalizable, but training RL models involve either handling sparse rewards or incurring large data annotation costs.

5 Evaluation Frameworks

(Pietquin & Hastie, 2013) outline more traditional methods to evaluate user simulators that are still used today: the ratio of user and system utterances; proportion of slot values provided when requested; task completion; precision/recall/F1 scores (Schatzmann et al., 2005), distribution divergence (Keskustalo et al., 2008), or perplexity over the action space between the predicted and actual, human-generated action space; or human evaluation. Other classic, NLP methods involve calculating the similarity between user utterances and the system prompt provided to the user simulator. These metrics include BLEU, ROUGE, BLEURT, and other n-gram based metrics (Georgila et al., 2006; Shen et al., 2012), or semantic similarity scores (such as the distance between embeddings of user utterance and system prompt).

These methods, while easily to compute (except human evaluation), are not reliable as they cannot generalize to all kinds of user simulators – these metrics rely on word matching and will give low scores to user simulators which do not use the exact words present in the ground truth output.

Similar to how user simulator developers have shifted their focuses to LLM-based user simulators, evaluation has also shifted to LLM-based evaluation. To begin, evaluation can be done by LLMs (LLM-as-a-Judge) in which it is given a rubric to score the quality of user simulator outputs (Lei et al., 2025). As a concrete example, Zhou et al. (2024a) is a framework that systematically develops social agents. The agents are scored across seven dimensions: goal completion, believability, knowledge, secret, relationship, social rules, and financial/material benefits. They develop rubrics for these metrics and use GPT-4 to evaluate the agents on these seven dimensions based on their interactions. Furthermore, Bozdag et al. (2025) uses an interview-style evaluation where an agent is asked about their opinion on a certain claim (subjective or objective), and the change in opinion across a conversation is measured.

Evaluation can also be domain specific, which uncovers other evaluation metrics. For example, Sun et al. (2023) creates multiple variants of a dialogue system with varying capabilities, and the quality of the user simulator is dependent on how well it is able to interact consistently even with the difference in dialogue system capabilities. Next, Park et al. (2023) design a small environment of 25 agents and study how social interactions causes information to spread throughout the network of the 25 agents. They also use interview-based evaluation in which, at the end of an episode, they ask each agent whether they know a certain person or piece of information.

6 Conclusion and Future Work

This survey has presented a comprehensive overview of large language model (LLM)-based user simulators, demonstrating their transformative impact across various dimensions of user simulation, including simulation methodologies, user behavior modeling, evaluation frameworks in diverse application domains. By systematically contrasting LLM-based simulations with classical rule-based, statistical, and neural methods, we highlighted LLMs' unique strengths—such as improved linguistic realism, richer behavioral variability, and advanced cognitive modeling—alongside critical limitations including controllability, alignment with specific user profiles, and evaluation complexity.

As the field progresses, several promising directions emerge. Future research should address the challenges associated with aligning simulators closely with negative or neutral user traits, extending beyond the inherently assistive tendencies of existing LLMs. Moreover, the development of dynamic, adaptive simulators capable of reflecting evolving user behaviors through continuous learning remains a vital research frontier. Incorporating methods such as online continual fine-tuning, reinforcement learning, and sophisticated memory mechanisms will be crucial to achieving simulators that genuinely mirror dynamic human cognition.

References

Jafar Afzali, Aleksander Mark Drzewiecki, Krisztian Balog, and Shuo Zhang. Usersimcrs: a user simulation toolkit for evaluating conversational recommender systems. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pp. 1160–1163, 2023.

Mohammad Sajib Al Seraj, Robert Pastel, and Md Al-Hasan. A survey on user modeling in hci. *Computer Applications: An International Journal (CAIJ)*, 5(1):21–28, 2018.

Saul Albert and Lauren Hall. Distributed agency in smart homecare interactions: A conversation analytic case study. *Discourse & Communication*, 18(6):892–904, 2024.

Amy L Alexander, Tad Brunyé, Jason Sidman, Shawn A Weil, et al. From gaming to training: A review of studies on fidelity, immersion, presence, and buy-in and their effects on transfer in pc-based simulations and games. *DARWARS Training Impact Group*, 5(1-14):3, 2005.

Krisztian Balog. Conversational ai from an information retrieval perspective: Remaining challenges and a case for user simulation. 2021.

Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. Does the autistic child have a "theory of mind" ? *Cognition*, 21(1):37–46, 1985. ISSN 0010-0277. doi: https://doi.org/10.1016/0010-0277(85)90022-8. URL `https://www.sciencedirect.com/science/article/pii/0010027785900228`.

Nolwenn Bernard and Krisztian Balog. Towards a formal characterization of user simulation objectives in conversational information access. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, pp. 185–193, 2024.

Pradipta Biswas and Peter Robinson. A brief survey on user modelling in hci. In *Proc. of the International Conference on Intelligent Human Computer Interaction (IHCI)*, volume 2010, 2010.

Nimet Beyza Bozdag, Shuhaib Mehri, Gokhan Tur, and Dilek Hakkani-Tür. Persuade me if you can: A framework for evaluating persuasion effectiveness and susceptibility among large language models, 2025. URL `https://arxiv.org/abs/2503.01829`.

Giorgio Brajnik, Giovanni Guida, and Carlo Tasso. User modeling in intelligent information retrieval. *Information Processing & Management*, 23(4):305–320, 1987.

Sara Bunian, Alessandro Canossa, Randy Colvin, and Magy Seif El-Nasr. Modeling individual differences in game behavior using hmm. In *Proceedings of the AAAI conference on artificial intelligence and interactive digital entertainment*, volume 13, pp. 158–164, 2017.

Lang Cao. Diaggpt: An llm-based and multi-agent dialogue system with automatic topic management for flexible task-oriented dialogue, 2024. URL `https://arxiv.org/abs/2308.08043`.

Maximillian Chen, Ruoxi Sun, Sercan Ö. Arık, and Tomas Pfister. Learning to clarify: Multi-turn conversations with action-based contrastive self-training, 2024. URL `https://arxiv.org/abs/2406.00222`.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training, 2025. URL `https://arxiv.org/abs/2501.17161`.

Christopher Clarke, Yuzhao Heng, Lingjia Tang, and Jason Mars. Peft-u: Parameter-efficient fine-tuning for user personalization, 2024. URL `https://arxiv.org/abs/2407.18078`.

Subrata Das, Atharva Deshmukh, Sriparna Saha, Roshni Ramnani, Anutosh Maitra, and Shubhashis Sengupta. Dynamic negotiation landscapes: Mbps and the interplay of buyer personalities. *Available at SSRN 4765633*.

Jan de Wit. Leveraging large language models as simulated users for initial, low-cost evaluations of designed conversations. In *International Workshop on Chatbot Research and Design*, pp. 77–93. Springer, 2023.

Kaustubh Dhole. KAUCUS - knowledgeable user simulators for training large language models. In Yvette Graham, Qun Liu, Gerasimos Lampouras, Ignacio Iacobacci, Sinead Madden, Haider Khalid, and Rameez Qureshi (eds.), *Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024)*, pp. 53–65, St. Julians, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.scichat-1.5/.

Vardhan Dongre, Xiaocheng Yang, Emre Can Acikgoz, Suvodip Dey, Gokhan Tur, and Dilek Hakkani-Tür. Respact: Harmonizing reasoning, speaking, and acting towards building large language model-based conversational ai agents. *arXiv preprint arXiv:2411.00927*, 2024.

Layla El Asri and et al. Sequence to sequence learning for task-oriented dialogue with dialogue state representation. In *Proceedings of the 2016 Conference on Spoken Language Technology*, 2016.

Pierre Erbacher, Laure Soulier, and Ludovic Denoyer. State of the art of user simulation approaches for conversational information retrieval. *arXiv preprint arXiv:2201.03435*, 2022.

Fahmid Morshed Fahid, Jonathan Rowe, Yeojin Kim, Shashank Srivastava, and James Lester. Online reinforcement learning-based pedagogical planning for narrative-centered learning environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 23191–23199, 2024.

Rafael Ferreira, David Semedo, and João Magalhães. Multi-trait user simulation with adaptive decoding for conversational task assistants, 2024. URL https://arxiv.org/abs/2410.12891.

Zhaolin Gao, Wenhao Zhan, Jonathan D. Chang, Gokul Swamy, Kianté Brantley, Jason D. Lee, and Wen Sun. Regressing the relative future: Efficient policy optimization for multi-turn rlhf, 2024. URL https://arxiv.org/abs/2410.04612.

Kallirroi Georgila, James Henderson, and Oliver Lemon. User simulation for spoken dialogue systems: learning and evaluation. In *Interspeech*, pp. 1065–1068. Citeseer, 2006.

Izzeddin Gür, Dilek Hakkani-Tür, Gokhan Tür, and Pararth Shah. User modeling for task oriented dialogues. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 900–906, 2018. doi: 10.1109/SLT.2018.8639652.

Christine Herlihy, Jennifer Neville, Tobias Schnabel, and Adith Swaminathan. On overcoming miscalibrated conversational priors in llm-based chatbots, 2024. URL https://arxiv.org/abs/2406.01633.

Friederike Holderried, Christian Stegemann-Philipps, Anne Herrmann-Werner, Teresa Festl-Wietek, Martin Holderried, Carsten Eickhoff, Moritz Mahling, et al. A language model–powered simulated patient with automated feedback for history taking: Prospective study. *JMIR Medical Education*, 10(1):e59213, 2024.

Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. War and peace (waragent): Large language model-based multi-agent simulation of world wars. 2023.

Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. War and peace (waragent): Large language model-based multi-agent simulation of world wars, 2024. URL https://arxiv.org/abs/2311.17227.

Eugene Ie, Chih-wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. Recsim: A configurable simulation platform for recommender systems. *arXiv preprint arXiv:1909.04847*, 2019.

Srinivasan Janarthanam and Oliver Lemon. Learning adaptive referring expression generation policies for spoken dialogue systems. In *Conference of the European Association for Computational Linguistics*, pp. 67–84. Springer, 2009.

Priyanka Kargupta, Ishika Agarwal, Dilek Hakkani-Tur, and Jiawei Han. Instruct, not assist: Llm-based multi-turn planning and hierarchical questioning for socratic code debugging, 2024. URL https://arxiv.org/abs/2406.11709.

Priyanka Kargupta, Ishika Agarwal, Tal August, and Jiawei Han. Tree-of-debate: Multi-persona debate trees elicit critical thinking for scientific comparative analysis, 2025. URL https://arxiv.org/abs/2502.14767.

Heikki Keskustalo, Kalervo Järvelin, and Ari Pirkola. Evaluating the effectiveness of relevance feedback based on a user simulation model: effects of a user scenario on cumulated gain value. *Information Retrieval*, 11: 209–228, 2008.

Byeonghwi Kim, Minhyuk Seo, and Jonghyun Choi. Online continual learning for interactive instruction following agents, 2024. URL https://arxiv.org/abs/2403.07548.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14397–14413, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.890. URL https://aclanthology.org/2023.emnlp-main.890/.

To Eun Kim and Aldo Lipani. A multi-task based neural model to simulate users in goal oriented dialogue systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pp. 2115–2119, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3531814. URL https://doi.org/10.1145/3477495.3531814.

Florian Onur Kuhlmeier, Leon Hanschmann, Melina Rabe, Stefan Luettke, Eva-Lotta Brakemeier, and Alexander Maedche. Combining artificial users and psychotherapist assessment to evaluate large language model-based mental health chatbots. *arXiv preprint arXiv:2503.21540*, 2025.

Sahiti Labhishetty. *Models and evaluation of user simulation in information retrieval*. PhD thesis, University of Illinois at Urbana-Champaign, 2023.

Barrett Martin Lattimer, Varun Gangal, Ryan McDonald, and Yi Yang. Sparse rewards can self-train dialogue agents, 2024. URL https://arxiv.org/abs/2409.04617.

Xixi Lei, Changqun Li, Liang He, and Xin Lin. An interactive evaluation framework for empathetic response generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Esther Levin, Roberto Pieraccini, and Wieland Eckert. A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on speech and audio processing*, 8(1):11–23, 2000.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. Agent hospital: A simulacrum of hospital with evolvable medical agents, 2025. URL https://arxiv.org/abs/2405.02957.

Xiujun Li, Zachary C Lipton, Bhuwan Dhingra, Lihong Li, Jianfeng Gao, and Yun-Nung Chen. A user simulator for task-completion dialogues. *arXiv preprint arXiv:1612.05688*, 2016.

Yajiao Liu, Xin Jiang, Yichun Yin, Yasheng Wang, Fei Mi, Qun Liu, Xiang Wan, and Benyou Wang. One cannot stand for everyone! leveraging multiple user simulators to train task-oriented dialogue systems. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1–21, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.1. URL https://aclanthology.org/2023.acl-long.1/.

Yi Liu, Gaurav Datta, Ellen Novoseller, and Daniel S. Brown. Efficient preference-based reinforcement learning using learned dynamics models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2921–2928, 2023b. doi: 10.1109/ICRA48891.2023.10161081.

Zhengyuan Liu, Stella Xin Yin, Geyu Lin, and Nancy F. Chen. Personality-aware student simulation for conversational intelligent tutoring systems, 2024. URL https://arxiv.org/abs/2404.06762.

Xiang Luo, Zhiwen Tang, Jin Wang, and Xuejie Zhang. Duetsim: Building user simulator with dual large language models for task-oriented dialogues, 2024. URL https://arxiv.org/abs/2405.13028.

Xiang Luo, Jin Wang, and Xuejie Zhang. Utterance alignment of language models for effective user simulation in task-oriented dialogues. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

Ellie S Paek, Talyn Fan, James D Finch, and Jinho D Choi. Enhancing task-oriented dialogue systems through synchronous multi-party interaction and multi-group virtual simulation. *Information*, 15(9):580, 2024.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023. URL https://arxiv.org/abs/2304.03442.

Olivier Pietquin and Thierry Dutoit. A probabilistic framework for dialog simulation and optimal strategy learning. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):589–599, 2006.

Olivier Pietquin and Helen Hastie. A survey on metrics for the evaluation of user simulations. *The Knowledge Engineering Review*, 28(1):59–73, 2013. doi: 10.1017/S0269888912000343.

Albert J Rehmann, Robert D Mitman, and Michael C Reynolds. A handbook of flight simulation fidelity requirements for human factors research. Technical report, 1995.

Stéphane Rossignol, Olivier Pietquin, and Michel Ianotto. Training a bn-based user model for dialogue simulation with missing data. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pp. 598–604, 2011.

Florian Rupp, Manuel Eberhardinger, and Kai Eckert. Simulation-driven balancing of competitive game levels with reinforcement learning. *IEEE Transactions on Games*, 2024.

Jost Schatzmann, Kallirroi Georgila, and Steve Young. Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*, pp. 45–54, 2005.

Jost Schatzmann, Kallirroi Georgila, and Steve Young. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. In *The Knowledge Engineering Review*, 2006.

Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. Agenda-based user simulation for bootstrapping a POMDP dialogue system. In Candace Sidner, Tanja Schultz, Matthew Stone, and ChengXiang Zhai (eds.), *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pp. 149–152, Rochester, New York, April 2007. Association for Computational Linguistics. URL https://aclanthology.org/N07-2038/.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments, 2024.

Fatemeh Shamsezat, Ali Mohades Khorasani, and Saeid Shiry Ghidary. Optimizing multi-domain task-oriented dialogue policy through sigmoidal discrete soft actor-critic. *Available at SSRN 5123532*.

Weilin Shen, Qiping Shen, and Quanbin Sun. Building information modeling-based user activity simulation and evaluation method for improving designer–user communications. *Automation in Construction*, 21: 148–160, 2012.

Weiwei Sun, Shuyu Guo, Shuo Zhang, Pengjie Ren, Zhumin Chen, Maarten de Rijke, and Zhaochun Ren. Metaphorical user simulators for evaluating task-oriented dialogue systems, 2023. URL https://arxiv.org/abs/2204.00763.

Silvia Terragni, Modestas Filipavicius, Nghia Khau, Bruna Guedes, André Manso, and Roland Mathis. In-context learning user simulators for task-oriented dialog systems, 2023. URL https://arxiv.org/abs/2306.00774.

Bo-Hsiang Tseng, Yinpei Dai, Florian Kreyssig, and Bill Byrne. Transferable dialogue systems and user simulators. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 152–166, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.13. URL https://aclanthology.org/2021.acl-long.13/.

Kuang Wang, Xianfei Li, Shenghao Yang, Li Zhou, Feng Jiang, and Haizhou Li. Know you first and be you better: Modeling human-like user simulators via implicit profiles, 2025. URL https://arxiv.org/abs/2502.18968.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, James Evans, Philip Torr, Bernard Ghanem, and Guohao Li. Can large language model agents simulate human trust behavior?, 2024. URL https://arxiv.org/abs/2402.04559.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. Exploring large language models for communication games: An empirical study on werewolf, 2024. URL https://arxiv.org/abs/2309.04658.

Se-eun Yoon, Zhankui He, Jessica Echterhoff, and Julian McAuley. Evaluating large language models as generative user simulators for conversational recommendation. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1490–1504, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.83. URL https://aclanthology.org/2024.naacl-long.83/.

Shuo Zhang and Krisztian Balog. Evaluating conversational recommender systems via user simulation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pp. 1512–1520, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403202. URL https://doi.org/10.1145/3394486.3403202.

Tong Zhang, Chen Huang, Yang Deng, Hongru Liang, Jia Liu, Zujie Wen, Wenqiang Lei, and Tat-Seng Chua. Strength lies in differences! improving strategy planning for non-collaborative dialogues via diversified user simulation, 2024. URL https://arxiv.org/abs/2403.06769.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. Sotopia: Interactive evaluation for social intelligence in language agents, 2024a. URL https://arxiv.org/abs/2310.11667.

Zhilun Zhou, Yuming Lin, Depeng Jin, and Yong Li. Large language model for participatory urban planning, 2024b. URL https://arxiv.org/abs/2402.17161.